OPEN ACCESS FULL TEXT. OPEN ACCESS EXCEL FILES.

Assessment of Treatment Plant Performance and Water Quality Data

A GUIDE FOR STUDENTS, RESEARCHERS AND PRACTITIONERS

Marcos von Sperling, Matthew E. Verbyla and Sílvia M. A. C. Oliveira

IWA PUBLISHING

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Assessment of Treatment Plant Performance and Water Quality Data

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Assessment of Treatment Plant Performance and Water Quality Data

A Guide for Students, Researchers and Practitioners

Marcos von Sperling, Matthew E. Verbyla and
Sílvia M. A. C. Oliveira

IWA PUBLISHING

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Published by **IWA Publishing**
Alliance House
12 Caxton Street
London SW1H 0QS, UK
Telephone: +44 (0)20 7654 5500
Fax: +44 (0)20 7654 5555
Email: publications@iwap.co.uk
Web: www.iwapublishing.com

First published 2020
© 2020 IWA Publishing

Apart from any fair dealing for the purposes of research or private study, or criticism or review, as permitted under the UK Copyright, Designs and Patents Act (1998), no part of this publication may be reproduced, stored or transmitted in any form or by any means, without the prior permission in writing of the publisher, or, in the case of photographic reproduction, in accordance with the terms of licenses issued by the Copyright Licensing Agency in the UK, or in accordance with the terms of licenses issued by the appropriate reproduction rights organization outside the UK. Enquiries concerning reproduction outside the terms stated here should be sent to IWA Publishing at the address printed above.

The publisher makes no representation, express or implied, with regard to the accuracy of the information contained in this book and cannot accept any legal responsibility or liability for errors or omissions that may be made.

Disclaimer
The information provided and the opinions given in this publication are not necessarily those of IWA and should not be acted upon without independent consideration and professional advice. IWA and the Editors and Authors will not accept responsibility for any loss or damage suffered by any person acting or refraining from acting upon any material contained in this publication.

British Library Cataloguing in Publication Data
A CIP catalogue record for this book is available from the British Library

ISBN: 9781780409313 (Paperback)
ISBN: 9781780409320 (eBook)
ISBN: 9781780409337 (ePub)

This eBook was made Open Access in January 2020.

© 2020 The Authors

This is an Open Access eBook distributed under the terms of the Creative Commons Attribution Licence (CC BY-NC-ND 4.0), which permits copying and redistribution for non-commercial purposes with no derivatives, provided the original work is properly cited (https://creativecommons.org/licenses/by-nc-nd/4.0/). This does not affect the rights licensed or assigned from any third party in this book.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

To

Vanessa and Bruno Guerra de Moura von Sperling, and
Roger and Roberta Verbyla and
Simão and Felipe Corrêa Oliveira

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Contents

Foreword ..... ***xvii***
Preface ..... ***xix***
Authors ..... ***xxi***

Chapter 1: Introduction ***1***
1.1 Concept of the Book 2
1.2 Structure of the Book 3
1.3 Why Should You Use this Book? 4
1.4 Who Should Use this Book? 6
1.5 Additional Information 6
1.6 Schematic Overview of the Book Chapters 8

Chapter 2: Flow data and the concept of loading ***21***
2.1 The Importance of Flow Data and the Concept of Load 22
2.2 Measuring Flow Rates and Analysing Data 24
2.2.1 Methods for measuring flow rates 24
2.2.2 Recording flow data 28
2.2.3 Flow variations 28
2.2.4 Flow equalization 28
2.2.5 Determining typical flow rates and distributions 30
2.2.6 Analysing flow data 33
2.3 Using Flow Rates to Assess Performance 36
2.3.1 Hydraulic retention time 36
2.3.2 Water losses and gains 37
2.4 Check-List for Your Report 38

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Chapter 3: Planning your monitoring programme. Sampling and measurements ***39***
3.1 Types of Monitoring Programmes and Studies 40
3.2 Quality Assurance and Quality Control 41
3.2.1 Introductory concepts 41
3.2.2 Scope of the study 41
3.2.3 Environmental samples, statistical samples, and populations 43
3.2.4 Measurements and anticipated use of data 45
3.2.5 Standard assessment thresholds and operating procedures 45
3.2.6 Quality control samples 48
3.2.7 Data management and analysis 50
3.3 Sample Collection 50
3.3.1 Spatial aspects of sampling 50
3.3.2 Types of samples 54
3.3.3 Need for a time delay to collect the downstream sample? 58
3.4 Sample Size, Containers, and Holding Times 62
3.5 Statistical Power and Number of Samples 63
3.6 Check-List for Your Report 67

Chapter 4: Laboratory analysis and data management ***69***
4.1 Raw Data, Calculated Values, and Statistics 70
4.2 Storing Data and Calculated Values 72
4.2.1 Where and how to store your data 72
4.2.2 Storing data in a spreadsheet (most datasets) 74
4.2.3 Storing data in a database (larger datasets) 75
4.3 Metadata 80
4.4 Accuracy and Precision 81
4.5 Uncertainty and Variability 82
4.5.1 Variability of a population 82
4.5.2 Uncertainty in our estimate of parameters 82
4.5.3 The central limit theorem and confidence intervals 83
4.5.4 Prediction intervals and confidence intervals 85
4.6 Detection Limits 87
4.6.1 Variability from instruments and sample processing 87
4.6.2 Limits of detection and quantification 88
4.7 Significant Figures 90
4.7.1 Significant figures for direct measurements from instruments that give live readings 90
4.7.2 Significant figures for direct measurements from instruments that do not give live readings 90
4.7.3 Significant figures for calculated values based on standard curves 91
4.8 Check-List for Your Report 94

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Chapter 5: Descriptive statistics: numerical methods for describing monitoring data ***95***
5.1 An Overview on Descriptive Statistics 96
5.2 Structuring Your Tables with Summary Descriptive Statistics 101
5.2.1 Different types of studies requiring different types of summary tables 101
5.2.2 Summary tables of studies in treatment plants 101
5.2.3 Summary tables of studies in water bodies 113
5.3 Missing Data 116
5.4 Censored Data 117
5.4.1 The concept of censored data 117
5.4.2 Treatment of left-censored data (below the DL) 118
5.4.3 Treatment of right-censored data (data above the DL) 122
5.5 Outliers 123
5.5.1 Concept of outliers and importance of their analysis 123
5.5.2 Determination of outliers 125
5.6 Measures of Central Tendency 128
5.6.1 Introduction 128
5.6.2 Mean 131
5.6.3 Median 134
5.6.4 Geometric mean 135
5.6.5 Weighted average 138
5.7 Measures of Variation 142
5.8 Measures of Relative Standing 148
5.9 Check-List for Your Report 150

Chapter 6: Descriptive statistics: graphical methods for describing monitoring data ***151***
6.1 Introduction 152
6.2 Time Series Graphs 154
6.2.1 Use of time series graphs 154
6.2.2 Connection of data points with lines 156
6.2.3 Missing data and days without monitoring in scatter charts and line charts 156
6.2.4 *Y*-axis scale in time series graphs 159
6.2.5 Graphs with two *Y* axes 160
6.2.6 Arithmetic and logarithmic scales 160
6.2.7 Moving averages 161
6.3 Frequency Distribution 165
6.3.1 Frequency distributions and frequency histograms 165
6.3.2 Frequency polygon 168
6.3.3 Percentile graphs 169
6.4 Box-and-Whisker Graphs (Box Plots) 172

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

6.5 Scatter Plots ... 175
6.6 Graphs for Qualitative (Categorized) Data ... 176
6.7 General Advices on Presenting Graphs ... 179
6.8 Check-List for Your Report ... 180

Chapter 7: Removal efficiencies ... ***181***
7.1 The Concept of Removal Efficiency ... 182
7.2 How to Calculate and Report Removal Efficiencies ... 182
7.2.1 Expressing removal efficiencies as relative values or percentages ... 182
7.2.2 Expressing removal efficiencies as logarithmic units removed ... 183
7.2.3 Relationship between removal efficiencies as percentages and log reduction values ... 184
7.2.4 Removal efficiencies (% and LRV) for units in series ... 186
7.3 Specific Aspects in the Calculation of Removal Efficiencies ... 188
7.3.1 The influence of water losses on the calculation of removal efficiencies ... 188
7.3.2 The influence of censored data on the calculation of removal efficiencies ... 188
7.3.3 Minimum and maximum possible values of removal efficiencies ... 191
7.3.4 Differences between removal and reduction ... 191
7.4 How to Interpret Values of Removal Efficiency ... 194
7.5 The Importance of Analysing Effluent Concentrations and Removal Efficiencies Together ... 195
7.6 Measures of Central Tendency for Removal Efficiencies ... 200
7.6.1 Two different ways of calculating central tendency of removal efficiencies ... 200
7.6.2 The case of missing data ... 201
7.6.3 The case of outliers ... 201
7.6.4 Mean efficiency versus mean of efficiencies: impact on results ... 201
7.6.5 Mean efficiency versus mean of efficiencies: which one to use? ... 203
7.7 Frequency Distribution of Removal Efficiencies ... 204
7.8 Check-List for Your Report ... 206

Chapter 8: Symmetry and asymmetry in monitoring data. Normal and log-normal distributions ... ***207***
8.1 Frequency Distributions of Monitoring Data ... 208
8.2 Normal Distribution ... 211
8.2.1 Basic concepts about the normal distribution ... 211
8.2.2 Influence of the mean and standard deviation on the normal distribution ... 212
8.2.3 Negative values for concentrations and values above 100% for removal efficiencies in normal distributions ... 213
8.2.4 Generation of values for the normal distribution ... 214
8.2.5 Standard normal variable (Z) ... 215
8.2.6 Skewness of a distribution ... 217
8.2.7 Fitting a normal distribution to your data ... 217
8.2.8 Tests for normality and goodness-of-fit tests for a normal distribution ... 219

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

8.3 Log-normal Distribution . . . 222
8.3.1 Basic concepts about the log-normal distribution . . . 222
8.3.2 Influence of geometric mean and geometric standard deviation on the log-normal distribution . . . 224
8.3.3 Generation of values for the log-normal distribution . . . 226
8.3.4 Fitting a log-normal distribution to your data . . . 226
8.3.5 Measures of central tendency and variation in the log-normal distribution . . . 229
8.3.6 Comparison between normal and log-normal distributions . . . 235
8.4 Moment Matching to Use Other Distributions . . . 235
8.5 Check-List for Your Report . . . 239

Chapter 9: Compliance with targets and regulatory standards for effluents and water bodies . . . ***241***
9.1 Regulatory Standards and Targets for Treatment Plant Effluents and the Quality of Drinking Water and Ambient Water Bodies . . . 242
9.2 Graphical Methods for Comparing Monitored Data with Quality Standards . . . 243
9.3 Evaluation of Compliance Based on Average Values . . . 247
9.3.1 Introductory concepts . . . 247
9.3.2 Fundamentals of hypothesis testing . . . 247
9.3.3 Different types of hypothesis tests . . . 249
9.3.4 Parametric one-sample test (*t*-test) . . . 250
9.3.5 Non-parametric one-sample test (sign test) . . . 251
9.3.6 Non-parametric one-sample test (Wilcoxon signed-rank test) . . . 254
9.3.7 Application of one-sample hypothesis tests to assess compliance . . . 255
9.4 Evaluation of Compliance Based on the Proportion of Non-conformity with Standard Using *Z*-test for Proportions . . . 257
9.5 Probabilities of Conformity or Non-conformity Obtained Directly from the Monitoring Data . . . 260
9.6 Estimation of Compliance with the Standard Based on Frequency Analysis Using Normal and Log-normal Distributions . . . 263
9.7 Reliability Analysis . . . 270
9.7.1 Reliability and stability . . . 270
9.7.2 Background concepts about reliability analysis . . . 271
9.7.3 The Coefficient of Reliability (COR) . . . 272
9.7.4 Expected probability of compliance with the standards . . . 276
9.8 Control Charts . . . 281
9.8.1 Introductory concepts on statistical process control . . . 281
9.8.2 Concepts behind a control chart for means . . . 282
9.8.3 Setting up a control chart for means (assumption of a normal distribution) . . . 291
9.8.4 Setting up a control chart for means (assumption of a log-normal distribution) . . . 300
9.8.5 Control chart for individual measurements (normal and log-normal distributions) . . . 306
9.8.6 Control chart for the proportion of failures (p-chart) . . . 310
9.9 Check-List for Your Report . . . 315

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Chapter 10: Making comparisons with your monitoring data. Tests of hypotheses ***317***
10.1 Introduction 318
10.1.1 Types of hypothesis tests 318
10.1.2 Decisions that need to be made before testing hypotheses 318
10.1.3 Summary of the different hypothesis tests 323
10.2 Inferences about Population Central Values 323
10.2.1 Establishing the test hypotheses 323
10.2.2 The four potential outcomes to a statistical test 327
10.2.3 One-tailed and two-tailed hypotheses tests 329
10.2.4 Rejection and non-rejection regions 330
10.2.5 Probability levels (*p*-values) and effect size estimates 336
10.3 One-sample Parametric Tests for a Population Mean (*Z* Test and *t* Test) 338
10.3.1 One-sample *Z* test (when σ is *known*) 338
10.3.2 One-sample *t* test (when σ is unknown) 340
10.3.3 Sample size and the *t* test 344
10.4 Inferences Comparing Two Population Central Values 348
10.4.1 Two-sample tests covered in this chapter 348
10.4.2 Inferences about the population means: parametric *t* test for two independent samples 349
10.4.3 Inferences about the population medians: non-parametric Mann Whitney *U*-test (Wilcoxon–Mann–Whitney *U*-test) for two independent samples 358
10.4.4 Inferences about the population means: parametric *t* test for two dependent samples (paired data) 363
10.4.5 Inferences about the population medians: non-parametric Wilcoxon signed-rank test for two dependent samples (matched pairs) 366
10.5 Comparing the Central Values of More Than Two Samples 371
10.5.1 Types of multiple-sample tests covered in this chapter 371
10.5.2 Parametric test for more than two population central values. ANOVA 371
10.5.3 *Post hoc* multiple comparison analysis following ANOVA: the parametric Tukey test 380
10.5.4 Non-parametric Kruskal–Wallis test for more than two population central values 385
10.5.5 *Post hoc* multiple comparison analysis following Kruskal–Wallis: the non-parametric Dunn test 390
10.6 Check-List for Your Report 395

Chapter 11: Relationship between monitoring variables. Correlation and regression analysis ***397***
11.1 Introduction 398
11.2 Correlation Coefficient 402
11.2.1 Pearson's linear correlation coefficient 402
11.2.2 Spearman rank correlation coefficient (non-parametric) 419
11.3 Correlation Matrix 424

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

11.3.1 Pearson correlation matrix . . . 424
11.3.2 Spearman rank correlation matrix (non-parametric) . . . 427
11.4 Cross-correlation and Autocorrelation . . . 429
11.4.1 Cross-correlation . . . 429
11.4.2 Autocorrelation . . . 436
11.5 Simple Linear Regression . . . 440
11.5.1 The simple linear regression equation . . . 440
11.5.2 Testing the significance of a regression . . . 445
11.5.3 Confidence intervals and prediction intervals . . . 450
11.5.4 Residual analysis . . . 452
11.5.5 The effect of influential observations and outliers in the regression analysis . . . 454
11.5.6 Data transformation . . . 455
11.5.7 Complete example of a simple linear regression . . . 455
11.5.8 Conceptual problems of a linear regression equation traditionally used in wastewater treatment design and evaluation . . . 468
11.6 Multiple Linear Regression . . . 470
11.6.1 Basics of multiple linear regression . . . 470
11.6.2 Potential problems or difficulties with multiple linear regression . . . 472
11.6.3 Graphical outputs for multiple linear regression . . . 472
11.6.4 Data transformations to linearize a model for using in a multiple regression model . . . 472
11.7 Non-linear Regression . . . 473
11.8 Check-List for Your Report . . . 476

Chapter 12: Water and mass balances . . . ***479***
12.1 Steady State and Dynamic State . . . 480
12.2 Water Balance . . . 481
12.3 Mass Balance . . . 487
12.4 Check-List for Your Report . . . 497

Chapter 13: Loading rates applied to treatment units . . . ***499***
13.1 The Different Types of Loading Rates . . . 500
13.2 Hydraulic Retention Time . . . 507
13.2.1 The general concept of hydraulic retention time . . . 507
13.2.2 Influence of the reactor dimensions on the theoretical hydraulic retention time . . . 509
13.2.3 Influence of internal recirculations on the theoretical hydraulic retention time . . . 509
13.2.4 Influence of a support medium on the theoretical hydraulic retention time . . . 511
13.2.5 Hydraulic retention time in tanks operated in batch mode . . . 512
13.2.6 Actual mean hydraulic retention time and departures from the theoretical behaviour . . . 513

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

13.3 Volumetric Hydraulic Loading Rate . . . 517
13.4 Surface Hydraulic Loading Rate . . . 518
13.5 Volumetric Mass Loading Rate . . . 520
13.6 Surface Mass Loading Rate . . . 521
13.7 Specific Surface Mass Loading Rate . . . 522
13.8 Food-to-microrganism Ratio (F/M) . . . 523
13.9 Sludge Age . . . 525
13.10 Check-List for Your Report . . . 529

Chapter 14: Reaction kinetics and reactor hydraulics . . . ***531***
14.1 Introduction . . . 532
14.2 Reaction Order . . . 533
14.2.1 Reaction orders – 0, 1, and 2 . . . 533
14.2.2 Zero-order reactions . . . 535
14.2.3 First-order reactions . . . 536
14.3 Experimental Determination of the Reaction Order and Kinetic Coefficient in Batch Reactors . . . 541
14.3.1 Estimation of the reaction order n and the reaction coefficient K . . . 541
14.3.2 Influence of a refractory fraction on the removal of a constituent (first-order reaction) . . . 549
14.3.3 First-order reaction with lag phase . . . 550
14.3.4 Influence of temperature on the reaction rate . . . 551
14.3.5 Time to reach a certain removal efficiency . . . 554
14.3.6 Applicability of reaction coefficients obtained from experiments done with continuous-flow reactors . . . 555
14.4 Idealized Flow Regimens in Continuous-Flow Reactors . . . 556
14.4.1 General concepts . . . 556
14.4.2 Idealized plug-flow reactor . . . 556
14.4.3 Idealized complete-mix reactor . . . 562
14.4.4 Deriving kinetic coefficients from existing continuous-flow reactors using idealized hydraulic models (plug-flow and complete-mix) . . . 566
14.5 Plug-Flow with Dispersion and Apparent Tanks-In-Series Models . . . 569
14.5.1 Conversion of the idealized hydraulic models to models that more closely represent reality . . . 569
14.5.2 Plug-flow with dispersion model . . . 571
14.5.3 Apparent number of tanks-in-series (NTIS) model . . . 574
14.5.4 Deriving kinetic coefficients from existing continuous-flow reactors using the plug-flow with dispersion and the apparent number of tanks-in-series models . . . 579
14.5.5 Applicability of kinetic coefficients derived under batch and continuous-flow experiments . . . 585
14.5.6 Utilization of the kinetic coefficient and hydraulic representation for the mathematical modelling of your reactor . . . 586
14.6 Check-List for Your Report . . . 592

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Chapter 15: Model application, calibration, and verification ***595***
15.1 Concepts Involved in Water Quality and Treatment Plant Modelling 596
15.1.1 A simple concept of mathematical models 596
15.1.2 A procedure for modelling 597
15.1.3 Definition of the model objectives 599
15.1.4 Model conceptualization 599
15.1.5 Selection of the model type 600
15.1.6 Properties required for mathematical models 602
15.2 Model Calibration 602
15.2.1 General aspects of model calibration 602
15.2.2 Calibration by minimization of the residuals 604
15.2.3 Evaluation of the goodness-of-fit of the model 605
15.2.4 Sensitivity analysis 616
15.3 Model Verification (Analysis of Residuals) 618
15.3.1 Required properties for the residuals 618
15.3.2 Assessing the normality of the distribution of residuals 620
15.3.3 Testing whether the residual mean is significantly different from zero 621
15.3.4 Checking whether the variance is constant (homoscedasticity of variance) 621
15.3.5 Evaluating the existence of autocorrelation in the residuals 621
15.4 Check-List for Your Report 630

References ***633***

Index ***639***

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Foreword

Over the past few decades technological developments have advanced enormously, even to the extent that they are often overwhelming, particularly for students and young water professionals entering the wastewater and water quality field. The quantity, handling, interpretation and understanding of water quality data generated in a wastewater treatment plant's lifecycle is becoming an increasing challenge, even to the most experienced users. The rapid developments in computational technology, combined with this deeper, fundamental understanding of the chemical, biological and physical processes involved in wastewater treatment and aquatic ecosystems, are causing this increased complexity in data management. Conversely, in many middle- and low-income countries, scientists and practitioners are regularly experiencing data scarcity and facing the challenge of how to interpret the data they do have to generate useful information that would lead to the creation of knowledge and ultimately to increased wisdom.

This book will make a major contribution to addressing these issues better and to bridging the gap between science and technology and their practical applications. The innovative 'alternative approach' that the authors of the book have consciously chosen to follow, starting with practice then moving to theory, and from application to fundamentals, will quickly attract many followers. Such an approach in our field is refreshing as it combines statistics, mathematics, modelling, process engineering, microbiology, physics and bio-chemistry in a balanced way, providing theoretical and fundamental information to the extent required for the solution of practical problems, regularly demonstrated by one or more examples. To many the final outcome may appear natural, and ultimately not even 'alternative'; however to get to that stage of practical simplification is an achievement in itself, and is thanks to the extensive experience and knowledge of the authors on this matter.

© IWA Publishing 2020. Assessment of Treatment Plant Performance and Water Quality Data: A Guide for Students, Researchers and Practitioners
Author(s): Marcos von Sperling, Matthew E. Verbyla and Sílvia M. A. C. Oliveira.
doi: 10.2166/9781780409320_xvii

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

I have known Professor von Sperling, the lead author, for over a decade and we have been working closely on a large research and capacity-building project for the developing world involving more than 90 PhD and MSc students and post-doctoral Fellows. When I read this book, I can hear him saying the words in his characteristic Brazilian-English accent, because that is exactly what he has been preaching for years to students and to all of us. I recall and am grateful for all the advice he has generously offered during our research encounters.

This book is a breath of fresh air in our field; the authors set the tone from the very first paragraph, their approach is surprisingly direct and transparent, their knowledge is genuinely shared, the book is open access, and the attached tools are accessible and changeable, giving the reader the feeling of 'what you see is what you get'. The usefulness of this book to all stakeholders in the field is undoubted; it will be used by its intended audience and will soon become a compulsory, 'must have', item in the collection of water scientists and professionals. I am delighted that the authors have made such a tremendous effort to create this book; I am looking forward to using it myself and to introducing it to a curriculum of programs I lead, and my students will use it too. I would like to take this opportunity to congratulate the authors on this great and unique piece of work.

Prof. Dr. Damir Brdjanovic
Professor of Sanitary Engineering
IHE Delft Institute for Water Education and
Delft University of Technology
The Netherlands
September 2019

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Preface

We, the three authors, have experience working as engineers in the private sector, but we all now work in the academic field. We feel very fortunate about the range of learning opportunities we have in our roles as professors. We are able to continue our own learning through our daily activities: by teaching and having direct interactions with students in the classroom; by supervising research students and participating in MSc and PhD examinations; by serving on the scientific committees for conferences and serving as peer-reviewers or editors for academic journals; by preparing research proposals, working on projects with colleagues, attending and presenting our work at national and international symposiums, conferences and congresses, and by submitting our own manuscripts for publication and receiving feedback from other peer reviewers.

We feel very indebted about this continuous learning opportunities available to us, and we strongly believe that knowledge needs to be shared in a way that is open and accessible to all. The knowledge we learn needs to be freely and openly passed on, so that others may build upon it, further develop on these concepts and ideas, and disseminate them to future generations of students and practitioners. In our experience, we have seen several cases of excellent water quality studies of natural systems and engineered treatment plants that involved a lot of hard work to obtain high-quality monitoring data, but unfortunately fell short in terms of the way the data were presented and analysed. In many cases, data were not presented in a way that was clear and transparent, the statistical methods used were limited or inappropriate, or the monitoring results were not fully integrated with the authors' knowledge of the processes associated with the system being studied. This leads to a situation where the knowledge generated from these excellent studies is limited and not very generalizable. Throughout all these years, we have been able to identify the major difficulties encountered by researchers and practitioners when processing and reporting their data and results. We realized some important gaps in the way that

© IWA Publishing 2020. Assessment of Treatment Plant Performance and Water Quality Data: A Guide for Students, Researchers and Practitioners
Author(s): Marcos von Sperling, Matthew E. Verbyla and Sílvia M. A. C. Oliveira.
doi: 10.2166/9781780409320_xix

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

we teach the analysis of data from water quality and treatment plants that needed to be filled in order to teach others how to allow the findings to become useful (i.e., making your findings generalizable so that they may be more useful to others who are working with similar systems in different environments).

This was our motivation for writing this book. We aim to guide you through the conceptualization of your research, the design of your experiment, the presentation of your experimental data, the use of basic descriptive statistics, as well as more advanced statistical analyses to interpret your data and integrate it with your knowledge of the processes and the governing principles of the system you are studying. Our subject matter is the analysis of monitoring data from water and wastewater treatment plants and water bodies. We believe that our book encompasses the following elements:

- A problem-oriented approach, working from practice to theory, in a clear and didactic way
- Innovative approach of combining process knowledge with statistical analysis
- Major concepts supported by fully worked-out examples and Excel spreadsheets
- Completely open-access material

We have the following target readership in mind and possible uses of the book:

- **Research students, postdoctoral scientists and professors** may find the book useful if they are assessing water quality or the performance of treatment systems or treatment technologies and they want to extract the most out of their data, to make findings that are both insightful and of broader interest.
- **Environmental engineers, water and wastewater sector practitioners, and environmental (water quality) policy makers** who use this book will develop a better understanding about how to set and ensure compliance with water quality norms, guidelines and regulations through the use of statistical inference.
- **Master's students, PhD students and upper-division undergraduate students** may utilize this book as support material for a course they are taking as part of an engineering degree program or another program that emphasizes the use of applied sciences to assess water quality.

The publication in open-access mode was made possible by the utilization of incentive funds from an international programme financed by the Bill and Melinda Gates Foundation for the project "Stimulating local innovation on sanitation for the urban poor in Sub-Saharan Africa and South-East Asia – SaniUp", under the coordination of UNESCO-IHE, Institute for Water Education, Delft, the Netherlands. Additional financial support to make this publication open access was also provided by the Department of Civil, Construction, and Environmental Engineering at San Diego State University and from a project entitled "Knowledge to Practice with the Global Water Pathogens Project," led by Michigan State University and funded by the Bill and Melinda Gates Foundation. This material is also based upon work supported by the National Science Foundation under Grant No. 1827251.

We would like to give thanks for the support received from the universities where we work (Federal University of Minas Gerais, Brazil, and San Diego State University, California, USA). We also would like to show our appreciation to IWA Publishing, for their incentive and patience in following the development of this book.

We hope you enjoy the book!

Marcos von Sperling
Matthew E. Verbyla
Sílvia M. A. Corrêa Oliveira
September 2019

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Authors

Marcos von Sperling Civil engineer, working for four decades in the field of wastewater treatment and water pollution control. Full professor at the Department of Sanitary and Environmental Engineering, Federal University of Minas Gerais (UFMG), Brazil. Fellow of the International Water Association (IWA). International Honorary Member of the American Academy of Environmental Engineers and Scientists, USA. Researcher level 1 of the Brazilian Research Council (CNPq). Former chair of the IWA Specialist Group on Wastewater Pond Technology. Editor of the IWA Journal on Water, Sanitation and Hygiene for Development. PhD in Environmental Engineering (Imperial College London), MSc in Sanitary Engineering (Federal University of Minas Gerais, Brazil). Author of several textbooks published in Portuguese, Spanish and English (the latter by IWA Publishing).

Matthew E. Verbyla Environmental engineer, originally from Connecticut, USA. Assistant Professor of Environmental Engineering at San Diego State University, California, USA. Recipient of a US Fulbright Fellowship (2007), US National Science Foundation Graduate Research Fellowship (2012), and the W. Wesley Eckenfelder Graduate Research Award (American Academy of Environmental Engineers and Scientists, 2016). Member of the editorial team for the Global Water Pathogens Project. PhD and MSc degrees in Environmental Engineering from the University of South Florida (2012 and 2015), and BS degree in Civil Engineering from Lafayette College (2006).

Sílvia Maria Alves Corrêa Oliveira Electrical engineer, with master's and doctorate in Sanitation, Environment and Water Resources at the Federal University of Minas Gerais (UFMG), Brazil. Associate Professor at the Department of Sanitary and Environmental Engineering at UFMG, and former coordinator of the Undergraduate Course in Environmental Engineering at UFMG. Researcher of the Brazilian Research Council (CNPq). Experience in the area of statistical treatment of environmental data, with emphasis on water, air and soil quality assessment; assessment and management of impacts and environmental risks and characterization, prevention and control of pollution.

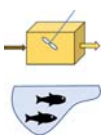
Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Chapter 1

Introduction

This chapter introduces our book to you, describing its approach, structure, applicability, and target readership. We also provide a schematic overview of each of the book chapters.



The contents in this chapter are applicable to both treatment plant monitoring and water quality monitoring.

CHAPTER CONTENTS

1.1 Concept of the Book **2**
1.2 Structure of the Book **3**
1.3 Why Should You Use this Book? **4**
1.4 Who Should Use this Book? **6**
1.5 Additional Information **6**
1.6 Schematic Overview of the Book Chapters **8**

© 2020 The Authors. This is an Open Access book chapter distributed under the terms of the Creative Commons Attribution Licence (CC BY-NC-ND 4.0), which permits copying and redistribution for non-commercial purposes with no derivatives, provided the original work is properly cited (https://creativecommons.org/licenses/by-nc-nd/4.0/). This does not affect the rights licensed or assigned from any third party in this book. The chapter is from the book Assessment of Treatment Plant Performance and Water Quality Data: A Guide for Students, Researchers and Practitioners, Marcos von Sperling, Matthew E. Verbyla and Sílvia M. A. C. Oliveira (Authors).

doi: 10.2166/9781780409320_0001

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

1.1 CONCEPT OF THE BOOK

The purpose of our book is to present the basic principles for evaluating water quality and treatment plant performance in a clear and didactic way using a combined approach that involves the interpretation of **monitoring data** associated with:

- the basic **processes** that take place in *water bodies* and in *water* and *wastewater treatment plants*
- data management and **statistical** calculations to allow a deep interpretation of the data.

This book does not purely contain math and statistics. There are already several excellent books that cover pure and applied statistics, including books with a focus on statistics for environmental problems. These other books generally follow a typical structure, first presenting the major statistical concepts and then building examples around them. Some of these books are great and many are extensively used in courses and as a supporting material for our research studies.

However, our book approaches these concepts from an alternative perspective. We made it *problem-oriented*, that is, we *start* with the **problems and needs regarding the assessment of water quality and treatment plants**. Then, we present the required statistical tools and process knowledge needed to assess treatment plant performance and water quality using **monitoring data**. As such, our proposal is not to work from theory to practice, but rather **from practice to theory** or from **application to fundamentals**, and to present theory in the simplest way possible. See Figure 1.1 for a summary of our approach for writing this book.

The book includes a vast number of summary **tables**, **illustrations**, **graphs**, and **examples**, related to processes taking place in water bodies and treatment plants, supported by statistical tools that assist in the interpretation of the monitoring data.

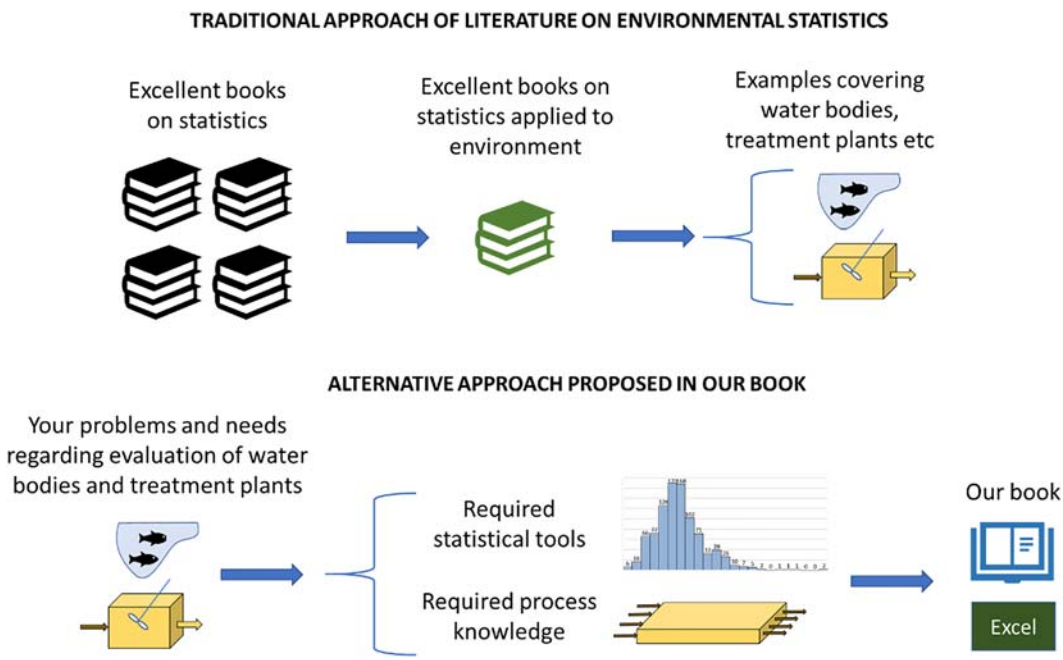


Figure 1.1 Traditional approach on the literature on environmental statistics and proposed approach for this book, combining process and statistical calculations.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

We strongly believe in **practical examples** as a means of consolidating theory. We want to have theory and practice presented and understood together. The examples are fully worked out in the book and supported with customized Microsoft® **Excel spreadsheets** that are freely available to the readers. We try to show how to do most of the calculations in the book, but we demonstrate how to also make good use of the built-in Excel functions.

We want to teach you to make the most of your monitoring data, using the values of **flows**, **concentrations**, and **loads** that you have obtained to create the most insight about the performance or condition of the water body or treatment plant you are studying. Therefore, *we start at the planning stages of your monitoring programme and then advance your knowledge, step by step, about the methods needed to interpret and present your data with the support of process and statistical calculations*.

The Excel spreadsheets are available for download through the IWA Publishing website (https://doi.org/10.2166/9781780409320)

1.2 STRUCTURE OF THE BOOK

Initially, a warning to you. It was not simple for us to devise a strategy in which the concepts should be presented in a reverse order (e.g., **from application to fundamentals** or **from practice to theory**), compared with traditional statistics and process books, which start with the theory and then present examples of applications. However, we do feel that our approach will resonate with many students and professionals who are very familiar with the engineering and water quality systems but who may have struggled in the past to understand concepts related to data management and statistical analysis.

Thus, in order to accomplish our goal, we had to explicitly **structure the chapters and sections starting with the problem or application** and then **including *only the necessary* theoretical background to be able to apply the method or solve the problem**. However, when doing this, we were sometimes forced to *split the statistical theory into complementing sections in different chapters*. To give one example, our chapter on hypothesis testing (Chapter 10) is presented after Chapter 9, which applies hypothesis tests to solve problems related to assessing compliance. We split the presentation of the fundamental theory and methodology for making statistical inferences into different chapters, in order to prioritize our focus on the application rather than the theory.

If you enjoy learning by direct application, then we feel that this structure will work well for you. The concepts presented along with the applications in this book are explained in sufficient detail for you to learn the fundamentals. However, if you want to further expand your depth of some of the statistical theory before completing the application, you may need to consult different sections, skipping forward or backward between the different book sections. Additionally, if you need to build a very strong background in the theoretical statistics, there are times when you should also consult classical statistics textbooks. In summary, we have:

- **Application of concept**. Direct use. Chapters and sections are self-contained and stand alone. Practical approach. Theoretical background is sufficient for the application.
- **Expanding theoretical knowledge**. To go deeper in the statistical theory, you will need to consult other sections that will complement your knowledge and allow you to get a broader view. You may need to return to the content you are covering for a full understanding of the procedures. You may also consult complementary information in textbooks or additional material available at the internet.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

To assist you on this, we tried to make our text **as didactic as possible**. Also, we make explicit references to complementary sections using symbols in the left-hand margin, which clearly indicate additional sections you may need to consult if you want to deepen your understanding of the theory, or see the theoretical concepts used in a different context or for a different application. For example:

C. 3

S. 4.5

… additional details can be found in Chapter 3 …
… this topic is further discussed in Section 4.5 …

Now, let us present the book structure, which is illustrated in Figure 1.2. There are four main parts, each of them comprised by individual chapters dealing with **process knowledge** and **statistical analysis**. The main concepts are built progressively throughout the book, but each chapter retains some independence and may be consulted individually if you are working on a specific topic. Several cross-references are made between the different chapters to help you review and delve deeper into a particular topic.

1.3 WHY SHOULD YOU USE THIS BOOK?

We started conceiving the book with the following question in mind: *how could the book be useful to you*? Our initial drive for creating this book was motivated by the following experience: we have observed many instances where so much effort is put into monitoring programmes – laborious hours, days, and months are spent in the field or in the laboratory to obtain important data – but in the end, the presentation and analysis of the data do not do justice to all of the efforts that went into collecting it. Sometimes, only mean values and simple bar charts are presented in the final reports, precluding the opportunity to make a whole lot more inference about the system!

With data in hand, we have a rich opportunity to understand important concepts about data variability, the relationships between variables, comparisons between samples, compliance with quality targets, the influence of loading rates, mass balances, attempts in deriving kinetic coefficients and a process model, and a whole set of other possibilities of casting a new light on the system you are investigating. We always have to keep in mind that **our study must be useful for others**, and extracting most of our data and presenting them in a clear way is an essential step in this direction.

In this book, we will push you to do more with your monitoring data! Initially, if you have not yet started monitoring, we will teach you how to plan your studies and how to organize the raw and processed data. After you have collected the monitoring data, we will teach you how to present basic descriptive statistics using summary tables and charts. Then, we will show you how to analyse the data distribution and make inferences about compliance with quality standards or targets. We will show you how to make meaningful comparisons between different water bodies or treatment units, between different operational phases and seasons, using hypothesis tests. We will show you how to investigate the relationship between variables, making use of correlation and regression analysis. And, if you want to delve even deeper to understand the behaviour of your system, we will teach you how to apply process knowledge to complete water and mass balances and investigate the influence of hydraulic and mass loading on performance. Finally, to make your results have broader impacts for people studying other systems with similar characteristics, we will show you how to characterize the hydraulic behaviour of your reactor, derive estimates for kinetic coefficients of reactions, incorporate them into a mathematical model, and see whether you can use this model to represent the system.

Therefore, let us **guide you through each of these steps** so that you can take your monitoring data and use it to produce the best possible report or publication. We recognize that each of these steps is extensively covered in the literature (statistical and process books, including several texts freely available in the internet). It is not our intention to duplicate this content. Rather, we aim at presenting the material in a

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

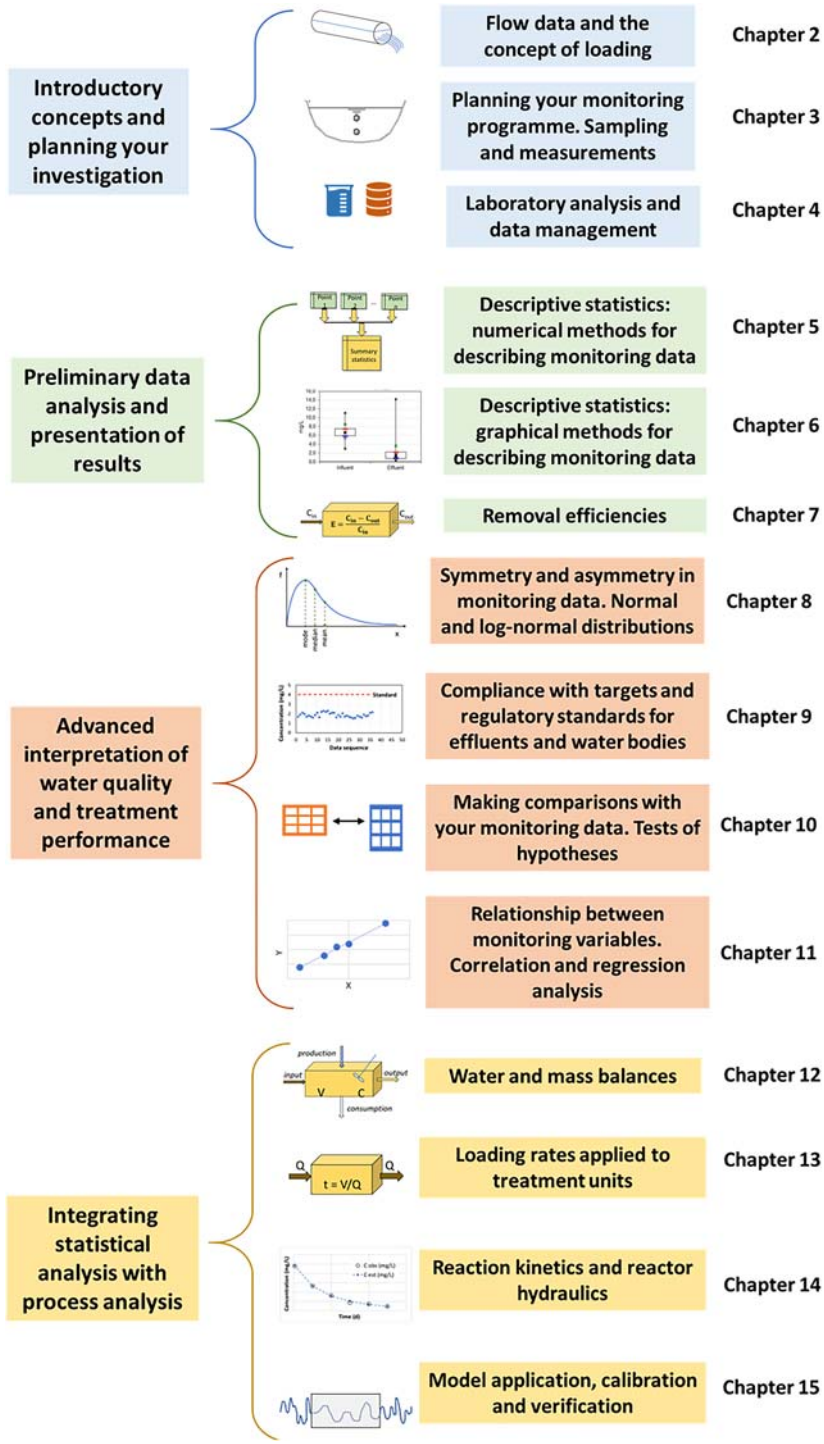


Figure 1.2 Main parts and chapters that comprise the book structure.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

way that focuses on the application while still teaching you the important fundamentals, starting with simple approaches in a structured way, so that you may be able to put the theory directly into practice and, if you like, expand your knowledge about the theory using other complementing literature. *You may perceive statistics to be difficult, but trust us, it is possible for you to learn it and even become an expert!*

1.4 WHO SHOULD USE THIS BOOK?

Some water and wastewater treatment engineers, students, and practitioners may be highly proficient with concepts related to water quality and the performance of treatment processes, but may lack familiarity with some basic statistical methods. If you are one of these people, do not worry. Statistics is a complex topic! On the other hand, some scientists and researchers may have a thorough background on statistical concepts but would benefit from seeing how they can be applied to real-life situations, such as the evaluation of water quality in a water body or a diagnosis of the performance of a water or wastewater treatment plant. If any of this sounds familiar to you, then this book is for you!

We hope that this book will enable **environmental, water and wastewater treatment engineers, practitioners, and policy makers** to have a better understanding about how to set and ensure compliance with water quality norms, guidelines and regulations through the use of statistical inference.

Research students, postdoctoral scientists, and professors are also an explicit target audience for our book, and they may also find it useful if they are completing projects that involve the assessment of water quality or the performance of treatment systems.

For **classroom use**, the academic level could be for upper-division **undergraduate students, Master's students or even PhD students**, especially from engineering programmes or other programmes that emphasize applied science. Some instructors may prefer to continue using one of the other excellent textbooks on statistics, water quality, and treatment processes that already exist (since they follow a more traditional structure). However, even these instructors may use our book as an additional reference or *supporting text*, given that our book is open access and that we incorporate several of these statistical concepts in an applied way. Our book's examples and the associated *Excel spreadsheets* may be used as bases for classroom *exercises*.

All Microsoft® **Excel spreadsheets** are open, do not use macros, and you can clearly see all the formulae employed and how the calculations are done. Therefore, these files become additional *learning tools*, and you are free and encouraged to modify them and adapt to your intended uses.

Although our book concentrates entirely on the evaluation of the water quality in water bodies and the performance of water and wastewater treatment plants, other readers may also find some basic concepts useful to *other field of studies, such as soil or air quality*, since the statistical tools are mostly the same.

Our readership does not need to have a background in advanced mathematics to be able to use this reference book.

The book is **open access** so that it can be accessed and used wherever and whenever you feel it could be useful for you! You may feel free to reuse, adapt or repurpose any of our materials, so long as you provide attribution and share alike (in an open access publication). *There are many important benefits to keeping educational materials in the public realm, accessible to all*!

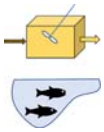
1.5 ADDITIONAL INFORMATION

In the book, we use a direct language with you, and we try to keep a simple and informal style. Of course, simplicity does not compromise the rigour we tried to keep in the methods we present.

In order to catch your attention to the main concepts and keywords in a paragraph, we make use of **bold** and *italics* in the text.

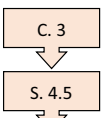
Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

We make use of the following **symbols**, which are presented at the left-hand margin of some paragraphs:



To explain whether the chapter contents are applicable to treatment plant monitoring and/or water quality monitoring.

To explain whether the contents in a particular section are at a basic or advanced level.



Indicates that additional information and theoretical background can be found in other chapters (e.g., Chapter 3) or in other sections (e.g., Section 4.5).

Indicates an example that is fully worked out in the book.

Indicates the availability of an Excel spreadsheet. In most of the cases, a spreadsheet is associated with an example and may be used for didactic or practical applications. In some cases, the spreadsheet is associated with a particular figure or table. Note that Microsoft and Excel are registered trademarks of the Microsoft Corporation. This book only uses the software and has not been sponsored by nor involves any responsibility from Microsoft.

Each chapter closes with a section entitled 'Check-List for your report'. We present bullet lists of points that you should check when preparing your technical report or scientific publication.

Please also note the following additional points:

- We are very conscious of the importance of reporting values with the correct number of **significant digits** (this is discussed in the book). However, in many cases, we show results of calculations with many decimal cases, just for you to be able to check the results of your own calculations.
- However, there may be some differences in the results from the calculations you do using **Excel** and using a **calculator**, if for the latter you are adopting rounded values. This will not affect the concepts and main results, but it is good that you know that in order not to be frustrated if you are not able to reproduce exactly the same values of the examples.
- We adopt the system of **separating thousands using a comma** (e.g., 1,000) or without comma (e.g., 1000). **Decimal cases are separated by a dot** (e.g., 1.45). However, in some graphs, because they have been produced using Excel in different languages, some values may appear separated by a comma (e.g., 1,45, but you should understand that they mean 1.45).
- The **Excel spreadsheets are available for downloading** together with the book DOI number. We also include **master spreadsheets**, for you to insert your own monitoring data and obtain, directly, the basic descriptive statistics and charts.
- **Excel** may vary with time, as new **versions** become available. Also, new functions are added and some functionalities may be expanded or removed. In principle, the Excel files provided here should work with moderately recent versions. If you encounter some problem with an add-in function, try to find

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

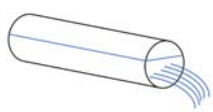
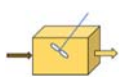
what is the closest one that can perform the calculations you intend to do. Search for information on the internet.

- Please note that **we are not software developers**. We tried to make the spreadsheets as didactic as possible, but you may find better ways of calculating or presenting the data, results, and graphs.

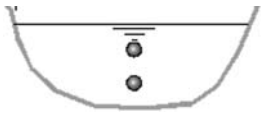
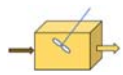
1.6 SCHEMATIC OVERVIEW OF THE BOOK CHAPTERS

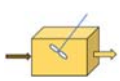
We present below a schematic overview of each chapter, including the following information:

- A title that shows to which of the four book parts the chapter belongs
- A short description of the chapter's contents
- A description of its applicability to water quality monitoring and/or treatment plant monitoring
- The overall level of the chapter contents (basic and/or advanced)
- The primary topics covered (i.e., the main chapter subsections)
- A description of content related to process knowledge (topics related to the analysis of the behaviour of the process of the water body or treatment plant)
- A description of content related to data analysis and statistics (data management and statistical tools used in the chapter)

<table>
<tr><th colspan="4">INTRODUCTORY CONCEPTS AND PLANNING YOUR INVESTIGATION</th></tr>
<tr><td colspan="4">CHAPTER 2
FLOW DATA AND THE CONCEPT OF LOADING</td></tr>
<tr><td colspan="2">  </td><td colspan="2">Description. How flow data are obtained and used in practice to support the assessment of water bodies or treatment plant performance.</td></tr>
<tr><td>  </td><td> </td><td colspan="2">Applicability. The contents in this chapter are applicable to both treatment plant monitoring and water quality monitoring.</td></tr>
<tr><td> </td><td> </td><td colspan="2">Level. Most of the concepts in this chapter are basic, but there are some advanced concepts.</td></tr>
<tr><td>Topics</td><td>Process knowledge</td><td colspan="2">Data analysis and statistics</td></tr>
<tr><td>• The importance of flow data
• Measuring flow rates
• Recording flow data
• Flow variations
• Flow equalization
• Typical flow rates and distributions
• Using flow rates to assess performance</td><td>• Concept of load
• Dosing of chemicals
• Structures for measuring flows
• Effect of flow equalization on pollutant concentrations. Analysing flow rate data (hourly and seasonal variations)
• Introduction to hydraulic retention time and water balance</td><td colspan="2">• Introductory concepts of the distribution of flow rates and their descriptive statistics</td></tr>
</table>

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

CHAPTER 3 PLANNING YOUR MONITORING PROGRAMME. SAMPLING AND MEASUREMENTS		
	**Description**. How to design research studies and establish monitoring programmes, with an emphasis on quality assurance, quality control, and the collection of representative samples.	
	***Applicability***. *The contents in this chapter are applicable to both treatment plant monitoring and water quality monitoring.*	
	Level. *Most of the concepts in this chapter are basic, but there are some advanced concepts.*	
Topics	**Process knowledge**	**Data analysis and statistics**
• Types of monitoring programmes and studies • Quality assurance and quality control • Sample collection • Sample size, containers, and holding times • Number of sample replicates	• Operational monitoring, compliance monitoring, research projects or special studies, and emergency studies • Types of measurements and anticipated use of data • Standard assessment thresholds and operating procedures • Quality control samples • Data management and analysis • Spatial aspects of sampling • Types of samples	• Power calculations to determine the appropriate sample size

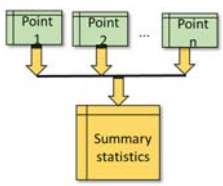
CHAPTER 4 LABORATORY ANALYSIS AND DATA MANAGEMENT	
	Description. Elements of importance when organizing, storing, reporting, publishing, and interpreting data obtained from laboratory analyses.
	***Applicability***. *The contents in this chapter are applicable to both treatment plant monitoring and water quality monitoring.*
	Level. *Most of the concepts in this chapter are basic, but there are some advanced concepts.*

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

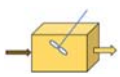
Topics	Process knowledge	Data analysis and statistics
• Raw data, calculated values, and statistics • Storing data and calculated values	• Types of replicates • Where and how to store your data • Storing data in a database (larger datasets)	• Storing data in a spreadsheet (most datasets) • Metadata • Accuracy and precision • Uncertainty and variability • Detection limits • Significant figures

PRELIMINARY DATA ANALYSIS AND PRESENTATION OF RESULTS

CHAPTER 5
DESCRIPTIVE STATISTICS: NUMERICAL METHODS FOR DESCRIBING MONITORING DATA



Description. How you should prepare and present the general results from your monitoring programme, in terms of flows, concentrations, removal efficiencies, and loads. Basic elements of descriptive statistics and covering simple numerical methods for describing your data are presented.



Applicability. *The contents in this chapter are applicable to both treatment plant monitoring and water quality monitoring. The exceptions are the mentions to 'removal efficiencies', which are applicable only to the assessment of treatment plants.*

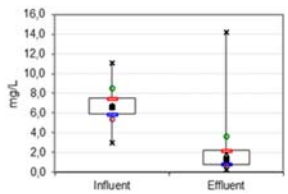
Level. *There is a balance between basic and advanced concepts.*

Topics	Process knowledge	Data analysis and statistics
• Overview of descriptive statistics • Structuring your tables with summary descriptive statistics • Missing data • Censored data • Outliers • Measures of central tendency • Measures of variation • Measures of relative standing		• Different types of studies requiring different types of summary tables • Summary tables of studies in treatment plants and water bodies • Handling missing data • Treatment of censored data (data below or above the detection limit) • Detection of outliers • Mean, median, geometric mean, mode, and weighted average

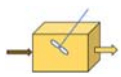
Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

		• Amplitude, variance, standard deviation, coefficient of variation, and geometric standard deviation • Percentiles

CHAPTER 6
DESCRIPTIVE STATISTICS: GRAPHICAL METHODS FOR DESCRIBING MONITORING DATA



Description. How to build and interpret the main types of charts used for describing your monitoring data: time series, frequency histograms, frequency polygons, percentile graphs, box plots and scatter plots for quantitative data, and bar/column charts and pie charts for qualitative data.



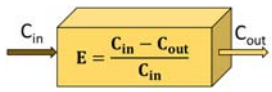
Applicability. *The contents in this chapter are applicable to both treatment plant monitoring and water quality monitoring. The exceptions are the mentions to 'removal efficiencies', which are applicable only to the assessment of treatment plants.*

Level. *There is a balance between basic and advanced concepts.*

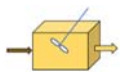
Topics	Process knowledge	Data analysis and statistics
• Main types of graphs for describing monitoring data • Time series graphs • Frequency distribution • Box-and-whisker graphs (box plot) • Scatter plots • Graphs for qualitative (categorized) data • General advices on presenting graphs		• Use of time series graphs, practical aspects in formatting time series graphs (connection with lines, missing data, *Y*-axis scale, and moving average) • Frequency distributions, frequency histograms, frequency polygons, and percentile graphs • Construction and interpretation of box-and-whisker plots • Construction of scatter plots • Graphs for categorized data (bar charts, column charts, and pie charts) • Useful hints on how to present charts

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

CHAPTER 7
REMOVAL EFFICIENCIES



Description. Descriptive statistics for removal efficiencies. Specificities on their calculation and interpretation. Different ways of presenting removal efficiencies (percentages or log reduction values). Influence of water losses, handling of censored data, and minimum and maximum possible values of removal efficiency. Joint interpretation of removal efficiencies and effluent concentrations. Measures of central tendency of efficiencies. Typical patterns of the associated frequency distributions.



–

Applicability. *The contents in this chapter are applicable only to treatment plant monitoring, since the concept of removal efficiencies does not apply to water quality monitoring in water bodies.*

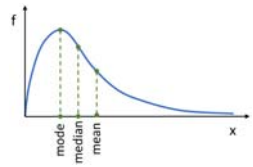
Level. *Most of the concepts in this chapter are advanced, but there are some basic concepts.*

Topics	Process knowledge	Data analysis and statistics
• The concept of removal efficiency • How to calculate and report removal efficiencies • Specific aspects in the calculation of removal efficiencies • How to interpret values of removal efficiency • The importance of analysing together effluent concentration and removal efficiency • Measures of central tendency of removal efficiencies • Frequency distribution of removal efficiencies	• Expressing removal efficiencies as relative values or percentages • Expressing removal efficiencies as logarithmic units removed • Influence of water losses and influence of censored data • Minimum and maximum values of removal efficiencies • Differences between removal and reduction • Concepts of good or poor removal efficiencies, and sufficient or insufficient removal efficiencies • Joint analysis of removal efficiencies and effluent concentrations (comparison of different treatment plants, comparison of different operational periods, and variations in influent concentrations)	• Measures of central tendency (mean of removal efficiencies, mean removal efficiency) • Frequency distributions of removal efficiencies and remaining fractions

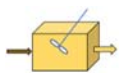
Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

ADVANCED INTERPRETATION OF WATER QUALITY AND TREATMENT PERFORMANCE

CHAPTER 8
SYMMETRY AND ASYMMETRY IN MONITORING DATA. NORMAL AND LOG-NORMAL DISTRIBUTIONS



Description. Symmetry and asymmetry in monitoring data. Foundations of two of the most important frequency distributions in environmental monitoring: normal and log-normal distributions. Main characteristics, properties, and parameters.



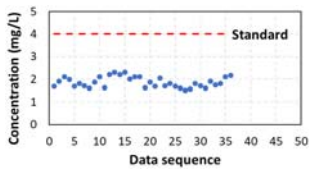
Applicability. *The contents in this chapter are applicable to both treatment plant monitoring and water quality monitoring.*

Level. *Most of the concepts in this chapter are advanced, but there are some basic concepts.*

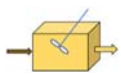
Topics	Process knowledge	Data analysis and statistics
• Frequency distributions of monitoring data • Normal distribution • Log-normal distribution		• Symmetry and asymmetry • Main types of frequency distributions • Basic concepts on normal and log-normal distributions • Influence of mean and standard deviation on the normal distribution and geometric mean and geometric standard deviation on the log-normal distribution • Negative values for concentrations; values above 100% for removal efficiencies in normal distributions • Generation of values for the normal and log-normal distributions • Standard normal variable (Z) • Skewness of a distribution • Fitting a normal distribution and a log-normal distribution to the data • Tests for normality and goodness-of-fit tests for a normal distribution • Comparison between normal and log-normal distributions

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

CHAPTER 9
COMPLIANCE WITH TARGETS AND REGULATORY STANDARDS FOR EFFLUENTS AND WATER BODIES



Description. How to assess conformity with targets established by managers or standards specified by regulatory agencies for the quality of water bodies or treatment plant effluents. Statistical tools for a broad view on compliance assessment. One-sample one-tailed parametric and non-parametric hypotheses tests. Frequency analysis, reliability analysis, and control charts under the assumptions of normal and log-normal distributions.



Applicability. *Most of the contents in this chapter are applicable to both treatment plant monitoring and water quality monitoring.*

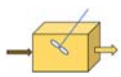
Level. *Most of the concepts in this chapter are advanced, but there are some basic concepts.*

Topics	Process knowledge	Data analysis and statistics
• Standards and targets for treatment plant effluents and water quality in water bodies • Graphical methods for comparing monitored data with quality standards • Evaluation of compliance based on average values • Evaluation of compliance based on the proportion of non-conformity (failures) • Probabilities of conformity obtained directly from the monitoring data • Estimation of compliance based on frequency analysis using normal and log-normal distributions • Reliability analysis • Control charts	• Quality standards and targets based on concentrations and removal efficiencies	• Time series graphs, box plots, and percentile graphs • Application of one-sample parametric and non-parametric tests • Parametric one-sample one-tailed test (*t*-test) • Non-parametric one-sample one-tailed test (sign test and Wilcoxon signed-rank test) • *Z* test for proportions, percentage of failure using Poisson distribution • Probability of conformity using percentiles from the monitored data • Probability models for assessing conformity based on normal and log-normal distributions • Reliability and stability, concept of reliability analysis, coefficient of reliability, expected percentage of compliance with the standards using normal and log-normal distributions • Statistical process control, control chart for means (normal and log-normal distributions), control chart for individual measurements (normal and log-normal distributions), and control chart for proportion of failures

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

CHAPTER 10
MAKING COMPARISONS WITH YOUR MONITORING DATA. TESTS OF HYPOTHESES

Description. How to compare two or more samples (different water bodies, treatment plants. or operating conditions) to infer whether there are significant differences between the means or medians of their underlying populations. Parametric and non-parametric two-sample tests followed by analysis of variance making multiple comparisons, also using parametric and non-parametric procedures.



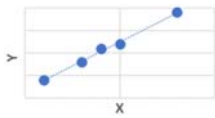
Applicability. *The contents in this chapter are applicable to both treatment plant monitoring and water quality monitoring.*

Level. *There is a balance between basic and advanced concepts.*

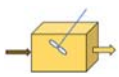
Topics	Process knowledge	Data analysis and statistics
• Inferences about population central values • Inferences comparing central values of two populations • Inferences comparing central values of more than two populations		• Introductory concepts of hypothesis testing • Parametric statistical test for a population mean; two-tailed test; *t*-test • Parametric tests for inferences about population means from independent samples (*t*-test) • Non-parametric tests for inferences about population medians from independent samples (Wilcoxon–Mann–Whitney U test) • Parametric tests for inferences about population means from dependent or paired samples (*t*-test) • Non-parametric tests for inferences about population medians from dependent or paired samples (Wilcoxon signed-rank test) • Analysis of variance • Multiple-comparison procedures. Parametric Tukey test, non-parametric Kruskal–Wallis test, and Dunn test

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

CHAPTER 11
RELATIONSHIP BETWEEN MONITORING VARIABLES. CORRELATION AND REGRESSION ANALYSIS



Description. How to analyse the relationship between two or more variables from your monitoring programme (influent and effluent concentrations, environmental conditions, removal efficiencies, applied loading rates, or others). Correlation between variables. Regression analysis, with emphasis on the linear regression model, which is fully analysed. Other regression models (multiple linear regression and non-linear regression).

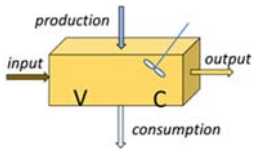
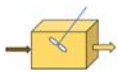


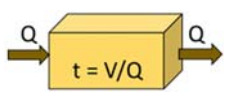
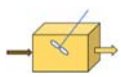
Applicability. *The contents in this chapter are applicable to both treatment plant monitoring and water quality monitoring.*

Level. *There is a balance between basic and advanced concepts.*

Topics	Process knowledge	Data analysis and statistics
• Correlation coefficient • Correlation matrix • Cross-correlation and autocorrelation • Simple linear regression • Multiple linear regression • Non-linear regression		• Differences between correlation and regression • Pearson and Spearman correlation coefficients (simple correlation and correlation matrices) • Cross-correlation between variables and autocorrelation of a single variable • Linear regression model (assumptions, regression coefficients, significance of the regression, coefficients of correlation and determination, confidence intervals, residuals analysis, influencing factors in the regression, and complete example) • Multiple linear regression model (structure, applicability, and interpretation) • Non-linear regression (non-linear multiple regression, polynomial regression, and other non-linear models)

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

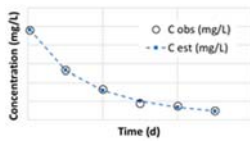
INTEGRATING STATISTICAL ANALYSIS WITH PROCESS ANALYSIS			
CHAPTER 12 WATER AND MASS BALANCES			
		**Description**. Basic elements of water and mass balances, important calculations for understanding the behaviour of a treatment plant. The concepts of steady state and dynamic state are also presented.	
	–	***Applicability***. *The contents in this chapter, in the way they have been structured, are mainly applicable to treatment plant monitoring. However, the overall concepts of steady and dynamic states, water balance, and mass balance are also applicable to water bodies.*	
		Level. *There is a balance between basic and advanced concepts.*	
Topics		**Process knowledge**	**Data analysis and statistics**
• Steady state and dynamic state • Water balance • Mass balance		• Steady state and dynamic state • Water balance around a treatment unit (input, output, gain, and loss) at steady state and dynamic state • Mass balance around a treatment unit (transport terms: input, output; reaction terms: production, consumption) at steady state and dynamic state	

CHAPTER 13 LOADING RATES APPLIED TO TREATMENT UNITS		
		**Description**. Different types of hydraulic and mass loading rates, and how to calculate and interpret them. Loading rates are used for the design of treatment units and for experimental studies that aim at investigating treatment performance under different loading conditions.
	–	***Applicability***. *The contents in this chapter are only applicable to treatment plant studies and not to the evaluation of water bodies.*
		Level. *There is a balance between basic and advanced concepts.*

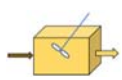
Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Topics	Process knowledge	Data analysis and statistics
• Hydraulic retention time • Surface and volumetric hydraulic loading rates • Surface and volumetric mass loading rates • Other types of loading rates	• Introductory concepts • Hydraulic retention time (HRT). General concept; theoretical HRT; influence of the tank dimensions, internal recirculations, and support medium; tanks operated in batch mode; and actual mean HRT (dead zones and short circuiting) • Volumetric hydraulic loading rate • Surface hydraulic loading rate • Volumetric mass loading rate • Surface mass loading rate • Specific surface mass loading rate • Food-to-microorganism ratio (F/M) • Sludge age	

CHAPTER 14
REACTION KINETICS AND REACTOR HYDRAULICS



Description. Main reaction orders (0, 1, 2) and how to derive them, with emphasis to first-order reactions. The determination of reaction coefficients based on batch experiments is detailed, and the precautions in their utilization for continuous-flow reactors are given. The determination of reaction coefficients at continuous-flow reactors is described, including the characterization of the hydraulics of the reactor (idealized plug-flow, idealized complete-mix, plug-flow with dispersion, and apparent tanks-in-series). Applications for steady-state and dynamic-state conditions are exemplified.



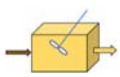
Applicability. *The contents in this chapter are applicable to both treatment plant monitoring and water quality monitoring. As the chapter is structured, most of the applications are for treatment plant reactors. However, we can also consider that water bodies are reactors, and several concepts presented here will also be applicable.*

Level. *Most of the concepts in this chapter are advanced, but there are some basic concepts.*

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Topics	Process knowledge	Data analysis and statistics
• Introductory concepts • Reaction order • Experimental determination of the reaction order and kinetic coefficient in batch reactors • Idealized flow regimens in continuous-flow reactors • Plug-flow with dispersion and apparent tanks-in-series models	• Reaction orders – 0, 1, and 2 • First-order reactions. Structure of a first-order reaction, interpreting the removal coefficient K, analytical integration, and numerical integration • Estimation of the reaction order n and the reaction coefficient K • Specific aspects: influence of a refractory fraction, lag phase, influence of temperature, and time to reach a certain removal efficiency • Applicability of coefficients from batch experiments to continuous-flow reactors • Idealized plug-flow reactor and idealized complete-mix reactor • Deriving coefficients from continuous-flow reactors using idealized hydraulic models • Non-idealized flow regimens: plug-flow with dispersion and apparent tanks-in-series models • Deriving coefficients from continuous-flow reactors using non-idealized hydraulic models • Applicability of kinetic coefficients derived under batch and continuous-flow experiments • Utilization of the kinetic coefficient and hydraulic representation for the mathematical modelling of the reactor (steady-state and dynamic-state conditions)	

CHAPTER 15
MODEL APPLICATION, CALIBRATION, AND VERIFICATION

		Description. Introductory concepts on water quality and treatment plant modelling, and specific coverage on model calibration, assessment of goodness-of-fit, model verification, and residuals analysis.
		***Applicability***. *The contents in this chapter are applicable to both treatment plant monitoring and water quality monitoring.*
		Level. *Most of the concepts in this chapter are advanced, but there are some basic concepts.*

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

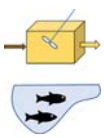
Topics	Process knowledge	Data analysis and statistics
• Concepts involved in water quality and treatment plant modelling • Model calibration • Model verification (analysis of residuals)	• Concept of mathematical models, a procedure for modelling, definition of the model objectives, model conceptualization, selection of the model type, and required properties	• Model calibration. General aspects and minimization of the residuals • Evaluation of the goodness-of-fit of the model (graphical visualization, coefficient of determination, root mean square residual, relative residual, and relation between estimated and observed values) • Sensitivity analysis • Model verification (analysis of residuals). Required properties for the residuals, assessing normality of the distribution, testing zero-mean, checking constancy of the variance, and checking autocorrelation

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Chapter 2

Flow data and the concept of loading

This chapter highlights the importance of having flow data from all lines in a treatment plant. The concept of load is introduced as an important element in the evaluation of the system. Examples demonstrate how flow data are used in practice to support the assessment of treatment plant performance.



The contents in this chapter are mainly applicable to treatment plant monitoring, but the main concepts are also applicable to water quality monitoring (discharge of effluents in water bodies).

CHAPTER CONTENTS

2.1 The Importance of Flow Data and the Concept of Load **22**
2.2 Measuring Flow Rates and Analysing Data **24**
2.3 Using Flow Rates to Assess Performance **36**
2.4 Check-List for Your Report **38**

© 2020 The Authors. This is an Open Access book chapter distributed under the terms of the Creative Commons Attribution Licence (CC BY-NC-ND 4.0), which permits copying and redistribution for non-commercial purposes with no derivatives, provided the original work is properly cited (https://creativecommons.org/licenses/by-nc-nd/4.0/). This does not affect the rights licensed or assigned from any third party in this book. The chapter is from the book Assessment of Treatment Plant Performance and Water Quality Data: A Guide for Students, Researchers and Practitioners, Marcos von Sperling, Matthew E. Verbyla and Sílvia M. A. C. Oliveira (Authors).
doi: 10.2166/9781780409320_0021

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

2.1 THE IMPORTANCE OF FLOW DATA AND THE CONCEPT OF LOAD

Basic

It is very important to collect data about flow rates at treatment plants! Flow data will help you assess treatment plant performance and pollution impact by allowing you to calculate loading rates. Treatment plant staff collect liquid or sludge samples and measure the **concentration** of a pollutant in that sample. But if you also know the flow rate at the location where the sample was collected, then you can calculate the **loading** of that pollutant.

What is the difference between concentration and loading? Figure 2.1 shows that **concentration** is the *amount of a pollutant in a volume of water*, while **loading** is the *amount of a pollutant that passes through a point during a given time duration*.

Formally speaking, we have

$$\boxed{\text{Load} = \text{Flow} \times \text{Concentration}} \tag{2.1}$$

Mass loads have the dimension of **mass per unit time** and are generally calculated as

$$\text{Load}\left(\frac{\text{g}}{\text{d}}\right) = \text{flow}\left(\frac{\text{m}^3}{\text{d}}\right) \times \text{concentration}\left(\frac{\text{g}}{\text{m}^3}\right) \tag{2.2}$$

Note $\text{g/m}^3 = \text{mg/L}$.

If you want to express loads as kg/d, as is usually done, the value calculated in Equation 2.2 should be divided by 1000 g/kg:

$$\boxed{\text{Load}\left(\frac{\text{kg}}{\text{d}}\right) = \frac{\text{flow}\ (\text{m}^3/\text{d}) \times \text{concentration}(\text{g/m}^3)}{1000(\text{g/kg})}} \tag{2.3}$$

Loads can also be expressed as kg/year, kg/h, g/h, g/min, or by any other suitable unit representing mass over time, provided consistency is given to all units in the calculation. Concentrations can also be expressed in other mass units, such as µg/L or ng/L, or even MPN/100 mL (MPN = most probable

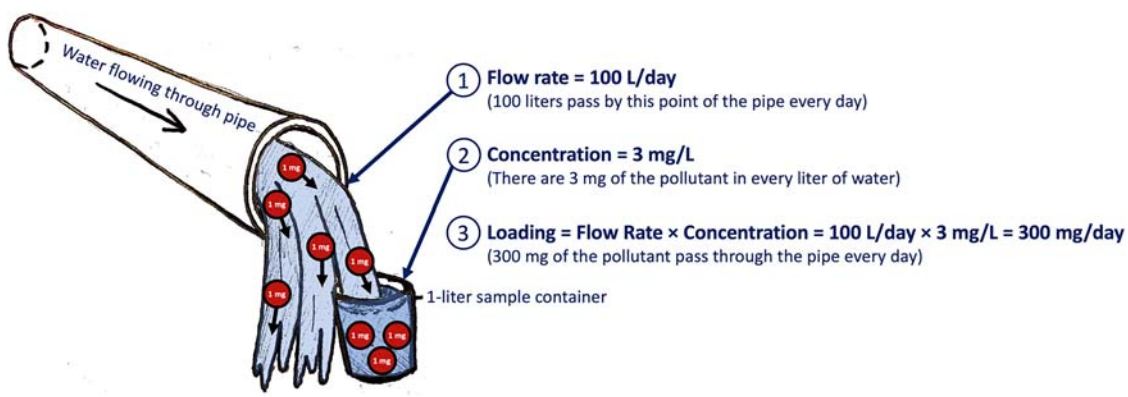


Figure 2.1 The difference between the concentration and the loading of a pollutant. Each circle contains a mass of 1 mg of the constituent.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

number), if we are dealing with microorganisms; eggs/L, if we are studying helminths; and so on. *The concept of load can be applied to the influent and to the effluent of a treatment unit and is essential in the evaluation of its performance*.

A treatment unit can be affected in a somewhat similar way if it receives a *small flow with a high concentration* or a *high flow with a small concentration*, provided the loads are the same. A comparable comment can be made regarding the pollution potential from wastewaters discharged into a river: sewage, with a high flow and low concentration, can have a similar impact of an industrial discharge, with a small flow and a high concentration, in case both of the loads are the same. Of course, there are hydraulic implications, directly associated with flow, but this general concept can be maintained when making an analysis of the behaviour of a treatment unit.

S. 12.3

In a treatment plant with several inputs and outputs in each treatment unit, it should also be understood that each concentration is directly associated with its respective flow. As will be seen in the section on mass balances (Section 12.3), *we can add or subtract flows and loads, but not concentrations*.

In a mass balance (see Section 12.3) of several units in a treatment plant, if the load and flow are known, the concentration can be estimated by simple rearrangement of Equation 2.3:

$$\text{Concentration}\left(\frac{\text{g}}{\text{m}^3}\right) = \frac{\text{load}\left(\text{kg/d}\right) \times 1000\left(\text{g/kg}\right)}{\text{flow}\left(\text{m}^3/\text{d}\right)} \tag{2.4}$$

Example 2.1 shows how to undertake the calculation of a load based on values of flow and concentration. For a more detailed description of mass loadings and some example problems, see Chapter 13 that deals with the loading rates applied to treatment units.

C. 13

Flow rates are also used to determine appropriate dosing rates of chemicals used in treatment processes such as coagulation and flocculation, as shown in Example 2.2.

Flow rate information can also let you know if the treatment system is operating under or over its design capacity.

Example

EXAMPLE 2.1 CALCULATING LOADING FROM A FLOW RATE AND A CONCENTRATION

(a) *Calculate the total load of a certain constituent in the influent to a treatment unit, given that*

- *concentration = 300 mg/L*
- *flow = 50 L/s*

Solution:

Expressing flow in m^3/d

$$Q = \frac{(50\,\text{L/s}) \times (86{,}400\,\text{s/d})}{1000\,\text{L/m}^3} = 4320\,\text{m}^3/\text{d}$$

The load is (Equation 2.3)

$$\text{Load} = \frac{(300\,\text{g/m}^3) \times (4320\,\text{m}^3/\text{d})}{1000\,\text{g/kg}} = 1296\,\text{kg/d}$$

(b) *In the same works, calculate the concentration of another constituent in the influent to a treatment unit, given that the influent load is 35 kg/d.*

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

From Equation 2.4, one has

$$\text{Concentration} = \frac{(35\,\text{kg/d}) \times (1000\,\text{g/kg})}{4320\,\text{m}^3/\text{d}} = 8.1\,\text{g/m}^3 = 8.1\,\text{mg/L}$$

Example

EXAMPLE 2.2 USING FLOW RATES TO DETERMINE DOSING FLOW RATES FOR COAGULANTS

Assume that a water treatment plant has determined that 15 mg/L of ferric chloride and 4 mg/L of polymer are required to optimize the coagulation–flocculation process. Industrial ferric chloride is supplied to the treatment facility in barrels at a concentration of 40% (40 g/100 mL or 400 g/L). Industrial stock polymer, likewise, is supplied at a concentration of 50% (500 g/L). If the flow rate of raw water coming into the system is constant at 300,000 m^3/d, what flow rates should be provided for ferric chloride and polymer?

Solution:

First, convert the units of the required concentrations of coagulants (ferric chloride and polymer) from mg/L to g/m^3 (remember, 1 mg/L = 1 g/m^3). Then, multiply the required coagulant concentrations by the design flow rate to get the loading of coagulant required. Then, divide that loading by the concentration of the coagulant stock to calculate the required flow rate of coagulant that should be dosed into the raw water.

- *Ferric chloride*
  15 g/m^3 × 300,000 m^3/d = 4,500,000 g/d
  (4,500,000 g/d)/(400 g/L) = 11,250 L/d = **7.81 L/min**
- *Polymer*
  4 g/m^3 × 300,000 m^3/d = 1,200,000 g/d
  (1,200,000 g/d)/(500 g/L) = 2400 L/d = **1.67 L/min**

2.2 MEASURING FLOW RATES AND ANALYSING DATA

2.2.1 Methods for measuring flow rates

Basic

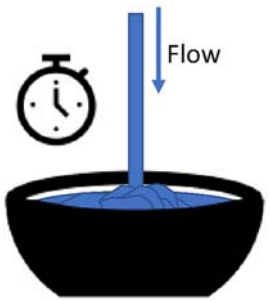
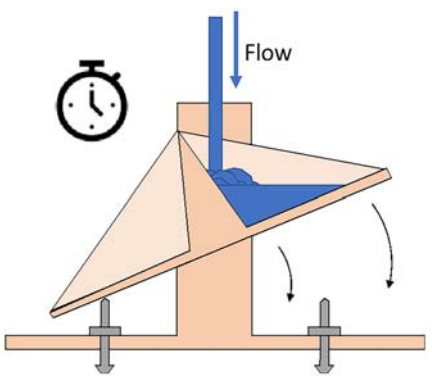
Okay, so it is clearly important to have data about the flow rate at a treatment facility, but how can you measure it? The methods used to measure flow rates depend on whether the water is moving through an **open channel** or a **closed conduit**, and also on the magnitude of the flow rate. An open channel is a structure that contains water on the bottom and on the two sides, with the surface of the water flowing free. Examples of open channel flow are when water moves through a concrete swale. A large drainage pipe, such as the kind used for sanitary or storm sewer networks, is also considered as open channel flow as long as the pipe is not flowing full or under pressure. Even the flow through rivers and streams can be approximated by open channel hydraulics. On the contrary, in a closed conduit, the water is completely contained. An example of a closed conduit is a pressurized pipe, such as the type used for potable water distribution systems.

To measure flow rates in open channels, you can use structures such as weirs and flumes. In closed conduits, flow rates can be measured using devices such as orifice plates, Venturi meters, magnetic and ultrasonic flow meters, or turbine and propeller flow meters. However, if you have very small flow rates, you may use simple procedures such as volumetric measurements and tipping buckets. Tables 2.1–2.3

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

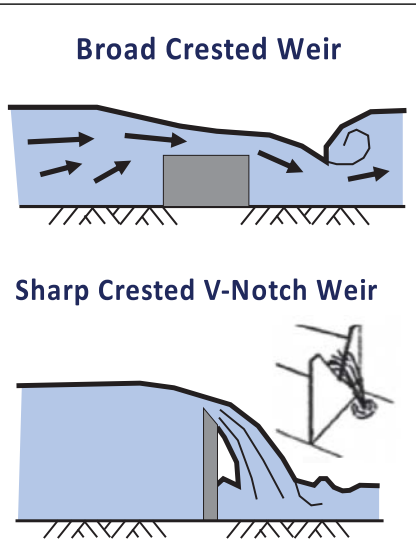
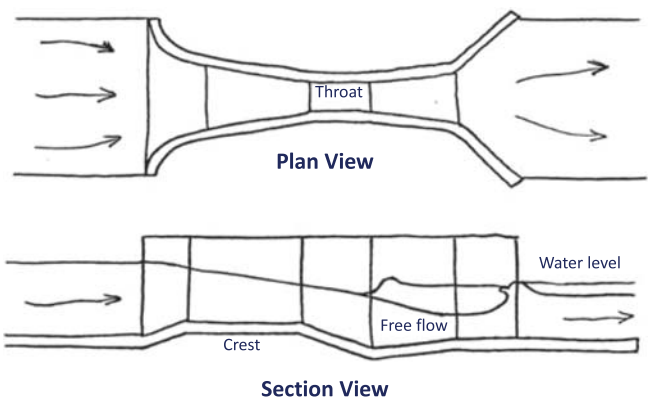
Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Table 2.1 Procedures commonly used to measure flow rates of water and wastewater for systems with small flows.

Structure	Image	Description
Bucket and stopwatch		This is the simplest method for measuring flow rates. It requires two people: one to fill a bucket of a known volume with flow from the system and the other to record the amount of time using a stopwatch. The flow rate is calculated by dividing the volume of the bucket by the amount of time required to fill it up. You need to repeat this measurement several times. *Advantages* • Simple operation • Low cost *Disadvantages* • Less accurate than other methods • Requires more labour, not automated • Requires a clearance for flow to fall into the bucket • Only works with small flows • Subject to user error (starting and stopping exactly at the right moment)
Tipping bucket flow gauge		The tipping bucket has two compartments, one on each side, and when the first compartment fills up, the device tips over, dumping out the water and rotating to let the other side start filling up. When that side fills up, the device tips over again, dumping out the water and rotating to let the first side start to fill up again. Each time the device tips, a count is registered by an electromagnetic sensor, allowing for the calculation of the volume flowing per unit time throughout the day. *Advantages* • Simple operation • Low cost • May be more accurate than flow measurement in channels or weirs in the case of very small flows *Disadvantages* • Requires a clearance for flow to fall into the device • Only works with small flows

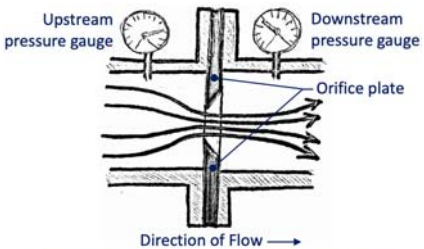
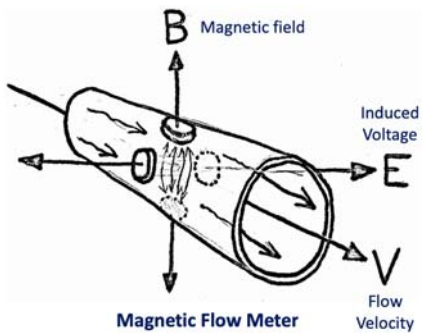
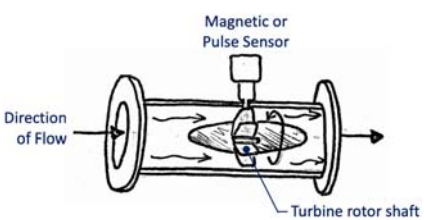
Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Table 2.2 Structures commonly used to measure flow rates of water and wastewater in open channels.

Structure	Image	Description
Weir		The cross-sectional area of flow above the crest of the weir is proportional to the flow rate, so the depth of water can be correlated with flow. Broad-crested weirs are typically used to measure very large flows, like in channelized rivers. *Advantages* • Simple operation • Easy construction • Appropriate for wastewater systems • Depth can be measured automatically with gauges, floats, or ultrasonic meters • V-notch weirs are less influenced by small differences in their level compared with continuous weirs (e.g., effluent weirs in sedimentation tanks) *Disadvantages* • Must calculate the structure's unique weir coefficient to correlate depth with flow rate • V-notch weirs may accumulate solids when receiving wastewater, what can affect the measurements
Parshall flume		Consists of a reduction in the channel width, and a drop in the bottom slope of the channel causes a correlation between the water depth and the flow rate. *Advantages* • Simple operation • Appropriate for wastewater systems, since the liquid is always flowing • Depth can be measured automatically with gauges, floats, or ultrasonic meters *Disadvantages* • Construction is more difficult • Must be precision-fabricated

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Table 2.3 Devices commonly used to measure flow rates of water and wastewater in closed conduits.

Device	Image	Description
Orifice plate and Venturi meters	 Section View of a Pipe with an Orifice Plate	Orifice plate has an opening that is narrower than the pipe diameter, producing a pressure drop that can be used to estimate the flow rate. The Venturi meter operates with the same principles but the convergence from larger to smaller diameter is less drastic, which reduces friction loss. *Advantages* • Simple and inexpensive *Disadvantages* • Medium to high friction losses (higher friction losses for orifice plates)
Magnetic and ultrasonic meters	 Magnetic Flow Meter	For magnetic meters, a voltage proportional to the flow rate is produced as the liquid moves through a magnetic field. For ultrasonic meters, the frequency of sound waves reflected by gas bubbles and dissolved solids is converted by a piezoelectric transducer into a flow velocity. *Advantages* • Lower friction loss (compared to orifice plates and Venturi meters) *Disadvantages* • Contaminants can coat electrodes, limiting suitability for wastewater • Doppler-type ultrasonic meters may work with wastewater, but transit-time meters are only applicable for measurements of flow rates in clean water sources • More expensive than other types of meters
Turbine and propeller meters	 Turbine or Propeller Flow Meter	The rotation frequency and voltage produced by spinning propeller blades or other rotating elements as water passes through the meter is proportional to the flow rate. *Advantages* • Inexpensive *Disadvantages* • Very high friction losses • Only appropriate for clean water, as particulates can cause bearings to fail

show the differences between these different flow measurement structures and devices and the typical applications for water, wastewater, and stormwater treatment systems.

2.2.2 Recording flow data

Basic

Flow rates must be recorded on a periodic basis to statistically analyse treatment plant performance. Flow rates can be recorded manually (e.g., by observing the depth of flow passing through a weir or a flume and by recording the amount of time required to fill a receptacle of a known volume). However, at larger facilities, it is often advantageous to record flow rates using online continuous measurement devices such as a data logger.

The use of flow rate data from treatment plant side streams is equally as important for the assessment of performance as the use of flow rate data from the treatment plant influent and effluent points. Some example side streams for which flow rate data are used to evaluate performance are filter backwashing lines, excess sludge wasting lines, sludge recycle lines, and return flows.

2.2.3 Flow variations

Basic

Flow rates at a treatment plant may vary considerably throughout the course of the year (for water and wastewater treatment plants) and even throughout the course of a single day (for wastewater treatment plants).

Peak flow rates at wastewater facilities are normally associated with the rainy season in combined sewerage systems, while peak demands at water treatment facilities normally occur during the summer season or holiday periods. The flow rate in stormwater collection systems varies on account of the rainfall intensity and duration. Knowing flow rates is extremely important for *sizing and designing treatment facilities*, and there are already excellent text references that cover the use of flow rate data to design water, wastewater, and stormwater management and treatment facilities (e.g., Hammer & Hammer, 2012). This chapter will focus on the use of flow rate data to **assess the performance** of treatment facilities.

Use the following statistics to understand and describe the variation of flow rates throughout the day and throughout the year:

- Use the **average seasonal flow** to compare pollutant loadings and treatment plant performance between different seasons
- Use the **average daily flow** to calculate daily loading rates and mass balances
- Use the **average hourly flow** to determine peaking factors and their impact on hydraulic retention time (HRT)

Daily and annual **hydrographs** can also be used to visualize the variation in flow rates with respect to the time of day or the time of year. Flow rate **peaking factors** are commonly used for design purposes, but the peak daily flow can also be used to assess treatment plant performance, for example, to predict the effect of an equalization basin on influent pollutant concentrations to a treatment process. Peaking factors can be calculated using the 95^{th} or 99^{th} percentile associated with the normal score of the plotting position (see Example 2.4 in Section 2.2.5).

S. 2.2.5

2.2.4 Flow equalization

Advanced

Flow equalization tanks or basins are frequently used in treatment systems to mitigate the effect of varying flow rates and make it easier to design and operate treatment unit processes. Engineered treatment systems simply work better and are easier to operate when the flow rate of the liquid being treated is stable. Another benefit of flow equalization is that the concentrations of pollutants in the water also become more stable.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

The larger the relative volume of the equalization tank or basin, the more stable the concentration of pollutants will be throughout the course of the day. Thus, when assessing treatment plant performance, it is often useful to be able to predict the *impact of flow equalization on the concentration of pollutants* (Example 2.3) (Metcalf & Eddy, 2003).

C. 12

Please note that in this example, we anticipate some concepts that will be detailed in Chapter 12, relative to water and mass balances.

Example

EXAMPLE 2.3 THE EFFECT OF FLOW EQUALIZATION ON POLLUTANT CONCENTRATIONS

Use the flow data in the Excel spreadsheet associated with this example. Calculate the effect of a 50,000 m^3 equalization basin on the following biochemical oxygen demand (BOD) concentrations:

Time Period During the Day	Average BOD Concentration Coming into the Basin (mg/L)	
	Dry Season	Rainy Season
0:00 – 1:00	146	110
1:00 – 2:00	126	132
2:00 – 3:00	101	109
3:00 – 4:00	42	110
4:00 – 5:00	50	100
5:00 – 6:00	56	95
6:00 – 7:00	101	85
7:00 – 8:00	132	116
8:00 – 9:00	171	180
9:00 – 10:00	200	195
10:00 – 11:00	227	233
11:00 – 12:00	235	220
12:00 – 13:00	244	215
13:00 – 14:00	225	225
14:00 – 15:00	201	188
15:00 – 16:00	160	150
16:00 – 17:00	150	153
17:00 – 18:00	144	178
18:00 – 19:00	177	195
19:00 – 20:00	209	200
20:00 – 21:00	288	255
21:00 – 22:00	314	240
22:00 – 23:00	252	186
23:00 – 0:00	180	141

Excel

Note: This example is available as an Excel spreadsheet.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

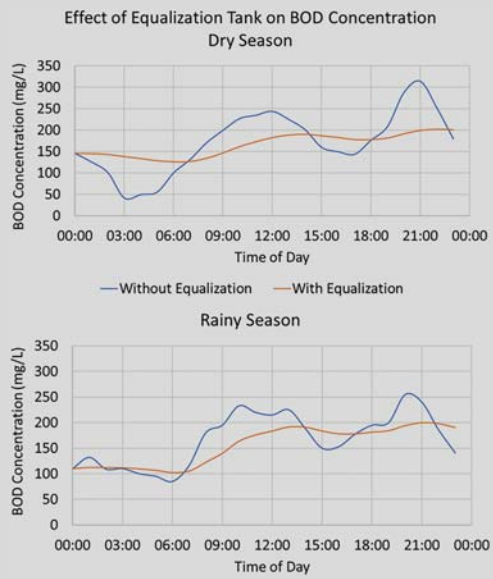
Solution:

First, calculate average hourly flow rates and use those to determine the average volume of flow entering the equalization basin each hour. The overall volume of water in the basin at any given time is then computed by subtracting the average daily flow rate from the fluctuating hourly flow rate. Finally, the BOD concentration leaving the basin (assuming the basin is well mixed) is calculated as follows, where BOD_{in} and BOD_{basin} are the influent and effluent concentrations of BOD to the equalization basin, V_{in} is the volume entering the basin within an hour, and V_{basin} is the volume of water stored in the basin at time t or $t-1$:

$$BOD_{basin,t} = \frac{BOD_{in,t} V_{in,t} + BOD_{basin,t-1} V_{basin,t-1}}{V_{basin,t-1} + V_{in,t}}$$

Because the data set is very large, we will not show the calculations here, and you should consult the Excel spreadsheet.

The results, shown in the plots below, demonstrate the smoothing effect of flow equalization on BOD concentrations. The minimum and maximum concentrations (dry season) without equalization are 42 and 314 mg/L; with equalization, the minimum and maximum concentrations are 126 and 202 mg/L.



2.2.5 Determining typical flow rates and distributions

Advanced

We anticipate here concepts that will be further detailed in other chapters of this book, but we present them so that you get the feeling of dealing with flow rate distributions. If you feel that not all concepts are entirely clear, refer to the sections we mention below for their detailed coverage, and then come back to this section.

C. 8

Flow rates are usually distributed *normally* or *log-normally* (see Chapter 8 for more information about normal and log-normal distributions). In order to determine which distribution your flow data follow, you should rank the measured flow rates from lowest to highest and then plot measured flow rates with respect to the normal score of the plotting position. This procedure is detailed in Section 9.6, which deals with frequency analysis using normal and log-normal distributions.

S. 9.6

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

S. 9.5

The **plotting position (PP)** is determined using Equation 2.5, where R is the rank of the data point and n is the total number of data points (this concept is further detailed in Section 9.5).

$$\mathrm{PP} = \frac{R}{n+1} \tag{2.5}$$

The normal score is calculated in Excel using the command **NORM.S.INV()** and then referring to the PP value. If the points connect to form a straight line, then the distribution may be considered to be normal. If the points form a curved line, then the distribution may be log-normal, but you need to verify by plotting the points on a log scale or calculating the log of the values and then plotting them on a normal scale. If log-transformed points form a straight line, then the flow data may be considered to be log-normally distributed. In Chapter 8, we will present in a more formal way the procedures for assessing the adherence of your data to a normal distribution and a log-normal distribution.

C. 8

Example

EXAMPLE 2.4 DETERMINING THE DISTRIBUTION OF FLOW DATA

Use the data shown in the spreadsheet associated with this example. Determine the distribution of the flow rate data collected daily over one year during wet and dry weather. Use the data to determine the typical (mean) flow rates during each season, as well as the peaking factor associated with the 95th percentile flow rates.

Excel

Note: This example is available as an Excel spreadsheet.

Solution:

Because the data set is very large, we will not show all the calculations here, and you should consult the Excel spreadsheet.

First, rank the values from 1 to 365. Then, use the rank to calculate the plotting position (Equation 2.5). The following tables show the first few rows of data and then the few rows of ranked data for each season with the calculated PPs and normal scores.

We have $n = 184$ data for the wet season and $n = 181$ data for the dry season.

=1/(184+1)

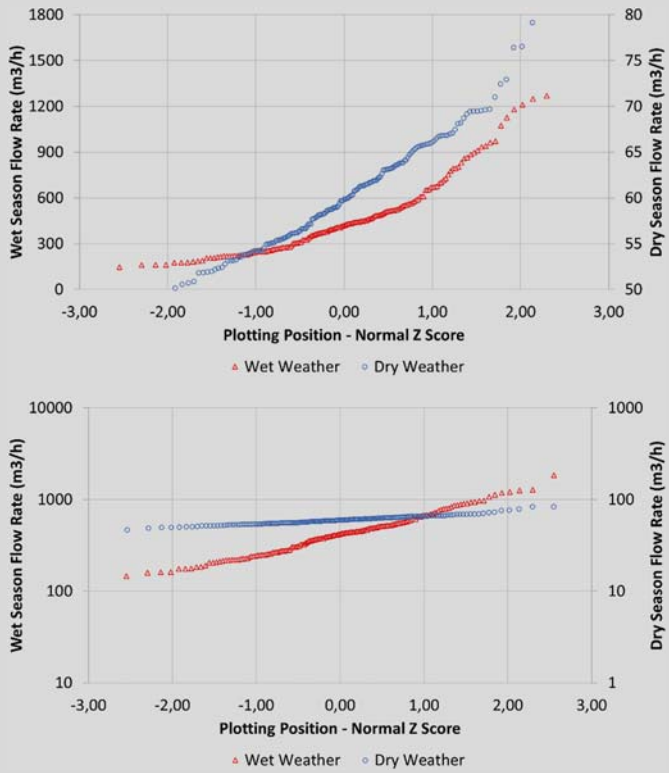
=1/(181+1)

Date	Season	Flow rate (m^3/h)	Wet Season: Flow rate (m^3/h)	Rank	Plotting Position PP	Normal Score (Z)	Dry Season: Flow rate (m^3/h)	Rank	Plotting Position PP	Normal Score (Z)
1/1/18	Wet	609	145	1	0.5%	-2.55	47	1	0.5%	-2.54
1/2/18	Wet	241	160	2	1.1%	-2.30	49	2	1.1%	-2.29
1/3/18	Wet	301	160	3	1.6%	-2.14	50	3	1.6%	-2.13
1/4/18	Wet	669	162	4	2.2%	-2.02	50	4	2.2%	-2.01
1/5/18	Wet	162	175	5	2.7%	-1.93	50	5	2.7%	-1.92
1/6/18	Wet	910	175	6	3.2%	-1.85	51	6	3.3%	-1.84
1/7/18	Wet	258	177	7	3.8%	-1.78	51	7	3.8%	-1.77
…	…	…	…	…	…	…	…	…	…	…

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Below is a plot of the measured flow rates versus the plotting position, first on an arithmetic scale and then on a logarithmic scale. The trend is curved on the arithmetic scale (top panel) and linear on the logarithmic scale (bottom panel), which indicates that the data are closer to a log-normal distribution. Therefore, the geometric mean is a better representation of the typical flow rates for each season (see Section 5.6.4 for the concept of geometric means).

S. 5.6.4



Plots of the measured flow rates with respect to the normal *Z* score associated with their plotting position on an arithmetic scale (above) and on a logarithmic scale (below). The shapes of the curves indicate that the data are closer to a log-normal distribution.

The typical flow rates are calculated using the geometric mean, since the data are log-normally distributed.

- Geometric mean wet weather flow rate: **410 m^3/h**
- Geometric mean dry weather flow rate: **60 m^3/h**

The peaking factors associated with the 95th percentile are determined using the plotting positions. To get the 95th percentile peaking factors, divide the flow rate associated with the plotting position of 0.95 by the geometric mean flow rate for each season.

- Wet weather 95th percentile flow rate: **939 m^3/h**
- Wet weather peaking factor = 95th percentile/geometric mean = 939/410 = **2.29**
- Dry weather 95th percentile flow rate: **70 m^3/h**
- Dry weather peaking factor = 95th percentile/geometric mean = 70/60 = **1.17**

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

2.2.6 Analysing flow data

Advanced

Flow data are also used to calculate **peaking factors** to anticipate future peak flows during rain events, for example. The calculation of peak flow rates can be useful when assessing treatment plant performance when considering 'worst case scenario' situations. See Example 2.5 (Metcalf & Eddy, 2003).

Example

EXAMPLE 2.5 ANALYSING TRENDS IN THE HOURLY FLOW RATES

Use the data shown in the spreadsheet associated with this example. The spreadsheet contains example flow rate measurements collected at the influent of a wastewater treatment facility during seven random days in the dry season and seven random days in the wet season.

(a) *Calculate the mean, minimum, and maximum daily flow rates and the mean, minimum, and maximum hourly flow rates.*
(b) *Plot daily hydrographs showing wet and dry season conditions using the mean hourly flow rate data from these seven random days.*
(c) *Calculate a flow rate peaking factor for wet conditions (compared to dry conditions) using the upper 99% prediction interval for data from the rainy season (assumed equal to the mean value plus three times the standard deviation).*

Excel

Note: This example is available as an Excel spreadsheet.

Solution:

Because the data set is very large, we will not show all the calculations here, and you should consult the Excel spreadsheet.

(a) Mean, minimum, and maximum values

The mean (min, max) **daily** flow rates are 87 (27, 157) and 170 (21, 500) m^3/h for dry and wet seasons, respectively.

The mean, minimum, and maximum **hourly** flow rates are shown in the following table.

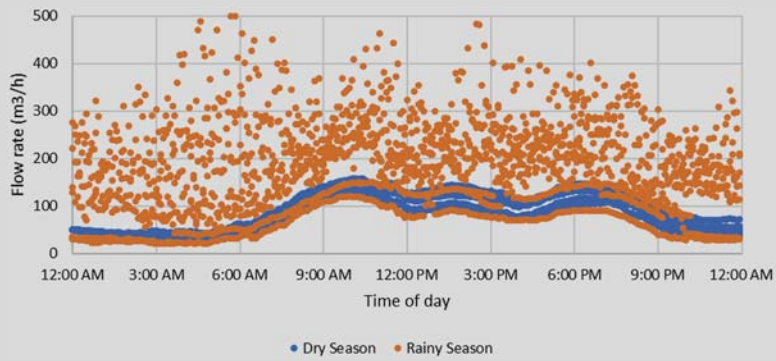
Time of Day	Hourly Flow Rates (m^3/h)					
	Dry Season			Rainy Season		
	Mean	Minimum	Maximum	Mean	Minimum	Maximum
0:00	47	39	51	117	23	322
1:00	39	39	47	105	25	311
2:00	37	39	48	112	22	351
3:00	37	39	48	130	21	418
4:00	38	39	47	150	21	489
5:00	44	39	63	158	25	500
6:00	57	39	80	160	32	463
7:00	83	39	111	174	48	451
8:00	114	39	141	186	79	369
9:00	137	39	156	210	109	369
10:00	140	39	157	233	108	432
11:00	126	39	148	207	76	464
12:00	113	39	136	189	76	355
13:00	123	39	145	195	84	383
14:00	116	39	142	198	77	483

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

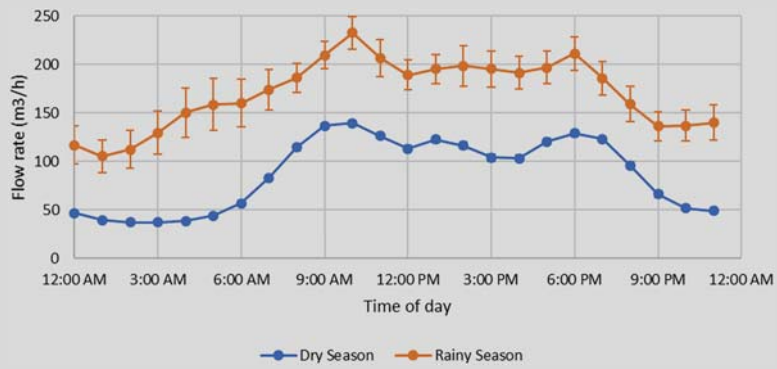
Time of Day	Hourly Flow Rates (m^3/h)					
	Dry Season			Rainy Season		
	Mean	Minimum	Maximum	Mean	Minimum	Maximum
15:00	104	39	132	195	70	401
16:00	103	39	122	191	70	409
17:00	120	39	145	197	74	386
18:00	129	39	147	211	90	402
19:00	123	39	144	186	74	357
20:00	96	39	126	159	45	374
21:00	66	39	103	136	30	304
22:00	52	39	79	137	28	264
23:00	49	39	74	140	28	343

(b) Daily hydrographs

The three figures below show (top) all flow rate data, (medium) mean hourly flow rates, and (bottom) the upper end of the 99% prediction interval.

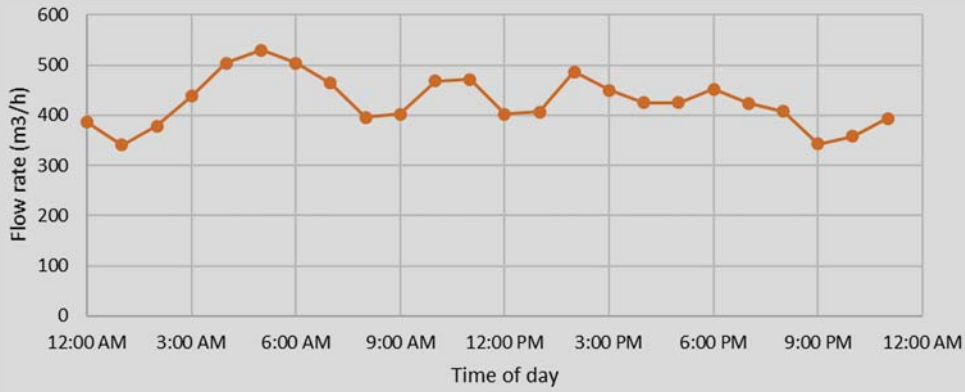


Flow rate data with respect to time of day for the dry and rainy seasons.



Mean hourly flow rates for the dry and rainy seasons, with error bars corresponding to the 95% confidence intervals.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest



Upper limit of the 99% prediction interval for the hourly flow rates during the rainy season.

(c) Flow rate peaking factor

The flow rate peaking factor can be calculated using the 99% prediction interval, which can be estimated, in a practical way, as the mean flow rate plus three standard deviations. This upper prediction interval value can be divided by the estimated mean hourly flow rate during the dry season to yield an estimated peaking factor. The estimated hourly peaking factors are shown in the following table. Note that for the design of wastewater treatment facilities, the use of peaking factors greater than 4:1 is not always cost-effective.

Hourly peaking factors calculated by dividing the upper 99% prediction interval for rainy season hourly flow rates by the mean dry season hourly flow rates are shown in the following table:

Time of Day	Flow Rate Peaking Factor	Time of Day	Flow Rate Peaking Factor
0:00	8.3	13:00	3.3
1:00	8.6	14:00	4.2
2:00	10.2	15:00	4.3
3:00	11.7	16:00	4.1
4:00	13.1	17:00	3.5
5:00	12.0	18:00	3.5
6:00	8.9	19:00	3.4
7:00	5.6	20:00	4.3
8:00	3.5	21:00	5.2
9:00	2.9	22:00	6.9
10:00	3.4	23:00	8.0
11:00	3.7	13:00	3.3
12:00	3.6	14:00	4.2

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

2.3 USING FLOW RATES TO ASSESS PERFORMANCE

2.3.1 Hydraulic retention time

Time is a very important factor for many treatment processes. The flow rate is related to the **mean theoretical HRT** (HRT) of a treatment process such as a reactor. The mean theoretical HRT is the amount of time that water stays within the reactor, before being discharged in the effluent. The theoretical mean HRT of a unit process is calculated as the reactor volume (V) divided by the average daily flow rate (Q).

$$\text{HRT} = \frac{V}{Q} \tag{2.6}$$

Thus, flow rate data are used to calculate daily and seasonal variations in the theoretical mean HRT of a treatment unit process. This can give you some insight regarding why the performance of a system may fluctuate throughout the year. Example 2.6 shows an example of monthly mean HRTs calculated for a wastewater treatment facility that utilizes waste stabilization ponds.

We present here only introductory concepts related to this highly important process variable. In reality, due to mixing, the true retention time in a reactor is a distribution, rather than a single value. Some water molecules move more quickly through the reactor, while others may stay around for longer before leaving in the effluent. The distribution of HRT can be estimated using data from a tracer study. It is important to note that the **actual mean HRT** (calculated using data from a tracer study) is **often different from the theoretical mean HRT** (e.g., V/Q). See Chapter 13 for more details on this regard. In Section 13.2, we cover the concept of HRT in a thorough way, including the factors that may lead to the actual mean HRT being different from the theoretical one, calculated by Equation 2.6.

C. 13

Example

EXAMPLE 2.6 USING FLOW RATES TO CALCULATE MEAN HYDRAULIC RETENTION TIME

A waste stabilization pond system has an overall volume of 15,000 m^3 and a flow rate that varies throughout the year between 280 and 659 m^3/d. Use the flow rate data in the associated spreadsheet to calculate the mean theoretical hydraulic retention time (HRT) and plot that with respect to the per cent BOD removal. Determine if the trend is for BOD removal to increase or decrease with respect to increasing hydraulic retention times.

Excel

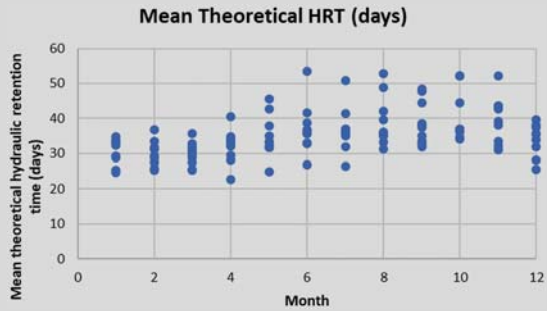
Note: This example is available as an Excel spreadsheet.

Solution:

The data are monthly averages and span a total period of 10 days. Because of this, they will not be shown here, and you should consult the Excel spreadsheet for further details.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

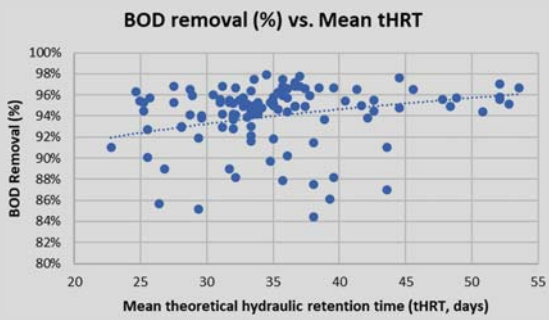
Using the flow rate data provided, the HRT ranged from 22.8 to 53.6 days, with lower retention times corresponding with the months of December through April (see figure).



Mean theoretical hydraulic retention time for a waste stabilization pond system with respect to month of the year.

When these retention times are plotted against the per cent BOD removal, there are some indications that higher retention times may correlate with higher BOD removal values, which would be expected. The more time wastewater stays inside the ponds, the more BOD degradation should occur. However, you can also see that the data points show a wide scatter, and therefore, it is difficult to conclude whether there is a significant correlation between HRT and BOD removal efficiency. This is a very important point, and it will be discussed in detail in Chapter 11 that deals with correlation and regression analysis.

C. 11



Per cent BOD removal versus mean theoretical hydraulic retention time for a waste stabilization pond system.

2.3.2 Water losses and gains

The **loss of water** due to *evaporation*, *evapotranspiration*, or *infiltration* and the **gain of water** due to *precipitation* (e.g., rain) can affect the flow rates coming into and going out of certain treatment facilities with long hydraulic retention times. Water and wastewater pass through many treatment facilities within a few hours; however, some facilities have retention times on the order of days or weeks. Similarly,

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

water storage reservoirs may have retention times on the order of months. Underground water aquifers may have retention times on the order of years or even decades.

Water levels in surface and groundwater reservoirs will often fluctuate throughout the year in a seasonal pattern, storing more water during the winter when the demand is low, and drawing down the additional storage during the summer months when the demand is high. *In cases where hydraulic retention times are measured on the order of days, weeks, or months, it may be necessary to account for water losses and gains in order to accurately assess the concentrations of pollutants going into or coming out of the facility*.

A mass balance approach can be used to balance the water in a treatment unit. To start, define the boundary of the system. Then, record flow rate measurements at all influent and effluent points of the system. A comparison of the recorded flow rates entering the system and the recorded flow rates exiting or withdrawn from the system over a long period of time will allow you to estimate net gains or losses of water due to evaporation or rainfall.

Influent flow rates are commonly used for design purposes; however, for performance assessment, the average influent and effluent flow rates should be used if available.

S. 12.2

The subject of water balance is very important in treatment plant assessment and is covered in detail in Section 12.2.

2.4 CHECK-LIST FOR YOUR REPORT

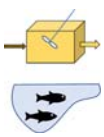
- ✓ Check that the flow rates have been measured using appropriate devices depending on whether the flow is through an open channel or a closed conduit.
- ✓ Flow rate data are collected either manually or using a data logger; verify whether it is important that raw flow rate data are included in the appendix of the report.
- ✓ Verify whether the distribution of flow rate data has been assessed.
- ✓ Typical seasonal flow rates, daily flow rates, and hourly flow rates are calculated using the arithmetic or geometric mean as necessary based on the assessment of the flow rate distribution.
- ✓ Hourly peaking factors are reported.
- ✓ Mean theoretical hydraulic retention times are calculated using the flow rates and the reactor volume.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Chapter 3

Planning your monitoring programme. Sampling and measurements

This chapter addresses how to design research studies and establish monitoring programmes, with an emphasis on quality assurance, quality control, and the collection of representative samples.



The contents in this chapter are applicable to both treatment plant monitoring and water quality monitoring.

CHAPTER CONTENTS

3.1 Types of Monitoring Programmes and Studies **40**
3.2 Quality Assurance and Quality Control **41**
3.3 Sample Collection **50**
3.4 Sample Size, Containers, and Holding Times **62**
3.5 Statistical Power and Number of Samples **63**
3.6 Check-List for Your Report **67**

© 2020 The Authors. This is an Open Access book chapter distributed under the terms of the Creative Commons Attribution Licence (CC BY-NC-ND 4.0), which permits copying and redistribution for non-commercial purposes with no derivatives, provided the original work is properly cited (https://creativecommons.org/licenses/by-nc-nd/4.0/). This does not affect the rights licensed or assigned from any third party in this book. The chapter is from the book Assessment of Treatment Plant Performance and Water Quality Data: A Guide for Students, Researchers and Practitioners, Marcos von Sperling, Matthew E. Verbyla and Sílvia M. A. C. Oliveira (Authors).

doi: 10.2166/9781780409320_0039

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

3.1 TYPES OF MONITORING PROGRAMMES AND STUDIES

Whether you are in charge of monitoring the performance of a treatment plant, monitoring an outfall for compliance with regulations, or completing a special study such as a thesis project, there are several important considerations that must be taken regarding:

- What is a sample?
- Where should I collect samples?
- When should I collect samples?
- How should I collect samples?
- How many samples should I collect?
- What measurements should I take in the field and in the laboratory?

A robust **operational monitoring** programme is essential for any water or wastewater treatment plant to evaluate the efficacy of the treatment system or a water body to assess the quality of its water. Monitoring programmes at some facilities (especially large facilities serving urban centres) or for some research projects or special studies might include continuous and real-time measurements by probes, sensors, and/or data loggers, or remote-operated controls to make operational changes to the system based on incoming data or system alarms. However, at a minimum, monitoring programmes and studies should include the following elements:

- Visually inspect different components of the treatment system periodically
- Measure flow rates in the system
- Collect and analyse liquid and/or solid samples for the concentrations of relevant contaminants
- Implement quality assurance and quality control measures and document them in a quality assurance project plan (QAPP) report

Compliance monitoring refers to monitoring activities that are intended to ensure compliance with laws and regulations, such as the maximum contaminant levels (MCLs) for drinking water or effluent discharge limits on the concentrations of certain pollutants for wastewater treatment facilities. These laws and regulations are often established by governmental environmental agencies and public health authorities, operating either at the national level and/or the regional (state, department, and province) level. Many of the specifics of compliance monitoring programmes are often driven by local laws and regulations. The contaminants or pollutants of interest for the study are often specified by the legislation. They may vary from site to site, depending on the water body in question, its beneficial use category, the characteristics of its watershed (land use and industrial activity), or the results obtained in previous monitoring efforts. You should conform with any specific details specified in the legislation, such as monitoring frequencies, detection limits for reporting, sample location, and sample type.

For **research projects** or **special studies**, the types of monitoring activities and elements must be chosen by the researcher or the director of the study. It is typically up to the researcher to decide which method to use for analysing flow rates (e.g., manual measurements or automated measurements) and water quality (e.g., collecting and analysing samples in the laboratory versus the use of probes and sensors with data loggers). If you are engaged in a study like this, often you must balance the precision and accuracy of the different methods with the cost of acquiring the equipment, materials, and supplies needed to use a particular method.

Emergency studies are often triggered by specific environmental accidents, public health emergencies, or hazardous weather events, such as chemical spills, disease outbreaks, algal blooms, hurricanes, wildfires, or other natural disasters. The parameters to be studied are typically associated with the nature and type of

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

disaster or accident, and the duration of the study is typically short and intensive, in order to obtain answers as quickly as possible.

3.2 QUALITY ASSURANCE AND QUALITY CONTROL

3.2.1 Introductory concepts

Basic

Quality assurance (QA) and **quality control (QC)** are essential in order for people to have trust in the results of your study. Furthermore, starting with a good quality assurance plan can help you ensure that you are collecting the right samples, processing and analysing them using appropriate methods, and getting enough results to make appropriate findings based on the project's goals. We recommend that you take the following five-step process, which is inspired by the California State Water Resources Control Board's Surface Water Ambient Monitoring Program Quality Assurance Plan, to document your monitoring assignment, project, or research study. You should summarize the following items in a quality assurance project plan (QAPP), which can be part of your overall report or a separate report from your main study report:

- Scope of the study
- Samples and populations
- Measurements and anticipated use of data
- Standard assessment thresholds and operating procedures
- Quality control samples
- Data management and analysis

3.2.2 Scope of the study

Basic

First, you must determine what you plan to study or monitor. Start by asking yourself the following three questions that define what will be addressed, as well as where and when the study or monitoring programme will happen.

- What question do you hope your study will answer?
- What are the boundaries and limits of the study in terms of its location?
- What is the general length and time frame of the study?

Let us discuss each of the questions individually.

What question do you hope your study will answer?

You should develop a guiding question (or a set of questions) for your study or programme. Your question (s) should be *specific* and *measurable*. The study or programme you are proposing should also be **FIRE** – that is, it should be **F**easible, **I**nteresting, **R**elevant, and **E**thical; and, if you are a thesis student or a research scientist, the research question should also be novel and have intellectual merit (Farrugia *et al.*, 2010).

A **feasible** study is one that (a) includes a sufficient number of samples, (b) utilizes methods that are standardized, recognized, or rigorously tested, and (c) can be completed for a cost that fits within the project or programme's budget. An **interesting** study is one that will be read and/or referenced by others. A **relevant** study is one that has direct and practical application to practice or policy. An **ethical** study is one that protects the rights, welfare, and well-being of participants or beneficiaries, ensures compliance with local, national, and international laws and regulations, and adheres to the principles

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

outlined in the Belmont report, specifically: *respect* for persons (treating people as autonomous agents and protecting individuals with diminished autonomy), *beneficence* (securing the well-being of people, doing no harm or maximizing possible benefits while minimizing possible harms), and *justice* (selecting participants equitably in terms of who receives the benefits of research studies and who bears their burden). In the case of water and wastewater treatment plants, ethics could pertain to the treatment plants selected for study and their beneficiary populations, as well as the people or organizations responsible for operating such facilities.

After stating your question(s), *write down a brief background or context of the problem(s) being addressed*. Even if there are currently no problems, write down a summary of the problem(s) that you are trying to avoid by executing the study. During the planning stage of a monitoring and sampling programme, it is often helpful to determine if there are synergistic or overlapping monitoring and evaluation efforts or other studies that have been previously completed or are currently in progress to avoid duplication of efforts. When planning a research project, for example, this can be done by conducting a literature review on the topic.

What are the boundaries and limits of the study in terms of its location?

Describe the water body(ies) or treatment plant(s) that will be the focus of the study. Provide a brief description of the study location(s) with a map, if available. Show a schematic of the water body or the treatment plant, with a visual indication of the location of all samples collected and analysed. For more information about where to collect samples, refer to Section 3.3.

S. 3.3

You should also take note of any **obstacles** that may interfere with collecting samples or obtaining a complete data set, for example, is the site bounded by fences, is access limited to daytime hours, are there safety concerns with going to the sampling site, is there a potential for dangerous weather conditions, is a permit required for accessing the sampling site, etc.

In terms of **ethical** considerations, think about who is potentially benefitting or putting themselves at risk as a result of the study being carried out at the chosen location. For example, if you are conducting a research study that documents the performance of a wastewater treatment plant at removing pathogens, what if the findings indicate that pathogens are not very effectively removed from the treatment plant. If these findings are made public and linked to the facility, will it put the manager or operator of the facility at risk of losing his or her job? Will there be a potential for public fear or outrage due to the findings? Will those findings benefit or harm the public in the long run? Are there certain populations who will gain economic or health benefits as a result of the knowledge being produced, and if so, are these populations the frequent recipients of such benefits (e.g., due to the treatment plant being located close to the university), and are there other more remote communities that will fail to benefit from the data being produced by the research? These are important considerations when choosing the location for a study.

What is the general length and time frame of the study?

Here, you should indicate if the project is intended to be **short term** or **ongoing**. If it is short term, indicate the day(s), month(s), and/or year(s) during which the study will take place. *The length and time of a study may be determined based on the research question* (sufficient to obtain a large enough sample size to answer the question or address the hypothesis), *the legislative authorities* (which may specify the length, frequency, or nature of sampling and measurements), and *budgetary limitations*. More samples, more information, and more data points will always be desirable and helpful to answer a particular research question, but this comes at a cost, and it is the researcher's job to determine cost-benefit trade-offs and establish a length and time frame that are appropriate for making cost-effective decisions or findings.

Planning an appropriate time frame for a study is an important consideration, especially if the research question addresses variations with **seasonality** or **time of day**. For example, if temperatures during winter

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

seasons cause lower efficiencies in terms of treatment plant performance, or the pollutant levels in a water body are of greatest concern during the rainy season, then the timeframe for the study should take place during the season of concern. For more information about the temporal aspects of sampling and sample collection, see Section 3.3.

S. 3.3

Table 3.1 shows three hypothetical sampling and monitoring programmes and their respective scopes of work, including a study/research question, a description for the general timeframe and length of the study, and whether or not the study is connected with another project.

Basic

3.2.3 Environmental samples, statistical samples, and populations

Your study will produce data and information that result from samples that are collected and analysed and measurements that are made in the field. Our purpose for collecting these samples and taking these measurements is to understand the quality of the system, the efficiency of the process, or the quality of the liquid, solid, or gas products emitted by our system. We often deal with different matrices, including liquids, solids, and gases. For the purpose of the explanation that follows below, we will talk about the mass of some constituent in a liquid volume of water. However, understand that this same concept applies to the mass, number, or amount of constituent in any matrix (e.g., mass of solid, volume of air, etc.).

Therefore, we want to know the quantity of some constituent in our system. For example, we might want to quantify the mass of suspended solids in treated wastewater effluent, or the mass of total nitrogen in drinking water, or the amount of dissolved oxygen in a river. However, these systems are often 'turned on' 24 h/d, 7 d/week, 52 weeks/year, so it is impossible for us to know the true total quantity of solids in all of the water discharged, the true total mass of nitrogen in all of the water at the drinking water plant, or the true total amount of dissolved oxygen in the entire river.

You can consider these true total amounts to be the **population** of the pollutants or constituents of interest. The population is sort of like if you were able to collect an infinite number of samples from the system. But, because we cannot collect an infinite number of samples, we will never know the true amount of a constituent in the system. Therefore, we collect **samples**, for instance, of a manageable volume of water, and we measure the quantity of the constituent (say, nitrogen) contained in these samples. We then use those measurements to make inferences (i.e., draw conclusions) about the amount of the constituent likely contained in the rest of the volume of water that we were not able to sample and analyse. The more volumes of water sampled, the more confidence we have about the true amount of nitrogen (or any other constituent) in our system.

Therefore, in summary, with respect to monitoring programmes, when we talk about the total quantity of pollutants or constituents in our system, we are referring to the population.

The **population** of the concentration of any given pollutant or constituent comprises the amounts of that constituent (e.g., mg, moles, colony-forming units, etc.) contained in many individual volumes of water – in fact, the amounts contained in so many volumes of water that they account for every single drop of water in your system.

A **sample** is the amount of constituent contained in a limited number of smaller volumes of water collected as a subset of the total amount of water in your system. For instance, if our system is a lake, then the *population* is all of the water in the entire lake and our *sample* is the small volume of water taken back to our laboratory.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Table 3.1 Example scopes for three hypothetical sampling/monitoring studies.

	Non-Point-Source Contamination Study	**Evaluation of Treatment Plant Performance**	**Drinking Water Source Compliance Programme**
Study/research question	During a rain event, how does the concentration of total suspended solids (TSS) in the New River change with respect to time, at locations upstream and downstream of the New River Condominiums construction site?	In a pilot-scale upflow anaerobic sludge blanket (UASB) reactor treating wastewater from a small city with low industrial activity, how does the reduction of chemical oxygen demand (COD) vary throughout the year and with respect to highly fluctuating ambient temperatures and variable influent organic and hydraulic loading rates?	In the Unionville Reservoir, which is a proposed source of raw water for a new drinking water treatment plant, are the concentrations of general physical contaminants, inorganic contaminants, nutrients, regulated synthetic organic compounds, and volatile organic compounds below the current maximum contaminant levels?
Location	The New River	A pilot-scale UASB reactor in a small city (with highly fluctuating ambient temperatures, variable organic and hydraulic loading rates)	The Unionville Reservoir
Background or context of the problem	Concentrations of TSS in a local river are not in compliance with regulatory standards, and the water management authority is suspecting that a construction site may be one source of the pollution. Therefore, the authority has arranged to carry out a special study to measure the construction site's impact on water quality during rain events	A thesis student is evaluating the performance of UASB reactors in a region with highly fluctuating seasonal temperatures and variable water usage throughout the year, affecting organic and hydraulic loading at the treatment plant. The goal is for this research to inform the design and operation of a proposed upgrade to the treatment plant	A county water authority is designing and planning a new drinking water treatment plant, and as such needs to evaluate the quality of a proposed source of raw surface water for the treatment facility. There is no inherent problem to be solved, but the monitoring programme is intended to avoid health problems associated with contaminants
Length and timeframe	This project will start in October 2019 (at the beginning of the rainy season) and is anticipated to end in April 2020 (at the end of the rainy season)	This project will begin in September 2019 and is anticipated to end in May 2021, with sampling carried out throughout the year, with samples collected during different seasons and for different water use scenarios	This project originally spanned from January until December 2007. However, the water body is now the main raw water source for the treatment facility, and the programme has continued as a long-term ongoing programme with no established end-date
Synergistic studies and programmes	Monthly historical TSS data at a downstream location is available since 2001 from an ongoing compliance programme. The construction firm also has a stormwater pollution prevention plan report, dated March 2019, which includes a peak runoff analysis and information about pollution prevention measures. Inspection reports are available from the local government authority	A previous student published a thesis on a study of the performance of UASB reactors at high temperatures and steady loading rates. There are also numerous academic journal articles focussed on the performance of these reactors under various conditions. A thorough literature review should be conducted by the student	There are several inputs to the water body from local rivers and streams, some of which are monitored regularly as part of a water quality compliance programme. There is also land use data for the watershed, available as geographical information system shapefiles

Finally, it is worth noting some additional details about the semantics of the word 'sample' that may cause confusion for some readers. In our discipline, and in all disciplines that deal with **water quality and treatment processes**, the word 'sample' refers to *the physical smaller volume or mass of water* (or another liquid, or a solid or gas, etc.) that is collected from a larger body of water (and typically analysed, for instance, in a laboratory).

However, in the field of **statistics**, the word 'sample' refers to *a smaller set of data collected from a larger population*. Therefore, our **statistical sample** would consist of the data obtained from several **water samples** collected from a larger body of water. A good way to avoid confusion in the terminologies is to distinguish a *statistical sample* from the *environmental samples* (e.g., water samples, sludge samples, biogas samples, etc.). In this chapter and in Chapter 4, we will discuss best practices for collecting and analysing environmental samples. You will learn more about statistical samples and distributions in Chapter 5.

C. 4

C. 5

3.2.4 Measurements and anticipated use of data

Characterize what type of data will be collected, determine what measurements and/or observations will be made (Table 3.2), and specify what type of **matrix** will be sampled, observed, or probed. It may include one or more of the following:

- Water (drinking water, environmental water, polluted (waste)water)
- Sludge/biosolids
- Soil/sediments
- Animal tissue/collection of organisms (e.g., for a bioassessment)

S. 3.1

It is helpful to **define the anticipated use of the data** prior to commencing the monitoring programme or study. This will typically depend on the type of data collection activity being conducted (see Section 3.1 – e.g., operational monitoring, compliance monitoring, emergency assessment, research project, etc.). If the programme's intent is to monitor *ambient water quality*, then the data might be used to characterize watershed health, support water quality control plans, develop policies, or address impacts to human and animal health (e.g., fishing, swimming, or drinking advisories). If the purpose of the study is purely to *advance science*, then the data might be used for a peer-reviewed journal article to elucidate a mechanism associated with a treatment process, to evaluate cutting edge methodologies, or to pilot-test innovative technologies. In some cases, the data might be used for *regulatory purposes* (e.g., issuing permits, investigative orders, waivers, or establishing maximum daily loads).

Then, determine *what kinds of decisions will be made from the study's results* and identify **possible actions** that may be taken, depending on the results obtained. For example, will a fine be applied if a discharge point to a water body is found to be not in compliance with regulations? Will a treatment process be implemented in full scale if it achieves a certain per cent removal of a contaminant at a pilot scale?

3.2.5 Standard assessment thresholds and operating procedures

Advanced

It is important to document and communicate any assessment thresholds needed for your project to ensure that the analytical results are fully supportive of your decision. Assessment thresholds may include any of the following:

- A **total maximum daily load (TMDL)** is defined as the maximum amount of a pollutant allowed to enter a waterbody in order to meet water quality standards. In the United States, the TMDL determines the target pollutant reduction and allocates load reductions necessary for any

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Table 3.2 Examples of measurements commonly used for monitoring programmes and research studies.

Type of measurement	Examples
Field measurements	• Dimensions of the treatment unit process • Temperature • Wind speed • Water depth
Bioassessment	• Benthic macroinvertebrate survey • Periphyton survey • Fish survey
Continuous data	• Flow rate • Turbidity • Temperature • Dissolved oxygen • Conductivity • Ammonia nitrogen • Nitrate • pH • Dissolved organic carbon (DOC)

Chemistry	• Conventional
	  ○ Alkalinity
	  ○ Hardness
	  ○ Biochemical oxygen demand (BOD)
	  ○ Chemical oxygen demand (COD)
	• Nutrients
	  ○ Organic nitrogen
	  ○ Ammonia nitrogen
	  ○ Nitrate
	  ○ Nitrite
	  ○ Phosphate and total phosphorus
	• Inorganics
	  ○ Trace metals
	  ○ Mercury
	• Organics
	  ○ Pesticides
	  ○ Fuels
	  ○ Surfactants
	  ○ Solvents

Microbiology	• Total heterotrophic count • Microscopic evaluation (e.g., of mixed liquor suspended solids) • Faecal indicator bacteria • Microbial source tracking markers • Pathogenic microorganisms
Solids	• Total solids • Volatile solids • Total suspended solids

(*Continued*)

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Table 3.2 Examples of measurements commonly used for monitoring programmes and research studies (*Continued*).

Type of measurement	Examples
	• Suspended sediment concentrations • Total dissolved solids
Algal bloom response	• Toxins • Microscopy • Chlorophyll-a
Toxicity	• Acute • Chronic
Other	• Satellite imagery • Remotely sensed data • Aerial drones • Cutting edge research methodology

source(s) of the pollutant. It is equal to the sum of all waste load allocations from point sources of pollution, plus the sum of all load allocations from non-point sources of pollution, plus a margin of safety to account for the uncertainty associated with predicting pollutant reductions (US EPA, 2018).

- A **maximum contaminant level goal (MCLG)** or **public health goal (PHG)** is defined as the level of a contaminant in drinking water that does not pose a significant risk to health (OEHHA, 2019). MCLGs and PHGs are not regulatory standards but instead are used to trigger risk communication activities. For example, in some jurisdictions, if the MCLG or PHG for a public water system is exceeded, a public notice must be distributed to all users of the water system, but no fine or penalty is imposed to the water authority. MCLGs and PHGs are established using rigorous methods. It starts with a compilation of relevant information about a contaminant from the scientific literature (e.g., studies of the contaminant's effects on laboratory animals and humans who have been exposed to the contaminant). The data from these studies are then used to perform a *chemical or microbial risk assessment* to determine the levels of the contaminant that could be associated with various adverse health effects. Certain thresholds have to be set in order to establish the MCLG or PHG – for example, in California, PHGs are calculated assuming a maximum one in 1,000,000 probability of adverse health effects for people who drink water every day for 70 years. This means that, on average, not more than one person in a population of 1×10^6 would be expected to develop cancer as a result of exposure to the particular pollutant. For microbial risk assessments, lower thresholds are often adopted, such as one in 10,000 or even as low as one in 100 in some countries.
- A **maximum contaminant level (MCL)** is the maximum permissible level of a contaminant in water delivered to any user of a public water system in the United States (U.S. Code, 1974). These levels are set as close to the MCLG or PHG as feasible. Other countries have adopted similar terminologies for such levels.

You should also define what **standard operating procedures (SOPs)** will be used for sample collection and field measurements. In many cases, if the programme is for compliance, the SOPs will be specified by the regulations. For research projects, the SOPs must be based on protocols recognized in the scientific literature or must be thoroughly tested against other standard methods for quality control.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

3.2.6 Quality control samples

Advanced

Establishing and maintaining a quality control is essential for any project or programme. A **quality control programme** should consist of an *initial demonstration of capability, ongoing demonstration of capability, method detection limit determination*, and *quality control sampling*, which consists of control and background samples (analysed to isolate background conditions and site-specific effects) and an *assortment of variability controls*, including sample spikes, sample blanks, inhibition controls, and field or laboratory replicates (APHA, 2017; US EPA, 2017).

First, the laboratory must demonstrate its initial capability of implementing each method. This typically requires analysing the following control samples, which are used to determine the precision (standard deviation) and accuracy (per cent recovery):

- Any required calibration standards
- At least one **reagent blank** (negative control), which should be free of the contaminant of concern
- ≥4 **spiked controls** (positive controls), which are like reagent blanks that are spiked with known concentrations of the contaminant of concern
  - *Accuracy*: From these samples, calculate the per cent recovery and ensure that it is within the specified acceptance criteria. In the absence of acceptance criteria, aim for a per cent recovery between 80% and 120% as a starting point (APHA, 2017).
  - *Precision*: As a measure of sample precision, calculate the coefficient of variation (CV) from the replicate spiked controls, which is equal to the standard deviation divided by the mean value. Ensure that the CV is within the specified acceptance criteria, but if none are provided, then aim to achieve a CV of ≤20% as a starting point (APHA, 2017).

The **method detection limit (MDL)** is defined as the concentration that produces a signal that is different from the blank with a probability of 99%. At a minimum, at least seven replicates of a **process blank** (also known as a **method blank** or a **reagent blank**) should be analysed. A process blank is a sample blank (typically reagent water), that is free from the contaminant of interest, and that is analysed *and processed* exactly the same way as the samples, using the same methods, and coming into contact with all other reagents in the complete procedure. This is distinct from an **instrument blank**, which is a sample blank that is only analysed in the instrument (but not processed). For more information about how to calculate the MDL and other detection and quantitation limits, see Chapter 4.

C. 4

After demonstrating initial capabilities, the laboratory should **continue to demonstrate ongoing capabilities** by analysing process blanks and spiked controls periodically and evaluating them to ensure continued precision and accuracy. The frequency of ongoing demonstration of capability should be as specified in the protocol or standard operating procedure but at a minimum should be conducted quarterly. If process blanks are reading concentrations below the MDL, then no qualification is needed in the results. If process blanks are above the MDL but below the limit of quantification (see Chapter 4), then a qualifying statement should be provided with the sample results to indicate a positive process blank. If the process blank is detected at a concentration above the limit of quantification, then corrective action is needed (APHA, 2017).

C. 4

A **background control sample** is one that is collected from a site that is not impacted by pollution or from a time when the level of the pollutant is at a stable 'background' level. This type of control sample is especially useful if you are trying to identify a source of contamination. Specifically, you should compare concentrations in this sample with concentrations in samples collected at sites suspected to be impacted from the pollution source to give you more confidence that the levels you detect in the sample are indeed elevated by the suspected source of pollution.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Table 3.3 Method for analysing field blanks, process blanks, and instrument blanks to determine the source of contamination.

Field blank result	Process blank result	Instrument blank result	Interpretation
Negative	Negative	Negative	No contamination occurred
Negative	Negative	Positive	Contamination occurred at the instrument or the instrument needs to be recalibrated
Negative	Positive	Positive	Contamination likely occurred during sample processing and may also have occurred at the instrument
Positive	Positive	Positive	Contamination likely occurred during sample collection and/or transportation and storage. It may also have occurred during sample processing or at the instrument

A **field blank** is a sample of reagent water that is taken out to the field during sample collection, stored along with the samples, transported to the laboratory along with the samples, and analysed along with the samples. The purpose of the field blank is to test for contamination that may have occurred during sample collection, storage, or transportation. If a contamination event is detected in the field blank, it can be compared with the process blank and the instrument blank to determine where the contamination happened (Table 3.3).

Other important variability controls include **field replicates** and **laboratory replicates**, which can be used to calculate coefficients of variation for losses of precision resulting from variation in the field or in the laboratory. These coefficients of variation can be compared to the coefficient of variation calculated for spiked laboratory controls. Field or laboratory replicates might be analysed for every one out of 10 or 20 samples.

Inhibition controls are a normal part of quality control sampling for certain protocols. Essentially, some environmental constituents may inhibit certain reactions that are necessary to produce a signal. Tests for inhibition can be done either by diluting the sample and measuring the resulting signal, which should be proportional to the dilution factor. Otherwise, samples can be spiked with a known concentration of the contaminant and the measured to see if the amount added corresponds to the increase in the signal (Table 3.4).

Table 3.4 Method for interpreting dilution or spike controls for inhibition.

Type of inhibition test	Result	Interpretation
Dilution (1:10)	Sample concentration in the dilution control is 10% of the undiluted sample	No evidence of inhibition
	Sample concentration in the dilution control is much greater than 10% of the undiluted sample	Inhibition likely occurred
Spiked sample (three times the sample concentration was spiked into a replicate sample and analysed)	Spiked sample concentration is four times as high as the un-spiked sample	No evidence of inhibition
	Spiked sample concentration is less than, equal to, or only slightly greater than the un-spiked sample concentration	Inhibition likely occurred

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

3.2.7 Data management and analysis

Basic

Data should be managed in a way that allows for it to be **archived in a common format** to data from other studies, and with appropriate metadata to describe the data set. More information about data management is provided in Chapter 4. In addition to *primary data* (e.g., data collected from laboratory analysis on collected samples), list what *other sources of data*, if any, will be used in the study or monitoring programme. This might include data provided by another agency or entity, complimentary data gathered from a weather station, or qualitative data gathered through surveys, interviews, observation, or a mixed methods approach.

C. 4

Before even starting to sample and collect data, you should determine which **statistical method(s)** will be used to analyse the data and what the **acceptable level of error** will be for the statistical test(s) being used. A common level of acceptable error for research studies is an *alpha error of 0.05* and a *beta error level of 0.20*. This is the probability you are willing to accept for making a **type I error (false positive)** or a **type II error (false negative)**, respectively. Further description of these types of errors is presented in Chapter 10.

C. 10

Table 3.5 shows a summary of **different types of studies** that are commonly completed for research projects and the **corresponding statistical test(s)** that should be used for such studies. It should be noted that this is not necessarily a comprehensive list, as there are many more statistical tests that are outside the scope of this book. Chapters 5 and 6 of this book cover descriptive statistics; however, an in-depth coverage of some of the more advanced statistical tests highlighted in Table 3.5 is beyond the scope of this book. There are many excellent text resources which cover these methods and others (e.g., Sokal & Rohlf, 2012).

C. 5

C. 6

For a useful analogy on understanding the meaning of *alpha* and *beta errors*, consider the penal system of a country. Suspects are considered innocent until proven guilty, just as two samples are considered equal until proven to be significantly different from each other. Accepting an alpha level of 0.05 (5%) is like accepting that you may erroneously convict an innocent person to be guilty 1 out of 20 times (5%) on average. Accepting a beta level of 0.20 (20%) is like accepting that you may fail to convict a guilty person (for lack of sufficient evidence) 1 out of every 5 times (20%) on average.

3.3 SAMPLE COLLECTION

3.3.1 Spatial aspects of sampling

Basic

To evaluate and monitor the efficacy of a treatment system, samples should be collected at the **influent** and **effluent** of the system. For compliance programmes, at a minimum, samples should be collected at the final effluent location (to demonstrate compliance with MCLs and effluent discharge limits). However, it is also useful to monitor the performance of a particular unit process. This requires collecting samples at the *influent and effluent point of that unit process*. In some cases, it might even be desirable to collect samples at *various locations within the unit process* (e.g., in a reed bed or horizontal constructed wetland, you might want to collect samples at intermediate points spatially distributed within the wetland bed). Also, *environmental variables* (e.g., temperature, dissolved oxygen, etc.) and *control variables* (e.g., mixed liquor suspended solids, sludge blanket levels) may need to be collected or measured inside the treatment unit. In many treatment systems, it is useful to collect data from samples collected in *side streams* or *waste streams* associated with the process.

Many researchers place a greater emphasis on **water samples**, but for many treatment systems, it is useful to collect **sludge samples** as well. For some processes, collecting **gas** might be necessary or beneficial.

It might be useful to think of sampling locations as being essential, important, and potentially useful. **Essential sampling locations** are the most important locations that you shall not do without. The *final effluent* point of water and wastewater treatment systems is an essential sampling point, because it allows

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Table 3.5 A guide for choosing the appropriate statistical test or procedure based on the purpose of the experiment, the number of sample groups, and notes about the type of data that can be used.

Purpose	Samples	Statistical tests	Types of data	Chapters
Descriptive: Describing the central tendency and variation of data (such as the mean, median, standard deviation, percentiles, confidence interval around the mean, etc.)	1 or more	• Confidence intervals ○ Real numbers ○ Proportions (percentages) ○ Binary parameters ○ Positive integers	• Continuous numbers • Proportions and percentages (binomial data from binary measurements) • Poisson counts (positive integers)	5 and 6
Inferential comparative (one sample): Comparing sample(s) against a threshold (e.g., regulatory limit or target value)	1	• One-sample *t*-test ○ One-sided or two-sided	• Continuous numbers • Proportions and percentages (binomial data from binary measurements) • Poisson counts (works best with large samples; often necessary to transform data)	9 and 10
Inferential comparative (two samples): Comparing treatment versus control differences Comparing two samples with different treatments applied	2	• Two-sample *t*-test ○ One-sided or two-sided ○ Independent or paired samples ○ Pooled or not pooled variance (homoscedastic/heteroscedastic) • Rank tests (non-parametric) such as Mann–Whitney–Wilcoxon for independent samples or Wilcoxon signed-rank test for dependent samples may be better for small data sets or with non-normally distributed data		10
Inferential comparative (two samples): Comparing proportions and percentages between a treatment and a control (2×2 contingency tables)	2	• Chi-square test (expected frequencies in contingency table must be >5) • Fisher's exact test (can be used when expected frequencies are <5)	• Proportions and percentages	Outside scope of this book
Inferential comparative (three+ samples): Multiple comparisons with one or more treatment factor	3 or more	• Analysis of variance (ANOVA) ○ One-way, two-way, 2^f factorial ○ Balanced versus unbalanced ○ Blocking factor versus no blocking	• Continuous numbers • Proportions (check the normality assumption for parametric methods)	10

(*Continued*)

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Table 3.5 A guide for choosing the appropriate statistical test or procedure based on the purpose of the experiment, the number of sample groups, and notes about the type of data that can be used (*Continued*).

Purpose	Samples	Statistical tests	Types of data	Chapters
		○ Multiple comparisons to control or multiple comparisons of all ○ Post hoc comparisons with adjustments to achieve the desired familywise error rate or control the false discovery rate (FDR) □ Bonferroni □ Least significant difference □ Dunnett □ Tukey–Kramer □ Benjamini–Hochberg (FDR) □ Storey (FDR) • Kruskal–Wallis non-parametric test followed by the post hoc Dunn test	• Poisson counts (works best with large samples; often necessary to transform data)	
Descriptive trends: Determine the relationship between explanatory variable(s) and a response variable	1	• Correlation • (Multiple) linear regression • Generalized linear models (GLMs) ○ Gaussian (continuous numbers; conventional linear regression – identity link) ○ Logistic Regression (binomial – logit link) ○ Gamma (positive continuous numbers – inverse link) ○ Poisson counts (positive integers – log link)	• Continuous numbers (regression) • Binomial proportions (GLM, use link function) • Poisson counts (GLM, use link function)	11
Inferential trends (two samples): Compare two regression slopes	2	• Regression or GLM with *F*-test		Outside the scope of this book
Inferential trends (three+ samples): Comparing more than two regression slopes	3 or more	• Regression or GLM with analysis of covariance		Outside the scope of this book

Note: Common underlying assumptions for the above-mentioned statistical tests include the following: homoscedasticity (constant variance of errors); non-stochastic explanatory variables (explanatory variables are accurately measured); normal distribution of residual errors; linearity (randomness of residuals with respect to the explanatory variables); no multi-collinearity (no significant correlation between explanatory variables); and independence of observations.

you to assess compliance and risk by comparing your measurements against some regulatory limit or desired level. However, the nature of essential sampling locations also depends on the aim or goals for the study. For example, if the purpose of the project is to evaluate the performance of a particular treatment technology, then samples should obligatorily be collected at both the *influent and effluent* points at a minimum. If you want to study the performance of each unit comprising the treatment plant, you need to collect samples *upstream and downstream each unit* (see Figure 3.1). In some cases, especially when evaluating wastewater treatment reactors using a mass balance approach, it is also important to collect samples of *sludge* and sometimes also *gas* emissions.

For studies related to **water quality in rivers and streams**, samples should be collected immediately upstream and downstream of the suspected point source of pollution. In addition, the point source of pollution should be sampled, and if the point source originates from a treatment plant, ideally samples should also be collected at the influent of the treatment plant, in order to evaluate the efficacy of the treatment process at eliminating the pollutant of concern. Finally, it may be desirable to collect several additional samples further downstream of the treatment plant, at different distances, to evaluate the degradation or further dilution of the pollutant in the water body (Figure 3.2).

You should avoid sampling in areas where water is stagnant or where reverse flow patterns occur. In addition, areas near the inner edge of curves in a river may not be representative due to the patterns of flow and turbulence at those locations. Samples are best collected below the surface to avoid the influence of surface boundary effects. Samples should also not be collected too close to the bottom of a river. However, if collected and analysed separately, samples collected at the bottom sediment of a river or another water body surface may help understand the evolution of pollution over time and the potential for accumulation of possible chemical substances in macrobiota. The sampling points should be representative, avoiding areas affected by atypical habitats, such as those under bridges (ABNT, 1987).

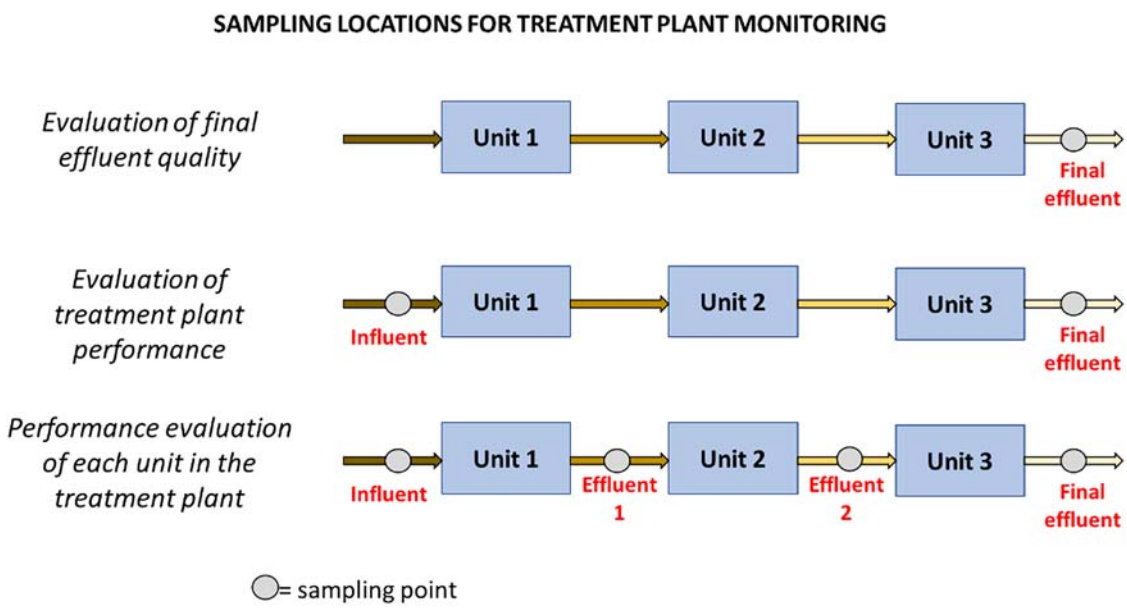


Figure 3.1 Recommended sampling points for different types of studies in a treatment plant.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

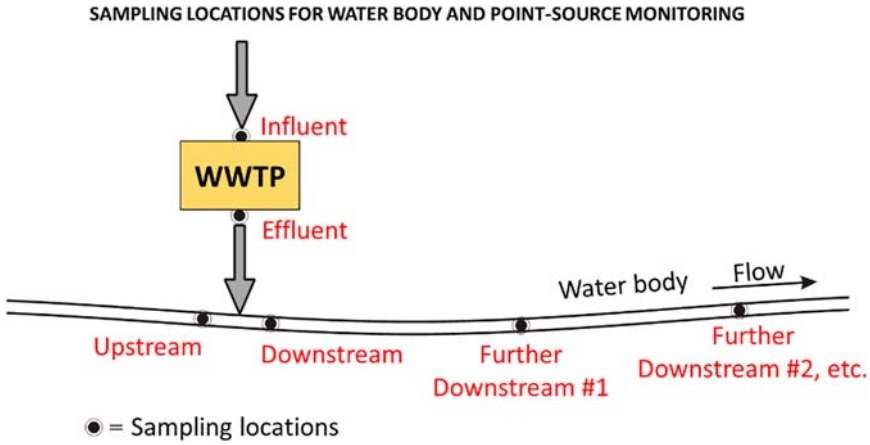


Figure 3.2 Recommended sampling points for a study of pollution in a water body receiving a point-source discharge from a wastewater treatment plant (WWTP).

Table 3.6 shows some example sampling locations and timeframes for our three hypothetical studies described in Table 3.1. Note that the frequency of sample collection should be determined after conducting a power analysis with the proposed alpha and beta error levels and desired effect size (see Section 3.5 for power analysis).

S. 3.5

3.3.2 Types of samples

Basic

Figure 3.3 illustrates different types of samples that we may collect, depending on the objective of our study and the available resources:

- Instantaneous conditions: **grab sample**
- *Approximation of average conditions*: **composite sample** (fixed volume and flow-proportional volume)
- *Concentration profiles over time*: **sequence of grab samples** or measurements by a **sensor**

- **Grab sample**

  A **grab sample** (Figure 3.3a) consists of a single sample of water collected at a given instant of time. It is the *easiest type* of sample to collect, but it may not be the most representative at locations where the quality of water changes throughout the day. This type of sample *does not take into account the potential variability of concentrations with respect to time*, and it may lead to the underestimation or overestimation of the true mean concentration, unless concentrations are relatively constant with respect to time. If you need to know the *variation in the concentrations over time*, several **sequential grab samples** must be collected individually and analysed separately (Figure 3.3d).

  Some types of analysis require the use of grab samples, since the samples cannot be stored for the period of time required for a composite sample (see below), rather they must be analysed or measured immediately after collection. Some examples include pH, temperature, and dissolved oxygen. *If using grab samples over a long period of time, it is important to ensure that samples are collected at approximately the same time of day for consistency*. Grab samples are

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Table 3.6 Example sampling locations and types for three hypothetical studies.

	Non-point-source contamination Study	**Evaluation of treatment plant performance**	**Drinking water source compliance programme**
Study/research question	During a rain event, how does the concentration of TSS in the New River change with respect to time, at locations upstream and downstream of the New River Condominiums construction site?	In a pilot-scale UASB reactor treating wastewater from a small city with low industrial activity, how does the reduction of COD vary throughout the year and with respect to highly fluctuating ambient temperatures and variable influent organic and hydraulic loading rates?	In the Unionville Reservoir, which is a proposed source of raw water for a new drinking water treatment plant, are the concentrations of general physical contaminants, inorganic contaminants, nutrients, regulated synthetic organic compounds, and volatile organic compounds below the current maximum contaminant levels?
Sampling locations and timeframe	This project will start in October 2019 (at the beginning of the rainy season) and is anticipated to end in April 2020 (at the end of the rainy season) Flow-proportional spatial and temporal composite samples will be collected in the New River during up to 10 rain events, at locations immediately upstream and immediately downstream from the construction site. In addition, grab samples will be collected prior to the start of each rain event to understand background concentrations during dry weather. Grab samples will also be collected after the hydrograph recedes back to background flowrates	This project will begin in September 2019 and is anticipated to end in May 2021, with sampling carried out throughout the year, with samples collected during different seasons and for different water use scenarios Samples will be collected at the following locations: • Influent wastewater • Effluent treated water • Waste sludge • Sludge blanket In addition, the influent flowrate will be measured daily, and the volume of waste sludge will be measured each time the reactor is de-sludged	This is a long-term ongoing programme with no established end-date. Spatial composite samples are collected from the reservoir on a monthly basis, with aliquots at five different depths that correspond to the intake depths used by the water treatment facility

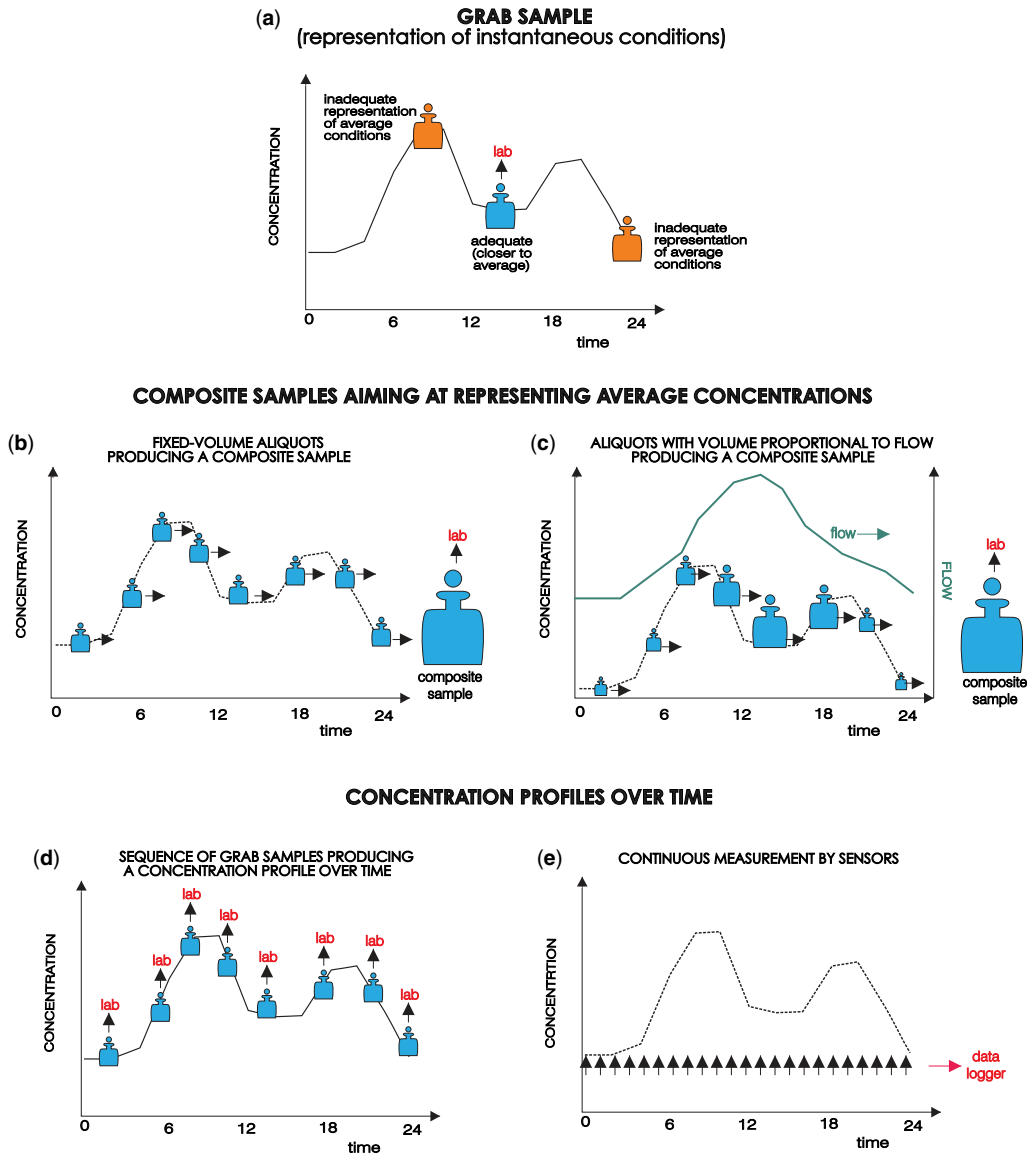


Figure 3.3 Different types of samples to be collected and analysed.

appropriate for the assessment of an effluent stream that *does not discharge on a continuous basis* and to provide information about the *concentration of a contaminant at a particular time of day*. Certain parameters, including pH, temperature, dissolved oxygen, and residual chlorine, cannot be analysed with composite samples due to short holding times (US EPA, 2017). In most other cases, composite samples are the most appropriate, especially when calculating loading rates as a mass per unit time.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

- **Temporal composite sample**

  A **temporal composite sample** is a mixture of smaller sub-samples (called aliquots), collected periodically throughout the day. This type of sample is more representative at locations where water quality changes throughout the day, as the composition of the sample *helps minimize the effects of variability in the concentrations over time*, giving a better representation of the true average concentration. It is especially useful at wastewater treatment plants, where the flow rate and quality of influent sewage can vary considerably throughout the day. Usually, a composite sample is collected over a 24-h period, and autosamplers can be programmed to collect composite samples for a period of 24 h. However, in some cases, 12-h or 8-h composite samples are used when autosamplers are not available, for convenience purposes (e.g., to avoid having to collect samples during the night time or during non-working hours). The frequency of collection for each aliquot is usually every 1 h, but may be higher or lower, depending on the expectation of variability of concentrations. You should ensure that the *aliquots collected at the beginning of sampling are well preserved* to prevent internal reactions that may affect the concentration of the pollutant of concern. For this reason, it is recommended to use a cooler to store samples or to use automatic samplers with ice space. When preparing the composite sample once all aliquots are collected, each container containing the aliquots should be thoroughly mixed, as sedimentation may have occurred.

  There are different types of temporal composite samples: the two most common are **fixed-volume** composite samples (aliquots each have equal volumes) (Figure 3.3b) and **flow-proportional** composite samples (aliquots have volumes that are proportional to the flow measured at the time they are collected – higher flow → higher volume; lower flow → lower volume) (Figure 3.3c). Flow-proportional samples are more representative of changing water quality conditions throughout the day, which is common with wastewater plants. Example 3.1 illustrates the differences between a fixed-volume composite sample and a flow-proportional composite sample. The associated Excel spreadsheet contains a worksheet that allows you to calculate aliquot volumes for your own flow-proportional composite sample.

- **Spatial composite sample**

  A **spatial composite sample** refers to the combination of individual samples collected at different geographical or physical positions. This type of sample is especially important to get representative estimates of *water and solid matrices in systems with poor mixing*. For example, when collecting samples from a mid-size or large river, it is recommended to collect samples at various points in the cross section of the *river* and mix them into a single sample. This way, you get an idea of the average concentration in the water passing through all points of the river. Spatial composite samples are also commonly used when collecting *sludge* or *sediments*.

  When collecting a spatial composite sample, it is important to note that the aliquots should be collected within a short time interval to minimize the influence of temporal variations. In rivers, concentrations of the constituents are rarely homogeneous throughout their cross section. In fact, the river cross section may have several stagnant zones, in which the concentrations may vary greatly. Furthermore, there may be differences with respect to depth. The Brazilian National Standards Organization (ABNT, 1987) recommends sampling aliquots at different locations along the cross section and at different depths, depending on the width and depth of the river (Figure 3.4). Sample aliquots will compose a spatial composite sample that accounts for variation throughout the cross section. In general, if the river or stream width is greater than 5 m, then your composite sample should include spatial variations; if the river or stream is deeper than 2 m, then a composite sample with aliquots collected at different depths should be collected.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

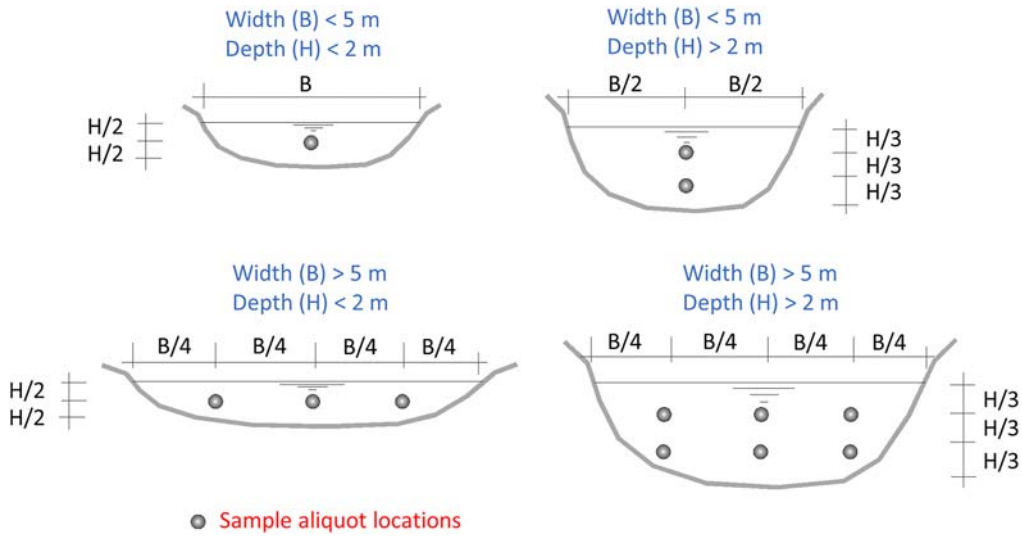


Figure 3.4 Recommended spatial composite sampling plan for a river or a stream.

- **Sensors**

  Sensors are used to collect real-time measurements of certain parameters, or surrogate measurements that correlate with the concentrations of certain pollutants. Sensors are commonly used in treatment plants, because they provide real-time information to operators, who may make operational changes based on the sensor readings. Sensors may collect single measurements at a time (e.g., if the sensor is manually inserted into the water body) or multiple measurements throughout the course of a day (e.g., if the sensor is installed in-line or connected to a data logger) (Figure 3.3e). There are sensors for various parameters of interest in water quality, such as temperature, pH, dissolved oxygen, and electrical conductivity.

3.3.3 Need for a time delay to collect the downstream sample?

Basic

There is some debate whether a downstream sample should be collected after a time interval from the collection of the upstream sample, with this time interval being equivalent to the hydraulic retention time (HRT) of the unit or system. Let us analyse Figure 3.5 and the following possibilities:

C. 14

- *Sampling in a river receiving a point-source pollution*. This first case (top figure) is slightly simpler. The river flows approximately like a plug flow (see Chapter 14 for the concept of plug flow), and then the time spent for the water to reach the downstream sampling location is approximately the travelling time dictated by the distance between the points and the mean flow velocity. You could then take this into account to collect a sample with a delay equivalent to the travelling time, in an effort to collect the same plug of water that received the discharge.
- *Sampling in a treatment unit*. The second case (middle and bottom figures) is a much debated one. Several people argue that we should collect the downstream sample with a delay equivalent to the hydraulic retention time of the treatment unit, treatment plant, or water body. The expectation is that we would be able to collect the same water that entered the unit, underwent treatment, and then left the unit. However, this will depend essentially on the hydrodynamic behaviour of the

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

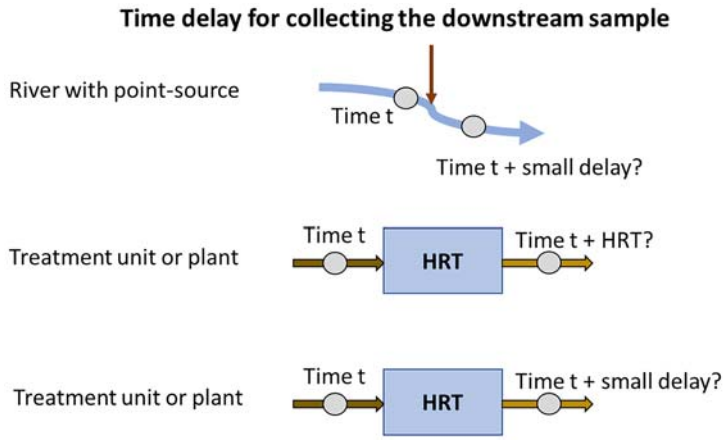


Figure 3.5 Possible time delays for collecting the downstream sample compared with the time for collecting the upstream sample.

unit. If our unit approaches plug flow, then the same considerations made above for a river would apply, but to a lesser extent. In this case, the travelling time through the unit could be close to HRT, if there is little dispersion in the unit. However, if the unit has some degree of mixing (as most units do), the contaminant is dispersed in the reactor volume, and any peak value in the influent would bring a response in the effluent at a faster time compared with HRT. The higher the degree of mixing, the faster the response in the outlet. In this case, implementing a delay equal to HRT does not assist us in obtaining the same fluid elements, before and after the unit.

An overall comment is that our monitoring programme should be established on a practical basis, according to the frequently difficult logistics on site. If HRT is 12 h and you collect the influent sample at 9 : 00 am, you would need to collect the effluent sample at 9:00 pm if you believe that the strategy of the delay equivalent to the HRT should be implemented. Ok, you could have an automatic sampler and solve this problem. But what if the HRT of the unit is 5, 10, 30, or 60 days, as some units in natural treatment processes have? Would you wait that long? Would it be meaningful? Would you still believe that you are sampling the same fluid elements, before and after treatment?

We believe not, and we think **you should be practical in your monitoring programme** and collect *as many samples as possible* from the influent and effluent locations (preferably *composite samples*). By analysing the time series of data, you will be able to draw conclusions about the performance of the unit. If you want to make more advanced analyses between the upstream and downstream data sets, you could study the *cross-correlation* between them (correlation with one of the series subjected to a lag – see comments in Chapter 11).

C. 11

Example

EXAMPLE 3.1 CALCULATE ALIQUOT VOLUMES FOR TEMPORAL COMPOSITE SAMPLES

Develop a plan to collect (a) a 1-L fixed-volume temporal composite sample and (b) a 1-L flow-proportional composite sample of wastewater at a treatment plant.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

The flow rate measured at two-hourly intervals is

Time of Day	*Flow (L/s)*	*Time of Day*	*Flow (L/s)*
00:00	*1.2*	*12:00*	*3.1*
02:00	*2.4*	*14:00*	*3.7*
04:00	*3.7*	*16:00*	*3.2*
06:00	*4.3*	*18:00*	*2.5*
08:00	*3.8*	*20:00*	*2.0*
10:00	*3.3*	*22:00*	*1.4*

Excel

Note: This example is also available as an Excel spreadsheet.

Solution:

(a) Fixed-volume composite sample

For fixed-volume composite samples, you can collect sub-samples every hour for 24 h. However, you can also collect sub-samples at any time interval (e.g., every 30 min, every 2 h, or every 3 h), as long as you continue for a period of 24 h. The representativeness of the sample increases for smaller intervals.

Sometimes, composite samples are only collected during daytime hours (for convenience); however, it should be noted that this may introduce a bias in the measurement of water quality. Flow rates are typically much lower during evening hours, and influent wastewater quality may be quite different during late evening and early morning hours (when users are sleeping) than it is during the day (when users are awake). Wastewater quality may also change drastically throughout the day as users engage in different activities (e.g., using the toilet versus showering and washing dishes). Industrial activities (which often only take place during working daytime hours) can also drastically change the quality of wastewater.

Suppose you choose to collect a fixed-volume composite sample of 1 L at intervals of 3 h, for a total of eight sub-samples ($24/3 = 8$). The sub-samples, with volumes of 125 mL ($1000/8 = 125$), could be collected as shown in the following table, then mixed to form a composite sample with a volume of 1 L.

Fixed-volume composite sample collection plan (eight aliquots) is shown in the following table:

Time of the Day	Volume of Sub-sample (aliquot) (mL)
06:00	125
09:00	125
12:00	125
15:00	125
18:00	125
21:00	125
00:00	125
03:00	125

If you had chosen to collect hourly sub-samples, the number of aliquots in a day would be 24, and the volume of each aliquot would be $1000/24 = 42$ mL.

(b) Flow-proportional composite sample

Like fixed-volume composite samples, for flow-proportional composite samples, the sampling interval can be anything (e.g., every 30 min, every 2 h, or every 3 h). The representativeness of

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

the sample likewise increases for smaller sub-sample collection intervals. Suppose you choose to collect a 1-L flow-proportional composite sample with 2-h intervals. A total of 12 sub-samples ($24/2 = 12$) may be collected as shown in the following table and then mixed together to form a composite sample with a volume of at least 1 L.

To determine the sub-sample volume, you need to

- assume an average daily flow rate;
- divide the desired sample volume by the number of sub-samples to get the sub-sample volume for a fixed-volume composite sample;
- measure the flow rate each time a sub-sample is collected;
- calculate a multiplier ratio by dividing the measured flow rate by half of the assumed average daily flow rate;
- multiply the multiplier ratio by the average sub-sample volume.

For the example shown below in the following table, assume that the average daily flow rate is expected to be approximately 2.9 L/s. As a matter of fact, we adopted here the average of the 12 flow measurements. However, in practice, you cannot anticipate the average flow you will have when collecting the sub-samples.

Later, determine the sub-sample volume for a fixed-volume composite sample (1000 mL/12 = 83.3 mL). Then, calculate the multiplier ratio by dividing the measured flow rates by the assumed average daily flow rate. Finally, calculate the sub-sample volume by multiplying the multiplier ratio by 83.3 mL. With these elements, you can construct the following table.

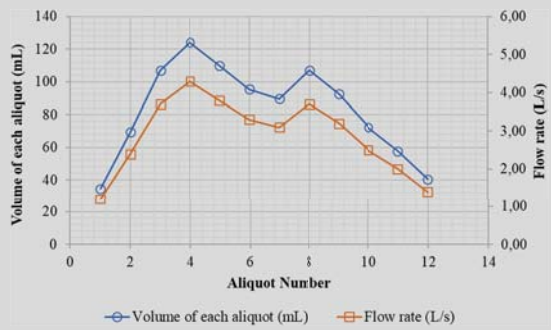
Flow-proportional composite samples (12 aliquots) are shown in the following table.

Aliquot number	Measured flow rate (L/s)	Ratio of the flow rate to the average flow rate	Volume of each aliquot (mL)
1	1.2	0.416	**35**
2	2.4	0.832	**69**
3	3.7	1.283	**107**
4	4.3	1.491	**124**
5	3.8	1.318	**110**
6	3.3	1.145	**95**
7	3.1	1.075	**90**
8	3.7	1.283	**107**
9	3.2	1.110	**92**
10	2.5	0.867	**72**
11	2.0	0.694	**58**
12	1.4	0.486	**40**
Average	**2.9**	**Total volume**	**1,000**

=1.2/2.9

=0.416×83.3

The profiles of flows and aliquot volumes over time is shown in the chart below. You can clearly see the relationship between flow rate and aliquot volume.



Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

It is important to note that collecting a flow-proportional composite sample requires 'guessing' what the average flow rate will be before you start collecting the first sub-sample. If you overestimate the flow rate, the total volume of sample collected will be less than you originally anticipated. If you underestimate the flow rate, your sample volume will be more than you anticipated. **It is smart to aim to collect a larger volume than you actually need for your laboratory analysis**. For example, you could also have half of the assumed average daily flow rate (i.e., $2.9\,L/s/2 = 1.45\,L/s$) for your sub-sample volume calculations, and this would have resulted in collecting a sample with a total volume of 2000 mL, more than you may have actually needed. You can always discard extra sample, but once you start the flow-proportional composite sampling you cannot go back and collect more volume if you come up short, in case the actual flow rate is less than the anticipated flow rate.

3.4 SAMPLE SIZE, CONTAINERS, AND HOLDING TIMES

Basic

The **size** of a sample (its volume or mass), the **type of container** used, the **length of time between sample collection and analysis**, and the **methods used to preserve the sample** prior to analysis all depend on the type(s) of analysis that will be conducted and are generally specified in the standard operating procedure or the standardized method. Every project quality assurance plan should include a table like that in Table 3.7, which outlines the parameters, methods, containers, preservation, and holding times for each analysis.

Table 3.7 Methods, containers, preservation, and holding times for a selection of analytical and field measurement parameters (adapted from US EPA, 2005).

Parameter	Method Number/ Reference	Maximum Holding Time	Container(s)	Preservation
Aluminium, arsenic, calcium, chromium, copper, iron, lead, manganese, magnesium, and zinc	EPA 200.7	6 months	1 × 1-L polyethylene bottle	HNO_3 to pH <2
Antimony, cadmium, and selenium	EPA 200.8			
Mercury	EPA 245.1	28 days		
Anions (Cl, NO_3, NO_2, PO_4, and SO_4)	EPA 300.0	48 h	1 × 1-L polyethylene bottle	Chill to 4°C
Total dissolved solids (TDS)	EPA 160.1	7 days		
Alkalinity	SM 2320B	14 days		
Total coliforms/*E. coli*	IDEXX Colilert	24 hours	1 × 500-mL polypropylene bottle, autoclaved	Chill to 4°C
Temperature, pH, and conductivity	Field probe	Immediate	1 × 250-mL mid-mouth glass bottle	None
Dissolved oxygen	Field probe	Immediate	None, *in situ* measurement	

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

3.5 STATISTICAL POWER AND NUMBER OF SAMPLES

Advanced

How many environmental samples should you collect (i.e., how many data points do you need for each statistical sample)? Often, practitioners, students, and scientists are not able to provide a good justification for their answer to this question. In some labs, it may be common to collect a sample size of $n = 3$ or $n = 10$ as a rule of thumb, but this may not always be the most appropriate sample size for every study. Furthermore, it might make more sense to spread out the samples temporally or spatially, depending on the study objectives, the project budget, and the desired statistical power.

If you want to conduct scientifically sound experiments and make the most use of limited time and funding, you need to use **power calculations** *to determine the appropriate sample size*. However, before a power test can be performed, you first need to define what type of comparison you want to make and what question you want to answer with your study. Only then can you determine which statistical test is the most appropriate to evaluate that comparison. The type of power calculation needed depends on the statistical test that will be used once you finish collecting your data. Table 3.8 shows three common

Table 3.8 Three common types of studies used to assess treatment plant performance and the corresponding statistical tests.

Study type	Description	Examples	Statistical Tests
Compliance monitoring	Comparing the average contaminant concentration with a target regulatory compliance limit	A wastewater treatment facility needs to evaluate if the average concentration of BOD_5 in the treated effluent is below a regulatory threshold of 30 mg/L. The regulatory guidelines state that the monthly average concentration should be significantly lower than the regulatory threshold	One-sample *t*-test Sign test Wilcoxon signed-rank test *Z*-test for proportions Poisson probability of failure/success Frequency analysis and reliability analysis
Evaluate alternative treatment processes	Comparing two parallel treatment trains (e.g., with different processes or operating conditions) to determine if one performs significantly better than the other with respect to the removal of some contaminant	An advanced water treatment facility utilizes a biological activated carbon filter followed by ultrafiltration and reverse osmosis. You are evaluating the impact of seeding the filter with different water sources to see its effect on downstream fouling	Two-sample *t*-test Wilcoxon rank sum or Mann–Whitney test Sign test
Evaluate performance with respect to different factors	This is the type of study you might perform if you want to see how different design, environmental, or other factors influence the performance of a treatment process with respect to the removal of some contaminant	You are developing a new treatment process for the removal of phosphorus, and you would like to better understand how well the system removes phosphorus at different temperatures, pH levels, and loading rates	ANOVA Factorial analysis Kruskal–Wallis test Regression Correlation

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

types of studies that are often performed for the assessment of treatment plant performance and describes the statistical test that should be used for each comparison.

C. 10

Power calculations to determine the appropriate sample size for any test start by defining a **level of acceptable error**. Convention is to use **0.05 for the alpha error** and **0.20 for the beta error** (i.e., **80% power**). Alpha and beta errors have been briefly mentioned in Section 3.2.6 and are further detailed in Chapter 10.

S. 3.2.6

Next, it is necessary to define the **desired standardized effect size**, also known as **Cohen's *d*** (Cohen, 1988). Cohen's *d* is calculated as the *difference that you desire to be able to detect (with significance) divided by the standard deviation of the sample mean*. Note that this difference is standardized by the precision with which you can measure the effect (i.e., it is divided by the standard deviation). *The smaller the difference you want to be able to detect with significance, the more samples you will need to analyse (i.e., the more data points you will need for your statistical sample)*.

Once you determine the desired standardized effect size for the experiment, the next step is to use a **non-central distribution** to calculate the **beta error** for a given sample size. For example, if you are doing a *t*-test to compare your samples, you will use a non-central *t*-distribution. *Central distributions describe the test statistic under the null hypothesis, but non-central distributions describe the test statistic when the null hypothesis is false.* To define a non-central *t*-distribution for a power analysis, use a *non-centrality parameter that is equal to Cohen's d multiplied by the square root of the sample size*. Evaluate the non-central distribution at the *critical statistic for your desired alpha level*. The *cumulative value of this distribution will be equal to the beta error*. Thus, the **power of the test is equal to 1 minus the beta error**.

Power calculations can be *easily performed in several statistical software packages* such as R, Minitab, etc. For a ***t*-test**, in order to calculate the **required sample size**, you generally need to provide the following inputs:

- Cohen's effect size
- desired alpha or type I error (typically 0.05)
- desired beta or type II error level (typically 0.20)
- type of test (one sample or two sample, paired or unpaired, one-sided or two-sided).

The *non-central t-distribution cannot be computed in Excel*, but the Excel spreadsheet for Examples 3.2 through 3.4 of this book contains a custom power calculator, which accesses the non-central *t*-distribution using a series of look-up tables. Practice using it to calculate statistical power for a given sample size:

- **Example 3.2**. To find the power associated with a particular sample size and a desired effect size.
- **Example 3.3**. To find the required number of samples to detect a desired effect size with a particular power (e.g., 80%).
- **Example 3.4**. To find the minimum effect size that can be detected with a particular power and a particular sample size.

The power calculation for the **two-sample *t*-test** is similar. The main parameter that changes is the effect size. Instead of being the difference between the sample mean and the regulatory limit, divided by the sample standard deviation, it is equal to the difference between the mean values of the two samples, divided by the pooled standard deviation.

We will show the examples here, but you will need to follow the calculations in the associated Excel spreadsheet, given their complexity and need to use look-up tables.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

This topic is an advanced one and uses concepts that are further discussed and detailed in other parts of the book. We opted to keep it here, because it is associated with the planning of your work.

S. 10.3.3

You might need to consult other sections in our book and come back here to have a full grasp of the concepts involved. In special, *in Section 10.3.3, we show an iterative procedure for estimating the required sample size for your studies, based on the concepts of hypothesis testing using the t-test. Both procedures lead to the same results*.

Example

EXAMPLE 3.2 DETERMINE POWER BASED ON EFFECT SIZE AND SAMPLE SIZE

The maximum contamination level goal (MCLG) for nitrate in drinking water is 10 mg/L. Suppose you measure the concentration of nitrate in a water source with $n = 5$ samples and record a mean concentration of 9.3 mg/L with a standard deviation of 0.5 mg/L.

C. 9

C. 10

Using a one-sample, two-sided t-test, a p-value of 0.035 is calculated. See Chapters 9 and 10 for more on how to do a t-test and why some people prefer to use a one-sided t-test. Chapter 9 presents several methods to analyse compliance with a regulatory standard, and Excel spreadsheet for Example 9.2 allows you to do a two-sided one-sample t-test and come to this value of $p = 0.035$. This p-value indicates that the measured mean concentration is significantly below the MCLG level (at the 0.05 significance level).

However, the p-value alone does not tell us anything about the beta (type II) error or the power of the analysis. Use the Excel spreadsheet for Example 3.2 to calculate the post hoc power of this statistical analysis.

Excel

Note: This example is also available as an Excel spreadsheet.

Solution:

The **beta error** is found to be equal to **33%** in this case, meaning that the test only had a **power of 67%**.

For this particular experiment, you might consider yourself to be 'lucky' to have found a significant difference, despite the low power of the experimental set-up. Remember, having a statistical power of only 67% means that you have a two out of three chance of finding a significant difference at the given effect level. This is like being a prosecution attorney and acknowledging that you only collect enough evidence to convict two out of every three guilty people on average.

In the future, it might be more prudent to collect more evidence (i.e., increase your sample size), so that your 'conviction success rate' (i.e., your statistical power) is **at least 80%**.

Example

EXAMPLE 3.3 DETERMINE SAMPLE SIZE TO ACHIEVE A DESIRED POWER

A wastewater treatment facility needs to determine how many samples need to be collected to determine if the average biochemical oxygen demand (BOD_5) concentration in a treated effluent is significantly below the regulatory threshold of 30 mg/L. Use the Excel spreadsheet for Example 3.3 to determine the minimum number of samples to ensure that the BOD_5 concentration is significantly below the regulatory threshold with 80% statistical power. Assume a significance level of 0.05, a standard deviation of 4.6 mg/L (this is the assumed standard deviation of repeated BOD_5 measurements in your laboratory from past experiments), and assume that you want to detect an effect size of 2 mg/L. If your desired effect size is 2 mg/L and the standard is set at 30 mg/L, then

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

the highest mean BOD_5 concentration you can measure in the sample and still detect a significant difference from the regulatory threshold is 30 – 2 = 28 mg/L.

Excel

Note: This example is also available as an Excel spreadsheet.

Solution:

First, Cohen's *d* is found to be equal to 0.43, calculated as the difference between the regulatory threshold and the mean BOD_5 concentration (30–28 = 2), divided by the standard deviation (4.6). Therefore, $d = 2/4.6 = 0.43$.

C. 9

This is a one-sample, two-sided *t*-test (we use a two-sided test because we assume that the BOD_5 concentration could be greater or less than the regulatory value). See Chapters 9 and 10 about whether to use a one-sided test versus a two-sided test.

C. 10

Determining sample size is generally a trial and error process. Let us start with a typical sample size of $n = 10$, which is used by default in some labs. The non-centrality parameter (δ) is calculated by multiplying Cohen's *d* by the square root of the sample size. Therefore, $\delta = 0.43 \times \sqrt{10} = 1.36$.

The type II (beta) error is calculated by looking up the value of the non-central *t*-distribution table for the critical value associated with the alpha level and sample size chosen, as well as the non-centrality parameter, equal to Cohen's *d* multiplied by the square root of the sample size.

For a sample size of $n = 10$, the statistical power is only 23%. In order to achieve a power of 80%, we need to increase the sample size to at least $n = 43$ data points. Therefore, environmental samples should be collected approximately weekly in order to acquire at least 43 data points throughout the year. However, using the Excel spreadsheet, you can see that if you increase the number of samples, you will see that the power also increases (e.g., for a sample size of $n = 100$, the power is 99%).

Example

EXAMPLE 3.4 DETERMINE THE EFFECT LEVEL GIVEN A SAMPLE SIZE AND A DESIRED POWER

A stormwater authority is investigating the contribution of agricultural runoff to phosphate pollution in a stream during storm events. To do this, they plan to collect samples upstream and downstream of the agricultural field and measure the concentration of phosphate in the upstream and downstream locations. If the difference between paired phosphate concentrations upstream and downstream is significantly greater than zero, they will determine that the runoff from the agricultural site is contributing phosphate pollution to the stream.

Assume a standard deviation of 0.44 mg/L (i.e., the standard deviation of the differences between paired phosphate concentrations). What effect size can be detected at a significance level of 0.05 with 80% power if the stream is only sampled during three storm events (i.e., experiment done in triplicate)?

Excel

Note: This example is also available as an Excel spreadsheet.

Solution:

This is a two-sample paired *t*-test. We can assume it is one-sided because our hypothesis is that the agricultural site will contribute phosphate to the river, making the concentrations relatively higher downstream, rather than the opposite.

With these assumptions and our standard deviation of 0.44 mg/L, we can determine that a sample size of only $n = 3$ only allows us to detect a difference of 1.0 mg/L between upstream and downstream concentrations with a power of 80%.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

If we want to detect a smaller difference between the upstream and downstream concentrations, the power of the test will be lower. For instance, if we change the effect size to 0.5 mg/L, we see that the power goes down to only 40%. To maintain a power of 80% and detect a difference of 0.46 mg/L between upstream and downstream concentrations, we would need to sample at seven different storm events.

3.6 CHECK-LIST FOR YOUR REPORT

- ✓ Check that quality assurance and quality control measures are summarized in a chapter of your report or as a separate, stand-alone report. In particular, make sure that you address the scope of the study, the type and anticipated use of the data, any relevant assessment thresholds, standard operating procedures, quality control samples, and data storage and management protocols.
- ✓ Confirm that quality control is demonstrated as acceptable precision and accuracy through an initial demonstration of capability and through ongoing demonstrations of capability, performed quarterly at a minimum.
- ✓ Verify that sample locations and sample types (e.g., grab versus composite) are described in detail, with appropriate consideration for anticipated temporal and/or spatial variabilities.
- ✓ Check that sample matrix, sample volume or mass, sample analysis methods, sample containers, sample preservation, and maximum holding times are defined for each parameter to be analysed and summarized (preferably in a table).
- ✓ Verify that acceptable type I (alpha) error and type II (beta) error levels are established.
- ✓ Confirm that the desired effect size has been established as well as the anticipated standard deviation between samples.
- ✓ Verify that the sample size has been determined using a power analysis for the desired alpha and beta error levels.

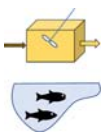
Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Chapter 4

Laboratory analysis and data management

This chapter discusses elements of importance when organizing, storing, reporting, publishing, and interpreting data obtained from laboratory analyses. The concepts of accuracy and precision, uncertainty and variability, and detection limits and significant digits are covered.



The contents in this chapter are applicable to both treatment plant monitoring and water quality monitoring.

CHAPTER CONTENTS

4.1 Raw Data, Calculated Values, and Statistics **70**
4.2 Storing Data and Calculated Values **72**
4.3 Metadata **80**
4.4 Accuracy and Precision **81**
4.5 Uncertainty and Variability **82**
4.6 Detection Limits **87**
4.7 Significant Figures **90**
4.8 Check-List for Your Report **94**

© 2020 The Authors. This is an Open Access book chapter distributed under the terms of the Creative Commons Attribution Licence (CC BY-NC-ND 4.0), which permits copying and redistribution for non-commercial purposes with no derivatives, provided the original work is properly cited (https://creativecommons.org/licenses/by-nc-nd/4.0/). This does not affect the rights licensed or assigned from any third party in this book. The chapter is from the book Assessment of Treatment Plant Performance and Water Quality Data: A Guide for Students, Researchers and Practitioners, Marcos von Sperling, Matthew E. Verbyla and Sílvia M. A. C. Oliveira (Authors)
doi: 10.2166/9781780409320_0069

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

4.1 RAW DATA, CALCULATED VALUES, AND STATISTICS

Basic

There are very important distinctions between data, calculated or computed values, and statistics. Many people often incorrectly refer to calculated values or statistics as 'data'. However, true **raw data** consist of direct observations.

As an example, let us consider a study of total solids concentrations at the influent and effluent of a sludge thickening unit process at a treatment facility, where samples are collected at the inlet and outlet in triplicate and analysed in duplicate. Briefly, the analysis requires weighing an empty dish, adding wet sludge (weighing the dish again), drying overnight at 103°C, then weighing the dish a third time. After analysis, solids concentrations are calculated as a mass/mass percentage, equal to m_{dry} (the mass of the solid residue remaining after drying) divided by m_{wet} (the sludge volume before drying). It is important to note that this percentage is **calculated** (see Table 4.2). The **original raw data** in this particular situation are the masses of the sludge before drying and the corresponding mass of the solid residue remaining after drying (Table 4.1). In order to measure sample-to-sample variability, samples are often collected and/or analysed in duplicate or triplicate, and **statistics** of those individual replicates are reported (Table 4.3). For example, the average (mean) per cent solids concentrations can be calculated from replicate measurements, and the standard deviations can help us understand variability between biological (field) and technical replicates. Other descriptive statistics can also be calculated, as shown in Chapter 5.

C. 5

What is the difference between biological (field) and technical replicates?

A **biological replicate** or a field replicate is when you collect more than one separate sample when you are out in the field. For example, if you were collecting sludge samples, you would fill multiple containers with sludge from the same sample collection point, one after another. The reason for analysing biological replicates is to understand the variability from a physical sample to a physical sample.

A **technical replicate** is when you analyse the same biological sample more than once in the laboratory. For example, if you had collected a single biological replicate of a sludge sample, once you return to the laboratory, you would perform the analysis more than one time for the same biological replicate. The reason for analysing technical replicates is to measure your variability from sample to sample based on the way you process the sample in the lab.

Finally, **replicate instrument readings** are when you take more than one reading for the same sample on the same instrument. For example, if you measure the pH of a water sample, it is common to take the reading three times, then compute the average of those three readings. The reason for doing this is to measure the precision of the instrument or device used to take the readings.

You should always **report the raw data** for laboratory analysis in the appendix of a report or in the supporting or supplemental information document of a publication. Ideally, you should also publish the raw data and archive it online with appropriate documentation. This way, if a reader or reviewer questions the calculated values being reported, they can always go back to the raw data and recalculate the values themselves. Other readers may want to complete a meta-analysis in the future – depending on the type of data you have and the type of statistical analysis being used to do the meta-analysis, they may need access to your raw data to do so.

Technically, **raw data** consist only of *direct observations*, and **calculated values** are *manipulations of raw data*. However, *some calculated values from the laboratory are often colloquially referred to as 'data'* even though they are calculated values, not direct observations. Some examples are total solids concentrations (the raw data are the weight measurements before and after drying) and nitrate

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Table 4.1 Example of raw data for calculating total solids concentrations in sludge samples.

Sample Location	Biological Replicate	Technical Replicate	Weight of Dish (g)	Weight of Dish + Sample (g)	
				Before Drying	After Drying
Inlet	1	1	30.124	74.077	31.048
Inlet	1	2	30.169	72.060	31.198
Inlet	2	1	30.183	72.059	31.221
Inlet	2	2	30.125	71.702	30.963
Inlet	3	1	30.155	71.292	31.153
Inlet	3	2	30.101	70.286	31.345
Outlet	1	1	30.151	69.303	31.697
Outlet	1	2	30.114	69.925	31.653
Outlet	2	1	30.148	67.713	31.728
Outlet	2	2	30.143	69.467	31.644
Outlet	3	1	30.163	71.058	31.582
Outlet	3	2	30.168	71.126	31.683

concentrations (the raw data are absorbance values, which are used to calculate estimated concentrations based on a standard curve produced from standard solutions). For the purposes of this book, *we will often use the word '**data**' even when referring to some calculated values of pollutant concentrations or loadings* (even though technically these do not constitute raw data).

Table 4.2 Example of calculated values of the total solids concentrations in sludge samples.

Sample Location	Biological Replicate	Technical Replicate	Weight of Sample Only[1] (g)		Per Cent Solids[2] (%)
			Before Drying	After Drying	
Inlet	1	1	43.953	0.924	2.10
Inlet	1	2	41.891	1.029	2.46
Inlet	2	1	41.876	1.038	2.48
Inlet	2	2	41.577	0.838	2.02
Inlet	3	1	41.137	0.998	2.43
Inlet	3	2	40.185	1.244	3.10
Outlet	1	1	39.152	1.546	3.95
Outlet	1	2	39.811	1.539	3.87
Outlet	2	1	37.565	1.58	4.21
Outlet	2	2	39.324	1.501	3.82
Outlet	3	1	40.895	1.419	3.47
Outlet	3	2	40.958	1.515	3.70

[1]Weight of sample only = (weight of dish + sample) – (weight of dish).
[2]Percent solids = weight of sample (after drying)/weight of sample (before drying).

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Table 4.3 Example statistics (means and standard deviations) for total solids concentrations in sludge samples.

Sample Location	Biological Replicate	Technical Replicate	Mean Values (%)		Standard Deviations (%)	
			Technical Replicates	Biological Replicates	Between Technical Replicates	Between Biological Replicates
Inlet	1	1	2.28	2.43	0.25	0.11
Inlet	1	2				
Inlet	2	1	2.25		0.33	
Inlet	2	2				
Inlet	3	1	2.76		0.47	
Inlet	3	2				
Outlet	1	1	3.91	3.83	0.06	0.11
Outlet	1	2				
Outlet	2	1	4.01		0.28	
Outlet	2	2				
Outlet	3	1	3.58		0.16	
Outlet	3	2				

4.2 STORING DATA AND CALCULATED VALUES

4.2.1 Where and how to store your data

The spreadsheet containing your data is likely to be large, and probably, because of this, it will not be included in the core of your report. If you are doing a thesis, dissertation, or technical report, a possibility is to include them in **Annexes** or **Appendices**. In the case of scientific publications in journals or conference proceedings, the available space is much more limited, and most likely it will not be included in the article, but often the publisher will allow you to publish the data as **supplementary material** (depending on its relative size), which is available to readers in the online version of some journals. Even if the publisher does not allow you to publish your data as supplementary material, you can still (and should still) publish it online and provide an **internet link** to your database. If you choose to publish your database on the internet, there are many repositories that you can choose from (for example, Zenodo, 4TU, Datacite, ICPSR, and DRYAD) that will publish your data online for free, and will even assign it a Digital Object Identifier (**DOI**), which is a unique alphanumeric string used to identify and provide a permanent link to your data on the web.

Therefore, when organizing your annexes, appendices, or supplementary material for a report or publication, it is important that you include **tables with your raw data** (direct observations) in addition to the calculated values.

Opening the access to your data can help avoid the duplication of efforts by others and can help disseminate scientific knowledge to those who otherwise would not have access to this type of information. Open data should ideally be stored according to the FAIR principles – that is, data should be **F**indable, **A**ccessible, **I**nteroperable, and **R**euseable (Wilkinson *et al.*, 2016). The box below provides a procedure for storing open data that satisfies the FAIR principles.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Advanced

Storing open data in accordance with the FAIR principles

Here is a recommended procedure that you can use to store your data for open access in accordance with the FAIR principles (Wilkinson *et al.*, 2016).

- Make your data ***findable*** by publishing it on an online open research repository (such as Zenodo, 4TU, Datacite, ICPSR, and DRYAD), including appropriate key words that facilitate searching, and accurately describing your data with rich metadata that have accurate and relevant attributes.
- Make your data ***accessible*** by providing them with a DOI, which is a unique alphanumeric string that provides a standard mechanism for accessing your data and retrieving metadata. Having a DOI associated with your published data also provides a permanent link to your data on the web. Many of the online open research data repositories listed above will automatically assign a DOI if you publish your data to their repository.
- Make your data ***interoperable*** by storing it as a .csv file with a single header row and using standard and broadly applicable language and vocabularies that describe the column headings and the data itself. For example, you should avoid the overuse of scientific jargon and include a description of technical terminologies used in the database and the metadata. Interoperability is defined by Wilkinson *et al.* (2016) as 'the ability of data or tools from non-cooperating resources to integrate or work together with minimal effort'. It refers to the ability of computer programs to work easily with your data. Storing data in a .csv file is easy if you are using Microsoft Excel. You simply choose the tab where your data are stored, click File – Save As, then click the dropdown box to select 'Comma Separated Values (.csv)' as the file format. Note that saving the data in this format causes you to lose some formatting (e.g., color, font, functions), and if you have multiple tabs in the Excel file, then each of them must be saved as their own .csv file. The advantages of using the .csv file format are that many programs recognize this format and can automatically import your data if it is stored in this format.
- Make your data ***reuseable*** by releasing it with an accessible data usage license that allows others to reuse your data for their own analysis. Some examples of licenses that enable this is Creative Commons (CC) BY-SA 4.0, which allows anyone to freely use your data as long as they attribute the authors (give appropriate credit and indicate if changes were made) and share alike (this means that anyone who uses your data must also use the same type of open license; they cannot claim copyright for themselves). More information about open access licensing can be found at creativecommons.org.

You also need to decide what **software program to use to store your data**. This partially depends on the size of the dataset. Data used in the assessment of treatment plant performance may come from the laboratory, or it may originate from online probes, sensors, and data loggers. In some cases, the dataset may become quite large. You need to organize the data in an appropriate way with an appropriate program that facilitates calculations, statistical analysis, and visualizations of the data. For most projects, you can use a spreadsheet software (such as Microsoft Excel). However, Excel will start to freeze up if you try to open very large data sets or spreadsheets with lots of calculations. In those rare cases, when you are working with some very large data sets, you may need to use a database software (such as MySQL), combined with some statistical software (such as R). The use of database software and advanced statistical software is beyond the scope of this book, and the subsequent chapters focus on analyses and procedures that can be done in a spreadsheet, such as Microsoft Excel.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

4.2.2 Storing data in a spreadsheet (most datasets)

Basic

If your dataset is small enough to store it and work with it in a **spreadsheet software** such as Microsoft Excel, then this is probably the most advantageous software tool to manage the data and complete the analysis. There are numerous advantages in using a spreadsheet software such as Microsoft Excel. Most importantly, most water engineers, scientists, managers, and practitioners are much more familiar and comfortable working with Excel than they are with databases and statistical analysis software. Excel comes with several built-in functions that enable the statistical analysis to be performed in the same spreadsheet file where the raw data are stored. Excel also comes with user-friendly plotting features, allowing the creation of graphs and figures (see Chapter 6). This allows you to store your data, calculated values, statistics, and plots to all be stored in a single file.

C. 6

If you are using a spreadsheet, in most cases, for each treatment plant or experimental unit you analyse, we recommend that you *organize your spreadsheet with the following fields (column headings)*:

- **Date**. The date for each measurement and sample collection. If samples were analysed on a different date than they were collected, then you should include two columns, one for 'Date of Sample Collection' and the other for 'Date of Analysis'.
- **Flow rate** (*measured*). This is an essential variable for treatment plant assessment. Influent flow is the most widely used one, but other relevant flows should also be included, if available (effluent flow, recycle flows, waste sludge flows, supernatant flows, etc.). See Chapter 2.

C. 2

- **Concentrations** (*measured or analysed*). The concentrations of major constituents or pollutants that need to be removed are an integral part of your work, and typically comprise values measured in the influent and effluent from your plant. If the treatment system is composed of units in series, ideally there should be data on the input and output from each stage. Of course, some constituents are more important than others, and may be monitored at higher frequencies. Depending on the question you are trying to answer, you may be measuring certain specific groups of pollutants which may require special methods of analysis. For certain analyses, additional raw data should be recorded in your spreadsheet in order to calculate the concentrations (e.g., absorbance values, fluorescence readings, etc.). In some cases, you may need to record the values of standards to generate a standard curve that is used to estimate the concentration in your samples. These essential raw data elements should be included in another tab of your spreadsheet.
- **Loads** (*calculated*). These are essential for mass balances and to analyse the loading conditions in your tank or reactor (surface and volumetric loading rates – see Chapter 13). Loads are the product of flow times concentration, as shown in Equation 4.1 (see also Section 2.1).

C. 13

S. 2.1

$$\text{Load} = \text{Flow} \times \text{Concentration} \tag{4.1}$$

- **Removal efficiencies of major constituents or pollutants** (*calculated*). This is also an integral part of your evaluation and is probably the most widely used variable for assessing treatment plant performance. Removal efficiency will be thoroughly discussed in Chapter 7, but its main concept is as simple as the one represented in Equation 4.2.

C. 7

$$\text{Efficiency} = \frac{\text{Influent concentration} - \text{Effluent concentration}}{\text{Influent concentration}} \tag{4.2}$$

- **Environmental and operational control variables** (*measured or analysed*). Environmental variables that may influence reaction rates, such as liquid temperature, pH, dissolved oxygen, alkalinity, and others should also be included, depending on the treatment process you are

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

investigating. Additionally, internal variables, specific for each tank or reactor (such as mixed liquor suspended solids, sludge blanket height, and chemical dosing rate), although not being part of the direct assessment of the effluent quality, are important elements in the performance of your treatment plant and should be incorporated in this database.

Your spreadsheet with the original measured data may be similar to the one shown in Table 4.4, with a simple structure. It may also contain calculated data, such as loads and concentrations, such as in the format exemplified in Table 4.5. Note that there are missing data, which is a very common situation in treatment plant monitoring (see Section 5.3).

S. 5.3

When making tables similar to Tables 4.4 and 4.5, pay attention to the following points:

- Make sure all **units** (e.g., m^3/d, g/m^3, kg/d, %) are included. You might find it helpful to recall that 1 mg/L is equivalent to 1 g/m^3. This can help you easily calculate loadings if your flow rates are measured in m^3/d.
- Report values with their suitable **significant figures** (number of **decimal places**). See Section 4.6.
- Leave **missing data as blanks** (do not put zero).
- Your **column for 'Date'** does not need to have all days of the year (01/01/2019, 02/01/2019, …), especially if monitoring is not carried out on a daily basis, and you would have a predominance of empty lines. If your data are collected on a daily frequency, then it is better to keep one line per day, that is, daily dates. If your frequency is on a weekly basis, then you should have one line per week, but always put the correct date. Similar comments are made for data obtained on a monthly or quarterly basis, or even on an hourly basis (in the latter case you will need one additional column for 'Hour of the day').
- If you also have measurements of the **flow rate at the effluent**, you can include a specific column for it, and also calculate output loads.
- If the effluent flow rate is substantially different from the influent flow rate, you should calculate **removal efficiencies based on input and output loads**, and not based on input and output concentrations (see Section 7.3.1). Make sure you make it clear which type of calculation you are doing.
- If you are analysing a single treatment unit (tank, reactor), instead of reporting loads, you can report **mass loading rates**, dividing the load by the surface area or volume of the unit (see Chapter 13).
- If your treatment plant or experiment has **units in parallel**, and if they are monitored separately, you will need one table like this one for each unit (alternatively, you can add more columns to the right and keep everything in the same table or spreadsheet, or you can reorganize your data table into a long format; see Section 4.2.3).
- If your plant has **units in series**, and if there is monitoring in-between the units, you will need one table like this one for each unit, knowing that the output from one unit is the input to the subsequent unit (alternatively, you can add more columns to the right and keep everything in the same table or spreadsheet or you can reorganize your data table into a long format; see Section 4.2.3).

S. 4.6
S. 7.3.1
C. 13
S. 4.2.3
S. 4.2.3

4.2.3 Storing data in a database (larger datasets)

Advanced

More and more, the availability of the internet, online data loggers, remote sensing, and other technological advances in data collection and management is creating the so-called **big data environment**, where datasets become extremely large and difficult to manage. If you are working with such large datasets that become cumbersome to work with in Excel, then some of the approaches and techniques described in this book may need to be implemented using a different more powerful computing software. Likewise, you might

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Table 4.4 Example of a simple spreadsheet for storing your measured data, organized in a chronological way.

Date	Flow (m^3/d) Influent	BOD (mg/L) Influent	BOD (mg/L) Effluent	COD (mg/L) Influent	COD (mg/L) Effluent	TSS (mg/L) Influent	TSS (mg/L) Effluent	Param 1 (units) Influent	Param 1 (units) Effluent	Param 2 (units) Influent	Param 2 (units) Effluent	Param 3 (units) Influent	Param 3 (units) Effluent
1/1/12	32,180												
1/2/12	32,470				35.2		4.8						
1/3/12	31,560			567	37.2	44	2.4						
1/4/12	37,486				46.5		4						
1/5/12	35,990	382	9	710	53	152	4.4						
1/6/12	39,398												
1/7/12	34,464												
1/8/12	32,522			356	34.4	44	4.4						
1/9/12	39,152				44.6		4.8						
1/10/12	38,489			517	49.1	72	5.2						
1/11/12	39,251				49.3		8.4						
1/12/12	36,934	231	23.8	513	48.7	62	15.6						
1/13/12	38,093												
1/14/12	32,149												
1/15/12	29,431			807	51.3	32	13.6						
1/16/12	37,785				57.1		9.6						
1/17/12	36,870			523	52.4	120	9.6						
1/18/12	37,872				49.8		5.6						
1/19/12	37,541	254	32.8	502	54.5	60	10.8						

Note: This table only shows flow rates and concentrations. Loads and removal efficiencies can be calculated from these concentrations and the flow rates using equations 4.1 and 4.2.

Table 4.5 Example of a simple spreadsheet for storing your measured data (flows and concentrations) and also your calculated data (loads and removal efficiencies).

Date	Inflow	Concentrations				Loads		Efficiency	
		Input Param 1	Output Param 1	Input Param *n*	Output Param *n*	Input Param 1	Input Param *n*	Param 1	Param *n*
	(m^3/d)	(g/m^3)	(g/m^3)	(g/m^3)	(g/m^3)	(kg/d)	kg/d)	(%)	(%)
dd/mm/yy									
dd/mm/yy									
dd/mm/yy									
…									
dd/mm/yy									

Param: parameter or constituent.

need to store your data in a slightly different format than the formats presented in the Excel worksheets associated with this book.

There are two fundamental ways to organize a data set containing concentrations of different constituents:

- The *first* is using the '**wide**' or '**short**' form, which is when multiple values for the same sample location are organized in a single row (Tables 4.6 and 4.8).
- The *second* way to organize a data set is using the '**long**' form, which is when there is only a single value reported in each row, and different sample dates or constituents are organized using column headings (Tables 4.7 and 4.9).

Table 4.6 Example of 'wide' data for BOD_5.

Regulated Discharge Point	Description	BOD_5 Concentration (mg/L)			
		JAN 2018	APR 2018	JUL 2018	OCT 2018
OUTFALL 1	Ocean outfall	10	13	12	13
OUTFALL 2	To reservoir	14	18	9	13

Table 4.7 Example of 'long' data for BOD_5.

Regulated Discharge Point	Description	Month	BOD_5 Concentration (mg/L)
OUTFALL 1	Ocean outfall	JAN 2018	10
OUTFALL 1	Ocean outfall	APR 2018	13
OUTFALL 1	Ocean outfall	JUL 2018	12
OUTFALL 1	Ocean outfall	OCT 2018	13
OUTFALL 2	To reservoir	JAN 2018	14
OUTFALL 2	To reservoir	APR 2018	18
OUTFALL 2	To reservoir	JUL 2018	9
OUTFALL 2	To reservoir	OCT 2018	13

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Table 4.8 Example of data that is 'long' with respect to the month but 'wide' with respect to the different constituents.

Regulated Discharge Point	Description	Month	Constituents		
			BOD_5 (mg/L)	TSS (mg/L)	NO_3-N (mg/L)
OUTFALL 1	Ocean outfall	JAN 2018	10	13	2.4
OUTFALL 1	Ocean outfall	APR 2018	13	13	3.3
OUTFALL 1	Ocean outfall	JUL 2018	12	14	2.7
OUTFALL 1	Ocean outfall	OCT 2018	13	13	2.8
OUTFALL 2	To reservoir	JAN 2018	14	11	3.0
OUTFALL 2	To reservoir	APR 2018	18	11	2.3
OUTFALL 2	To reservoir	JUL 2018	9	11	3.5
OUTFALL 2	To reservoir	OCT 2018	13	10	2.6

Table 4.9 Example of 'long' data with respect to the month and the parameters, but it is still not 'tidy' because there are multiple observation units in a single table.

Regulated discharge point	Description	Month	Parameter	Concentration
OUTFALL 1	Ocean outfall	JAN 2018	BOD_5 (mg/L)	10
OUTFALL 1	Ocean outfall	APR 2018	BOD_5 (mg/L)	13
OUTFALL 1	Ocean outfall	JUL 2018	BOD_5 (mg/L)	12
OUTFALL 1	Ocean outfall	OCT 2018	BOD_5 (mg/L)	13
OUTFALL 2	To reservoir	JAN 2018	BOD_5 (mg/L)	14
OUTFALL 2	To reservoir	APR 2018	BOD_5 (mg/L)	18
OUTFALL 2	To reservoir	JUL 2018	BOD_5 (mg/L)	9
OUTFALL 2	To reservoir	OCT 2018	BOD_5 (mg/L)	13
OUTFALL 1	Ocean outfall	JAN 2018	TSS (mg/L)	13
OUTFALL 1	Ocean outfall	APR 2018	TSS (mg/L)	13
OUTFALL 1	Ocean outfall	JUL 2018	TSS (mg/L)	14
OUTFALL 1	Ocean outfall	OCT 2018	TSS (mg/L)	13
OUTFALL 2	To reservoir	JAN 2018	TSS (mg/L)	11
OUTFALL 2	To reservoir	APR 2018	TSS (mg/L)	11
OUTFALL 2	To reservoir	JUL 2018	TSS (mg/L)	11
OUTFALL 2	To reservoir	OCT 2018	TSS (mg/L)	10
OUTFALL 1	Ocean outfall	JAN 2018	NO_3-N (mg/L)	2.4
OUTFALL 1	Ocean outfall	APR 2018	NO_3-N (mg/L)	3.3
OUTFALL 1	Ocean outfall	JUL 2018	NO_3-N (mg/L)	2.7
OUTFALL 1	Ocean outfall	OCT 2018	NO_3-N (mg/L)	2.8
OUTFALL 2	To reservoir	JAN 2018	NO_3-N (mg/L)	3.0
OUTFALL 2	To reservoir	APR 2018	NO_3-N (mg/L)	2.3
OUTFALL 2	To reservoir	JUL 2018	NO_3-N (mg/L)	3.5
OUTFALL 2	To reservoir	OCT 2018	NO_3-N (mg/L)	2.6

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

When working with **very large datasets**, the **'long' format** is the preferred way to store data, as it is more compatible with the use of advanced statistical computing software, facilitating your ability to manipulate, model, and visualize the data. If your **dataset is small** enough that you can store it all in a single Excel tab without slowing down the program too much, then it really *does not matter as much if you use the wide or long data format*. The **'long' data format** (e.g., Table 4.9) is also the preferred format for storing **raw data** in a .csv file for subsequent archiving or publication to an online repository.

If the **data set is very large**, it will need to be *stored in separate .csv files* or in a *linked database* (such as SQL), where the fields (column headings) from one table are linked in some way with other tables in the database. In this case, there are considerable advantages associated with cleaning up the data so that it is 'tidy' (Wickham, 2014) (Table 4.10). Tidy data saves storage space on the hard drive or server where it is located. For data to be 'tidy', the following three conditions must apply:

Table 4.10 Data that is 'tidy' because there is only one observation unit per table.

Regulated Discharge Point	Description
OUTFALL 1	Ocean outfall
OUTFALL 2	To reservoir

Regulated Discharge Point	Month	Parameter	Concentration
OUTFALL 1	JAN 2018	BOD_5 (mg/L)	10
OUTFALL 1	APR 2018	BOD_5 (mg/L)	13
OUTFALL 1	JUL 2018	BOD_5 (mg/L)	12
OUTFALL 1	OCT 2018	BOD_5 (mg/L)	13
OUTFALL 2	JAN 2018	BOD_5 (mg/L)	14
OUTFALL 2	APR 2018	BOD_5 (mg/L)	18
OUTFALL 2	JUL 2018	BOD_5 (mg/L)	9
OUTFALL 2	OCT 2018	BOD_5 (mg/L)	13
OUTFALL 1	JAN 2018	TSS (mg/L)	13
OUTFALL 1	APR 2018	TSS (mg/L)	13
OUTFALL 1	JUL 2018	TSS (mg/L)	14
OUTFALL 1	OCT 2018	TSS (mg/L)	13
OUTFALL 2	JAN 2018	TSS (mg/L)	11
OUTFALL 2	APR 2018	TSS (mg/L)	11
OUTFALL 2	JUL 2018	TSS (mg/L)	11
OUTFALL 2	OCT 2018	TSS (mg/L)	10
OUTFALL 1	JAN 2018	NO_3-N (mg/L)	2.4
OUTFALL 1	APR 2018	NO_3-N (mg/L)	3.3
OUTFALL 1	JUL 2018	NO_3-N (mg/L)	2.7
OUTFALL 1	OCT 2018	NO_3-N (mg/L)	2.8
OUTFALL 2	JAN 2018	NO_3-N (mg/L)	3.0
OUTFALL 2	APR 2018	NO_3-N (mg/L)	2.3
OUTFALL 2	JUL 2018	NO_3-N (mg/L)	3.5
OUTFALL 2	OCT 2018	NO_3-N (mg/L)	2.6

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

- *Each variable should form a column*. In the example shown in Table 4.10, the three independent variables are sample location, month, and parameter, and the dependent variable is concentration.
- *Each observation should form a row*. In the example shown in Table 4.10, each observation is a measurement of a concentration made in the laboratory. Some measurements were for BOD_5, some were for total suspended solids (TSS), others were for NO_3-N. But each measurement is organized in its own row.
- *Each type of observational unit forms a table*. In the example shown in Table 4.10, the descriptions of the regulated discharge points are organized in a different table, which is linked to the concentration table by the discharge point field (e.g., OUTFALL 1 or OUTFALL 2). Unlike Table 4.9, the description of the regulated discharge point does not have to be repeated in multiple rows.

There is always an *opportunity to make a dataset more* 'tidy' (and thus use less storage space) whenever you see the same pair of values from two different columns being repeated for all rows in the spreadsheet (e.g., in Table 4.9, OUTFALL 1 always has the description 'Ocean outfall' and OUTFALL 2 always has the description 'To reservoir'). In addition to saving storage space, another advantage of having tidy data is that when data is stored in databases in a 'tidy' format, the processing speeds for web applications that draw from the data can be much faster.

4.3 METADATA

Advanced

Your data need to be *well organized and described*. It is essential to provide sufficient documentation of your data so that it can be easily understood by someone who is not familiar with the project or the monitoring activity. In addition to your data spreadsheet or database, you should also produce **metadata** and a **data dictionary**. These two resources help describe your data set and provide documentation to others who might be interested in using your data.

Metadata is a resource that provides information about other data. It should be prepared by an information technology specialist, as it requires some computer coding. There are several different types of metadata, such as descriptive metadata, structural metadata, administrative metadata, reference metadata, and statistical metadata. The data you collect, store, and archive for the study of water or wastewater treatment processes should include **descriptive metadata**, which is a type of metadata that describes a resource (such as your data set) to help other people discover and identify it. Descriptive metadata includes elements such as the title of your data set, an abstract that describes the project or purpose for collecting the data, the author(s) of the data set, and some keywords. So, if you do not have a title for your data set, you should create one! Table 4.11 shows an example of information that would be included in metadata for a spreadsheet containing data on total suspended solids (TSS), biochemical oxygen demand (BOD), and chemical oxygen demand (COD) loadings at a wastewater treatment facility.

You need also to make sure that your data have sufficient documentation of **Quality Assurance and Quality Control (QA/QC)** measures used in the study. If you are using a spreadsheet, you should be **very explicit regarding the units of measurement** for each data element (e.g., ppm, mg/L, µg/L, mg/kg, meq/L, percentage by volume, percentage by mass, SI units, etc.). For example, for the hypothetical data set described in Table 4.11, you might include a tab at the beginning of the spreadsheet like a header page that contains information about the standard laboratory methods used for TSS, BOD, COD, and thermotolerant coliforms (TTC) analysis, units reported, and QA/QC measures such as positive and negative controls, and limits of detection.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Table 4.11 Example metadata for a data on a study of TSS, BOD, COD, and coliform loadings at a wastewater treatment facility.

Metadata element	Value
Title	Solids and Organic Loadings at the XYZ Wastewater Treatment Facility
Creator(s)	Marcos von Sperling, Silvia Oliveira and Matthew Verbyla
Keywords	Wastewater treatment Total suspended solids (TSS) Biochemical oxygen demand (BOD) Chemical oxygen demand (COD) Thermotolerant coliforms (TTC)
Description	This spreadsheet contains measured influent flow rates and concentrations of TSS, BOD, COD, and TTC at the influent and final effluent points of the XYZ Wastewater Treatment Facility in Brazil. Loadings are also calculated from the flow rate and concentration data.
Publisher	Universidade Federal de Minas Gerais (UFMG)
Contributors	Jane Doe and John Doe contributed by collecting the samples and completing the laboratory analyses which generated the data. Marcos von Sperling, Silvia Oliveira and Matthew Verbyla designed the study, supervised the lab work, and reviewed the final spreadsheet of raw data and calculations.
Date Published	November 1, 2019
Type	Dataset
Format	Spreadsheet (.xlsx)
Source	Original laboratory analysis
Language	Brazilian Portuguese
Coverage	January 1, 2014 through December 31, 2018
Rights	CC BY-SA 4.0

4.4 ACCURACY AND PRECISION

Basic

It is not uncommon to hear people use the terms '**accurate**' and '**precise**' incorrectly, and sometimes people even (incorrectly) use them interchangeably! These two terms have very specific definitions and they mean very different things. Both are important for you to know when collecting and reporting your data (see Figure 4.1 and Example 4.1).

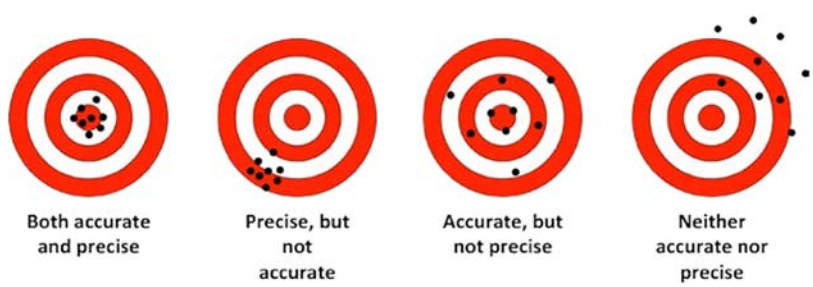


Figure 4.1 Illustration of accuracy and precision.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Accuracy is a *measure of how close the measured values of water quality parameters are to the true value* in the entire treatment system or water body over a defined time period. For example, suppose that over the course of 24 hours, a wastewater facility receives 10,000 m^3 of sewage containing 2500 kg of suspended solids. We collect a 1-litre, 24-hour composite sample of water and analyse it for TSS. If our sample was completely representative and our measurement was perfectly accurate, we should measure a concentration of 250 mg/L (2500 kg/10,000 $m^3 = 0.25$ kg/$m^3 = 250$ g/m^3). In practice, there is often no good way to measure the accuracy of a laboratory measurement, especially for analyses that require the use of a standard curve.

Precision is a *measure of whether repeated samples collected will show the same results, assuming that conditions are the same*. For example, if we collected a 5-litre, 24-hour composite sample, mixed it, then split it into 5 equal parts of 1 litre each and analysed each litre separately for TSS, we should get 250 mg/L in each of the 5 samples to have perfect precision. However, in reality there is some variability in our methods, and we may record slightly greater than 250 mg/L in some samples and slightly less than 250 mg/L in others. The closer the values are to each other, the more precise they are. In practice, the precision of measurements is assessed by performing repeated measurements on sample replicates, and calculating the standard deviation, variance, and standard error.

4.5 UNCERTAINTY AND VARIABILITY

4.5.1 Variability of a population

Water bodies and water and wastewater treatment systems often operate as continuous flow systems, meaning that water is always flowing in and always flowing out. The quality of the water is constantly changing throughout the course of a day (diurnal fluctuations) and throughout the year (seasonal fluctuations). This is called the natural variation in water quality. Even if we had a perfect sensor to detect the exact concentrations of water quality constituents at all times of the day and throughout the entire year, we would see these changes in concentrations. This is known as the **variability of the population** (where the population in this case is the entire body of water flowing through a system). So, we can say that **variability** describes the natural temporal and/or spatial changes in water quality within a system. Variability is measured by the **standard deviation**. Understanding the variability of a water or treatment system can help us to predict the range of probable values obtained from the analysis of future samples. This range is known as the **prediction interval**.

4.5.2 Uncertainty in our estimate of parameters

It is *impossible to measure the exact mass of pollutants in all of the water that passes through a given system*. Therefore, *we collect **samples** of water* and analyse them, in order to *gain some insight* or *make inference* about what the true value might be. Suppose you want to quantify the mass of some pollutant in a water system. So, you collect samples, measure their concentrations, then calculate the average value. However, the average you calculate is going to be different than the average calculated by another person who collects the same number of samples from the same system and completes the same analysis. In fact, if many people collect samples from the same system and perform the same analysis and compute average values from their own data sets, each person will find slightly different estimates of the average value, due to the fact that their estimates have uncertainty. Thus, **uncertainty** describes our lack of knowledge about the true value of a parameter calculated from our data, due to the fact that *our data were generated from samples collected from a larger population*.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

The samples we collect from a water system only represent a small fraction of the water that is passing through. For example, if you are assessing the performance of a system that is continuously receiving, treating, and discharging wastewater, you may collect samples daily or weekly, but *you are missing all of the water that flows through the system in between sample collection*. Even for a batch reactor, if you collect a sample from every batch, you are only able to collect a small volume of water from the entire reactor. Because of this, we can never be 100% positive that our estimate of that average concentration is exactly equal to the true concentration (which will always remain unknown to us). There is always some degree of uncertainty, partly because of the natural variability of the population (e.g., Section 4.5.1), but also due to *the indiscriminate nature of random sampling and the inherent limitations associated with methods used to measure water quality constituents*. Uncertainty is not limited to our estimate of the mean; it is also true of the standard deviation and other statistics such as percentiles (see Chapter 5). When we calculate these statistics using our data set, it is only an estimate of the true values of the population.

S. 4.5.1

C. 5

C. 10

C. 11

We can measure uncertainty in our estimate of the mean using the following statistics: the **standard error of the mean,** the **margin of error**, and the **confidence interval**. You will learn more about these concepts in Chapters 10 and 11.

4.5.3 The central limit theorem and confidence intervals

The **central limit theorem** tells us that if many people were to collect random samples from the same population, they would all get different values for the average (due to randomness), but these **average values would follow a normal distribution with the mean being equal to the true population mean**. This allows us to calculate **confidence intervals** which tell us the *probability that the true mean of the population is within a certain distance of our calculated sample average*.

This is a challenging concept to understand, so consider this analogy. Imagine there are 100 students in an environmental engineering laboratory class. The instructor is teaching the students how to measure the concentration of total dissolved solids in a water source. For the laboratory, the instructor prepares a synthetic water source that the students can use to practice their analysis, by adding 1000 g of dissolved solids to 10,000 L of deionized water (thus, the *true* and *exact* average concentration of dissolved solids in the water is $1000\ \text{g}/10{,}000\ \text{L} = 0.1\ \text{g/L} = 100\ \text{mg/L}$). Then, the instructor asks each student to complete exactly the same experiment using exactly the same water source, collecting 25 water samples each and measuring the concentration of dissolved solids. Assume, for this analogy, that the laboratory measurement method is perfectly accurate. However, because of natural randomness, natural variability, and the inherent imprecision associated with measurements, each student would produce a data set with slightly different values. Likewise, when each student calculates the average of the 25 values in their data set, they will each get slightly different average values. However, most students' calculated averages would cluster close to the true average of 100 mg/L. In fact, the students' calculated ***average values*** *would follow a normal distribution according to the central limit theorem*.

If the students each calculate **confidence intervals** (say, *95% confidence intervals*) around their estimates of the average concentration, then the confidence intervals of 95 of the students (on average 19 out of every 20 students) would include the true mean concentration of 100 mg/L. The confidence intervals of five students (on average 1 out of every 20) would not include this value (these students would experience what is known as **α** or **type I error**). If the students calculated 99% confidence intervals (instead of 95%), then only one student's confidence interval would not include the true mean concentration of 100 mg/L (on average). *The width of confidence intervals is dependent on the variability in the population and on the sample size*. In fact, the **confidence interval is directly proportional to the standard error,**

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

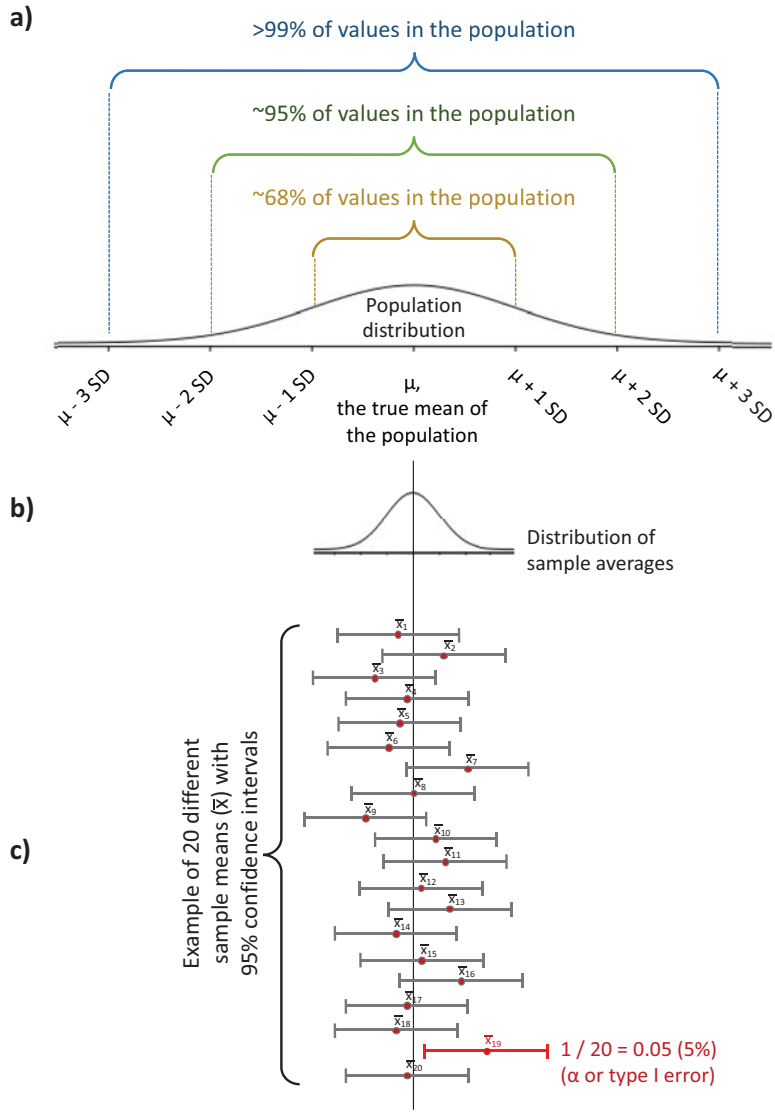


Figure 4.2 Graphical depiction of the difference between the population distribution, the distribution of sample averages, and confidence intervals: (a) the 68–95–99 rule states that for a normally distributed population, ~68% of the values are within one standard deviation from the mean, ~95% are within two standard deviations from the mean, and >99% are within three standard deviations from the mean; (b) the central limit theorem tells us that if many people were to randomly sample the same system, they would calculate slightly different average values, and the distribution of those average values follows a normal distribution centred on the true population mean (even if the population distribution is not normal); and (c) if 20 experiments are performed and 20 different data sets are collected, the 95% confidence intervals around the average values of those data sets will include the true population mean 19 out of 20 times on average ($19/20 = 95\%$).

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

which is equal to the standard deviation divided by the square root of the sample size. So, the standard error (and the confidence interval) always gets smaller and smaller as our sample size increases. So, *if we increase our sample size, we become more and more confident about the range of the true mean of the population*.

Figure 4.2 shows the *relationship between the population distribution, the distribution of the sample averages, and an illustration showing how the 95% confidence intervals from 20 different experiments overlap the true population mean*. The nice thing about confidence intervals and the **central limit theory** is that *even if the population distribution is **not normal**, as long as the **sample size is large (>30)**, the sampling **distribution of the mean** will still follow a **normal distribution*** (for an excellent and concise discussion that expands upon this topic, see Krzywinski & Altman, 2013).

4.5.4 Prediction intervals and confidence intervals

A *prediction interval* is *different from a confidence interval:*

- A **confidence interval** tells you the *probability of calculating an average value from your sample that includes the true mean value of the population*.
- A **prediction interval** tells you the *probability of your next sample producing a value in-between a given range*.

If we know that our population distribution is normal, then we can use the prediction interval to monitor and manage the quality of treatment systems (see Section 9.8 Control Charts for more detail and examples). The **68–95–99 rule** states that if the population is *distributed normally*, we can assume that ~68% of the values of future samples will fall within one standard deviation of the mean, ~95% will fall within two standard deviations of the mean, and >99% will fall within three standard deviations of the mean (see Figure 4.2).

S. 9.8

The use of prediction intervals is especially beneficial if a system is being regulated based on the concentration of a pollutant measured *in any single sample*. For example, if a drinking water treatment facility is required to analyse samples monthly and ensure that *none of the samples* has a benzene concentration above 0.010 mg/L, then we should make sure that the average benzene concentration over the course of the year is at least two or three standard deviations lower than this threshold. Whether you use **two standard deviations** (i.e., the *95% prediction interval*) or **three standard deviations** (i.e., the *99.7% prediction interval*) is a question of how big of a risk you are willing to take to comply with the regulatory standards. See Sections 9.3 (Compliance with standards) and 9.8 (Control charts) for a more detailed discussion about this concept with respect to monitoring compliance with standards, regulations, or target values.

S. 9.3

S. 9.8

The use of confidence intervals (instead of prediction intervals) is beneficial if a system is being regulated based on the *average* concentration of a pollutant measured in a set of samples. For example, if a drinking water treatment facility is required to ensure that the *mean* benzene concentration in a water supply system from 12 annual samples is significantly below 0.005 mg/L, then we should use hypothesis testing to make sure that our sample average is significantly below the regulatory limit. We can also calculate the confidence interval and make sure that the upper limit of that interval is below the regulatory threshold. Whether you use the 95% confidence interval, the 99% confidence interval, or some different confidence interval (i.e., 99.9%) is a question of how big of a risk you are willing to take to comply with the standards. See Chapter 10, and especially Section 10.2, for a background discussion of these topics and their application for hypothesis testing.

C. 10

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Example

EXAMPLE 4.1 APPLYING CONCEPTS OF VARIABILITY AND UNCERTAINTY: PREDICTION AND CONFIDENCE INTERVALS

You are managing a wastewater treatment facility that has to comply with a discharge permit that specifies a maximum limit for the effluent concentration of total suspended solids (TSS) at 50 mg/L. The regulation specifies two ways with which you must comply with the limit:

- *Samples must be collected once per month.*
- *The **mean annual** TSS concentration in the effluent must be significantly below 50 mg/L.*
- *The TSS concentration on **any single monthly sample** must not exceed 75 mg/L.*

The table below shows results from samples collected from effluents from two alternative treatment processes, Process A and Process B. To study the two processes and collect a large number of data points, you collect samples weekly, even though once you select a process and move forward with the implementation, the permit will only require you to monitor it on a monthly basis. Based on these results, which treatment process would you recommend using if you wanted to comply with the permit requirements? Assume the measured TSS concentrations are normally distributed.

Excel

Note: this example is also available as an Excel spreadsheet.

Data:

ID	Process A	Process B	ID	Process A	Process B	ID	Process A	Process B
1	42.1	25.1	19	39.1	37.8	37	40.3	25.3
2	42.9	46.0	20	44.4	24.2	38	38.2	15.8
3	46.5	21.3	21	35.9	33.6	39	45.7	49.1
4	52.4	42.2	22	50.2	33.7	40	45.5	7.8
5	40.7	49.2	23	44.8	21.3	41	37.8	45.5
6	35.7	38.7	24	45.5	15.7	42	52.3	17.9
7	37.3	37.8	25	46.2	47.9	43	47.8	44.9
8	44.9	65.6	26	38.7	74.9	44	50.9	50.7
9	55.4	14.2	27	42.2	52.6	45	39.4	45.7
10	41.2	73.4	28	43.4	24.7	46	42.5	36.9
11	48.6	46.3	29	39.8	63.2	47	52.8	24.6
12	50.7	73.3	30	50.1	29.6	48	39.4	35.0
13	53.4	27.4	31	44.9	72.1	49	53.5	27.4
14	50.4	49.4	32	39.1	62.4	50	47.4	32.0
15	32.2	38.0	33	43.9	71.7	51	48.3	38.2
16	48.0	51.9	34	47.9	13.6	52	39.3	58.0
17	44.8	13.5	35	45.0	28.4			
18	39.7	57.2	36	43.5	53.8			

Solution:

This is a problem of different levels of **precision and variability** between the two data sets, and our uncertainty in the average value of the data set. We should first recognize the two different types of

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

regulations, the first which is based on the **average annual concentration** and the second which is based on **a single sample**.

To evaluate our data against the *first type of regulation*, we calculate the average of each sample and the **95% confidence interval**, and compare the upper confidence limit to the standard value of 50 mg/L.

To evaluate our data against the *second type of regulation*, we calculate the values of the **95% prediction interval** and compare the upper prediction limit against the standard value of 75 mg/L. The results are shown below.

C. 9
C. 10

Note: this is a very quick introduction to the concept of confidence intervals and prediction intervals as they relate to the application of assessing compliance. You may have to review Chapters 9 and 10 for a much more detailed overview on these topics and concepts.

	Process A	Process B
Average	44.5	40.1
Standard deviation	5.34	17.9
Sample size	52	52
Standard error	0.740	2.48
Confidence level	95%	95%
Lower confidence limit	43.0	35.3
Upper confidence limit	45.9	45.0
Prediction level	95%	95%
Lower prediction limit	33.8	4.4
Upper prediction limit	55.1	75.8
Complies based on AVERAGE?	TRUE	TRUE
Complies based on MAXIMUM?	TRUE	FALSE

These results indicate that although Process B produced a lower average concentration, it produced results with more variability (as seen by the higher standard deviation). The upper confidence limits for the estimate of the averages were below the threshold of 50 mg/L for both processes, but Process B had an upper prediction limit that exceeded the threshold of 75 mg/L (even though no single sample from the set of 52 produced a value above that limit). The prediction interval of Process A is entirely below the limit of 75 mg/L. This evidence should lead us to choose Process A over Process B, as it will have a higher probability of complying with both types of regulatory limits.

As mentioned previously, this is a very quick introduction to an application of assessing compliance based on precision, uncertainty, and variability. For a more in-depth discussion on these topics, see Chapters 9.

C. 9

4.6 DETECTION LIMITS

4.6.1 Variability from instruments and sample processing

Basic

Treatment plant engineers have as a major objective the removal of pollutants from the water. As a result, they naturally encounter very low concentrations of pollutants in water and other environmental samples,

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

particularly the effluent samples from a treatment facility. The types of laboratory analyses used to study, monitor, and evaluate treatment processes often involve analytical chemistry or microbiological methods. *These methods have inherent limitations when concentrations are low in the sample being collected.*

A **detection limit** is the *lowest quantity or concentration of a pollutant that can be reliably measured in a sample and distinguished from a sample with an absence of that pollutant*. There are many different types of detection limits and unfortunately, there is often much confusion about the meaning of these limits. For analytical chemistry methods, often a sample has to be extracted and processed using some procedures, and then fed into the instrument to collect a reading. All of the relevant detection and quantification limits come down to manipulations of one of the following two standard deviations (Sawyer *et al.*, 2003):

- The **instrument blank standard deviation** is the standard deviation of repeated measurements taken from directly feeding the instrument a series of blanks (typically DI water). Let us call this s_b.
- The **process blank standard deviation** is the standard deviation of repeated measurements taken after a blank sample is processed (e.g., extracted). Let us call this s_p. In almost all cases, s_p will be larger than s_b (there is more variability due to the multiple steps involved in sample processing).

4.6.2 Limits of detection and quantification

The following definitions are similar to those provided by the *Standard Methods for the Examination of Water and Wastewater* (APHA *et al.*, 2017), with some additional explanations about the statistics behind them. These definitions mostly pertain to water and wastewater analysis methods that are based on analytical chemistry (there are slight differences in the way limits of detection are calculated for microbiology samples).

- The **instrument detection limit (IDL)**, also called the *limit of the blank* is the concentration of the pollutant that produces a signal that is five times the signal-to-noise ratio of the instrument. It is often estimated by adding the average instrument blank signal to the product of 1.645 by the standard deviation of instrument blanks, where 1.645 is the inverse of the standard normal distribution for a probability of 95%. See Example 4.2.
- The **lower limit of detection (LLD)** is equal to the concentration of the pollutant in reagent water that produces a signal that is equal to twice the IDL (i.e., 3.29 standard deviations above the average of instrument blank signals).
- The **method detection limit (MDL)** is the concentration of the pollutant that produces a signal that is different from a blank signal with 99% probability. The norm is to calculate a prediction interval based on the standard deviation from seven process blank replicates (i.e., $n = 7$), the prediction interval is calculated as the product of the process blank standard deviation and the left-tailed inverse of the Student's t-distribution with a probability of 99% and 6 degrees of freedom – the number of degrees of freedom for the t-distribution is equal to the number of process blank replicates minus one (i.e., $n–1 = 7–1 = 6$). If more or fewer process blank replicates are run, then the calculation of the MDL is adjusted accordingly by choosing a different number of degrees of freedom. See Example 4.3.
- The **limit of quantification/quantitation (LOQ)** is defined as the lowest concentration of a pollutant that produces a signal that can not only be reliably detected but that can also meet predefined goals for precision and accuracy. These predefined goals are commonly taken as having a coefficient of variation (CV) equal to or below 20% (Armbruster & Pry, 2008). It is often estimated as ten standard deviations above the average signal from blanks, but it should be verified using positive controls (spiked controls) at low levels and calculating the CV from replicate measurements.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Other terminologies that are used in some contexts are the *reporting limit (RL)* or the *practical quantification limit (PQL)*, which are taken to be the lowest level that can be quantified during normal operations.

Example

EXAMPLE 4.2 INSTRUMENT DETECTION LIMIT

Consider the following results from analysing 7 instrument blanks:

−0.7 1.2 −0.4 −0.1 0.6 0.2 1.3

What is the instrument detection limit (IDL) at 95% confidence?

Excel

Note: this example is also available as an Excel spreadsheet.

Solution:

Start by calculating the instrument blank standard deviation. In Excel, this can be done using the STDEV function.

$= \text{STDEV}(-0.7, 1.2, -0.4, -0.1, 0.6, 0.2, 1.3) = 0.77$

Then, calculate the inverse of the normal distribution (with a probability of 95%). In Excel, this can be done using the NORM.S.INV function:

$= \text{NORM.S.INV}(0.95) = 1.645$

Calculate the average of the instrument blank readings:

$= \text{AVERAGE}(-0.7, 1.2, -0.4, -0.1, 0.6, 0.2, 1.3) = 0.3$

Add the product of those two values to the average instrument blank reading to get the IDL at a confidence level of 95%.

$\text{IDL}_{95\%} = 0.77 \times 1.645 + 0.3 = \mathbf{1.6}$

Example

EXAMPLE 4.3 METHOD DETECTION LIMIT

Consider the following results from analysing 7 process blanks:

4.2 4.9 5.2 4.7 4.5 4.5 4.3

What is the method detection limit (MDL) at 99% confidence?

Excel

Note: this example is also available as an Excel spreadsheet.

Solution:

Start by calculating the process blank standard deviation. In Excel, this can be done using the STDEV function.

$= \text{STDEV}(4.2, 4.9, 5.2, 4.7, 4.5, 4.5, 4.3) = 0.35$

Then, calculate the left-tailed inverse of the *t*-distribution (with $7 - 1 = 6$ degrees of freedom and a probability of 99%). In Excel, this can be done using the T.INV function:

$= \text{T.INV}(0.99,6) = 3.143$

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Calculate the average of the process blank readings:

$= \text{AVERAGE}(4.2, 4.9, 5.2, 4.7, 4.5, 4.5, 4.3) = 4.6$

Add the product of those two values to the average process blank reading to get the MDL at a confidence level of 99%.

$\text{MDL}_{99\%} = 0.35 \times 3.143 + 4.6 = \mathbf{5.7}$

4.7 SIGNIFICANT FIGURES

4.7.1 Significant figures for direct measurements from instruments that give live readings

Basic

The number of significant figures you report for your laboratory analytical results should reflect the level of precision you have in your laboratory measurements. For **direct observations**, this is as simple as determining the number of stable digits provided by the machine being used to analyse the sample.

For example, suppose you are measuring the mass of a sludge sample using a precision balance (which gives a live reading), this is as simple as determining for how many digits the reading on the balance stays stable. If the balance reads 1.3537 g, but then the reading jumps down to 1.3534 g, and then jumps up to 1.3539 g, then back down to 1.3536 g, then you should report the value as 1.353 g (with four significant digits).

4.7.2 Significant figures for direct measurements from instruments that do not give live readings

Basic

If an instrument does not give live readings, but instead provides a **static number** each time you analyse the sample, there still might be too many digits in the number being reported. If that is the case, then you can determine how many digits to report based on the per cent change that occurs between different measurements made on the same sample using the same equipment with the same settings. Assuming your data are normally distributed, here is a step-by-step protocol that you can use to determine the number of significant figures that should be reported in this case. See Example 4.4.

- Start with the standard deviation (s) of replicate measurements of the same sample on the same piece of equipment and determine the variability associated with that measurement. For example, you can use the instrument blank standard deviation.
- Multiply that standard deviation by 3 (i.e., $3s$, where s_b is the standard deviation from Step 1). This is one-half of the >99% prediction interval for determinations made using the equipment in question.

Why do we use 3s? Based on the 68–95–99 rule, we can estimate that almost all values reported will occur within a range that is no lower than three standard deviations below the mean and no higher than three standard deviations above the mean.

- Subtract $3s$ from the value that you want to report and take note of how many digits remain the same.
- Add $3s$ to the value that you want to report and take note of how many digits remain the same.
- Report all of the digits that remain the same in steps 3 and 4, plus one additional digit.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Example

EXAMPLE 4.4 SIGNIFICANT FIGURES

When using methods that require the use of a machine, such as spectrophotometry, it is common to take replicate readings for a single sample and use the average of those replicate readings as your 'data point' for that single sample. Consider the following results that are obtained from taking a reading of the same standard sample vial a total of 5 times using the same machine and the same settings.

3.2156 3.2159 3.2160 3.2161 3.2155

Determine how many significant figures should be reported for this single reading.

Excel

Note: this example is also available as an Excel spreadsheet.

Solution:

The average value of those five readings is 3.215820000… and the standard deviation is 0.00025884358211… (the digits will go on forever if we do not round).

Using the >99% prediction interval, if we add 3*s* to the average value, we get an upper limit of 3.21659653075… and if we subtract 3*s* to the average value, we get a lower limit of 3.21504346925… The first three digits of these upper and lower limits are the same (i.e., 3.21, and then the digit after that changes.

Therefore, we can report a total of four significant figures for this reading (i.e., we round 3.215820000… to the fourth digit to obtain a value of **3.216**).

4.7.3 Significant figures for calculated values based on standard curves

Advanced

For other analytical methods of water analysis, the value you are reporting may not be a direct observation, but a **calculated value**. For example, suppose you are analysing water samples for nitrate concentrations using a method that uses a spectrophotometer to measure absorbance. You might report the absorbance in the appendix of the report, but what you are really interested in finding out is the nitrate concentration, which is obtained from a standard curve that is constructed by analysing standard solutions of known concentrations. In this case, you should use the linear regression equation for the standard curve to calculate the concentration (e.g., nitrate) based on the analytical measurement (e.g., absorbance).

This calculated value may have an endless number of digits. However, you can determine the upper and lower prediction limits of this calculated value using the prediction interval of the regression curve. If you compare these upper and lower limits to see how many digits are the same, then you can report those digits plus one additional digit after them as significant digits. See Example 4.5. Also, Chapter 11 has more information about how to calculate a regression curve with a confidence interval and a prediction interval.

C. 11

Example

EXAMPLE 4.5 STANDARD CURVE

Suppose you are measuring total nitrogen in a wastewater sample, and you measure an absorbance value of 0.202. Suppose you also analysed (in triplicate) standard solutions of 10, 20, 30, 40, and 50 mg/L of total nitrogen, and got the following absorbance values:

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Standard solution Total N

Absorbance	Replicate	Concentration (mg/L)
0.0810	*1*	*10*
0.0757	*2*	*10*
0.0696	*3*	*10*
0.1294	*1*	*20*
0.1151	*2*	*20*
0.1207	*3*	*20*
0.1782	*1*	*30*
0.1640	*2*	*30*
0.1546	*3*	*30*
0.2045	*1*	*40*
0.2056	*2*	*40*
0.2095	*3*	*40*
0.2481	*1*	*50*
0.2415	*2*	*50*
0.2469	*3*	*50*

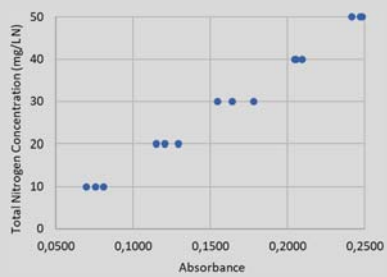
First, plot the standard curve, using the absorbance readings as x-values and the total nitrogen concentrations as y-values. Then, determine the corresponding regression equation and R^2 value. Use the regression equation to calculate the concentration of total nitrogen in the wastewater. Finally, calculate the confidence interval of the regression for the standard curve, and use it to determine the number of significant digits that should be reported for the total nitrogen concentration in the wastewater sample.

Excel

Note: this example is also available as an Excel spreadsheet.

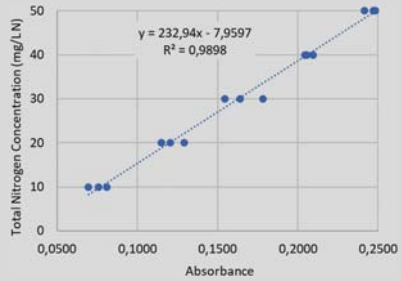
Solution:

First, we start by plotting the data for the standard solutions, with the nitrogen concentrations on the *Y*-axis and the absorbance on the *X*-axis. The reason for plotting the nitrogen concentrations on the *Y*-axis instead of the *X*-axis is because once we develop a regression curve for these standards, we will use the absorbance values as inputs to the model to estimate the nitrogen concentration in the unknown sample. If the nitrogen concentration is treated as the response variable, then we can also estimate the confidence and prediction intervals for the estimated concentration.



Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Now, use the Excel 'Add Trendline' feature to find the best fit linear regression curve for the data, display the equation and R^2 value on the chart.

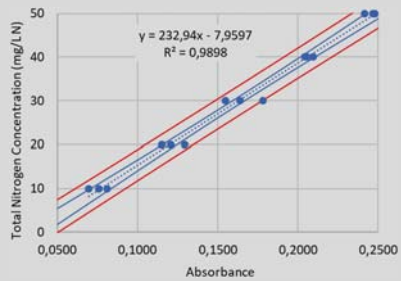


This equation can now be used to calculate the concentration of total nitrogen in the wastewater, based on the absorbance value of 0.202.

$$Conc_{TN} = 232.94 \times 0.202 - 7.9597 = 39.09393$$

C. 11

Using the methods described in Chapter 11, we can calculate the 95% confidence and prediction intervals and plot them on the graph.



Note: the *inner* lines are the **confidence intervals** and the *outer* lines are the **prediction intervals**.

We then find that the 95% confidence interval for the estimated total nitrogen concentration is [38.07668, 40.11119]. This means that based on our standard curve data, we have 95% confidence that the true concentration of total nitrogen in the wastewater sample is between 38.07668 and 40.11119.

Reporting the estimated concentration from the regression above as 39.1 (i.e., three significant figures) is an adequate reflection of the uncertainty associated with the estimated mean concentration. If we were to only use two significant figures (i.e., report a concentration of 39 instead of 39.1), then this would not reflect enough precision since only two possible values (39 and 40) would fit within our confidence interval (38 would be outside of the confidence interval). When we report three significant figures (39.1), now we have a total of possible 21 values rounded to that many digits that fall within our confidence interval (i.e., 38.1, 38.2, 38.3, …, 40.0, and 40.1). If we were to increase the number of significant figures to four (i.e., 39.09), now we will have more than 200 possible values fitting between the confidence limits.

We recommend that you choose a level of significant figures so that when you round your estimated value to that number of significant figures, you would have ideally somewhere between 10 and 100 possible values that fall within the 95% confidence interval.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

4.8 CHECK-LIST FOR YOUR REPORT

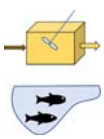
- ✓ Check that raw data are stored separately from calculated values and statistics; preferably the raw data are printed in the appendix or supporting information document, while the calculated values and statistics are summarized in the main report.
- ✓ Make efforts that your data are ideally published online in a way that is both open access and FAIR: findable, accessible interoperable, and reusable.
- ✓ Confirm that metadata is populated and stored appropriately.
- ✓ Verify that the limits of detection and quantification are reported along with other laboratory quality assurance and quality control data.
- ✓ Check that the correct number of significant figures is reported for all data.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Chapter 5

Descriptive statistics: numerical methods for describing monitoring data

This chapter describes how you should prepare and present the general results from your monitoring programme, in terms of flows, concentrations, removal efficiencies, and loads. We cover the basic elements of descriptive statistics, covering simple numerical methods for describing your data. Initial elements for data handling are described, such as the preparation of summary tables and the analysis of missing and censored data and outliers. After that we advance on descriptive statistics, covering measures of central tendency (mean, median, geometric mean, and weighted averages), variation (standard deviation and coefficient of variation), and relative standing (percentiles). The graphs associated with descriptive statistics are presented in the next chapter.



The contents in this chapter are applicable to both treatment plant monitoring and water quality monitoring. The exceptions are the mentions of 'removal efficiencies', which are applicable only to the assessment of treatment plants.

CHAPTER CONTENTS

5.1 An Overview on Descriptive Statistics **96**
5.2 Structuring Your Tables with Summary Descriptive Statistics **101**
5.3 Missing Data **116**
5.4 Censored Data **117**
5.5 Outliers **123**
5.6 Measures of Central Tendency **128**
5.7 Measures of Variation **142**
5.8 Measures of Relative Standing **148**
5.9 Check-List for Your Report **150**

© 2020 The Authors. This is an Open Access book chapter distributed under the terms of the Creative Commons Attribution Licence (CC BY-NC-ND 4.0), which permits copying and redistribution for non-commercial purposes with no derivatives, provided the original work is properly cited (https://creativecommons.org/licenses/by-nc-nd/4.0/). This does not affect the rights licensed or assigned from any third party in this book. The chapter is from the book Assessment of Treatment Plant Performance and Water Quality Data: A Guide for Students, Researchers and Practitioners, Marcos von Sperling, Matthew E. Verbyla and Sílvia M. A. C. Oliveira (Authors).
doi: 10.2166/9781780409320_0095

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

5.1 AN OVERVIEW ON DESCRIPTIVE STATISTICS

Basic

You have already collected your data, either directly, from your experimental system (**primary data**), or from a database from one or more existing treatment systems or water bodies (**secondary data**). You have done a major task, especially if there was lab or field work involved. Every single data you collected were fruit of a lot of work (you know it!), and the next logical way to compensate for this is to **make the best use of them**. Now comes a very noble part of your work that is to extract all the insight that your data can provide. We have seen many well-conducted experiments that never reach their full potential because after collecting all the data, the researcher did not spend enough time or effort to analyse the data and did not archive the data in a way that made them available for others who might want to analyse them in different ways.

Sometimes researchers collect loads of valuable data but then only present the results as simple averages, thus not disclosing all the potential information contained within the data, which shows its variability and its relationship with other variables. Yes, it is understandable: statistical analysis of data is not easy work (and it is sometimes lonely!), especially after you have dedicated so much of your energy in collecting the data. Indeed, presenting your results and (especially) discussing them within the context of existing knowledge and literature is by no means a trivial task.

Treatment plant and water quality data can be considered as environmental data and, as such, they have specific characteristics that require a specially dedicated analysis, such as:

- missing and censored data
- presence of outliers
- a large quantity of data with little information
- important variables that are not measured
- a large quantity of sampling and error analyses
- non-symmetrical (non-normal) distributions
- serial correlation

In this chapter, we will walk you through the different steps related to the calculation and interpretation of descriptive statistics, taking into account these elements.

Remember: your involvement with statistics should have started well before what we will cover in this chapter, at the planning stage of your experiments, when you are using concepts such as power analysis to determine the appropriate sample size for a given effect size and error levels (see Section 3.5).

S. 3.5

After you have obtained the data, the traditional starting point for your analysis is the calculation and presentation of **descriptive statistics**, with the support of summary tables and graphs. Descriptive statistics are *an integral part of all quantitative studies evaluating treatment plant performance and water quality in water bodies*. They are the foundation of your work, frequently presented at the beginning of the Results section of your report.

After calculating descriptive statistics, you can then move into **more advanced analyses**, putting together your knowledge of the treatment plant or water body and the processes involved in it. But note this important point: descriptive statistics will show a good overview of the performance of the system you are studying but will be of *limited use for helping other people to improve design and operational criteria if they are not complemented by additional analyses*, such as the interpretation of the influence of environmental conditions, loading rates (see Chapter 13), hydraulic behaviour, and removal coefficients (see Chapter 14).

C. 13
C. 14

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

This chapter deals with descriptive statistics, *specifically applied to the evaluation of treatment plants and water quality in water bodies*. In this chapter, *we will go into more detail about the statistical methods* since they are relatively simple compared to more advanced methods. Therefore, you should really understand the meaning of the statistical analyses presented here, since descriptive statistics of treatment plant performance and water quality monitoring are an important foundation of your study. Also note that *descriptive statistics are covered in virtually all basic books on statistics*. Tutorials and the 'help' features on any statistical software will also provide additional resources about descriptive statistics to help fortify your understanding. Therefore, we will assume that you will be able to consult good sources to expand your knowledge, if required.

Central tendency and variation

The two most fundamental types of descriptive statistics are **measures of central tendency** and **measures of variation**. Perhaps the most familiar measure of central tendency is the arithmetic mean (the average). Likewise, the standard deviation is one of the more familiar measures of variation. However, in addition to those two, there are other statistics that are sometimes used to measure central tendency and variation, and in some cases, it might be more useful for you to report alternative measures *in addition to or instead of the mean and the standard deviation*.

When to use central tendency measures other than the arithmetic mean?

The arithmetic mean may be the most commonly reported statistic for central tendency, but it is certainly not the only one, and in fact sometimes other statistics such as the median or the geometric mean may be more appropriate! A detailed overview of the arithmetic mean, the geometric mean, the median, or other measures of central tendency is provided in Section 5.6. Furthermore, the distribution of the data may indicate which measure of central tendency is most appropriate. Chapter 8 discusses data with normal versus log-normal distributions.

S. 5.6
C. 8

When to use variation measures other than the standard deviation?

The standard deviation is probably the most common measure of variation reported in the studies of treatment processes and water quality. Indeed, it is a very useful statistic to report when you are trying to communicate to your reader how much variation was encountered in your results from one sample replicate to another. The variance is another common measure of variation, and it communicates essentially the same information as the standard deviation (it is more common to report standard deviations than variances) – the variance of a data set is simply equal to the squared value of the standard deviation. However, there are times when you might be more interested in communicating to your reader the *uncertainty* you have in a particular estimate, such as the mean. In this case, it would be more useful to report the standard error, the margin of error, or the confidence interval associated with the mean of your sample. Section 5.7 contains more detailed information about the different measures of variation.

S. 5.7

Use …	When …
… standard deviation or variance …	… you want to show how much values vary from sample to sample or from replicate to replicate
… standard error, margin of error, or confidence interval …	… you want to show what is the level of uncertainty in your estimate of the mean

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

The main **numerical descriptive statistics** covered in this chapter are

Type	Statistics
Sample characterization	Number of data points (sample size)
Measures of central tendency	Arithmetic mean Median Geometric mean Weighted averages
Measures of variation	Minimum value Maximum value Standard deviation Variance Coefficient of variation (=standard deviation ÷ mean)
Measures of relative standing	10 percentile (or 5 percentile) 25 percentile (=first quartile) 50 percentile (=median = second quartile) 75 percentile (=third quartile) 90 percentile (or 95 percentile)

C. 6

The **graphical methods** for describing your monitored data will be covered in Chapter 6. *Numerical and graphical methods go hand-in-hand*, and you should incorporate both in your descriptive statistics analysis.

The topics covered in this chapter should be followed more or less in a sequence when calculating and presenting descriptive statistics for the treatment plant or water body you are studying:

Sequence for calculating and presenting your descriptive statistics

(1) Structure your original database with flows, concentrations, loads, and removal efficiencies
(2) Evaluate the impact of missing data
(3) Verify possible difficulties associated with censored data
(4) Detect outliers and decide about their maintenance or exclusion
(5) Calculate measures of central tendency
   - Mean
   - Median
   - Geometric mean
   - Weighted averages

(6) Calculate measures of variation
   - Minimum, maximum, and range
   - Standard deviation
   - Variance
   - Coefficient of variation

(7) Calculate measures of relative standing (percentiles)
(8) Prepare summary statistics tables with the basic descriptive statistics
(9) Prepare graphs for the quantitative representation of your data
   - Time series
   - Histograms
   - Frequency distribution (percentile graphs)

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

- Box-plot
- Scatter plot
- Bar/column and pie charts

As supporting material for this book, we have prepared general Excel spreadsheets for you to put your monitoring data and extract basic summary statistics, including graphs. These spreadsheets can be used for you to go into more detail into the analyses and have a broader view of the results. The example of the wastewater treatment plant is based on a well-monitored system (almost daily frequency of data collection), with data from influent flow and influent and effluent concentrations of biochemical oxygen demand (BOD), chemical oxygen demand (COD), total suspended solid (TSS), total Kjeldahl nitrogen (TKN), total phosphorus (P), and thermotolerant coliforms spanning a period of four years.

Excel

- **Spreadsheets with data**
  - Treatment plant (wastewater)
- **Spreadsheet with blank cells** (to be used with your own data)
  - Treatment plant (water or wastewater; you can insert the constituents you are monitoring)
  - Water body

The master spreadsheet for monitoring a treatment plant includes the following *worksheets* listed below. The spreadsheet for monitoring a water body is similar but does not have calculations and worksheets on input loads (loading rates) and removal efficiencies.

Worksheet	Comment
Data	Cells for you to fill-in with your data on date, flow, and concentrations (input and output) of the main constituents of interest. In the prepared sheet, the following constituents are included (but can be easily changed): BOD, COD, TSS, TKN, P total, and *E. coli*
Efficiency	Based on the input and output values entered in sheet 'Data', removal efficiencies are calculated for all dates and constituents
Input loads	Based on the flows and input concentrations, loads are calculated. If you provide the surface area or volume of the unit you are studying then applied mass loading rates are calculated
Stats on concentrations	Major descriptive statistics are calculated for the input and output concentrations
Stats on efficiencies	Major descriptive statistics are calculated for the removal efficiencies
Stats on input loads	Major descriptive statistics are calculated for the input loads or applied mass loading rates
Time series	Time series graphs are plotted based on your original data on flow and input/output concentrations and the calculated data on removal efficiencies

(*Continued*)

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Worksheet	Comment
Histograms	Frequency histograms are plotted based on your original data on flow and input/output concentrations and the calculated data on removal efficiencies
Box plots	Box-and-whisker plots are made based on your original data on flow and input/output concentrations and the calculated data on removal efficiencies
Frequency distribution of concentrations (percentile graphs)	Cumulative frequency distribution graphs of concentrations are plotted, showing percentiles in the *X*-axis and absolute cumulative frequency distributions of input and output concentrations on the *Y*-axis
Frequency distribution of efficiencies (percentile graphs)	Cumulative frequency distribution graphs of removal efficiencies are plotted, showing percentiles in the *X*-axis and absolute cumulative frequency distributions of removal efficiencies on the *Y*-axis
Monthly concentrations	Calculates the mean concentrations from each month of your entire series and plots time series graphs of monthly concentrations
Monthly efficiencies	Calculates the mean removal efficiencies from each month of your entire series and plots time series graphs of monthly efficiencies.
Monthly averages	Calculates and plots time series graphs of the mean concentrations and removal efficiencies in each of the 12 months of the year (January, February, …, December)
Yearly averages	Calculates and plots the mean concentrations and removal efficiencies from each year of your entire time series, starting from the first year and finishing on the last year of your data set
Standards	Based on the values you provide on existing discharge standard or desired targets for effluent concentrations and removal efficiencies of the main constituents you are analysing, calculates the percentage of compliance for each constituent and plots a summary bar-chart

You are highly encouraged to use **these spreadsheets for your monitoring data and modify them to your needs**, including other analyses, graphs with two or more interrelated variables together, new graphs such as scatter plot between two variables, new formats for your graphs (different markers and different styles), incorporate new Excel functions to your calculations (there are so many useful functions!), etc. After all, it is always better when we have our own spreadsheet, because we become more confident and have a more direct relationship with our data. Therefore, see the sheets provided as a *simple starting point* from where you can build your own analytical tools.

Furthermore, if you have access to **any statistical software**, even a very simple one, it will be able to provide you with good tools for carrying out descriptive statistics analyses.

Other applications of descriptive statistics, for instance, for helping you to *compare the performance of different treatment plants, compare the performance of different unit processes arranged in parallel, or compare the performance of a system between different operational phases* will also be covered here (see Section 5.2). However, they are analysed in more detail in other parts (especially Chapter 10), in which they have the support of comparative statistical analyses with hypothesis testing.

S. 5.2
C. 10

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

5.2 STRUCTURING YOUR TABLES WITH SUMMARY DESCRIPTIVE STATISTICS

5.2.1 Different types of studies requiring different types of summary tables

Basic

In the subsequent sections, we will present the basic concepts of descriptive statistics. However, since we have shown examples of tables or spreadsheets for your data (Chapter 4), we will also take the opportunity of presenting here examples of **possible structures for tables in which you will include your descriptive statistics**. Note that, most likely, the summary tables will be part of the core of your report and, as such, will need to be interpreted and discussed. Only if the tables are very large, they should be then incorporated in an Appendix. Also see our suggestions regarding the placement of the original raw data versus the calculated values (Chapter 4).

C. 4

When you prepare a summary table with descriptive statistics, you should structure this table *in such a way that it matches with the objectives of your study and with the rest of the information presented in the Results section of your report*. If you have very limited space, probably you will not need to present all the descriptive statistics and will include only the most important ones, such as a measure of central tendency and a measure of variation (see Sections 5.6 and 5.7). The number of data points (n) is also an important piece of information that should always be included. If there is no room for it in the table itself, you can provide it as a footnote, especially if 'n' is the same for all sample locations and all parameters summarized in the table.

S. 5.6

S. 5.7

Starting from the beginning of your Results section, give your reader a **general overview of the descriptive statistics, with all important variables shown together in a single summary table**, if possible. After presenting this general information, you can move on to a more detailed analysis of each variable, probably in subsequent sections of your report.

> *You should always start your description of the results by giving the reader a general overview before you move into specific details.*

Figure 5.1 presents examples of **different types of studies, each of them requiring different types of summary tables** with descriptive statistics, according to the examples listed below.

5.2.2 Summary tables of studies in treatment plants

Basic

From Figure 5.1 (top), we detect the following types of studies that need descriptive statistics and summary tables for your studies of treatment plant performance.

(a) One plant (input and output values)

This is the example that will be analysed in more detail in this chapter. You have data on input and output flows and concentrations of the treatment plant you are studying, and so, you have a simple structure for your summary table.

(b) One plant composed of treatment units in series

You have data on the inlet and outlet of each unit, and you need to prepare a summary table with the statistics of each stage of the treatment line and of the overall plant.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

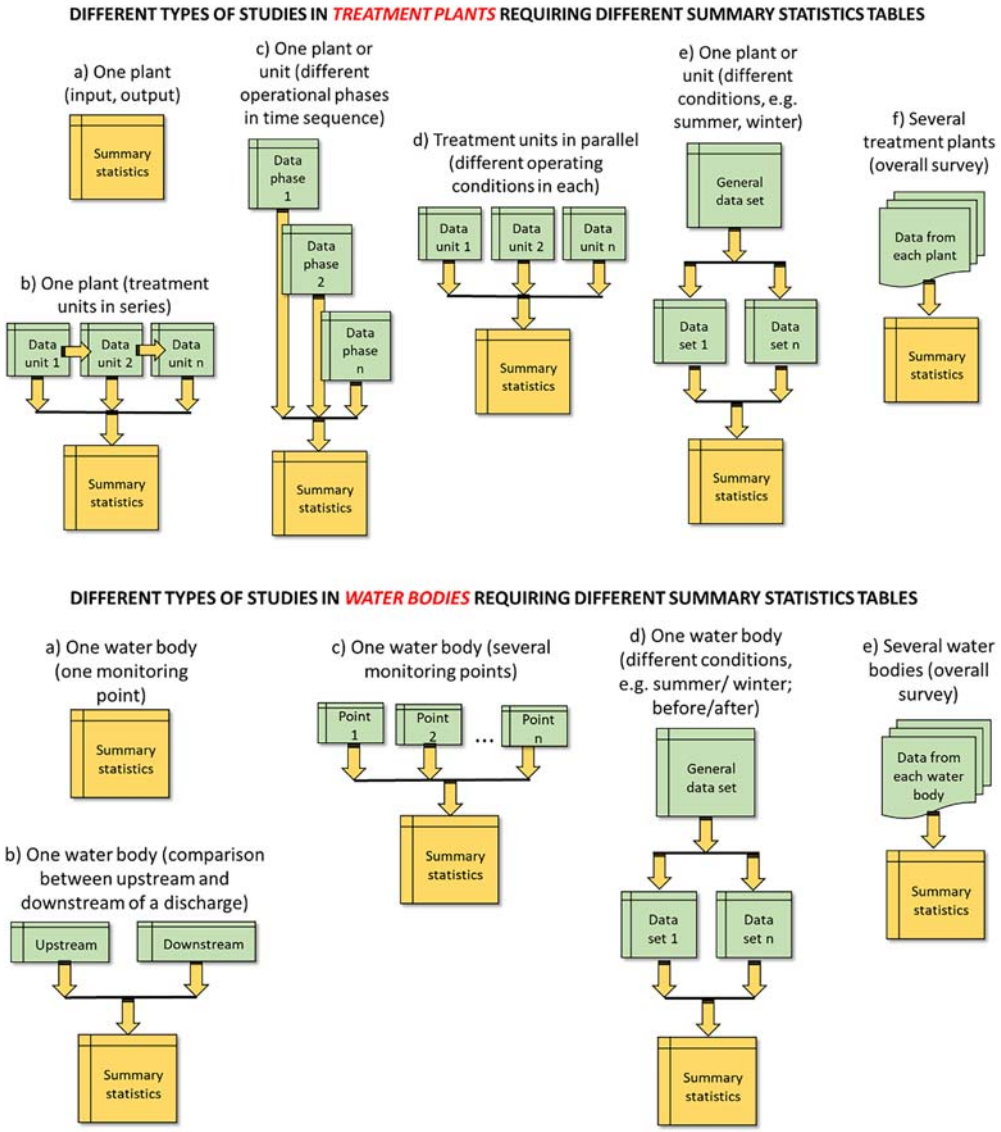


Figure 5.1 Examples of different types of studies in which different types of summary tables with descriptive statistics need to be prepared (top, studies in treatment plants; bottom, studies in water bodies).

(c) One plant or treatment unit, subjected to different operational phases, each one in a different time period

You have structured your experiments such that you apply different loading rates (or different operational conditions) to your plant or to a single treatment unit. Since you have only one plant or unit, the operational phases are in time sequence, one after the other, so that you obtain the data for each phase and prepare a summary table that shows the statistics for each phase. After that you evaluate the influence of the operating conditions on treatment performance.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

(d) Different plants or treatment units in parallel, each subjected to different operational conditions

You have more than one treatment unit, with similar characteristics, operating in parallel. As part of your experiment, you impose to each of the units different operating characteristics, such as applied loading rates. The experiments are run in parallel, at the same time, so that the influent to all the lines is the same, and differences in the effluent quality will be possibly associated with the applied operating conditions in each line. You obtain the data for each line and prepare a summary table that shows the statistics for each. After that you evaluate the influence of the operating conditions on treatment performance.

(e) One plant with *a posteriori* segregation of data from different time periods or operating conditions

In possession of the historical data from your treatment plant, you perform an analysis of the influence of different operating conditions. However, you decide to perform this analysis *a posteriori*, which means that you did not control the operating conditions or time periods prior to the data collection, but divided up the data in retrospect. For instance, you may wish to divide the whole data set into two sets, one for winter months and the other for summer months. Other options are to analyse dry versus wet periods, tourism season versus non-tourism season, etc. Your summary table presents the statistics for each operating condition, and you subsequently evaluate their influence on the treatment performance.

(f) Survey on the performance of several treatment plants

You obtain monitoring data from several treatment plants and wish to compare their performance and obtain general statistics for the entire collection of treatment plants. You prepare the summary statistics for each plant and then structure a general summary table, with the overall statistics of the set of plants evaluated.

For each of the different types of studies mentioned above, we present below possibilities and **suggestions for summary tables with descriptive statistics**. Notice that, in all of them, we give emphasis for reporting both concentrations and removal efficiencies, because they are essential elements in the evaluation of the performance of a treatment plant. In summary:

- Always try to report descriptive statistics for **concentrations** and **removal efficiencies** of the major pollutants of interest.
- For the constituents and variables that represent internal operational conditions, and are *not expected to be removed at your plant, you do not need to present statistics for removal efficiencies*. Some examples of these parameters might include pH, temperature, etc.

Basic

(a) One plant (input and output values)

Tables 5.1 and 5.2 present suggestions for this simple case in that only one treatment plant is investigated and in which a simple performance evaluation based on input–output data is undertaken. Table 5.1 presents a simpler structure, focussed on concentrations and removal efficiencies, while Table 5.2 presents a more complete version, with flows and applied mass loading rates.

In the tables, we present concentrations as **g/m^3**, because this facilitates the *calculation of loading rates*, when you multiply flow (m^3/d) times concentration (g/m^3). But we could have also used mg/L, and as a matter of fact, **mg/L** is *more common for reporting concentrations in summary tables* (unless you need to use other units, such as µg/L, MPN/100 mL, etc.). In summary, do not

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Table 5.1 Example of a simple summary table with descriptive statistics for concentrations and removal efficiencies for a single plant or a single treatment unit (i.e., two sample collection points at the input and output locations).

Statistics	**Constituent 1**			**Constituent 2**			**Constituent *n***		
	Input Concent (g/m^3)	**Output Concent (g/m^3)**	**Effic (%)**	**Input Concent (g/m^3)**	**Output Concent (g/m^3)**	**Effic (%)**	**Input Concent (g/m^3)**	**Output Concent (g/m^3)**	**Effic (%)**
n									
Mean									
Median									
Min									
Max									
St. Dev.									
CV									
…									
…									

Notes: n, number of data; Min, minimum; Max, maximum; st. dev., standard deviation; CV, coefficient of variation (standard deviation ÷ mean).
n and CV are dimensionless. '*n*' is an integer number, and CV is usually reported with two decimal cases or as percentages.
The order of the rows with the descriptive statistics may vary, according to the emphasis you want to put in the interpretation of the table. For instance: mean close to standard deviation (adjacent rows), mean close to median, etc. Usually the number of data (*n*) is in the first line.
You may want to incorporate other statistics, such as geometric means and percentiles (e.g., 10%, 25%, 75%, and 90%).
Alternatively, you may want to supress the presentation of some statistics, in order to decrease the number of rows in the table.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Table 5.2 Example of a simple summary table with descriptive statistics for flow, concentrations, loading rates, and removal efficiencies for a single plant or a single treatment unit (i.e., two sample collection points at the input and output locations).

Statistics	**Inflow (m^3/d)**	**Concentrations**				**Applied Surface Mass Loading Rate**		**Efficiency**	
		Input Param 1 (g/m^3)	**Output Param 1 (g/m^3)**	**Input Param *n* (g/m^3)**	**Output Param *n* (g/m^3)**	**Input Param 1 [$(kg/d)/m^2$]**	**Input Param *n* [$(kg/d)/m^2$]**	**Param 1 (%)**	**Param *n* (%)**
n									
Mean									
Median									
Min									
Max									
St. Dev.									
CV									
…									
…									

C. 13

Note: For the concept of applied loading rate, see Chapter 13.

forget that

$$\mathbf{g/m^3 = mg/L}\ (\text{one gram contains } 1000\,\text{mg; one } \text{m}^3 \text{contains } 1000\,\text{L})$$

S. 5.2

The following comments apply to all summary tables in Section 5.2 and the rest of the book, and *we recommend that you take them into account*.

Beware of significant figures! In principle, **the mean, median, geometric mean, and standard deviation have the same number of decimal places of the original data**. For instance, if the original data are for a constituent that is recorded as *integer numbers*, also these statistics should be reported as integer values. The same applies to original data with one decimal place, in that these statistics need to be reported with one decimal case, and so on. However, most statistical textbooks make some allowances to present these statistics with **one decimal place more** than the original data, especially if the measured values of the variable are small. For instance, if the original values of the variable are in the order of hundreds or more, no **additional** decimal cases will be necessary. However, if the values are in the order of a few tens or less, adding an **additional** decimal case can make the statistics clearer. We leave this to your own judgement and common sense. However, note that the incorporation of a large number of decimal cases (much higher than the accuracy afforded by the measurement) is a very common mistake found in summary tables in reports, because calculators and spreadsheets provide results that do not incorporate the concerns of significant figures, and it is up to you to adjust this in your report. Example:

- Original data: 8 5 6 9 4 5 (all integer values)
- Calculated and reported mean value: 6.1666666667 (**wrong!**)
- **Correct** mean value to be reported: 6 (integer value) or 6.2 (incorporation of an additional decimal place)

***Note**: **In this book**, for didactic purposes, in many places we do not follow this rule, when we are showing you how to do a certain calculation. In this situation, we show more decimal cases than necessary, so that you can **check that your calculations are correct**.*

S. 4.7

See Chapter 4, Section 4.7, for more information about significant figures.

Inconvenience of reporting values of mean and standard deviation as mean ± standard deviation. Frequently, in technical reports and scientific publications, summary tables are reported with mean ($\bar{x}$) and standard deviation (s) in the form of $\bar{\boldsymbol{x}} \pm \boldsymbol{s}$. The comment here applies to the symbol '±'. First of all, using the symbol '±' is vague – it does not communicate to the reader whether the number after the '±' is the standard deviation, the variance, or some other measure of variability or uncertainty such as the confidence interval or the prediction interval. Second, this practice can be misleading – when the symbol '±' is used like this, it implies that the distribution of data is symmetrical around the mean, and that the standard deviation is an indicator of the variability of data in a symmetrical way, below and above the mean. As will be discussed in Section 6.3 and Chapter 8, **treatment plant and water quality data are frequently not symmetrically distributed**, and thus, it is misleading to suggest that the variability will be symmetrical around the mean, as would occur with a bell-shaped (normal) distribution curve. Mean and standard deviation will be discussed in Sections 5.6 and 5.7, but we can make this comment here based on your prior knowledge of both concepts. Example:

S. 6.3
C. 8
S. 5.6
S. 5.7

- Original data: 7.3 5.2 6.4 9.1 4.2 5.3
- Mean ($\bar{x}$): 6.3
- Standard deviation (s): 1.8

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

- Common (but misleading) way to report the values of median and standard deviation: $\tilde{x} \pm s = 6.3 \pm 1.8$ (**misleading** in that it suggests implicit symmetry of data variability around the mean)
- **Suggested alternative** for treatment plant and water quality data: $\tilde{x}(s)$ or 6.3(1.8). Alternatively, you can put the **mean and the standard deviations in different cells** (rows or columns) of your summary table, as exemplified in Tables 5.1 and 5.2.

Basic

(b) One plant (treatment units in series)

Your treatment plant has several **units in series in the treatment line**. This is a *common situation and typical of most systems*: aeration tank followed by sedimentation tank, trickling filter followed by sedimentation tank, upflow anaerobic sludge blanket (UASB) reactor followed by post-treatment unit, anaerobic pond followed by facultative pond followed by maturation ponds in series, septic tank followed by horizontal flow wetland, coagulation/flocculation units followed by sedimentation tanks followed by filters followed by disinfection in a water treatment plant, and so on.

In this case, you assume that you are fortunate enough to have data from the input and output of all relevant units so that you can analyse the relative performance of each stage of your treatment line and its contribution to the overall treatment performance. This is an advancement in comparison with the situation described in item (a), in which there were data only on the influent and final effluent of a single unit process or a single treatment plant, and nothing could be said about the performance of each individual stage.

In terms of monitoring, in most cases, you can consider that **the output from one unit is the input to the next unit**, and thus reduce the number of sampling points. You should not do this if each stage is composed of units in parallel that work under different conditions.

The **challenge** here is to prepare a **single summary table that covers the descriptive statistics of each treatment stage and the constituents of interest**. If you do not have enough space in your report (apart from the Appendices, Annexes, and Supplementary Material that we discussed before), you will probably need to be selective and include only the most important statistics in your summary table.

In this case, we suggest that you include (i) a measure of central tendency and (ii) a measure of variation. An example can be given in Tables 5.3 and 5.4, one for the concentrations and the other for removal efficiencies. Depending on the page format of your report, you can merge both tables into a single one, with one part for concentrations and the other for efficiencies.

(c) One plant or treatment unit, subjected to different operational phases in a time sequence

Advanced

S. 5.2.2

C. 10

This is similar to the first situation (a), in which you investigate your system for a certain period. The difference now is that you deliberately decided to test the behaviour of your system when subjected to different operational conditions, such as applied loading rates for instance. Because you have only one treatment plant, you cannot run experiments in parallel. Therefore, you plan your experiment with, say, three different phases in a time sequence. See comments made in Section 5.2.2 for this type of experiment and how to use statistics for comparing the results in the different phases in Chapter 10.

In your summary tables, you have to present information on the operational conditions imposed for each phase, the results in terms of concentrations and removal efficiencies for each phase, and especially, if you have treatment units in series and need to analyse them individually. Therefore, you can end up with a large summary table if you have to include all descriptive statistics. These

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Table 5.3 Example of a simple summary table with mean and standard deviation of the concentrations at the effluent of each stage of the treatment plant (stages in series).

Constituent	Influent	Effluent Stage 1	Effluent Stage 2	…	Effluent Stage *n*
Param 1 (g/m^3)	Mean (st. dev.)				
Param 2 (g/m^3)	Mean (st. dev.)				
Param 3 (–)	…				
Param 4 (MPN/100 mL)	…				
…					
…					

Notes: st. dev., standard deviation.
Outside parentheses, mean value; inside parentheses, standard deviation. Fill in all cells with their respective values.

Table 5.4 Example of a simple summary table with median and standard deviation of the removal efficiency of each stage of the treatment plant and the overall efficiency.

Constituent	Stage 1	Stage 2	…	Stage *n*	Overall Removal
Param 1 (%)	Median (st. dev.)				
Param 2 (%)	Median (st. dev.)				
Param 3 (%)	…				
Param 4 (log units)	…				
…					
…					

Notes: Overall removal: global removal of the treatment plant, from its influent to the final effluent.
St. dev., standard deviation.
Values in each cell are the median, and inside parentheses, the standard deviation.

large tables can go to Appendices or Supplementary Material, and the more concise tables can stay in the main body of your report.

Table 5.5 gives an example of a possible summary table for the descriptive statistics of your operational phases. Table 5.6 presents the associated descriptive statistics of treatment performance (concentrations and removal efficiencies). In this example, we included the major descriptive statistics, but you could select only the most important ones (e.g., mean or median and standard deviation) if you need more concise tables.

(d) Different plants or treatment units in parallel, each subjected to different operational conditions

Advanced

This is similar to situation (c), with the difference that you run all experiments at the same time, in parallel. You have more than one treatment unit, each of them with similar characteristics, all operating in parallel. As part of your experiment, you impose each of the units to different operating characteristics, such as applied loading rates. The experiments are run at the same time so that the influent to all the lines is the same and differences in the effluent quality will be possibly associated with the applied operating conditions in each line.

The tables will be similar to those presented at situation (c), substituting 'phase' by 'unit' (or line, or plant) and taking into account that the influent will be the same, since the units are operated in parallel. Table 5.7 presents an example for the descriptive statistics of the applied

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Table 5.5 Example of a summary table with descriptive statistics of the operating conditions implemented in each phase of the experimental period.

Item	Phase 1	Phase 2	Phase *n*
	Low Organic Loading Rate	**Medium Organic Loading Rate**	**High Organic Loading Rate**
Target organic loading rate (gBOD/m^2)/d	2.0	4.0	6.0
Period	February 2017 to September 2017	October 2017 to June 2018	July 2018 to January 2019
Duration (months)	8	9	7
Statistics of the actual applied surface organic loading rate			
n	32	36	28
Mean (gBOD/d)/m^2	1.8	4.1	5.7
Median (gBOD/d)/m^2	1.7	3.9	5.5
Minimum (gBOD/d)/m^2	…	…	…
Maximum (gBOD/d)/m^2	…	…	…
Standard deviation (gBOD/d)/m^2	…	…	…
CV	…	…	…
…	…	…	…

loading rates in each treatment unit, and Table 5.8 shows the descriptive statistics related to the performance (concentrations and efficiencies) of each treatment unit.

(e) One plant with existing monitoring data, in which you analyse different time periods or operating conditions

Advanced

From the existing monitoring records from your treatment plant, you decide to analyse the influence of different operating conditions that took place during the operational period. For instance, you may want to compare the performance during winter months with that in the summer months, or dry months versus wet months. Or you know that the treatment plant had an expansion some years ago and want to compare the efficacy of the expansion by analysing data from the period before it against data after it.

The situation here is similar to (c), in that you have distinct operational phases. The difference is that you make the analysis *a posteriori*, which means that you do the analysis in retrospect, without controlling operating conditions during the experiment. What you need to do is segregate your data into subsets (e.g., summer versus winter, rainy versus dry, etc.), with each subset containing the data associated with your selection criterion. The summary table will be similar to Table 5.8, and each phase corresponds to one of the selected conditions.

(f) Survey on the performance of several treatment plants

Advanced

This is a distinct type of study compared with situations (a) to (e). Now, you contact water and sanitation companies, environmental agencies, and other institutions and obtain **monitoring data from several treatment plants**. You then **separate the plants into categories**, for instance, by treatment process employed. Ultimately, you want to report what is the general performance of the plants operating in a certain country or region, or using processes '*x*', '*y*', and '*z*'. '*What is the process offering the best performance*' is a typical question frequently asked by practitioners.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Table 5.6 Example of a simple summary table with descriptive statistics for concentrations and removal efficiencies in each phase of the experimental period.

Constituent/Statistics	Influent Concentrations			Effluent Concentrations			Removal Efficiencies		
	Phase 1	Phase 2	Phase *n*	Phase 1	Phase 2	Phase *n*	Phase 1	Phase 2	Phase *n*
Constituent 1									
n									
Mean (g/m^3)									
Median (g/m^3)									
Minimum (g/m^3)									
Maximum (g/m^3)									
St. dev. (g/m^3)									
CV									
…									
Constituent n									
n									
Mean (g/m^3)									
Median (g/m^3)									
Minimum (g/m^3)									
Maximum (g/m^3)									
St. dev. (g/m^3)									
CV									
…									
…									

Note: Phase 1, low surface organic loading rate – median 1.7 $(gBOD/d)/m^2$; phase 2, medium surface organic loading rate – median 3.9 $(gBOD/d)/m^2$; phase 3, high surface organic loading rate – median 5.5 $(gBOD/d)/m^2$; see Table 5.5 for more information on the operational phases.

Now, your strategy for **manipulating the data should be different**. Initially, you will separate the treatment plants into the category you want to analyse (e.g., treatment process). For instance, your entire database set is composed of 70 treatment plants that can be divided into the following three categories: 28 plants using process '*x*', 22 plants using process '*y*', and 20 plants using process '*z*'. After that, for each treatment plant, you calculate the descriptive statistics of the constituents you are analysing, in terms of concentrations and removal efficiencies, in a customary way, as in situation (a) described above.

Note that you cannot put all the data from process '*x*' together and obtain, for instance, the mean value of the effluent BOD concentration. This is because each of the 28 plants comprising process '*x*' has a different number of data, and we cannot put all of them together and extract an overall mean, because this mean value would be much influenced by the plants that have more monitoring data. Take Example 5.1, in which there are four treatment plants. The example uses few data to make it simpler to undertake the calculations and get the results.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Table 5.7 Example of a summary table with descriptive statistics of the operating conditions implemented in each of the units running in parallel.

Item	**Unit 1**	**Unit 2**	**Unit *n***
	Low Organic Loading Rate	**Medium Organic Loading Rate**	**High Organic Loading Rate**
Target organic loading rate (gBOD/m^2)/d	2.0	4.0	6.0
Statistics of the actual applied surface organic loading rate			
n	*32*	*36*	*28*
Mean (gBOD/d)/m^2	1.8	4.1	5.7
Median (gBOD/d)/m^2	1.7	3.9	5.5
Minimum (gBOD/d)/m^2	…	…	…
Maximum (gBOD/d)/m^2	…	…	…
St. dev. (gBOD/d)/m^2	…	…	…
CV	…	…	…
…	…	…	…

Table 5.8 Example of a simple summary table with descriptive statistics for concentrations and removal efficiencies in each of the units running in parallel.

Constituent/Statistics	**Influent Concentrations**	**Effluent Concentrations**			**Removal Efficiencies**		
		Unit 1	**Unit 2**	**Unit *n***	**Unit 1**	**Unit 2**	**Unit *n***
Constituent 1							
n							
Mean (g/m^3)							
Median (g/m^3)							
Minimum (g/m^3)							
Maximum (g/m^3)							
St. dev. (g/m^3)							
CV							
…							
Constituent n							
n							
Mean (g/m^3)							
Median (g/m^3)							
Minimum (g/m^3)							
Maximum (g/m^3)							
St. dev. (g/m^3)							
CV							
…							
…							

Note: Unit 1, low surface organic loading rate – median 1.7 (gBOD/d)/m^2; unit 2, medium surface organic loading rate – median 3.9 (gBOD/d)/m^2; unit 3, high surface organic loading rate – median 5.5 (gBOD/d)/m^2; see Table 5.7 for more information on the operational conditions of the treatment units.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Example

EXAMPLE 5.1 MEAN VALUES FROM SEVERAL TREATMENT PLANTS

Calculate the mean of the effluent concentration of a certain constituent, obtained from monitoring data from four treatment plants. The data are shown in the following table.

Data:

Effluent concentration values (g/m^3) for a certain constituent

Plant 1	***Plant 2***	***Plant 3***	***Plant 4***
10	*35*	*15*	*13*
12	*32*	*18*	*18*
8	*38*	*17*	*15*
13	*41*		*16*
9	*34*		
	36		
	37		
	31		
	39		
	37		
	32		
	39		
	41		
	37		
	44		
Mean plant 1 = 10	*Mean plant 2 = 37*	*Mean plant 3 = 17*	*Mean plant 4 = 16*

Discussion:

The bottom row at the table shows the mean values for each treatment plant. We see that Plant 2 has more data and, for some reason, a worse performance, because of the higher effluent concentration values (mean = 37 g/m^3), while the other plants have less data but also lower effluent concentrations (mean values of 10, 17, and 16 g/m^3).

If we calculate the mean of the four means, we obtain

$$\text{Mean of the means} = \frac{10 + 37 + 17 + 16}{4} = 20 \ \frac{g}{m^3}$$

However, if we calculate the overall mean, putting together all the 27 values of the four treatment plants, we obtain

$$\text{Overall mean} = 27 \ \frac{g}{m^3}$$

The overall mean of 27 g/m^3 is higher than the mean of the means (20 g/m^3). The mean of the means, in this case, is likely to be a better representation of the central tendency of the effluent data from this category, since three of the four plants have good effluent quality. The overall mean

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

($27\ g/m^3$), putting together all data, is very much influenced by Plant 2, with their larger number of data and higher effluent concentrations. Therefore, the overall mean does not seem to be a good descriptor of the central tendency of this category, because this value is much higher than the mean values of three of the four plants.

Of course, this is just a simple example, with few data, to facilitate calculations and interpretations. In your survey, we would expect to have much more data for each treatment plant, in order to give more confidence to the results.

In summary, we have

In surveys with several treatment plants, it is probably better to work with the **mean of the means** (or **median of the medians**) from the plants, instead of putting all data together and calculating a single overall mean (or median).

Tables 5.9 and 5.10 show examples of tables reporting surveys of treatment processes (adapted from tables presented in survey works by Oliveira and von Sperling, 2011, and von Sperling, 2005). Table 5.9 presents influent and effluent concentrations, together with removal efficiencies for several constituents. Because of this, it needs to be concise and concentrates only on presenting mean or median values. Table 5.10 presents the full descriptive statistics for only one constituent and for only removal efficiencies. You may select the format that best suits your interest or even a combination of both formats.

5.2.3 Summary tables of studies in water bodies

Basic

From Figure 5.1 (bottom), we exemplify the following types of studies that need descriptive statistics and summary tables when you are studying the water quality of water bodies. See Section 5.2.2, which

Table 5.9 Example of a summary table showing median concentrations and median removal efficiencies, according to the three treatment processes investigated in a survey.

Constituent	Processes		Process *x*	Process *y*	Process *z*
	Number of treatment plants evaluated		…	…	…
Constituent 1	Influent (raw) Effluent (treated) Removal efficiency	(g/m^3) (g/m^3) (%)	… … …	… … …	… … …
Constituent 2	Influent Effluent Removal efficiency	(g/m^3) (g/m^3) (%)			
Constituent *n*	Influent Effluent Removal efficiency	(g/m^3) (g/m^3) (%)			

Note: Descriptive statistics are calculated based on the median values from each treatment plant in a certain category (treatment process).

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Table 5.10 Example of a summary table showing descriptive statistics of removal efficiencies (%), according to the three treatment processes investigated in a survey.

Statistics	Process *x*	Process *y*	Process *z*
Number of data, *n*	…	…	…
Mean			
Median			
St. dev.			
Minimum			
Maximum			
CV			
…			

Note: Descriptive statistics are calculated based on the median values from each treatment plant in a certain category (treatment process).

presented several summary tables for treatment plant monitoring, for which many comments also apply here (the exceptions are removal efficiencies and loading rate conditions, which are not incorporated in water quality monitoring). We list below typical types of studies and possible examples of summary tables.

S. 5.2.2

(a) One water body (one monitoring point)

This is a simple situation, in which you have data on several water quality constituents, collected over a certain time in one sampling point from one water body. The structure of the summary table is simple. An example can be found in Table 5.11.

(b) One water body (comparison between upstream and downstream of an effluent discharge)

Your water body receives the discharge of an effluent, and you have monitoring data on two locations, one upstream of the discharge and the other downstream, so that you can compare the impact of the discharge in the water quality of the receiving body. In order to facilitate visualization of the results, you place in adjacent columns the values 'upstream' and 'downstream'. A possible summary table is exemplified in Table 5.12.

(c) One water body (several monitoring points)

You follow the profile of concentrations and environmental conditions along a river to analyse the conversion processes that take place or the influence of discharges along its course. Alternatively, you monitor a lake in several places spread in its surface area (and possibly in different depths of the water column). The structure could be similar to Table 5.12, in which you have two monitoring points. However, if you have several monitoring points and you still want to put the values of a same constituent in adjacent cells, you may want to invert the position of rows and columns, such as exemplified in Table 5.13. If you feel that your table is getting too large to enter in the main text of your report, you can put it in an Appendix and present a shorter version, with only, say, mean or median and standard deviation in the report.

(d) One water body with *a posteriori* segregation of data from different time periods or environmental conditions

In possession of the historical data from your water body, you decide to analyse (in retrospect) the influence of different environmental conditions or the effect of interventions in the catchment area. For instance, you may wish to divide the whole data set into two sets, one for winter months

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Table 5.11 Example of a simple summary table with descriptive statistics for monitoring of water quality in a water body (one monitoring point).

Statistics	Unit	Constit 1	Constit 2	Constit 3	Constit 4	Constit 5	Constit …	Constit $n-1$	Constit n
Number of data									
Mean									
Median									
Minimum									
Maximum									
St. dev.									
CV									
…									
…									

Notes: Constit, water quality constituent.
St. dev., standard deviation; CV, coefficient of variation (standard deviation ÷ mean).
Unit: mg/L, μg/L, MPN/100 mL, etc. Number of data and CV are dimensionless. '*n*' is an integer number, and CV is usually reported with two decimal cases or as percentages.
The order of the rows with the descriptive statistics may vary, according to the emphasis you want to put in the interpretation of the table. For instance, mean close to standard deviation (adjacent rows), mean close to median, etc. Usually the number of data (*n*) is in the first line.

Table 5.12 Example of a summary table with descriptive statistics for monitoring of water quality in a water body (upstream and downstream of an effluent discharge).

Statistics	Unit	Constituent 1		Constituent 2		Constituent $n-1$		Constituent n	
		Up	Down	Up	Down	Up	Down	Up	Down
Number of data									
Mean									
Median									
Minimum									
Maximum									
St. dev.									
CV									
…									
…									

Notes: See notes on Table 5.11.
Up, upstream of discharge; down, downstream of discharge.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Table 5.13 Example of a summary table with descriptive statistics for monitoring of water quality in a water body along four sampling points.

Constituent	Sampling Point	*n*	Mean	Median	Minimum	Maximum	St. dev.	CV	…
BOD (mg/L)	1								
	2								
	3								
	4								
Dissolved oxygen (DO) (mg/L)	1								
	2								
	3								
	4								
…	…								
…	…								
E. coli (MPN/100 mL)	1								
	2								
	3								
	4								

Notes: See notes on Table 5.11.
n, number of data.

and the other for summer months, or dry/wet periods. You can also analyse the influence of interventions, such as impact of the beginning of operation of a new industry, or benefits from the implementation of a new wastewater treatment plant (comparisons between 'before' and 'after'). The structure is similar to Table 5.12, but instead of having upstream/downstream, you have winter/summer, wet/dry, before/after, etc.

(e) Survey on the water quality of several water bodies

You obtain monitoring data from several water bodies and wish to compare their water quality. You prepare the summary statistics for each water body and then structure a general summary table, with the overall statistics of the set of water bodies evaluated. See comments on Section 5.2.2.f.

5.3 MISSING DATA

In treatment plant and water quality monitoring, **it is typical to have missing data**. After all, taking samples and measurements on site or *in situ* on a pilot- or full-scale system does not always work out. The field work is subject to challenges due to equipment failure and weather conditions, which can affect your ability to effectively collect, transport, and preserve samples. Furthermore, there can be problems with laboratory analysis that provide inconclusive results for a given set of samples. When working with real treatment plants and water systems, the samples often cannot simply be replaced, so these situations may result in missing data points. The many elements that comprise a monitoring programme are not trivial, and there are always chances that some data collection or measurement will not be done or some lab analysis will be unsuccessful.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

We have to live with this situation, recognizing that it is typical, and *use our own judgement to see whether the quantity of missing data will affect the monitoring results substantially*. In the spreadsheets where you store your data, there will be typically blank cells corresponding to the missing data (see example in Table 4.4). As mentioned in Section 4.2.2, *you need to leave the spreadsheet cells with missing data as 'blanks' or empty cells and do not put 'zero' values in them*.

S. 4.2.2

The number of blank cells in your spreadsheet with the monitoring data depends on how you organize it. If you collect samples on a weekly basis and your spreadsheet is structured for inputting daily values, you will have six blank lines (six days without monitoring) for each filled-in line (one day per week with monitoring). The cells in the days without monitoring are not considered missing data, because no data were obtained in those days. Therefore, if you have weekly monitoring, it is better that you organize your spreadsheet for receiving weekly data. A similar comment could be done for other time intervals, such as months, quarter, etc.

Usually your missing data can be left as such, and you will **use only the available data for your performance assessment** of the treatment plant or water body. However, if some of the monitored variables are essential input variables for a dynamic mathematical model (e.g., inflow, influent COD), for which you need complete time series in order to predict the output variables (e.g., effluent COD), you will need to fill-in the gaps. There are several ways of **imputing data** to replace missing cells, but these are outside the scope of this book. Good information can be found in books on hydrology.

5.4 CENSORED DATA

5.4.1 The concept of censored data

Advanced

Censored data are different from missing data. Missing data are when you do not conduct the analysis – censored data are when you conduct the analysis but you do not obtain a quantifiable result. In monitoring programmes focussing on treatment plant performance and water quality evaluation, the true concentrations of a sample may be very low, close to zero, and in this case, the measured value may be **below the method detection limit (MDL)** (see Section 4.6.2). This stems from the limitations inherent to analytical methods and laboratory analyses, and usually the result is reported as a '**non-detect**' or with the value of the detection limit (DL) preceded by the sign of '**less than (<)**'. As you will see below, we also have cases of non-detect results that are above a particular detection limit. In both of these cases, we are not able to report the results in the same way we do for the other values that are within the detection range, and we say that these values are '**censored data**'.

S. 4.6.2

There are two types of censored data (see Figure 5.2):

- **Left-censored data**. The non-detects are **below** the detection limit DL and should be reported as 'less than MDL' or '<**MDL**'. This is the most common type of censored data in studies of treatment plant performance and water quality.
- **Right-censored data**. The non-detects are **above** the limit of quantifiable values and should be reported as 'greater than [a particular value]' or using the '>' sign. The case of right-censored data usually results from insufficient dilutions of the original sample; the concentration is still too high and the result to be read is above the maximum capacity of the method. For microbiological analyses involving plate counts, this result is also often reported as 'too numerous to count' (TNTC).

Censored data interfere in the calculation of descriptive statistics. If you treat censored values inappropriately, it can lead to biased estimates of measures of central tendency and variability, and it can potentially cause you to have misleading results for statistical tests of the difference between groups or

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

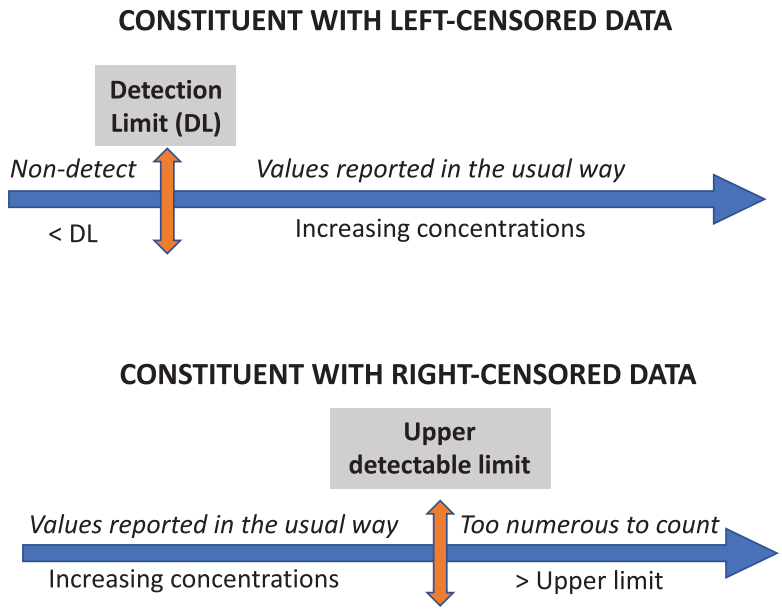


Figure 5.2 Representation of the two types of censored data: left-censored data (top) and right-censored data (bottom).

the development of regression models. However, these problems can be mitigated using appropriate techniques to handle censored data. **In particular, censored data should not be eliminated from the data set – deleting censored values will distort the results of your descriptive statistics and statistical analyses**. Treatment of censored data is a *topic widely covered in the statistical field* and in applications related to environmental and water quality data. There are sophisticated methods, but the approach adopted here is for a simple treatment of data.

Note that the way we treat the censored data will affect not only the measures of central tendency (mean and median) but also the measures of variability (standard deviation and relative standing). Also, the way censored data are treated will also affect estimated removal efficiencies.

Some researchers do not pay much attention to the considerations surrounding censored data, probably due to a generalized scepticism about the validity of the information contained in these observations. However, a lot of information is available in censored data, provided that appropriate methods for its extraction are used (Oliveira, 2017).

5.4.2 Treatment of left-censored data (below the DL)

Advanced

Basically, you have several options for treating left-censored data (non-detect values that are below the method detection limit MDL) – some are better than others, depending on the situation:

- **Option 1. Eliminate all non-detect values from your database.** This is an *incorrect approach*. By eliminating the low values from your data set, the descriptive statistics calculated from the remaining data will suggest higher values (e.g., a higher mean) than what was actually occurring.
- **Option 2. Substitute all non-detect values with zero.** This option is better than option 1, but it is also an incorrect approach, as it will also introduce a bias to your interpretation of the data. If you have a combination of detected values and non-detected values, most likely at least some of the non-detect

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

values were situations where the constituent was present in the sample, but at a concentration that was too low to be detected by the method used. In this case, by replacing all of these values with a value of zero, the resulting descriptive statistics will present lower values (e.g., a lower mean) than those actually occurring.

- **Option 3. Substitute the non-detects by the value of the MDL.** This is simply done by not taking into account the sign '<' that precedes MDL, and the value of the non-detect is kept as the MDL value. However, it will also introduce bias, because the resulting descriptive statistics will present higher estimated mean values than those actually occurring.
- **Option 4. Substitute the non-detects by a fraction of the MDL.** A value commonly used is **½MDL** (50% of the interval between zero and the detection limit). For instance, if the MDL = 0.10 mg/L, all non-detects are replaced by $0.10/2 = 0.05$ mg/L. This is a *good and simple approach*, but it still has limitations. For example, if the data are log-normally distributed (as environmental and water quality data frequently are), then using this substitution will still result in an overestimation of the mean value (though not as drastic of an overestimation as using option 3).
- **Option 5. Use more sophisticated statistical methods to impute non-detect values.** There are a number of more sophisticated and more accurate ways to calculate summary statistics for data sets that are censored, such as the use of Kaplan–Meier, maximum likelihood estimation (MLE), and regression on order statistics (ROS). A good review of these methods is provided by Helsel (2012).

It is interesting to note that the practice of *replacing censored data by any value between zero and the detection limit* is operationally simple and can be adequate, in practical terms, when the *percentage of censored data is low*. The following comments can be made (Oliveira & Gomes, 2011; Oliveira, 2017):

- **When proportion of non-detects is less than 20%.** *Substitution methods can be applied* when the proportion of censored data in terms of the whole data set is less than 20%.
- **When proportion of non-detects is less than 25%.** When less than 25% of the data are censored, the *interquartile range (IQR) (percentile 75% – percentile 25%)* may still be determined.
- **When proportion of non-detects is less than 50%.** When less than 50% of the data are below the detection limit, it is still possible to calculate some percentiles, such as the *median and the 25th percentile*.
- **When proportion of non-detects is high.** Unfortunately, for calculating the *arithmetic mean and standard deviation, the considerations above cannot be made*. In general, for data sets that present a high percentage of observations below the detection limit, the substitution of the censored data should be avoided. For these cases, there are other alternatives that can be selected and the correct choice of the method to be used depends both on the degree of censorship, which directly interferes in the results, and the type of application (descriptive statistics, confidence intervals, hypothesis tests, fitting to probability distributions, correlations, regression analyses, and trend analyses). Depending on the method used in the censored data treatment, the results may undergo substantial alterations, and their interpretation is impaired.
- **All measurements are non-detects.** In some situations, all measurements can be found below the detection limit of the analytical method, which still does not preclude the use of such data. Methods based on the binomial probability distribution can be used to extract important information from these data. Among them, we highlight the determination of confidence intervals, hypothesis tests for comparison between groups considering proportion, and calculation of the probability of violation of discharge standards.

Further information on statistical techniques for the treatment of censored data can be found in Cohen (1991), Helsel (2004, 2012), and Klein and Moeschberger (2005).

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Example

EXAMPLE 5.2 WORKING WITH LEFT-CENSORED DATA

You obtained monthly data on the concentration of a certain constituent in the effluent from a treatment plant (or in the water body you are studying). In total, there are 12 data, but you verify that 4 of them are below the method detection limit, which, in this case, is 0.10 mg/L. The data you obtained are presented below. Analyse the possibility of the utilization of substitution techniques for replacing the non-detects and also more advanced approaches.

Excel

Note: This example is also available as an Excel spreadsheet.

Date	Measured Concentration (mg/L)	Date	Measured Concentration (mg/L)
01/01/2018	0.20	01/07/2018	<0.10
01/02/2018	<0.10[1]	01/08/2018	0.21
01/03/2018	0.15	01/09/2018	0.19
01/04/2018	0.12	01/10/2018	0.12
01/05/2018	<0.10	01/11/2018	<0.10
01/06/2018	0.16	01/12/2018	0.11

[1]0.10 mg/L is the method detection limit.

Solution:

The proportion of non-detects is high: 4 out of 12 measurements (33.3%) are censored. Therefore, simple substitution methods may have strong limitations. Nevertheless, they will still be tried. Four simple substitution methods will be used: (i) substitute the non-detects by a blank value (remove the non-detects), (ii) substitute the non-detects by zero, (iii) substitute the non-detects by the value of the method detection limit (MDL), and (iv) substitute the non-detects by half the value of the detection limit (MDL/2). The following table can be produced, knowing that the detection limit MDL is 0.10 mg/L:

Date	Measured Concentration (mg/L)	Exclusion of <MDL Values (mg/L)	<MDL Values Substituted by Zero (mg/L)	<MDL Values Substituted by MDL (mg/L)	<MDL Values Substituted by MDL/2 (mg/L)
01/01/2018	0.20	0.20	0.20	0.20	0.20
01/02/2018	***<MDL***		***0.00***	***0.10***	***0.05***
01/03/2018	0.15	0.15	0.15	0.15	0.15
01/04/2018	0.12	0.12	0.12	0.12	0.12
01/05/2018	***<MDL***		***0.00***	***0.10***	***0.05***
01/06/2018	0.16	0.16	0.16	0.16	0.16
01/07/2018	***<MDL***		***0.00***	***0.10***	***0.05***
01/08/2018	0.21	0.21	0.21	0.21	0.21
01/09/2018	0.19	0.19	0.19	0.19	0.19
01/10/2018	0.12	0.12	0.12	0.12	0.12
01/11/2018	***<MDL***		***0.00***	***0.10***	***0.05***
01/12/2018	0.11	0.11	0.11	0.11	0.11

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

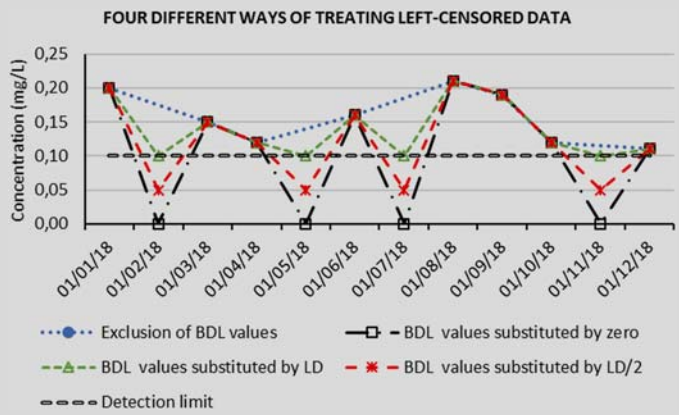
The descriptive statistics of the four data sets produced using the substitution methods are shown as follows:

Statistics	Exclusion of <MDL Values (mg/L)	<MDL Values Substituted by Zero (mg/L)	<MDL Values Substituted by MDL (mg/L)	<MDL Values Substituted by MDL/2 (mg/L)
Mean	**0.16**	**0.11**	**0.14**	**0.12**
Standard deviation	0.04	0.08	0.04	0.06
CV	0.25	0.80	0.30	0.51
25 percentile	0.12	0.00	0.10	0.05
50 percentile	0.16	0.12	0.12	0.12
75 percentile	0.19	0.17	0.17	0.17
Interquartile range	0.07	0.17	0.07	0.12

As was advocated before, the technique of replacing non-detects by half the value of the detection limit (MDL/2) is, among the simple substitution methods, the one likely to best allow further statistical treatment of the data. In this case, the mean was 0.12 mg/L, and also the median. The median of 0.12 was equal to those using other substitution techniques. But notice that the CV (=standard deviation ÷ mean) is very different in all situations. However, any conclusions are associated with this particular application. If we had a higher or a lower proportion of non-detects, the comments could be different. Also, if the detected values were much higher than the detection limit, we could have a distinct interpretation (in that latter case, it is possible that the data do not follow a normal distribution; see Chapter 8).

C. 8

The graph below shows the time series plot considering the four different treatments of non-detects. We can clearly see that different outcomes are obtained, depending on the substitution technique employed. Excluding the non-detects and also considering them equal to zero will produce time series that, on visual analysis, may leave you uncomfortable. Considering that the non-detects are equal to the method detection limit (MDL) leads to a more common type of graph while considering that the values of the non-detects are equal to half of the detection limit and will produce a time series that probably looks more reasonable to you.



Our example continues with a more advanced approach.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Advanced

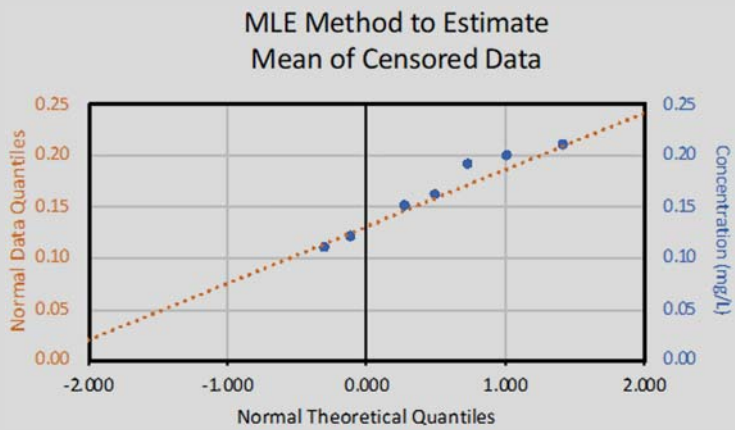
Gilbert (1987) describes a Maximum Likelihood Estimation (MLE) method that can be used to estimate the mean and standard deviation of a censored data set. We will not describe it here, but will exemplify it, and you can use the associated spreadsheet to obtain the necessary results and see how the calculations proceed.

Excel

Note: This example is also available as an Excel spreadsheet.

This method works by assuming a normal distribution of the data or the log-transformed data and then constructing a Quantile–Quantile (Q–Q) plot (see below) with the censored samples shown as omitted values at the lower end of the theoretical quantiles. Hypothetically, if our detection limit was lower, we would have detected those non-detect samples, and their values would have fallen along the line of the Q–Q plot. Using this approach, the mean is estimated to be **0.13** and the standard deviation is estimated to be **0.06**. In this particular example, these values are not much different from those calculated using the substitution techniques (see above). When using this approach, it is important that the data above the detection limit indicate a normal distribution. If the Q–Q plot shows a curved trend, it may be necessary to use a log transformation with the data. Further details on Q–Q plots will be presented in Section 8.2.8.

S. 8.2.8



5.4.3 Treatment of right-censored data (data above the DL)

Advanced

Adequate treatment of **right-censored data** (data above detection limit) is a more complex issue that may require advanced approaches, outside the scope of this book.

This is the case, for instance, in the determination of **coliforms** in water or wastewater samples. To comply with the detection limits of the laboratory method, we need to make *dilutions* to our original sample, because the actual values may be too high. If the dilutions we make are insufficient, we will not be able to come to a specific value, and thus need to report as '≥DL' (detection limit).

How to calculate measures of central tendency with right-censored data becomes a complex matter. A value above the maximum limit of quantification, for colony counts often reported as 'too numerous to count', could be any value, having 'infinite' as the upper boundary. Many researchers, in this case, adopt a *simple and practical approach* of estimating descriptive statistics **replacing the right-censored data by the value of the upper limit of quantification**. This approach will produce an average that is lower than the actual measure, but this is better than simply excluding the censored data. In the best case, if possible, you should repeat the analysis with a greater dilution factor to obtain quantifiable results.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

5.5 OUTLIERS

5.5.1 Concept of outliers and importance of their analysis

Basic

An **outlier**, as the name implies, is an observation that *lies outside* the values of the usual other observations in your sample. In other words, we can put this in a simple way (Mendenhall & Sincich, 1988):

> An **outlier** is an observation that is unusually large or small relative to the other values in the data set.

Outliers *can originate from problems or* ***errors*** *in your sample collection, sample preservation, laboratory analysis, transcription to the database, or any problem that may affect the reliability of your data*. After you detect that the value is anomalous, you should go back to the whole procedure used to obtain it and verify whether there have been problems that may cause this observation to be reported as a wrong value. Even if you are not able to identify the problems that caused this non-typical value, you may still consider that it is wrong, based on your pre-existing knowledge of treatment processes and methods of analysis. For instance, there are some circumstances where you measure one parameter that is essentially a subset of another parameter, and it would not make sense for the subset value to be larger than the overall value: for example, if you obtain a BOD value that is higher than the COD, or a volatile suspended solids (VSS) that is higher than the total suspended solids (TSS), or a soluble COD that is higher than the total COD, or a high TSS value in a sample in which the turbidity was very low, you know that something is wrong, and you may suspect of the values involved in this analysis. In this case, **if you identify errors, you have reasons to exclude the anomalous observations from your data set.**

But beware of a *very important statement related to treatment plant and water quality monitoring*. **Treatment plants and water bodies are highly dynamic** in their behaviour and frequently produce values that are not typical or not expected as part of their usual performance, but that, indeed, in that particular moment, reflect **a real phenomenon** that took place. This can happen in the influent and effluent concentrations, as well as in the inflow and in measurements of variables inside the tanks or reactors. Therefore, outliers can be a *very important element in the analysis of your plant dynamics*, and as such **should be thoroughly investigated**. We can learn a lot by trying to understand what caused such an unexpected value and, by digging into more data and information, you enhance your knowledge of the treatment plant or water body you are studying.

For instance, let us imagine that you obtained *monitoring data from the influent to a water treatment plant* (raw water). You have monthly measurements (a single measurement per month), and you notice that in October, the turbidity was unusually high (see Figure 5.3, left). You could have hastily considered this value to be an outlier and could have excluded it from your database. But you know that turbidity can be related to the run-off of suspended solids from the catchment area, especially during rainfall events. You then obtained data from precipitation, plotted it together with influent turbidity (see Figure 5.3, right), and saw that in October there had been high precipitation levels. Therefore, this could have been the reason for the unusually high turbidity value, and you then decide that it is worth to keep the outlier, unless additional information suggests that it is really a wrong value.

Now, let us analyse one example from the *effluent from a wastewater treatment plant*, also monitored with one sample per month. You obtained COD concentration values and clearly identified an anomalous observation in April (Figure 5.4, left). You knew you could not discard this value without

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

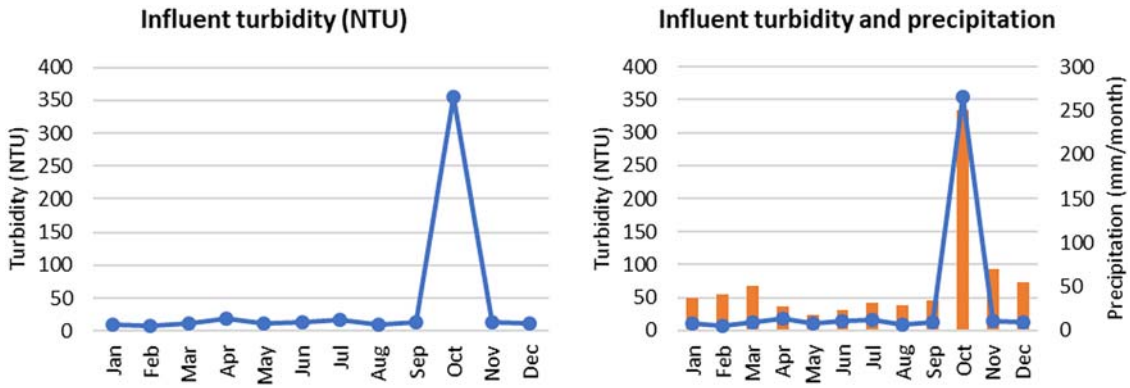


Figure 5.3 Time series of turbidity values, with an outlier in October (left). Plotting of turbidity and precipitation, and identification of a possible reason for the high turbidity value in October (right).

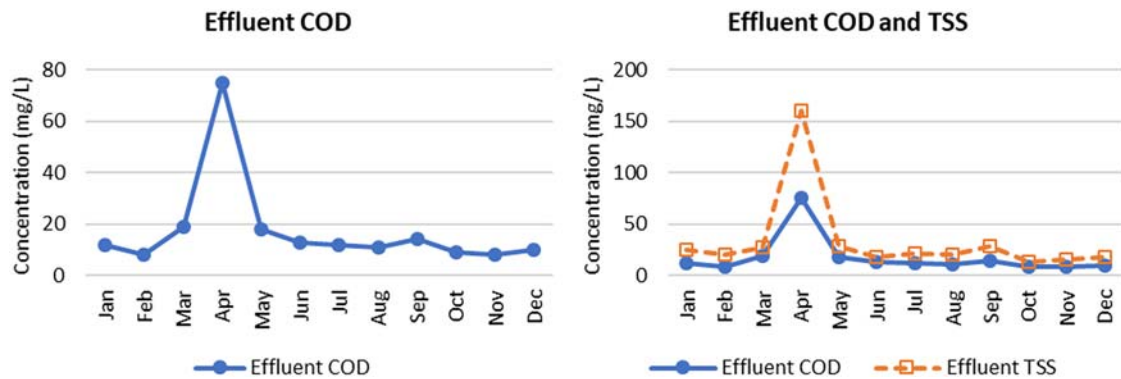


Figure 5.4 Time series of effluent COD concentrations, with a peak value in April (left). Plotting of COD and TSS, and identification of a possible reason for the high COD value in April (right).

further investigation. You then obtained data from TSS and plotted it together with COD (Figure 5.4, right) and noticed that also in April there was a peak value in TSS. Then, you got the logbook of the operator from the treatment plant and found the observation that in April there was a pump failure, and settled sludge could not be removed from the secondary clarifier, what caused solids loss in the effluent. You found a reasonable explanation and decide to keep both values.

Now, we will move into a different example highlighting the importance of due consideration of outliers before simply discarding them. Let us assume you are using a **dynamic mathematical model** of your plant. If your model is dynamic and is considered a good model, it should be able to pick up the plant instabilities, and the simulated values should show the main ups and downs of your measured concentrations (provided they are not associated with errors, as discussed above). Let us take the example shown in Figure 5.5 (left). You are trying to model a plant that is relatively stable, and your model systematically underestimates the observed values (all simulated values are lower than the measured values). If you carry out an analysis of the goodness-of-fit of your model (see Chapter 15),

C. 15

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

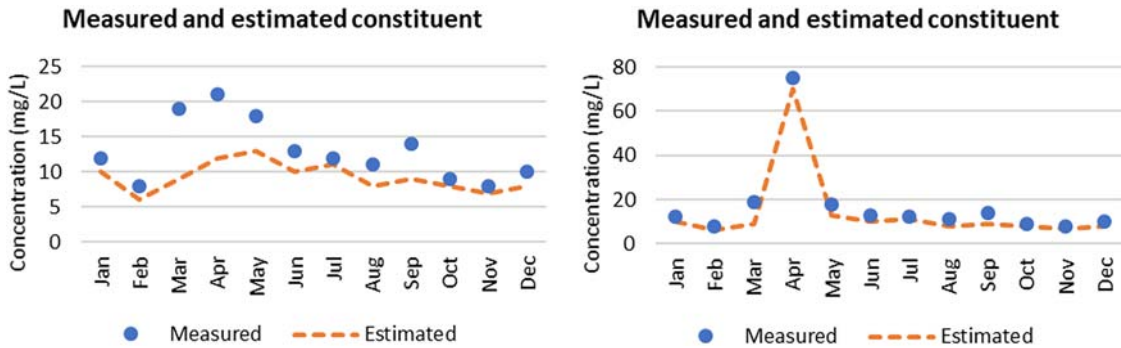


Figure 5.5 Measured and estimated values for a certain treatment plant constituent. Poor simulation of a stable time series (left) and good simulation of an unstable time series (right).

you will probably get disappointing indicators of model performance. But now let us analyse the situation in which all the measured values were the same, with the exception of the April value, which was exceptionally high. You run your model and celebrate the fact that it was able to pick up the peak value (Figure 5.5, right). Even though all your simulated values are below the measured ones (as a matter of fact, equal to those in the left graph, with the exception of the April value), your model was able to reproduce the main trend, and now you should get much better goodness-of-fit statistics.

5.5.2 Determination of outliers

Advanced

In the preceding section you analysed the possible occurrence of outliers by visual inspection of data plotted in a graph, together with your personal interpretation of non-typical values. In many cases, this should be sufficient, but in other cases, you need to apply a formal procedure for the detection of outliers. There are different formal ways of identifying outliers, but we will see here a simple but widely used method.

S. 5.8

In Section 5.8, we will cover the concept of percentiles in more detail (such as the definition of the first and third quartiles and the interquartile range). However, here is a brief description of these concepts for now:

- The **first quartile ($Q1$)** corresponds to the **25 percentile**, meaning that 25% of the data have a value that is less than or equal to $Q1$.
- The **third quartile ($Q3$)** corresponds to the **75 percentile**, meaning that 75% of the data have a value that is less than or equal to $Q3$.
- The **difference between the third quartile ($Q3$) and the first quartile ($Q1$)** is the so-called **interquartile range IQR** ($=Q3 - Q1$), meaning that $75 - 25 = 50\%$ of the data lies in the interval between $Q3$ and $Q1$.

Based on these statistics, we can define the lower and upper limits for the detection of outliers (Mendenhall & Sincich, 1988; Naguettini & Pinto, 2007; Oliveira, 2017), represented in Equations 5.1 and 5.2 and illustrated in Figure 5.6.

$$\text{Lower limit for outliers (LL)} = \text{Q1} - 1.5 \times \text{IQR} \tag{5.1}$$

$$\text{Upper limit for outliers (UL)} = \text{Q3} + 1.5 \times \text{IQR} \tag{5.2}$$

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

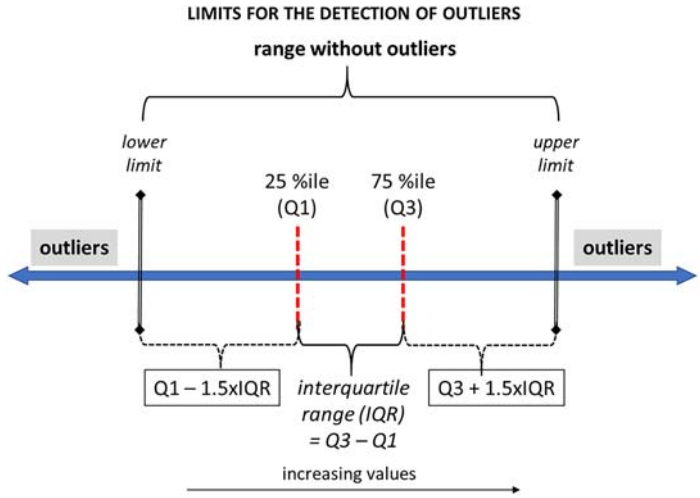


Figure 5.6 Scheme for the detection of outliers based on the interquartile range (IQR).

Example

EXAMPLE 5.3 DETECTION OF OUTLIERS

You obtained data on the concentration of COD in the effluent from the treatment plant (or in the water body) you are studying. In total, there are 20 data collected over a month (there were some days without sampling). Analyse the presence of outliers in your data set.

Excel

Note: This example is also available as an Excel spreadsheet.

Data:

Date	Effluent COD (mg/L)	Date	Effluent COD (mg/L)	Date	Effluent COD (mg/L)
10/04/2013	63	20/04/2013		30/04/2013	
11/04/2013	37	21/04/2013	62	01/05/2013	81
12/04/2013		22/04/2013	53	02/05/2013	134
13/04/2013		23/04/2013	50	03/05/2013	
14/04/2013	50	24/04/2013	61	04/05/2013	
15/04/2013	44	25/04/2013	66	05/05/2013	104
16/04/2013	51	26/04/2013		06/05/2013	142
17/04/2013	49	27/04/2013		07/05/2013	95
18/04/2013	57	28/04/2013	73	08/05/2013	79
19/04/2013		29/04/2013	83	09/05/2013	

Note that there are missing data, which are common in monitoring programmes. The information on 'date' is not necessary for this example, but it is included to allow the making of a time series graph, which will illustrate the results.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Solution:

Using the Excel function PERCENTILE for a range, with the percentile value (K) of 0.25, we obtain the value of the first quartile ***Q1*** (25^{th} percentile) equal to **51**.

Similarly, using the Excel function PERCENTILE for a range, with the percentile value (K) of 0.75, we obtain the value of the third quartile ***Q3*** (75^{th} percentile) equal to **82**.

Therefore, **IQR** is $Q3 - Q1 = 82 - 51 = \mathbf{31}$.

According to Equation 5.1, the lower limit (LL) for outliers is

$$\text{Lower limit for outliers (LL)} = Q1 - 1.5 \times IQR = 51 - 1.5 \times 31 = \mathbf{5}$$

According to Equation 5.2, the upper limit (UL) for outliers is

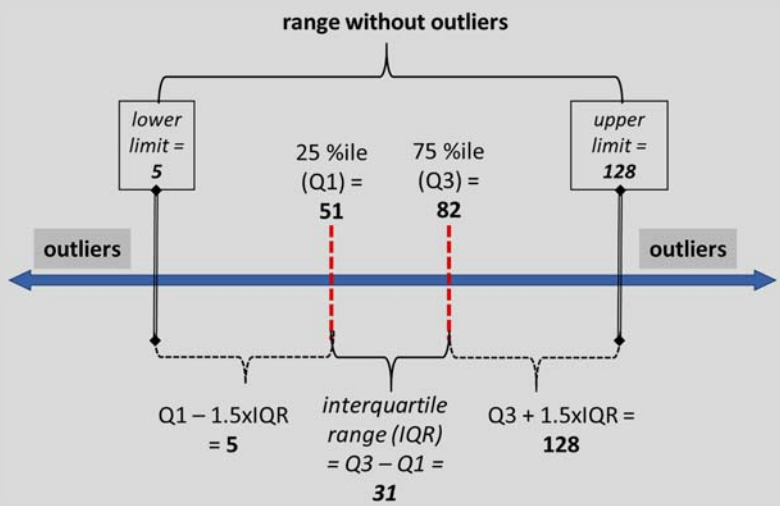
$$\text{Upper limit for outliers (UL)} = Q3 + 1.5 \times IQR = 82 + 1.5 \times 31 = \mathbf{128}$$

Based on your data set and the calculated lower and upper limits for outliers, you obtain the following summary:

Item	Absolute Values	Relative Values (%)
Total number of data	20	100
Number of outliers below the lower limit	0	0
Number of outliers above the upper limit	2	10
Total number of outliers	2	10

Therefore, you detected the presence of two outliers, based on the criterion used for outlier detection. This corresponds to 10% of your data set. The two values are related to data above the upper limit for outliers. No outliers below the lower limit were found (the minimum value in your data set is 37 mg/L, which is above the lower limit of 5 mg/L). From this, you will now investigate what may have caused the occurrence of these two outliers, and whether they should be maintained or excluded.

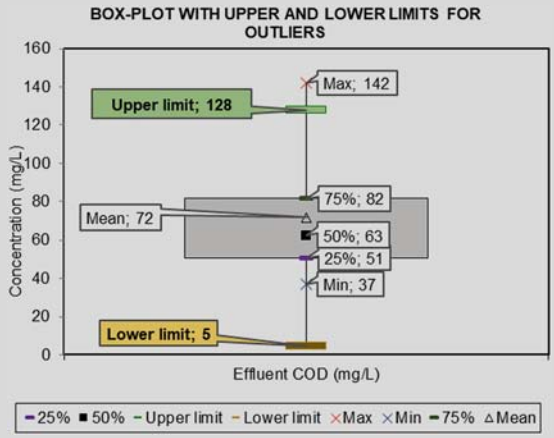
Your scheme looks like this



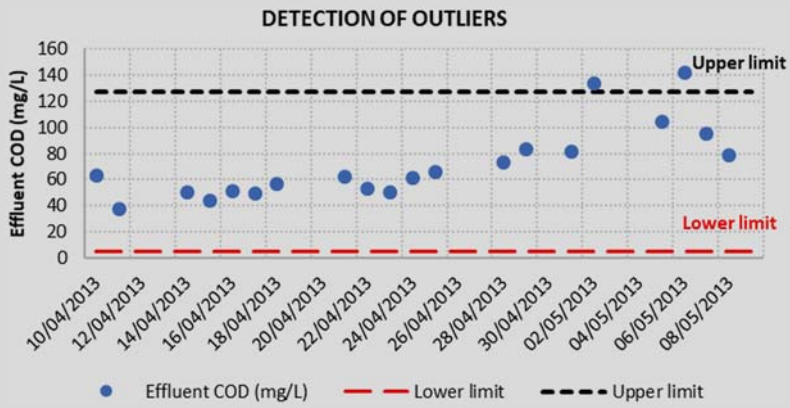
Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

S. 6.4

Your box-plot, with the indication of the 25 and 75 percentiles, together with the lower and upper limits for outliers, plus additional information, is shown as follows (see Section 6.4 for learning how to construct and interpret a box-plot graph):



The time series graph of your data, together with the lower and upper limits for outliers, is shown as follows:



You can easily identify the location of the two outliers above the upper limit. Although they have been identified as outliers, they are not very far away from the last values of your monitoring, which seemed to indicate an increasing trend. You could consider this in your analysis of possible explanations of the outliers.

5.6 MEASURES OF CENTRAL TENDENCY

5.6.1 Introduction

Basic

When analysing your data, you frequently need to calculate and interpret the **measures of central tendency** of the data. They are important for virtually all evaluations you make on flows, concentrations, loads, and

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

removal efficiencies, and they are an integral part of a large number of statistical analyses, several of them included in this book. The most widely used measures of central tendency are

- Mean
- Median
- Geometric mean
- Mode
- Weighted average

S. 6.3
C. 8

Mean is the most extensively used measure of central tendency and will for sure be part of any report you do on monitoring data. We will also emphasize the importance of the **median** in the case of treatment plant and water quality data, due to the fact that the distribution of data usually is not symmetrical (this will be analysed in detail in Section 6.3 and Chapter 8). The **geometric mean** is also very important in the case of treatment plant and water quality data, especially when the range of values varies by orders of magnitude, which is the case of coliforms and many environmental contaminants. **Mode** is not frequently used in our case and will be only mentioned briefly. **The weighted average** is widely used in treatment plant practice (even though we may even not notice it), every time we sum up loads and divide by the total flow (the loads are the concentrations multiplied by a weighting factor, which, in this case, is the flow associated to each measured concentration).

The most important concepts for these five measures of central tendency are presented below and they are further explained in the following sections of this chapter.

Main measures of central tendency of interest for monitoring data (partially adapted from Mendenhall & Sincich, 1988)

Mean. The arithmetic mean $\bar{x}$ of a sample of n measurements x_1, x_2, …, x_n is the average of the measurements

$$\bar{x} = \frac{(x_1 + x_2 + \cdots + x_n)}{n}$$

Median. The median of a sample of n measurements x_1, x_2, …, x_n is the middle number when the measurements are arranged in ascending (or descending) order, i.e., the value of x located so that half the area under the frequency distribution curve lies to its left and half the area lies to its right.

Geometric mean. The geometric mean $\bar{x}_g$ of a sample of n measurements x_1, x_2, …, x_n is the product of the measurements raised to the power $1/n$

$$\bar{x}_g = (x_1 \times x_2 \times \cdots \times x_n)^{(1/n)}$$

Mode. The mode of a sample of n measurements x_1, x_2, … , x_n is the value of x that occurs with the greatest frequency, that is, the peak point in the frequency distribution graph.

Weighted average. The weighted average $\bar{x}_w$ of a sample of n measurements x_1, x_2, … , x_n is the sum of the measurements multiplied by their respective weights w_1, w_2, … , w_n divided by the sum of the weights

$$\bar{x}_w = \frac{(x_1 w_1 + x_2 w_2 + \cdots + x_n w_n)}{(w_1 + w_2 + \cdots + w_n)}$$

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

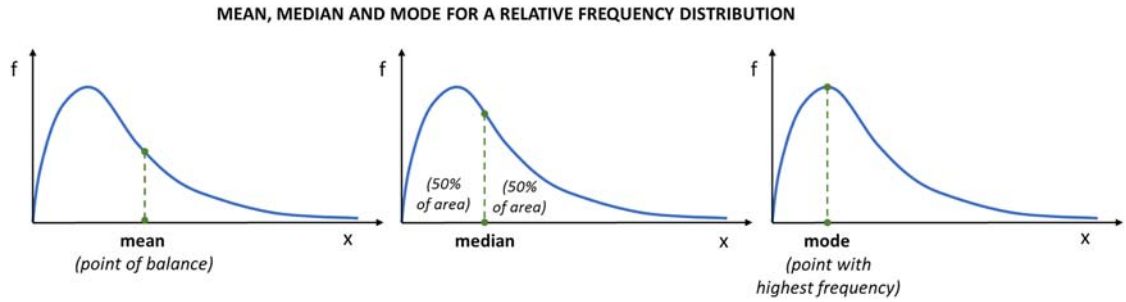


Figure 5.7 Interpretation of the mean, median, and mode for a typical frequency distribution found in treatment plant and water quality monitoring.

S. 6.3

We have not yet discussed frequency histograms or frequency distributions, but they will be covered in Section 6.3. Still, you may already have some knowledge about the meaning of frequency histograms or distributions, and we can use this to illustrate the relationship between these measures of central tendency and the shape of the distribution. In Figure 5.7, you can see the interpretation of the mean, median, and mode for a typical frequency distribution found in water bodies and treatment plant monitoring. The concepts of *point of balance* (for **mean**) and *percentage of areas* (for **median**) are also illustrated in the figure. *For a perfect log-normal distribution (see Section 8.3), the geometric mean (not shown in the figure) will be equal to the median.*

S. 8.3

A comparison of the relative positions of the mean, median, and mode for different forms of the frequency distribution (symmetrical, skewed-to-the-right, and skewed-to-the-left) is illustrated in Figure 5.8. The distributions shown are unimodal, that is, have only one mode. From the figure, you can make the following inferences:

- **Perfectly symmetrical** distribution: ***mean*** = ***median*** = ***mode***.
- Distribution **skewed-to-the-right** (tail on right side): ***mean*** > ***median*** > ***mode***.
- Distribution **skewed-to-the-left** (tail on left side): ***mean*** < ***median*** < ***mode***.

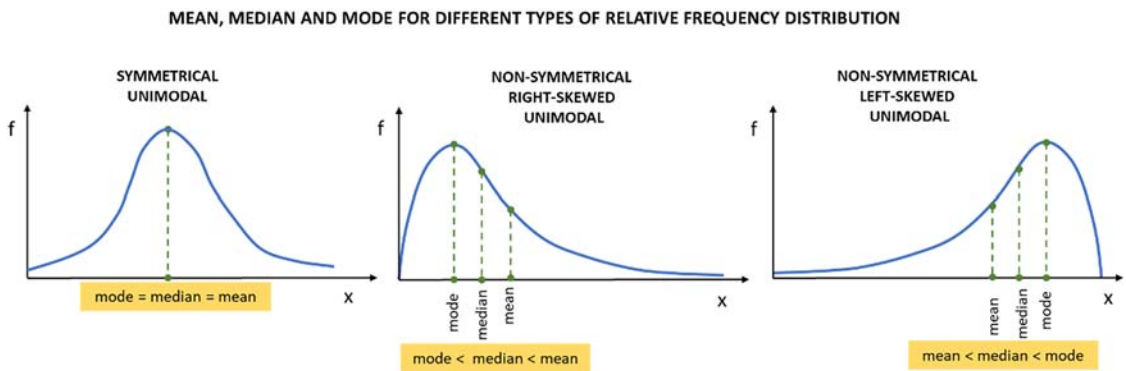


Figure 5.8 Relative position of the mean, median, and mode in different types of frequency distributions.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

S. 7.7

Skewed-to-the-right distributions are frequently found for concentrations (influent and effluent) in a treatment plant and also in water bodies, whereas *skewed-to-the-left distributions are common for removal efficiencies* (see Section 7.7). *As mentioned above, for a theoretical log-normal distribution* (see Section 8.3), *the geometric mean will be equal to the median.*

All these measures of central tendency have specific **Excel functions** to allow direct calculation after selecting your data range.

After this analysis, what is the best measure of central tendency? In this book, we will recognize the **importance of presenting means** in your reports, because of their widespread use, and will emphasize the **appropriateness of incorporating medians**, given the nature of most data involved in monitoring of treatment plants and water bodies. Furthermore, we will stress the **convenience of including geometric means** for variables that have a wide variability range. **Modes** will be mentioned only in very specific situations.

5.6.2 Mean

Basic

You are probably already very familiar with the concept of arithmetic mean and have likely used it several times. The arithmetic mean is sometimes referred to as the 'average' value of a data set. However, we use the term arithmetic mean here specifically to distinguish it from the geometric mean, which is different but also commonly used for treatment plant and water quality data sets. We will include here some concepts to reinforce your understanding of this very important measure of central tendency.

The arithmetic mean $\bar{x}$ of a sample of n measurements $x_1, x_2, \ldots, x_n$ is given as follows:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{5.3}$$

In **Excel**, you can use the function **AVERAGE** and some of its variations to calculate directly the mean of your data set.

The arithmetic mean works like a point of balance of your data, and you can look at it as a scale, with the left and right arms perfectly at equilibrium around the mean value. To illustrate this point, let us see Example 5.4, applied for a general constituent. After that we will see an example for coliforms, which are known for their wide variability (Example 5.5).

Example

EXAMPLE 5.4 CALCULATION OF MEAN AND ANALOGY WITH A POINT OF EQUILIBRIUM

You obtained data on five measurements of a variable. Calculate the mean and make the analogy of the point of equilibrium in a scale.

Excel

Note: This example is also available as an Excel spreadsheet.

Data:

4 2 5 1 7

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

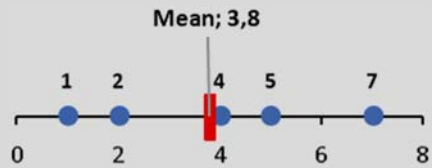
Solution:

The mean is simply calculated using Equation 5.3.

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{4 + 2 + 5 + 1 + 7}{5} = 3.8$$

Note: The mean should have the same number of decimal places as the original data. For instance, in this example, the data are integer values, and so should be the mean (and the other values of central tendency). However, to make our calculations clearer, we will keep the decimal value calculated.

Now, analyse the drawing below, representing the concept of a scale. To the left of the scale are two values that are lower than the mean (3.8), and to the right, three values that are higher than the mean.



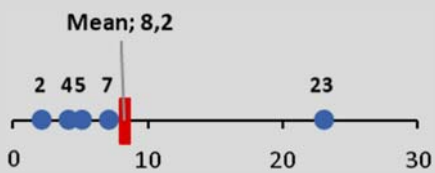
If we calculate the differences between each value and the mean, we get the following table:

x	**Difference from Mean** $x_i - \bar{x}$
4	$4 - 3.8 = 0.2$
2	$2 - 3.8 = -1.8$
5	$5 - 3.8 = 1.2$
1	$1 - 3.8 = -2.8$
7	$7 - 3.8 = 3.2$

The sum of the negative values (left arm of the scale) is $(-1.8) + (-2.8) = -4.6$.
The sum of the positive values (right arm of the scale) is $0.2 + 1.2 + 3.2 = +4.6$.

As expected, from the concept of point of balance, the sum of the negative values (left arm of the scale), -4.6, is equal to the sum of the positive values (right arm of the scale), $+4.6$. Therefore, all sums lead to zero or a perfect balance of the scale.

Now, let us analyse a different situation. Let us imagine that one of the measurements, say, the penultimate one, instead of being '1', had a much higher value (23). The mean will now be 8.2, of course, much higher than the previous value. The new scheme of the scale is presented as follows:



Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

The concept of equilibrium point holds, and we still have the balance of values around the mean. But what is now noteworthy is that the number of points to the left and to the right are completely different. The high value (23) pushes the mean to a much higher value, in this case, greater than the other four measurements. Now there are four points to the left of the mean, and only one point to the right. Because of this you may now consider that this arithmetic mean cannot be considered a good measure of central tendency anymore and will look for additional statistics to fulfil this role, one of them being the **median** (we will discuss it in Section 5.6.3).

S. 5.6.3

Example

EXAMPLE 5.5 CALCULATION OF MEAN AND ANALOGY WITH A POINT OF EQUILIBRIUM. THE CASE OF COLIFORMS

You obtained data on four analyses of coliforms (say, E. coli). Calculate the mean and make the analogy of the point of equilibrium in a scale.

Excel

Note: This example is also available as an Excel spreadsheet.

Data:

50 400 2000 3000 MPN/100 mL

Solution:

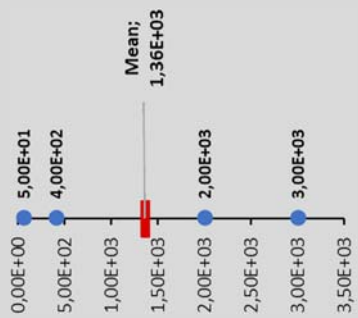
In scientific notation, which is the notation usually adopted for coliforms, your data are expressed as

5.00×10^1 4.00×10^2 2.00×10^3 3.00×10^3 MPN/100 mL

or in Excel format

5.00E + 01 4.00E + 02 2.00E + 03 3.00E + 03 MPN/100mL

The arithmetic mean is calculated using Equation 5.3, and you obtain the value of **1.36×10^3 MPN/100 mL**. You make the scale-plot and obtain the graph below. You feel comfortable because the mean seems to represent well a measure of central tendency: there are two points to the left and two points to the right. Although the points have different distances to the mean, the scale is balanced, and the sum of the negative and positive distances is, as expected, equal to zero.



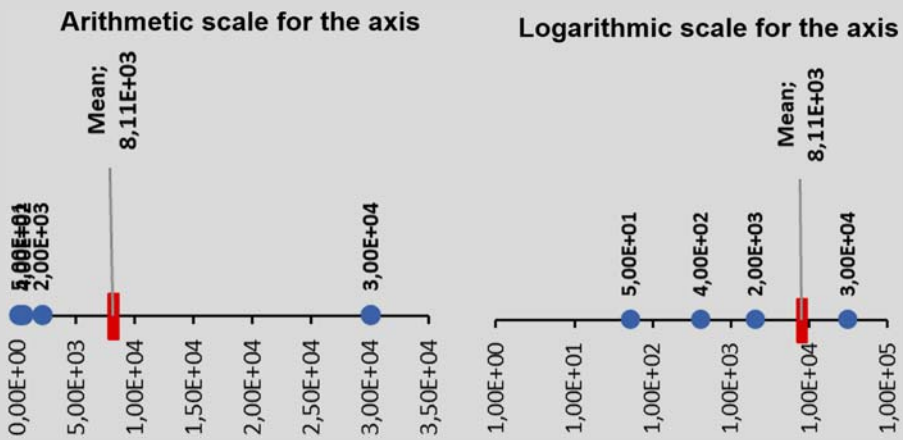
But now let us imagine that your last data, instead of being 3.00×10^3 MPN/100 mL, was 10 times higher, that is, 3.00×10^4 MPN/100 mL. This is indeed a possibility with coliforms, whose concentrations may vary widely, covering different orders of magnitude. The mean of the data set is

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

now **8.11 × 10^3 MPN/100 mL**, approximately six times the previous value. You make the scale-plot and obtain the graph shown at the end of the example (on the left).

Your graph now seems confusing, with data overlapping on the left-hand side and a far away value on the right-hand side of the mean. You then *question yourself whether this mean is really a good representation of central tendency*, because of the uneven distribution of points around the mean (in spite of the fact that the sum of negative and positive differences is still zero). Because of this, we will discuss later on other types of central tendency statistics, with special attention to **geometric means**, *which are the preferred central tendency measure for* ***coliforms*** (and other microbial pollutants) (see Section 5.6.4).

You may improve the plot by selecting a log scale for the axis. This is easily done in Excel, just by ticking on the option of 'logarithmic scale' after you select your axis in the graph. The graph now looks like the one shown below (on the right). This graph is much clearer than the previous one, and you can visualize all points on both sides of the mean. *Graphs with* ***logarithmic scale*** *are the preferred choice for plotting* ***coliform*** *data*.



Concluding remarks about the arithmetic mean. *Very important and widely used measure of central tendency, but substantially influenced by extreme values, even if they are present in lower frequencies, which is a typical situation for treatment plant and water quality monitoring data*.

5.6.3 Median

Basic

In the description of the mean, you saw its advantages but also realized that it is affected by extreme values, what frequently happens with treatment plant and water quality monitoring data. The **median** is another *widely used measure of central tendency, which is robust to the interference of extreme values*. The following concepts apply (partially adapted from Mendenhall and Sincich, 1988; Levine *et al.*, 1998):

- The median of a sample of n measurements $x_1, x_2, \ldots, x_n$ is the **middle number** when the measurements are arranged in an ordered sequence (ascending or descending).
- If there are no repeated values, *half of the observations will be lower than the median and half the values will be higher*.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

- The median is the value of x located so that *half the area under the frequency distribution curve lies to its left and half the area lies to its right*.
- If the number of measurements in a data set is **odd**, the median is the measurement that falls in the *middle* when the data are arranged in an ordered sequence. For example, in the data set from Example 5.4, we had five (odd number) measurements: 4, 2, 5, 1, 7. If we put them in increasing order, we have 1, 2, **4**, 5, 7. The value in the middle is 4, and so the median of the measurements is 4.
- If the number of measurements is **even**, the median is computed as the *mean of the two middle measurements* in the ordered sequence. For example, in the data set from the second part of Example 5.5, we had four (even number) measurements, listed in increasing order: 50, 400, 2000, 30,000. The two middle measurements are 400 (second observation) and 2000 (third observation). The mean of both observations is $(400 + 2000) \div 2 = \mathbf{1200}$. Therefore, the median of this data set is 1200.
- The median is *not affected by extreme values in your data set*. In the observations above, the median (1200) represents well the central tendency, since, in this case, two values are lower than it and two values are higher. The extreme value of 30,000 had no influence on the median. If it were only 3000, as in the first part of Example 5.5, the median would still be the same (1200).
- If the *frequency distribution is perfectly symmetrical and unimodal, then the median is equal to the mean*. If the distribution is not symmetrical and is *skewed-to-the-right*, the *median will be smaller than the mean*. If the distribution is not symmetrical and is *skewed-to-the-left*, the *median will be greater than the mean*. See Figure 5.7 for the illustration of these three situations. Therefore, the ratio mean/median for a *perfect normal distribution* is 1, and you will see in Section 8.3.5 how to estimate the ratio mean/median for a *perfect log-normal distribution*.
- The median is the same as the **percentile 50%** or **mid-quartile** (see Section 5.8, related to the measures of relative standing).
- In **Excel**, you can use the function **MEDIAN** to directly calculate the median of your data set. You can also use the **PERCENTILE** function, having $k = 0.5$ (since the median is the 50th percentile).

S. 8.3.5
S. 5.8

S. 5.5.2

Now, we have another important factor in the **comparison between mean and medians**. As pointed out in Section 5.5.2, **outliers** *may be key elements in the performance of your treatment plant or in the behaviour of your water body*. A statistic that is not influenced by extreme values, such as the **median**, is good for giving you an idea of the central tendency of your data and allowing some additional statistical tests. However, the extreme values that were not taken into account in the **median**, but were considered in the calculation of the **mean**, may be very important in your treatment plant, representing peak values in the influent or values not complying with the discharge standards in the final effluent. **Mass balances** (see Chapter 12) should take into account all masses entering and leaving your system, including the extreme values. Therefore, *for mass balances, mean values are more adequate*.

C. 12

The example of the calculation of the median will be included in Example 5.7, which also determines the mean and the geometric mean of the data set presented in Example 5.3.

5.6.4 Geometric mean

Advanced

For variables whose *values vary within several orders of magnitude*, it is more convenient to utilise the **geometric mean** (M_g) instead of the arithmetic mean. This is the case, for instance, in the monitoring of coliforms and other microorganisms in water and wastewater systems, which typically vary within a very wide range. As an example, coliform concentrations in raw sewage may vary between 10^6 and 10^9 MPN/100 mL. The higher values have a great weight on the arithmetic mean, distorting the concept of

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

the mean as a measure of central tendency, as seen in Example 5.5. In the range cited, the upper value of 10^9 MPN/100 mL is 1000 ($=10^3$) times the lower value of 10^6 MPN/100 mL, what highlights the very large width of the range and the span of different orders of magnitude. The calculation of the geometric mean is presented and discussed below (von Sperling & Chernicharo, 2005).

The geometric mean is given by the n root of the product of the n terms

$$\text{Geometric mean, } M_g = (x_1 \times x_2 \times \cdots \times x_n)^{1/n} \tag{5.4}$$

Now compare this equation for calculating geometric means (Equation 5.4) with the equation for calculating arithmetic means (Equation 5.3) and notice the similarities in their structure. In the calculation of **arithmetic means**, the *terms are additive* ($x_1 + x_2 + \cdots + x_n$), whereas in **geometric means**, the *terms are multiplicative* ($x_1 \cdot x_2 \ldots . x_n$). In **arithmetic means**, you *multiply the sum of the terms by 1/n* (or divide the sum by n), while for **geometric means**, you *raise the product of the terms to the power 1/n*.

Geometric means are also related to the logarithm of the original values. As such, the geometric mean can also be calculated by

$$\text{Geometric mean, } M_g = 10^{(\text{arithmetic mean of the log 10 of the original values})} \tag{5.5}$$

The following statement is also important and easily obtainable from the considerations above:

$$\text{Log}_{10} \text{ of the geometric mean} = \text{arithmetic mean of the } \log_{10} \text{ values} \tag{5.6}$$

We have seen the relationship between geometric means and arithmetic means. What about the relationship between geometric means and medians? This will depend on the distribution of your data. For a **perfect log-normal distribution**, the *geometric mean is equal to the median*.

A practical aspect that you need to take into account is that for the calculation of **geometric means you cannot have any value equal to zero**, otherwise you will get an **error** message in your calculation (you cannot calculate the $\log_{10}$ of zero). This may be the case, for instance, with some specialized or non-standardised laboratory analyses where the method detection limit is not always reported. One example is with the analysis of helminth eggs. The detection limit for the analysis of helminth eggs is frequently 1 egg/L, but may vary because it is related to the volume of sample concentrated and the fraction of the concentrated sample that is analysed on the microscope. Some labs do not always report their detection limits for this method, often reporting non-detect values as a 'concentration of 0 eggs/L'. Suppose you obtain the following results from five different samples: 8, 14, 3, 0, 5 eggs/L. Because one of your values has been reported as zero, you cannot calculate the geometric mean. However, you can calculate the arithmetic mean and the median in the usual way, and you will obtain mean = 6 egg/L and median = 5 eggs/L. If you verify the detection limit with the laboratory, instead of reporting a zero value, you could consider reporting it as a *value below the detection limit* (see Section 5.4.2). A similar comment could be made for coliforms in drinking water samples (where the detection limit is frequently one MPN or CFU per 100 mL).

S. 5.4.2

Example 5.6 shows you how to calculate geometric means. However, you can make the calculations directly by using the **Excel** function **GEOMEAN**. Note that, if you have a very large data set, with very high values in your measurements, such as those for coliforms in sewage, the Excel function may give an error, because the multiplication of all values may lead to an extremely high value, outside the allowable calculation range in Excel. Take the case of only these three data: 10^8, 10^7, and 10^9. Their

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

product will be $10^8 \times 10^7 \times 10^9 = 10^{(8+7+9)} = 10^{24}$, which is a high value. Now imagine a very large data set, with hundreds of values – their product will be an extremely high value. In this case, if you get an error in the calculation, use the method given in Equation 5.5.

You should bear in mind that the geometric mean is **useful** in the context we have explained, and it should be used instead of the arithmetic mean whenever you have data for microorganisms (e.g., coliforms, *E. coli*), but it can be **difficult to convey** the concept of the geometric mean in a short oral presentation or in a simple report for an audience or readership not familiar with the concept. However, this does not mean you should not use it!

Example

EXAMPLE 5.6 CALCULATION OF THE GEOMETRIC MEAN

You obtained data on four analysis of E. coli. Calculate their geometric mean.

Data *(same data as the second part of Example 5.5)*

50 400 2000 30,000 MPN/100 mL

Solution:

In scientific notation, which is the notation usually adopted for coliforms, your data are expressed as

5.00×10^1 4.00×10^2 2.00×10^3 3.00×10^4 MPN/100 mL

or in Excel format

5.00E + 01 4.00E + 02 2.00E + 03 3.00E + 04 MPN/100mL

You then calculate their $\log_{10}$ values and express this in a table format.

Coliform data (original data and log transformation).

Data	*E. coli* (MPN/100 mL)	Log_{10} (*E. coli*)
1	5.00×10^1	1.699
2	4.00×10^2	2.602
3	2.00×10^3	3.301
4	3.00×10^4	4.477

Applying Equation 5.4

Geometric mean, $M_g = (x_1 \cdot x_2 \cdot x_3 \cdot x_4)^{1/4} = (50 \times 400 \times 2000 \times 30{,}000)^{1/4} = 1047$

$= \mathbf{1.047 \times 10^3\ MPN/100\ mL}$

The geometric mean can also be calculated using Equation 5.5. In the example, the arithmetic mean of the $\log_{10}$ of the *E. coli* values presented in the table is

Arithmetic mean of the logarithms $= (1.699 + 2.602 + 3.301 + 4.477)/4 = 3.020$

Hence,

Geometric mean, $M_g = 10^{(3.020)} = 1047 = \mathbf{1.047 \times 10^3\ MPN/100\ mL}$

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

The value found is, of course, equal to the one obtained from Equation 5.4.
The calculation using Equation 5.6 is another way of obtaining the arithmetic mean of the logarithms

$$\text{Log}_{10}(1.047) = 3.020$$

In case the **arithmetic mean** of the original coliform data had been calculated, the following value would have been obtained: 8113 MPN/100 mL = **8.113×10^3 MPN/100 mL**. As discussed in Example 5.5, this value is much higher than that found through the geometric mean, being greater than three out of the four data available, and not giving, therefore, a good indication of the central tendency of the data.

S. 5.6.3

Now let us compare these values with the **median**. Using the instructions given in Section 5.6.3, if the number of measurements is even, the median is computed as the *mean of the two middle measurements* in the ordered sequence. In our case, we have four (even number) measurements, which, listed in increasing order, are: 50, 400, 2000, and 30,000. The two middle measurements are 400 (second observation) and 2000 (third observation). The mean of both observations is $(400 + 2000) \div 2 = 1200$. Therefore, the median of this data set is $1200 = \mathbf{1.200 \times 10^3}$ **MPN/100 mL**. This value is close to the geometric mean value and is also a better representation of central tendency compared with the arithmetic mean.

Example

EXAMPLE 5.7 CALCULATION OF MEAN, MEDIAN, AND GEOMETRIC MEAN USING EXCEL FUNCTIONS

Using the same data from Example 5.3 (effluent COD), compute mean, median, and geometric mean of the data set.

Data:

63	37	50	44	51	49	57	62	53	50
61	66	73	83	81	134	104	142	95	79

Solution:

Using the Excel functions MEAN, MEDIAN, and GEOMEAN, we obtain the following results:

- Mean = 72 mg/L
- Median = 63 mg/L
- Geometric mean = 67 mg/L

The mean and the median had already been presented in a box-plot in Example 5.3.

5.6.5 Weighted average

Basic

The weighted average (or weighted arithmetic mean) is similar to the arithmetic mean. However, the major difference is that, instead of each data point contributing equally to the average, each data point has a weight associated with it. The conventional arithmetic mean is a particular case, in which all weights are equal to 1.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

The weighted average $\bar{x}_w$ of a sample of n measurements $x_1, x_2, \ldots, x_n$ is the sum of the measurements multiplied by their respective weights $w_1, w_2, \ldots, w_n$, divided by the sum of the weights, given as follows:

$$\bar{x}_w = \frac{(x_1 w_1 + x_2 w_2 + \cdots + x_n w_n)}{(w_1 + w_2 + \cdots + w_n)} = \frac{\sum_{i=1}^{n} x_1 w_i}{\sum_{i=1}^{n} w_i} \tag{5.7}$$

If we divide each weight (w_i) by the sum of the total weights ($\sum w_i$), we obtain the relative participation of each weight (varying from 0 to 1). For instance, if we have $w_1 = 2.0$, $w_2 = 3.5$, and $w_3 = 1.5$, then $\sum w_i = 2.0 + 3.5 + 1.5 = 7.0$, and the relative participation of each weight is $w_1 = 2.0/7.0 = 0.29$, $w_2 = 3.5/7.0 = 0.50$, and $w_3 = 1.5/7.0 = 0.21$. The sum of each relative weight is, of course, 1.0 (in the current example, $0.29 + 0.50 + 0.21 = 1.0$).

In treatment plant studies, weighted averages are much more used than we normally notice. They are implicit in the calculations of loads, because loads are the product concentration × flow and, in this case, flow values act as the weighting factor of each concentration. Therefore, Equation 5.7 can be adapted to represent the case of concentrations (C_i) weighted by flows (Q_i).

$$\bar{C}_w = \frac{(C_1 Q_1 + C_2 Q_2 + \cdots + C_n Q_n)}{(Q_1 + Q_2 + \cdots + Q_n)} \tag{5.8}$$

S. 2.1

C. 12

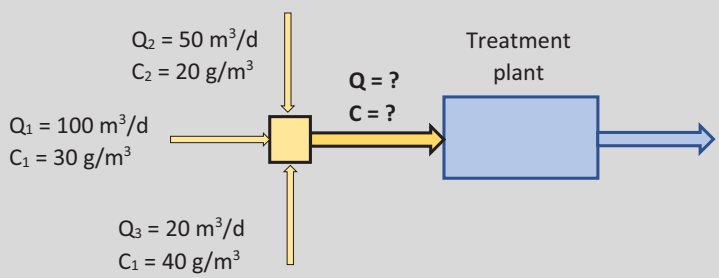
C. 13

The concept of the load is introduced in Section 2.1, explored in terms of mass balances in Chapter 12 and studied in terms of mass loading rates in Chapter 13.

Example

EXAMPLE 5.8 CALCULATION OF WEIGHTED AVERAGES WITH CONCENTRATIONS AND FLOWS

A treatment plant receives inputs from three different sources, each one with different flows and concentrations of the constituent you are investigating, according to the scheme as follows:



What are the total inflow and the average concentration at the inlet of the treatment plant?

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Solution:

The total inflow is the denominator of Equation 5.8 and is simply the sum of the three flow components

$$Q = Q_1 + Q_2 + Q_3 = 100 + 50 + 20 = 170\,\mathrm{m^3/d}$$

The total load is the numerator of Equation 5.8, corresponding to the sum of the individual loads or the concentrations multiplied by their respective weights (flows)

$$\sum Q_i C_i = 100 \times 30 + 50 \times 20 + 20 \times 40 = 3000 + 1000 + 800 = \mathbf{4800\,g/d}$$

The resulting concentration in the input to the treatment plant is the weighted average of concentrations by flows (Equation 5.8) or the division of total load by total flow:

$$C = \frac{4800\,\mathrm{g/d}}{170\,\mathrm{m^3/d}} = \mathbf{28\,g/m^3}$$

Example

Advanced

Excel

EXAMPLE 5.9 CALCULATION OF FLOW-WEIGHTED AVERAGE

You monitored your treatment plant over a period of 24 h, obtaining average hourly values of inflow and influent ammoniacal nitrogen ($N-NH_4^+$). Calculate the simple arithmetic mean of the 24 measurements of ammonia-N and also a flow-weighted average of ammonia-N.

Note: This example is also available as an Excel spreadsheet.

Data:

Hour of the Day	Flow, Q_{in} (m^3/h)	Concentration, C_{in} (g/m^3)	Hour of the Day	Flow, Q_{in} (m^3/h)	Concentration, C_{in} (g/m^3)
1	110	40	13	312	68
2	101	38	14	189	58
3	91	37	15	178	53
4	98	38	16	161	50
5	114	37	17	168	48
6	130	38	18	180	43
7	163	39	19	194	42
8	184	42	20	177	40
9	222	50	21	168	40
10	298	56	22	129	39
11	394	69	23	113	38
12	388	70	24	109	39

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Solution:

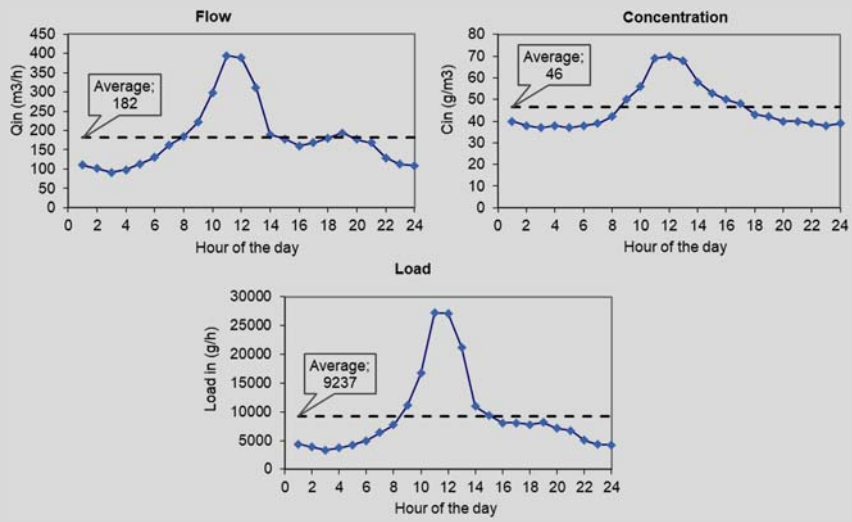
Let us calculate the ammonia-N load in each hour, knowing that load = flow × concentration. For this, we set up the following computational table:

110/4371

Hour of the day	Flow Q_{in} (m^3/h)	Concentration C_{in} (g/m^3)	Load $Q_{in}.C_{in}$ (g/h)	Relative weight of each flow value (Q_{in}/Q_{total})
1	110	40	110 x 40 = 4400	110 / 4371 = 0.03
2	101	38	3838	0.02
3	91	37	3367	0.02
4	98	38	3724	0.02
5	114	37	4218	0.03
6	130	38	4940	0.03
7	163	39	6357	0.04
8	184	42	7728	0.04
9	222	50	11100	0.05
10	298	56	16688	0.07
11	394	69	27186	0.09
12	388	70	27160	0.09
13	312	68	21216	0.07
14	189	58	10962	0.04
15	178	53	9434	0.04
16	161	50	8050	0.04
17	168	48	8064	0.04
18	180	43	7740	0.04
19	194	42	8148	0.04
20	177	40	7080	0.04
21	168	40	6720	0.04
22	129	39	5031	0.03
23	113	38	4294	0.03
24	109	39	4251	0.02
SUM	**4371 m^3**		**221,696 g**	**1.00**

The last line in the table shows the sum of the 24 values. Note that **we can sum flows and loads, but not concentrations**. The total volume reaching the treatment plant in that day was 4371 m^3 and the total mass entering it was 221,696 g. The sum of the relative weights (last column) is, as expected, 1.00.

The plots of the 24-h values of the computational table are shown as follows.



Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

If we calculate the averages from the 24 values of flow, concentration, and load, we obtain

- Mean $Q_{in} = 182\ m^3/h$
- Mean $C_{in} = \mathbf{46\ g/m^3}$ (simple arithmetic mean of concentrations)
- Mean $Load_{in} = 9237\ g/h$

But if we calculate the flow-weighted average of the concentration, we obtain the following value, knowing that concentration = load ÷ flow:

$$\bar{C}_w = \frac{\text{mass}}{\text{volume}} = \frac{221,696\ g}{4371\ m^3} = \mathbf{51\ g/m^3}$$

This calculation is the same as dividing the average load by the average flow

$$\bar{C}_w = \frac{\text{mean load}}{\text{mean flow}} = \frac{9237\ g/h}{182\ m^3/h} = \mathbf{51\ g/m^3}$$

Note that the mean concentration weighted by flow, or weighted average (51 g/m^3), in this example, is greater than the arithmetic mean (46 g/m^3). The arithmetic mean of the concentrations does not account for variations in flow and, in this case, underestimates the true mean concentration. You should take this into account when doing mass balances and other analyses that are based on loads and mean concentrations.

Similarly, the estimation of the mean ammonia-N load entering the plant should be given by the mean flow times the weighted average of concentration, or 182 × 51 (actually 182.13 × 50.72, to give exact values) = 9237 g/h. This is the same value as the mean load of the 24 values of load calculated above (9237 g/h). If we estimate the mean load and the total mass using the concentration of 46 g/m^3, we will underestimate the genuine value.

It is up to you, based on the knowledge of the system, to interpret whether these differences are important and may affect the calculations based on loads and mean concentrations. Note that we can make an **analogy with composite samples**. If we take *24 fixed-volume samples* and analyse the resulting composite sample after mixing all aliquots, we obtain a composite sample that resembles the *simple arithmetic mean* (in this case, with a concentration of 46 g/m^3). However, if we prepare our composite sample with *24 flow-proportional aliquots*, we obtain a concentration of the mixture that resembles the *flow-weighted average* (in this case, with a concentration of 51 g/m^3), which is more representative of the actual conditions. For more detail on composite sampling, see Section 3.3 and Example 3.1 in particular.

S. 3.3

5.7 MEASURES OF VARIATION

Basic

The degree of variability of your data points around the central value of your sample is given by the **measures of variation** or *dispersion*.

(a) Amplitude

Among those, the most intuitive one is given by the interpretation of the **minimum** and **maximum** values in your sample. The difference between them is the **amplitude**

$$\text{Amplitude} = \text{maximum} - \text{minimum} \tag{5.9}$$

However, consider the previous material presented regarding outliers and censored data. You can imagine that the magnitude of the amplitude of the values in your data set will depend

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

highly on the number of measurements you have. Therefore, you can imagine that this measure is *too fragile* to be a good measure of the variation in your data, since it depends on only two values, which may not be representative of the actual behaviour of your treatment plant or water body. Its use is not recommended without also including other measures of variation as well.

S. 5.6.2

(b) Variance

In Section 5.6.2, we discussed that the arithmetic mean was an equilibrium point, in an analogy with a scale, with points to the left and to the right of the mean. We saw that the sum of the distances from each point to the mean $(x_i - \bar{x})$ was equal to zero. The values that were lower than the mean (left-hand side of the scale) resulted in a negative sum, while the values that were greater than the mean (right-hand side of the scale) led to a positive sum, and both sums were equal, apart from the difference in sign (− and +). We can extend this concept to interpret the value of these sums: the greater they are, the more dispersed your data will be around the mean. We could use this to create a measure of dispersion.

But because we have positive and negative differences $(x_i - \bar{x})$, the negative and positive values will cancel out, and we will not be able to make inferences. A good solution to that is to raise all differences to the power 2, thus transforming everything into positive values. If we sum up all of them and divide by the number of data (as a matter of fact, $n - 1$), we will get a good measure of variation. This is the concept of **variance** that is defined in Equation 5.10.

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots (x_n - \bar{x})^2}{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{5.10}$$

(c) Standard deviation

By calculating the variance, we ended up with a value that is difficult to interpret, because it is based on squared values. If now we take the square root of the variance, we will obtain a value that has the same dimensions and units of the original values and of the mean. This is called the **standard deviation**, and it is given by Equation 5.11.

$$\begin{aligned} s &= \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1}} \\ &= \sqrt{\frac{1}{(n-1)}\sum_{i=1}^{n}(x_i - \bar{x})^2} \end{aligned} \tag{5.11}$$

Standard deviation is the **most widely used measure of variation**, and *you should include its value in your reports*, together with the measure(s) of central tendency.

In **Excel**, you can use the function **STDEV** and its variants to calculate the **standard deviation** and the function **VAR** and some of its modifications to calculate the **variance**. Furthermore, you can see that if you calculate standard deviation in Excel using STDEV and raise that value to the power of 2, you will get the same value as you do if you use the function VAR.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

S. 8.2

In Section 8.2, when we discuss the **normal distribution**, you will see that the *mean* and the *standard deviation* are *both* essential for defining the following intervals with specified values of data frequency in the *normal probability density function*:

- Interval between $\bar{x} \pm 1s \simeq 68\,\%$ of the data points are present in this interval.
- Interval between $\bar{x} \pm 2s \simeq 95\,\%$ of the data points are present in this interval.
- Interval between $\bar{x} \pm 3s \simeq 99\,\%$ of the data points are present in this interval.

For instance, if you have a mean of 100 and a standard deviation of 20 of a normally distributed variable, approximately 68% of the data will be inside the interval of 80 and 120, since $100 - 1 \times 20 = 80$ and $100 + 1 \times 20 = 120$. In the same way, around 95% of the data will be inside the interval of 60 and 140, since $100 - 2 \times 20 = 60$ and $100 + 2 \times 20 = 140$.

(d) Coefficient of variation

When comparing the variability or dispersion of two or more samples (same or different variables), it is common to use the so-called **coefficient of variation (CV)**, a result of the quotient between the standard deviation s and the mean $\bar{x}$

$$\mathrm{CV} = \frac{s}{\bar{x}} = \frac{\text{standard deviation}}{\text{mean}} \tag{5.12}$$

CV can be expressed as a **relative value** (as given by Equation 5.12) or multiplied by 100 to be expressed as **percentage** (**%**).

Note that *CV can be greater than 100%*, in case the standard deviation is larger than the mean, what is not infrequent in treatment plant and water quality data (especially for variables that have a wide range of variation, such as coliforms).

The coefficient of variation is a **positive** dimensionless number, and it should be applied only in cases where the means are different from zero and the values are always positive. If the values are always negative, CV should be calculated based on the absolute value of the mean (Oliveira, 2017).

Example 5.10 uses the same data from Example 5.4, in which the calculation of the mean was made, and the concept of the equilibrium point and differences from the mean were illustrated. We will take this further and from these, we will calculate the variance and standard deviation. Example 5.11 uses Excel functions and calculates directly the mean, standard deviation, and CV from the four data sets from Examples 5.4 and 5.5.

Example

Excel

EXAMPLE 5.10 CALCULATION OF STANDARD DEVIATION AND CV

Using the same data from Example 5.4 (calculation of mean), now calculate their standard deviation and coefficient of variation.

Note: This example is also available as an Excel spreadsheet.

Data:

4 2 5 1 7

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Solution:

The number of data is $n = 5$. In Example 5.4, the mean $\bar{x}$ was calculated as 3.8. Based on this, we can set up a simple computational table as follows:

x_i	Difference from Mean $x_i - \bar{x}$	$(x_i - \bar{x})^2$
4	$4 - 3.8 = 0.2$	0.04
2	$2 - 3.8 = -1.8$	3.24
5	$5 - 3.8 = 1.2$	1.44
1	$1 - 3.8 = -2.8$	7.84
7	$7 - 3.8 = 3.2$	10.24
Sum	0	22.80

From Equation 5.10, we can calculate the **variance** as

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{22.80}{5-1} = 5.7$$

Knowing that the **standard deviation** is the square root of the variance, we have

$$s = \sqrt{\text{variance}} = \sqrt{5.7} = 2.4$$

The coefficient of variation CV is the quotient of the standard deviation and the mean (Equation 5.12)

$$\text{CV} = \frac{\text{standard deviation}}{\text{mean}} = \frac{2.4}{3.8} = 0.63 = 63\%$$

Example

EXAMPLE 5.11 CALCULATION OF STANDARD DEVIATION AND COEFFICIENT OF VARIATION OF DIFFERENT DATA SETS USING EXCEL FUNCTIONS

With the same four data sets from Example 5.4 and 5.5, calculate the mean, standard deviation, and coefficient of variation using Excel functions directly and discuss the results.

Data:

The four data sets are

(a)	4	2	5	1	7
(b)	4	2	5	23	7
(c)	50	400	2000	3000	
(d)	50	400	2000	30,000	

Solution:

Using the Excel functions COUNT for n, MEAN for mean, and STDEV for the standard deviation, fill in columns 2, 3, and 4. CV is calculated by dividing the standard deviation by the mean.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Data Set	*n*	Mean	Standard Deviation	Coefficient of Variation (CV)
4, 2, 5, 1, 7	5	3.8	2.4	0.63
4, 2, 5, 23, 7	5	8.2	8.5	1.03
50, 40, 2000, 3000	4	1363	1383	1.01
50, 40, 2000, 30,000	4	8113	14,616	1.80

The second data set is similar to the first one, with the replacement of 1 by 23. You can see that the values of all statistics increased, and the data variation is higher, as indicated by the CV.

The third data set was meant to represent coliforms (it was presented in scientific notation in Example 5.5). The fourth data set is similar but substituting the value of 3000 by 30,000. We can see that, as expected, all statistics had a substantial increase, and now we have a CV that is well above 1, indicating the wide variability of the data.

Advanced

S. 5.6.4

(e) Geometric standard deviation

We have already presented you the concept of *geometric mean* (M_g) in Section 5.6.4 and highlighted that it is an important measure of central tendency for variables whose variability spans different orders of magnitude or that are not symmetrically distributed around its arithmetic mean.

Now we need to introduce you to the associated measure of variation, represented by the **geometric standard deviation** (s_g), which is given by (Metcalf & Eddy, 2003)

$$\log s_g = \text{standard deviation of } \log(x_i) = \sqrt{\frac{\sum(\log x_i - \log M_g)^2}{n-1}} \tag{5.13}$$

$$s_g = 10^{(\log s_g)} \tag{5.14}$$

where

s_g = geometric standard deviation
x_i = variable being analysed
$\log x_i$ = $\log_{10}$ of x_i
n = number of data
M_g = geometric mean of the variable
$\log M_g$ = $\log_{10}$ of the geometric mean M_g.

Note that Equation 5.13 is based on a *simple transformation of the original values into their log_{10} values (log-transformed data)*. The term inside the square root is the variance of the log-transformed data (same structure as Equation 5.10), and the term on the right-hand side, including the square root, is the standard deviation of the log-transformed data (same structure as Equation 5.11). In the same way that we calculated the *geometric mean as 10 raised to the mean of the log-transformed data* (see Equation 5.5), the **geometric standard deviation is 10 raised to the standard deviation of the log-transformed data**.

In summary, we have

$$\textbf{Geometric mean}, M_g = 10^{(\text{mean of the log 10−transformed data})} \tag{5.15}$$

$$\textbf{Geometric standard deviation}, s_g = 10^{(\text{standard deviation of the log 10-transformed data})} \tag{5.16}$$

The **geometric mean** has values that are **greater than 0** and the **geometric standard deviation** has values that are **greater than 1**.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Example

EXAMPLE 5.12 CALCULATION OF THE GEOMETRIC STANDARD DEVIATION

Calculate the geometric standard deviation (s_g) of the data from Example 5.6, in which you have already calculated the geometric mean (1047 MPN/100 mL).

The data are ($n = 4$)

50 400 2000 30,000 MPN/100 mL

Excel

Note: This example is also available as an Excel spreadsheet.

Solution:

In Example 5.6, we set up a computational table (first three columns of the following table), which we can now expand to include other columns for the calculation of s_g.

Data	*E. coli* (MPN/100 mL)	Log_{10} (x_i)	$\text{Log}_{10}(x_i) - \text{Log}_{10}(M_g)$	$[\text{Log}_{10}(x_i) - \text{Log}_{10}(M_g)]^2$
1	5.00×10^1	1.699	$1.699 - 3.020 = -1.321$	$(-1.321)^2 = 1.745$
2	4.00×10^2	2.602	$2.602 - 3.020 = -0.418$	$(-0.418)^2 = 0.175$
3	2.00×10^3	3.301	$3.301 - 3.020 = 0.281$	$(0.281)^2 = 0.079$
4	$3{,}00 \times 10^4$	4.477	$4.477 - 3.020 = 1.457$	$(1.457)^2 = 2.123$
Sum				4.122

Note: $M_g = 1047$.
$\text{Log}_{10} M_g = \log_{10}(1047) = 3.020$.

Applying Equations 5.13 and 5.14, we can calculate the geometric standard deviation s_g

$$\log s_g = \sqrt{\frac{\sum (\log x_i - \log M_g)^2}{n-1}} = \sqrt{\frac{4.122}{4-1}} = 1.1722$$

$$s_g = 10^{(\log s_g)} = 10^{1.1722} = 14.87$$

We can also calculate the geometric standard deviation as simply 10 raised to the standard deviation of the log-transformed data (Equation 5.16). *The standard deviation of the log-transformed data is 1.1722*, calculated above, or using the Excel function STDEV (for the values in the third column of the table). Raising 10 to the power 1.1722 leads to **14.87**, that is the geometric standard deviation (obviously, same value as above).

S. 8.3

In Section 8.3, when we discuss the **log-normal distribution**, we will explain that the *geometric mean* (M_g) and the *geometric standard deviation* (s_g) are important in defining the following intervals with specified values of data frequency in the *log-normal probability density function*:

- Interval between $\boldsymbol{M_g \div / \times (s_g)^1} \simeq$ **68%** of data points is within this interval.
- Interval between $\boldsymbol{M_g \div / \times (s_g)^2} \simeq$ **95%** of data points is within this interval.
- Interval between $\boldsymbol{M_g \div / \times (s_g)^3} \simeq$ **99%** of data points is within this interval.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Note the nature of the relationship between M_g and s_g for a **log-normal distribution**. For the conventional arithmetic mean and standard deviation in the normal distribution, the relationship was $-$ or $+$ (minus or plus), and now, for the log-normal distribution, it is $\div$ or $\times$ (divide or times). For instance, if you have a geometric mean of 100 and a geometric standard deviation of 2.0, approximately 68% of the data of a log-normally distributed variable will be inside the interval of 50 and 200, since $100 \div (2.0)^1 = 50$ and $100 \times (2.0)^1 = 200$. In the same way, around 95% of the data will be inside the interval of 25 and 400, since $100 \div (2.0)^2 = 25$ and $100 \times (2.0)^2 = 400$. Finally, about 99% of the data will be inside the interval of 12.5 and 800, since $100 \div (2.0)^3 = 12.5$ and $100 \times (2.0)^3 = 800$.

As mentioned previously, the most common use of the geometric mean and geometric standard deviation is when dealing with microbiological constituents such as coliforms, *E. coli*, enterococci, etc. Other constituents in environmental systems such as rivers, streams, and aquifers may also present log-normally distributed data (see Section 8.3). When dealing with data for these constituents, if you simply calculate the **$\log_{10}$-transformed value** of each data point, then you will be able to treat the *$\log_{10}$-transformed values using the arithmetic mean and standard deviation*. Just keep in mind that you will have to *transform the values back to their original units* (e.g., using Equations 5.5 and 5.16).

S. 8.3

5.8 MEASURES OF RELATIVE STANDING

Basic

Measures of **relative standing** are a way of describing the location of an observation *relative* to the other observations in your data set. The most usual way of reporting the relative positions are using the concept of **percentiles**, already explored in several sections of this book.

A percentile is a measure of the **relative position of an observational unit in relation to all others**. The p^{th} *percentile* has at least p% of the values below that point and at least $(100 - p)$% of the values above it.

For instance, if your data set has 75 observations ($n = 75$), then we can calculate the order of the percentiles following these simple calculations:

- The **25th percentile** is $X_{[(75)\times(0.25)]} = X_{[18.75]} \Rightarrow X_{[19]}$ (after rounding up 18.75 to 19, which is the *next integer above*), that is, the 19th lowest observation, after placing all the values in an increasing order.
- The **40th percentile** is $X_{[(75)\times(0.40)]} = X_{[30]} \Rightarrow (X_{(30)} + X_{(31)})/2$, that is, the mean of the 30th and 31st observations after sequential ordering. This calculation involving the mean of two consecutive observations (30 and 31) is because we obtain an *integer* value ($30 = 75 \times 0.40$), and thus, we need to involve the next observation in the sequence.

S. 5.5.2

As already emphasized in this book (e.g., in Section 5.5.2), the most frequently used percentiles in the characterization of a sample are:

- **25 percentile**, meaning that 25% of the data have a value that is less than or equal to it. It is also called the **first quartile (*Q*1)**.
- **50 percentile**, meaning that 50% of the data have a value that is less than or equal to it. It is also called the **second quartile (*Q*2)**. It is the same as the **median**.
- **75 percentile**, meaning that 75% of the data have a value that is less than or equal to it. It is also called the **third quartile (*Q*3)**.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

S. 6.4

These three percentiles are an integral part of the box-and-whisker plots, already shown in other previous sections, and presented in more detail in Section 6.4. The box plots we use in this book also include the **10 and 90 percentiles**, in order to allow you to have a better visualization of the distribution of your data.

In treatment plant practice and water quality studies, percentiles are also used in the **assessment of the compliance** with standards stipulated in legislations or target values defined by your company. For instance, a legislation may specify that 90% of the samples of the effluent from a treatment plant should have a **concentration** equal to or lower than X mg/L. In this case, *your* 90^{th} *percentile must be less than or equal to* X *mg/L*. Alternatively, the specification may be that 90% of the values of **removal efficiency** should be equal to or above Y%. In this case, *your* 10^{th} *percentile* $(= 100 - 90)$ *should be less than or equal to* Y%. Other percentages of compliance commonly specified in legislations are 80% and 95%.

We should *not confuse percentiles with percentages*. A percentile is only related to the relative position of an observation when compared with the other values.

Percentiles can be calculated in a simple and direct way by using the **Excel** function **PERCENTILE**. There are variations in the syntax of this function, and you should consult the manual of the software. Percentiles will also be a standard function for any other statistical software you might use.

Example

EXAMPLE 5.13 CALCULATION OF PERCENTILES

Using the same data from Example 5.3 (effluent COD), compute the 25, 50, and 75 percentiles of your data set.

Data:

63	37	50	44	51	49	57	62	53	50
61	66	73	83	81	134	104	142	95	79

Solution:

The first thing to do is to order the values in an increasing way. This is very easily done in Excel, and the ordered values are shown in the following table:

Order	Value	Order	Value	Order	Value	Order	Value
1	37	*6*	51	*11*	63	*16*	83
2	44	*7*	53	*12*	66	*17*	95
3	49	*8*	57	*13*	73	*18*	104
4	50	*9*	61	*14*	79	*19*	134
5	50	*10*	62	*15*	81	*20*	142

The sample has $n = 20$ observations.

The calculations of the percentiles are as follows:

- **Percentile 25%**. $n \times$ percentile $= n \times p = 20 \times 0.25 = 5$ (because $n \times p$ is an integer, take the mean of this observation and the next one). Thus, 25^{th} percentile $= (5^{th} + 6^{th})/2 = (50 + 51)/2 = 51.5 =$ **52**.
- **Percentile 50%**. $n \times$ percentile $= n \times p = 20 \times 0.50 = 10$ (because $n \times p$ is an integer, take the mean of this observation and the next one). Thus, 50^{th} percentile $= (10^{th} + 11^{th})/2 = (62 + 63)/2 = 62.5 =$ **63**.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

- **Percentile 75%.** $n \times \text{percentile} = n \times p = 20 \times 0.75 = 15$ (because $n \times p$ is an integer, take the mean of this observation and the next one). Thus, 75^{th} percentile $= (15^{th} + 16^{th})/2 = (81 + 83)/2 = \mathbf{82}$.

Of course, you do not need to make these calculations; the idea here was just to show you the background behind the determination of percentiles. You can simply select your data range and use the Excel function PERCENTILE, with k assuming the values of 0.25, 0.50, and 0.75 for the percentiles 25, 50, and 75.

5.9 CHECK-LIST FOR YOUR REPORT

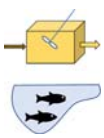
- ✓ Verify that you have selected a clear structure for the summary table that will be part of the body of your report, incorporating the information that you judge to be most important. Usually, these will minimally include mean and/or median and the standard deviation, for the constituents you are analysing. Verify which is the best way of reporting the number of data (n) of your samples.
- ✓ If there is more space in the body of your report, consider including a full descriptive statistics table.
- ✓ If the full table with descriptive statistics becomes very large, consider the possibility of incorporating it in an Appendix or as Supplementary Material. Also remember to publish your complete data set online or as Supplementary Material.
- ✓ In the case of summary tables for studies of treatment plants, try to include not only concentrations but also removal efficiencies if possible. Also consider incorporating loading rates and other information that may concisely describe the prevailing operating conditions you had.
- ✓ If your study covers different treatment units or operational phases, confirm that your summary table includes the descriptive statistics of each unit or phase.
- ✓ Check that all values in your table have units (e.g., mg/L, m^3/d, $(g/d)/m^2$, %, etc.).
- ✓ Make sure that your table is self-sufficient (meaning it can be read and understood by itself without the reader having to consult the main text of your report); include footnotes if necessary.
- ✓ Confirm that the number of significant decimal points for your mean, median, and standard deviation values are the same as those of your original data.
- ✓ Although common, avoid representing the values of mean and standard deviation as mean $\pm$ standard deviation, because your data may not be symmetrically distributed around the mean. Consider using mean (standard deviation).
- ✓ Verify that you have analysed all possible explanations for potential outliers in your data set based on the knowledge you have of the treatment system or water body; do not make these decisions purely based on mathematical criteria.
- ✓ State in your report how you have treated censored data (left-censored data: below lower detection limits; right-censored data: above upper detection limits).
- ✓ If possible, in the representation of the measures of central tendency, try to include the mean and the median of your data, because they may be different. In the case of variables that vary within orders of magnitude, consider reporting geometric means. Specify clearly which measure of central tendency you are using.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Chapter 6

Descriptive statistics: graphical methods for describing monitoring data

This chapter shows you how to build and interpret the main types of graphs and charts used for describing your monitoring data. Most of the graphs are for quantitative data, but charts for qualitative data are also covered. The graphs included are time series, frequency histograms, frequency polygons, percentile graphs, box plots, and scatter plots for quantitative data and bar/column charts and pie charts for qualitative data. Advice on the presentation of graphs is provided.



The contents in this chapter are applicable to both treatment plant monitoring and water quality monitoring. The exceptions are the mentions of 'removal efficiencies', which are applicable only to the assessment of treatment plants.

CHAPTER CONTENTS

6.1 Introduction **152**
6.2 Time Series Graphs **154**
6.3 Frequency Distribution **165**
6.4 Box-and-Whisker Graphs (Box Plots) **172**
6.5 Scatter Plots **175**
6.6 Graphs for Qualitative (Categorized) Data **176**
6.7 General Advices on Presenting Graphs **179**
6.8 Check-List for Your Report **180**

© 2020 The Authors. This is an Open Access book chapter distributed under the terms of the Creative Commons Attribution Licence (CC BY-NC-ND 4.0), which permits copying and redistribution for non-commercial purposes with no derivatives, provided the original work is properly cited (https://creativecommons.org/licenses/by-nc-nd/4.0/). This does not affect the rights licensed or assigned from any third party in this book. The chapter is from the book *Assessment of Treatment Plant Performance and Water Quality Data: A Guide for Students, Researchers and Practitioners*, Marcos von Sperling, Matthew E. Verbyla and Sílvia M. A. C. Oliveira (Authors).
doi: 10.2166/9781780409320_0151

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

6.1 INTRODUCTION

Basic

The knowledge and interpretation of your data can be greatly facilitated through visual analysis of graphs. The graphs to be employed depend on whether the data are (a) **qualitative** (*categorized*) or (b) **quantitative** (*numerical*). The basic graphs for quantitative data are related to the descriptive statistics presented in Chapter 5. Table 6.1 and Figure 6.1 show the main graphs covered in this chapter for descriptive analysis of monitoring data.

C. 5

In our book, we deal mainly with **quantitative data**. If they can be represented only by *integer numbers*, they are called **discrete variables**. Examples are variables that can be counted, for instance, number of samples per year complying with discharge standards or number of treatment plants using activated sludge. In this case, only integer numbers can be used to represent the quantity under analysis. When the quantitative data are expressed as numbers that can be measured or represented at any point along a *numerical scale* (including decimal numbers between integers), they are **continuous variables**. These are the majority of data we cover in this book, such as flows, concentrations, loads, and removal efficiencies.

It is also worth pointing out that the **majority of the data covered in this book are non-negative continuous variables by definition** – that is, they must be greater than or equal to zero. For example, it is impossible to have a negative flow rate, a negative concentration, or a negative load. However, it is possible for removal efficiency to be negative, which would indicate an increase in the concentration through the unit process (if you are dealing with a pollutant, this means that you have problems of malfunctioning in your treatment plant).

When working with continuous variables for treatment plant or water quality data, we also try to stay away from reporting or plotting values of zero, due to the limitations associated with sampling (e.g., there are limitations with regard to the volume of sample you can collect and process for analysis) and sample analysis (e.g., the readings from instrument blanks typically produce a signal corresponding to 'background noise'), resulting in limits of detection for a given method. Instead of reporting a value of zero, you should report that the value is below the method limit of detection (see Sections 4.6 and 5.4).

S. 4.6

S. 5.4

Contrary to quantitative data, **qualitative data** are those that cannot be measured or expressed on a quantitative scale. They represent *categories* that can be characterized by codes, names, letters, or numbers, and because of this they are also **categorized** or **categorical** data. Examples are the categories of treatment processes in a survey (stabilization ponds, treatment wetlands, activated sludge, upflow anaerobic sludge blanket (UASB) reactors, etc.), experimental phases in your study (phase 1, phase 2, phase 3, etc.), or the location of treatment plants (city A, city B, etc.).

Although there are recommendations for the selection of the right type of graph to be included in your report, you should try different ones, change formats, reflect about the scales of the axes, whether the correct font size is being used, where the legend should be placed, if the symbols you selected for the

Table 6.1 Main types of graphs for describing monitoring data covered in this chapter.

Quantitative (numerical) data	• Time series graphs • Frequency histograms • Frequency polygons • Frequency distribution (percentiles) • Box plot • Scatter plot
Qualitative (categorized) data	• Column/bar chart • Pie chart

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

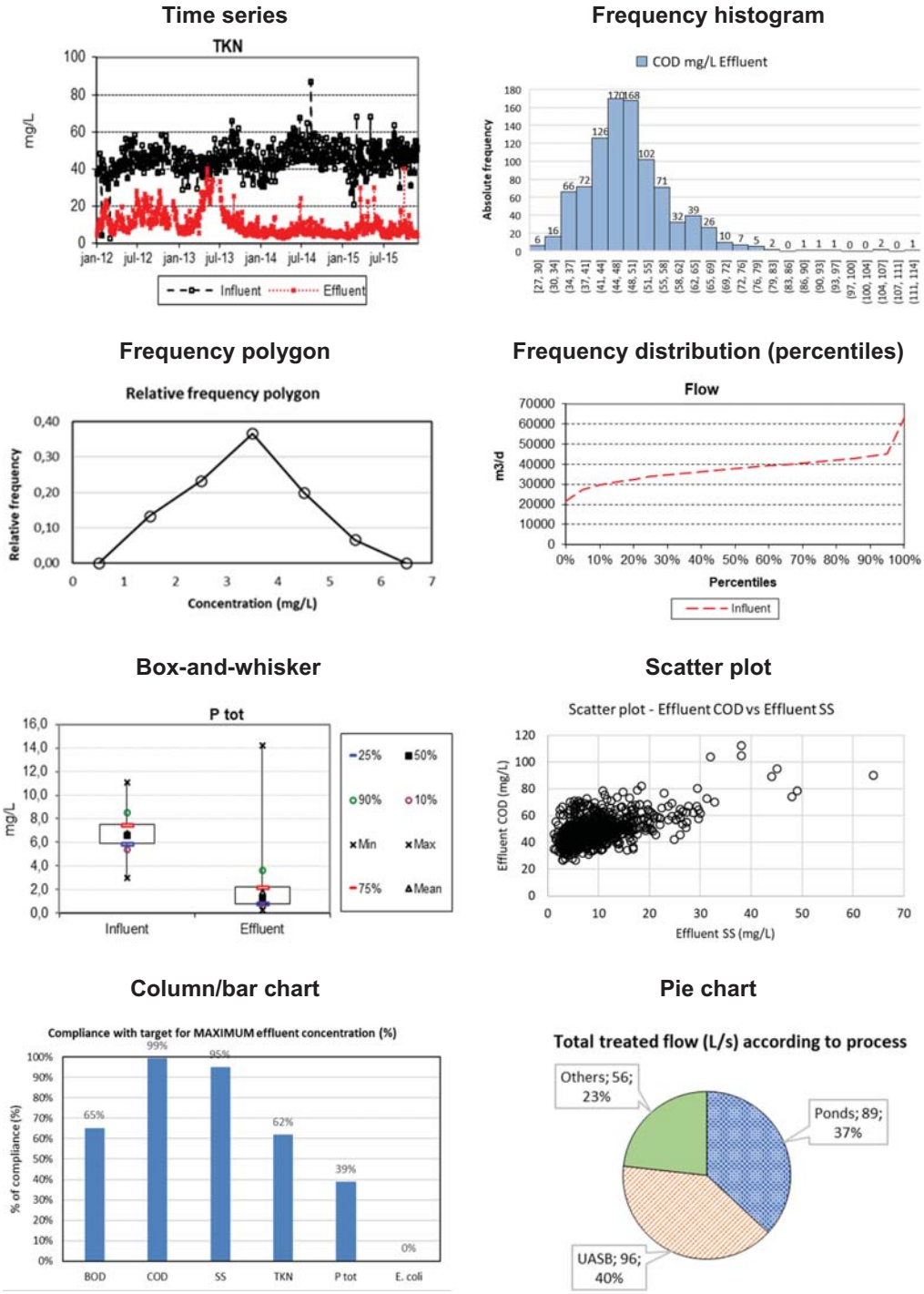


Figure 6.1 Examples of descriptive statistics graphs used for describing monitoring data.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

markers are clearly visible, whether the type and colour of the lines are easily distinguishable, and other points you might remember. *The best graph for your data is a matter of trial and error*. You spent much time and effort for obtaining the data that now you should spare some additional noble time for conceiving *the best possible visual communication of the message you want to convey through the graphs*.

In addition to the descriptive graphs listed above, there is also a wide variety of other more specific graphs, related to various statistical analyses (e.g., regression analysis graphs, multivariate analysis, etc.). Some of them are covered in other parts of this book, and others are outside the scope of this book.

The **balanced use of graphs** in your report can greatly enrich its content. However, you should take into account that the ease in the elaboration of graphs in spreadsheets and word processors does not justify the presentation of an *excess of visual information* in the report, often incompatible with the very nature of your data and the possible limitations of your sample.

It is also noteworthy that any graph presented in a technical or scientific report *should be referenced (called out) and **interpreted** in the text*. The graph should not be merely included in the work, detached from the text, just because it was easy to elaborate it on the computer. The position of the graphs in your report can be in the **body of the text** or at its end (**appendices**). Usually, figures are included in the body of the report, shortly after they are cited in the text, as this is a more effective way to communicate with your readers. The graphs placed sequentially at the end of the work are justified if they are in large quantities or relate to raw data graphs and not summary graphs associated with relevant descriptive statistics. If this is the case, their inclusion in appendices can be interesting to avoid breaking the text and the line of thought of the reader (von Sperling *et al.*, 1996; Nascimento *et al.*, 1996).

Excel

S. 5.1

Several examples of graphs shown in the sections to follow have been extracted from the master Excel spreadsheet on descriptive statistics mentioned in Section 5.1.

6.2 TIME SERIES GRAPHS

6.2.1 Use of time series graphs

Basic

You organized your data on a chronological way in your spreadsheet, and thus, it is a natural way to present it as a sequence in time. **Time series graphs** are widely used in performance assessment, typically presenting your raw or worked out data on flows, concentrations (influent and effluent), loads, and removal efficiencies on the *Y*-axis and time on the *X*-axis.

S. 6.2.3

In **Excel**, time series graphs can be produced using **line charts** or **scatter charts**. See Section 6.2.3 for a discussion on the difference between the two options. All general statistical software also produces time series graphs. Depending on the frequency you collect your data, your *X*-axis can be in *minutes*, *hours*, *days*, *weeks*, *months*, or *years*.

Time series graphs can be very informative on the **trends** of your data or on their **dynamics**, and also on the **relationship** between variables. However, they will not be of much use if they show only ups and downs without any possible hint that can lead to the interpretation of the plant behaviour. Let us take the example of the time series of influent and effluent chemical oxygen demand (COD) from the treatment plant of the master spreadsheet, spanning a period of almost four years (2012–2016), with an almost daily frequency, with hundreds of data points (Figure 6.2a). Basically, what you can see is that there are a lot of ups and downs, especially in the influent concentration, and that the effluent concentration values are much lower than the influent ones. At this stage, what you can conclude is that your treatment plant is probably doing a good job in removing COD, and that possibly the effluent series is more stable than the influent one.

But even about this last inference you cannot be sure, because the scale of the *Y*-axis does not let you see well the effluent data. Therefore, if you want to have a better interpretation of your effluent data, you can exclude the influent from the graph, and you will be able to see that the effluent also presents variability

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

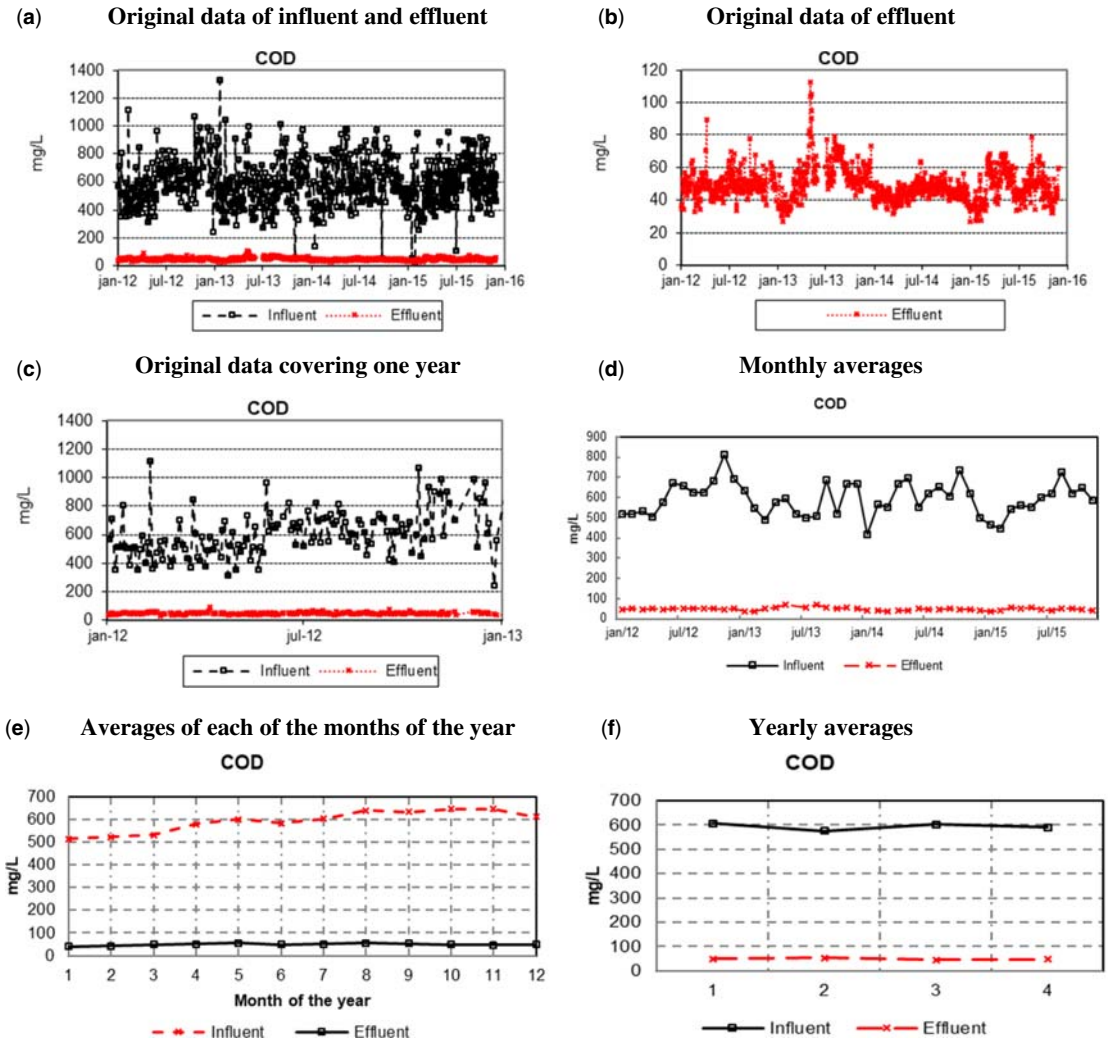


Figure 6.2 Example of the time series plot of influent and effluent COD from a treatment plant with daily data over a period of four years. (a)–(f) Different forms of presenting the data.

(Figure 6.2b). You can get a good idea about the scale for both the influent and the effluent by plotting the influent on a left *Y*-axis and the effluent on a right *Y*-axis, each one with its own scale. However, both series will probably be mixed together, and readers may have difficulty interpreting the graph.

To have fewer data points in your graph, you may select a shorter period. In Figure 6.2c, we changed the *X*-axis scale to cover only the year 2012. It is now slightly easier to identify the behaviour of your treatment plant in specific periods.

The master spreadsheet calculates monthly averages for all the months included in your time series (Figure 6.2d). Clarity of information is improved, but at the expense of the loss on the system dynamics.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Still another way of seeing the variation of your data as a function of time is to present the averages of each of the 12 months of the year. For instance, you get the average of all data collected in the month of January (month 1) in the four years, and then the average of all data from February (month 2), and so on (Figure 6.2e). This is also done automatically in the master spreadsheet.

Another possibility, especially if you have a very long time series, covering several years, is to present yearly averages (Figure 6.2f). In the current example, there are only four years in the series, so not much can be inferred in this particular situation.

In Figure 6.2, we have shown examples of time series graphs involving only one variable (COD). But this type of graph is also **very useful when you plot two variables that may be interconnected** in the treatment plant performance. For instance, you could plot in the *Y*-axis *effluent COD and effluent total suspended solids* (*TSS*), and you could observe if they have similar trends and whether the peaks in one variable are followed by peaks in the other variable, perhaps because you have prior knowledge that suspended solids are associated with particulate COD. Alternatively, for instance, if you plot *nitrification efficiency and alkalinity*, you could check whether a peak (high value) in nitrification occurs simultaneously with a valley (low value) in alkalinity based on your knowledge that nitrification consumes alkalinity.

6.2.2 Connection of data points with lines

Basic

In your time series graphs plotting experimental, measured, or observed data, **markers** representing these data are **compulsory**: you have to *clearly indicate which points correspond to your data*. The question now is whether you should connect the markers with lines. Some people prefer not to connect the points, because of the feeling that the lines will give readers the false idea that your variable follows exactly the trajectory of the straight lines, which may not be the case if there would have been a peak value between markers that was missed because a sample was not collected at that particular time. Nevertheless, we are of the opinion that the readers of your graphs understand this implicit limitation of a time series plot, and that *there are more advantages to connecting the markers* with **lines** because it will *allow readers to see possible trends in the time series more clearly*. After all, one of the objectives of a time series graph is to allow for the visualization of temporal trends.

In time series graphs, it is better to **connect the data points with straight lines, not smoothed curves**. Smoothed curves will make an automatic fitting to the data points, making *trajectories with what you cannot defend* and maybe even *passing through negative values* or other values without physical meaning. Take the example in Figure 6.3, which plotted the following points on days 1, 2, 3, 4, and 5: 30, 2, 1, 1, and 25. The left-hand side of the graph shows the connection of data points with straight lines, as advocated here. However, when you make the plot with smoothed curves (as shown in the right-hand side of the graph), you get two reaches that are showing negative values, which is a problem when dealing with concentration data. *Your plot should not show lines that imply negative concentrations or removal efficiencies greater than 100%.*

6.2.3 Missing data and days without monitoring in scatter charts and line charts

Basic

A certain problem may arise if you have periods without data, and two distant markers are connected by a straight line spanning a long time interval. In this case, you may choose to have your lines plotted as *dots* or *dashes* to give the reader the feeling that there is no actual continuity between the data points.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

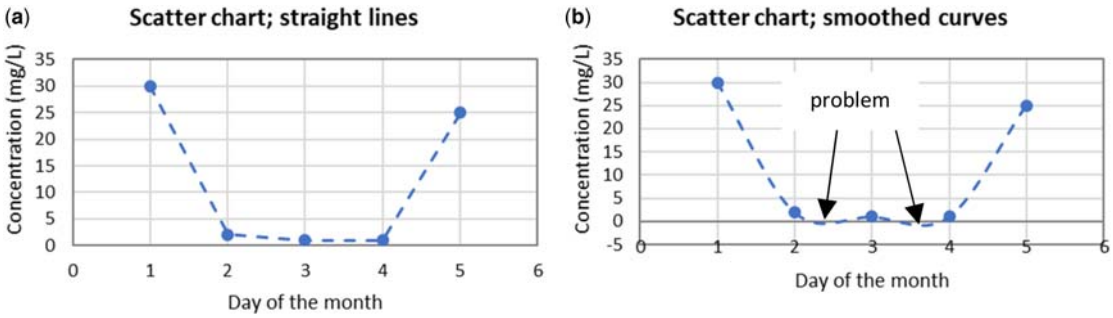


Figure 6.3 Example of two different ways of connecting data points. (a) Straight lines, which are the proposition we make. (b) Smoothed curves, which may generate reaches without physical meaning, such as the two reaches with negative values.

You should check how your statistical software produces time series graphs. Note that Excel has different options on how to treat empty cells in line and scatter charts, and also how to plot your time variable in the *X*-axis. *It is important that your X-axis is a faithful reproduction of your actual time intervals*. For instance, the distance between days 5 and 6 (1 day of difference) should be proportionally smaller than the distance from days 6 to 19 (13 days of difference) – both cannot be represented with the same time interval. We will illustrate this in Example 6.1.

Example

EXAMPLE 6.1 SELECTION OF THE *X*-AXIS ON A TIME SERIES GRAPH

You monitored effluent COD in a treatment plant on six days (1, 3, 5, 6, 19, and 20) of a certain month. Unfortunately, the sample from day 3 was suspicious, and you decided to discard it. Prepare a time series graph and analyse the best way of presenting it.

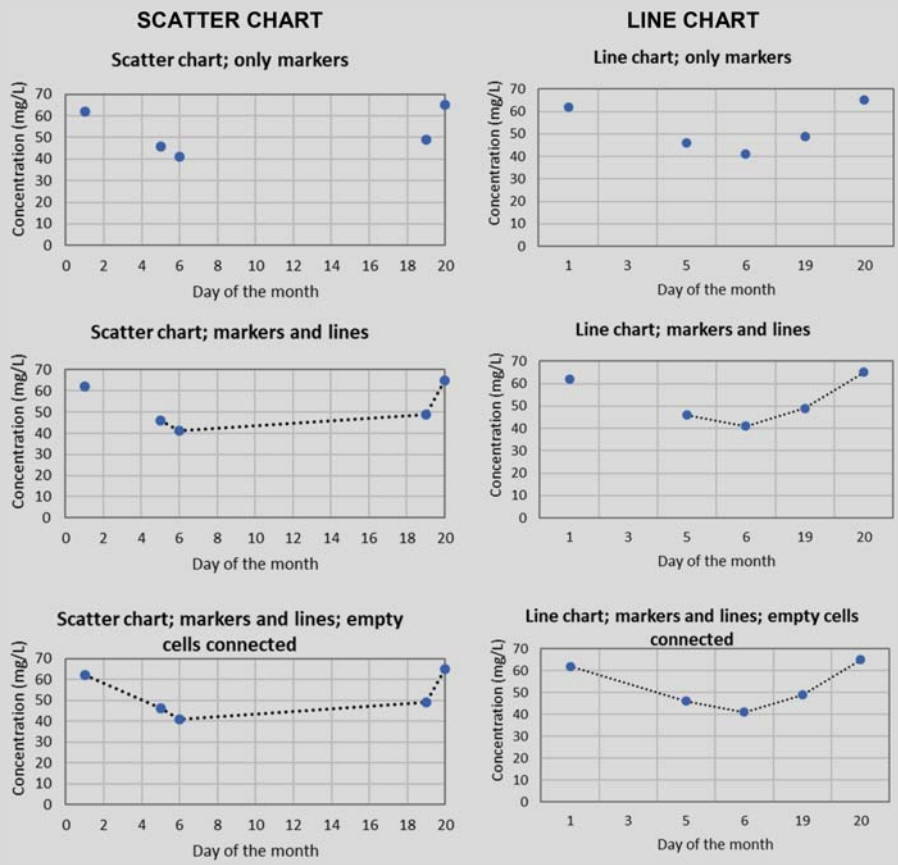
Data:

Day of the Month	Concentration (mg/L)
1	62
3	
5	46
6	41
19	49
20	65

Solution:

Using Excel, you prepared time series graphs, using the options of scatter charts and line charts. You then tried different ways of connecting or not your data points. The results you obtained are shown as follows:

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest



The first column presents variants of the scatter chart and the second column presents variants of the line chart. Note that the **scatter chart preserves the correct scale for time**: each day of the month has its own position, regardless of whether or not there was monitoring in that day. The **line chart plots only the days in which there was monitoring**. The space between days 5 and 6 is the same as days 6 and 19. This of course alters the visualization of trends, which is an important objective of time series graphs. In the current example, *the line chart seems to indicate a much stronger rate of decrease or increase in the concentrations as a function of time*, since the lines are steeper. However, when you look at the scatter plot, which has the correct time scale, you see that the rates are not so strong (the connecting lines have a smaller slope in the scatter chart than they do in the line chart).

A second observation relates to **whether or not you should connect your data points with lines**. In the first line of the figure, there are only markers, and points are not connected. In the second and third lines of the figure, there are lines connecting your points, and it is now easier to evaluate the existence of possible trends. You observe that there is a downward trend in the first days, and then concentrations increase again in the last days. Because there are long time intervals between days with monitoring, you decided to *connect the points with dotted lines*, giving the indication that there is no guarantee that the trajectory of your variable will follow the straight line.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Now you have to think about **missing data**. On day 3, there are no data, and so the cell is empty. You decided to keep day 3 in the table, because it had data on other variables (not shown here). Note that this is different from days 7 to 18, because on these days simply there was no monitoring, and your table with data does not include these days (therefore, you indicate that you do not consider that there are missing data in these days). There are different ways of handling this in Excel. You should consult the manual of the Excel version you have, but you could try selecting the chart, then Select Data, and then click on Hidden and Empty Cells. You will see that there are options for connecting or not the points in empty cells: Gap, Zero, and Connect. In the second line of the table, we did not connect the points in day 3, and in the third line, we connected them. You need to think and decide what best reflects the message you want to convey.

As a general conclusion with the data in this example, we could say that it is better to use scatter charts compared to line charts, and that connecting data points with lines may improve the visualization of possible trends.

Basic

6.2.4 *Y*-axis scale in time series graphs

The selection of the scale of the *Y*-axis of your graphs is very important, because it can influence their interpretation:

- **Scales with wide variation** between minimum and maximum tend to smooth the perception of variations and trends.
- **Scales with little variation** between minimum and maximum tend to amplify the perception of variations and trends.

Figure 6.4 presents two graphs with the same data but with different scales in the *Y*-axis. The left-hand side graph emphasizes the trend in the growth of values, while the right-hand side gives the impression of a relatively stable series. It is up to you to *highlight what you think is important and is a better representation of the message you want to pass*, and try to avoid misleading the reader into a wrong interpretation. Most of us are used to see *Y-axis scales starting from zero* as minimum value, but of course there are several situations in which we need to alter this configuration.

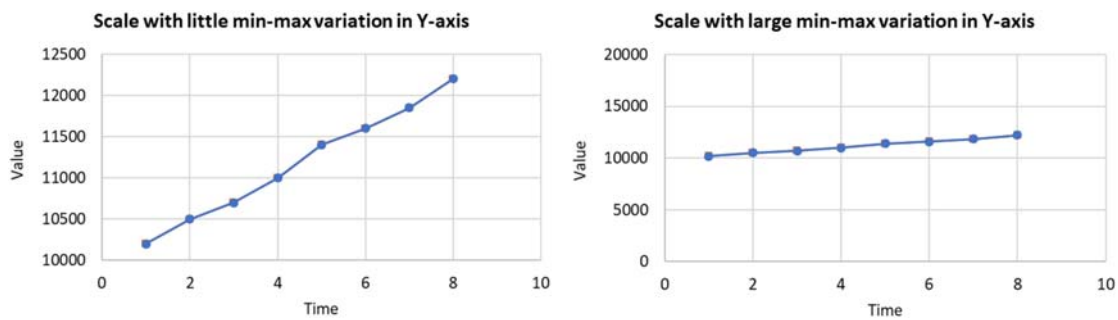


Figure 6.4 Comparison between two graphs representing the same data, emphasizing the importance of an adequate selection of the axis scale.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

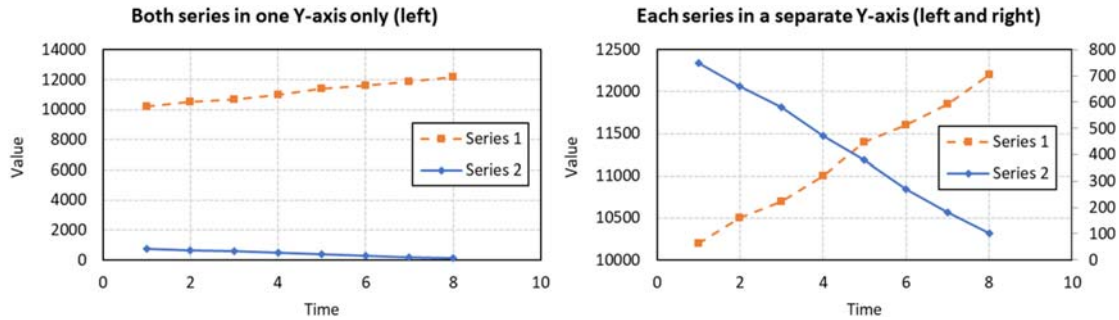


Figure 6.5 Comparison between two charts representing the same data. The left chart represents the two series on a single *Y*-axis (left), while the right chart represents each series on separate *Y*-axis with different scales (series 1 on the left axis and series 2 on the right axis).

6.2.5 Graphs with two *Y* axes

Basic

If the chart has to represent *series with very different values*, you can **use two different scales for the *Y*-axis**, one for the left axis and one for the right axis. Some points to keep in mind:

- Do not use this feature in excess, especially in oral presentations, because their interpretation by the audience takes longer.
- Limit the graph to only a few data series, preferably two (one series for each *Y*-axis). Indicate clearly in the chart which series corresponds to which axis.

Figure 6.5 shows two graphs representing the same data series. The chart on the left allows the immediate interpretation that the values of series 1 are greater than the series 2 values. On the other hand, the chart on the right shows more clearly both series and their trends, represented in the left and right *Y* axes, but the message that series 1 values are greater than series 2 values is not directly evident.

6.2.6 Arithmetic and logarithmic scales

Basic

If the *magnitude of the values to be represented varies widely*, a **logarithmic scale** can be adopted as a way to condense the values on the axis, highlighting each order of magnitude. Excel allows you to change your axis to log scale very easily. Notice the following points:

- If two series are represented, both having different orders of magnitude values, the logarithmic scale can be useful, allowing a clearer visualization of the series with lower values.
- The logarithmic scale can affect the visualization of a series with a high growth rate, because it dampens the highest values.
- A graph can have arithmetic scales (conventional scale), semi-log scales (e.g., *X*-axis, arithmetic scale; *Y*-axis, logarithmic scale), or log-log scales (*X* and *Y* axes: logarithmic scale).
- In oral presentations, mention when the scale of an axis is logarithmic.

Figure 6.6 shows two graphs representing the same data. The *Y*-axis on the left-hand chart is in arithmetic scale: the strong increasing trend of the upper series is evident, but the values of the lower series are somewhat masked. The graph on the right allows better visualization of the lower series, as well as the interrelationship between the two series, but loses in terms of showing the big difference between the absolute values of series 1 and 2.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

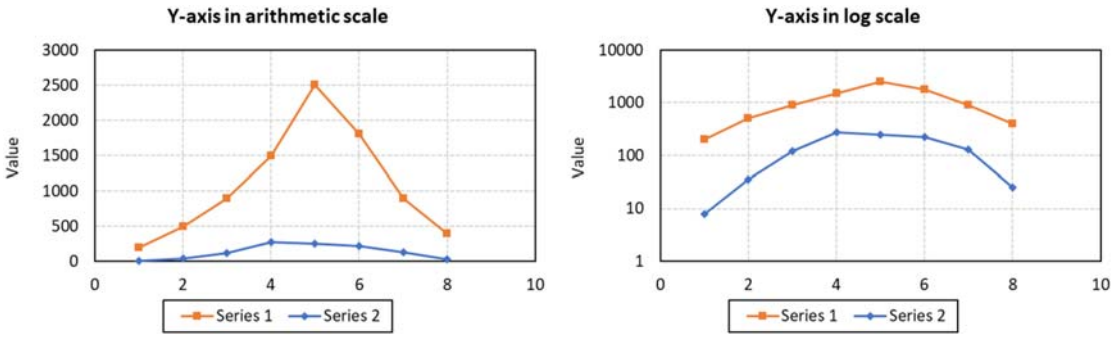


Figure 6.6 Comparison between two charts representing the same data. The chart on the left represents the two series in arithmetic scale in the *Y*-axis, while the graph on the right represents the two series on a logarithmic scale in the *Y*-axis.

Concentrations of **coliforms** and other microorganisms (e.g., *E. coli*, enterococci, etc.) should always be plotted on a **log scale**, with the numbers usually presented in **scientific notation** (unless you are dealing with very low concentrations). Otherwise, the $\log_{10}$-transformed concentrations should be plotted on an arithmetic scale (but the label of the axis should specify that the values are $\log_{10}$-transformed concentrations, not absolute concentrations). Figure 6.7 (left) shows how plotting coliform data using a conventional arithmetic scale conceals the values of the effluent concentrations from a wastewater treatment plant (WWTP), since they appear very close to the chart base. Because the effluent values are around five orders of magnitude lower than the influent ones, they simply do not appear well on the graph. However, when we use a log scale (right chart), we can see the effluent values and notice that they are in the order of 10^3 MPN/100 mL.

6.2.7 Moving averages

Advanced

In treatment plant monitoring, we frequently obtain data that fluctuate widely, and it is sometimes difficult to analyse if there are trends or any specific pattern in them. This is specially the case when we are not

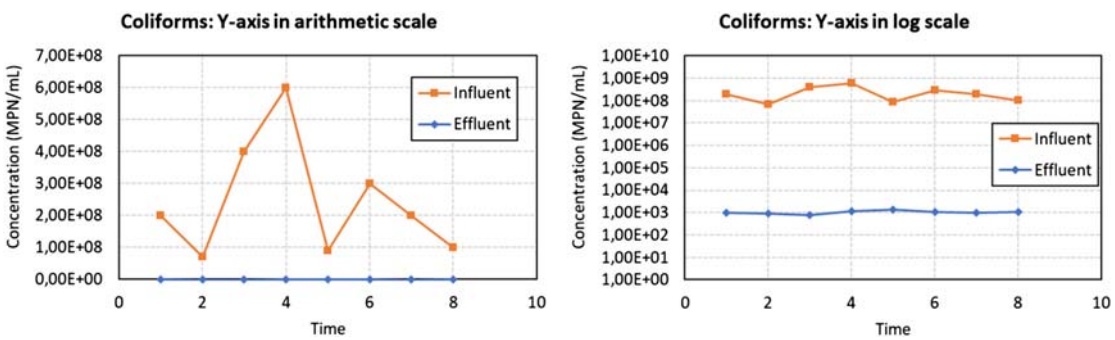


Figure 6.7 Comparison between two charts representing the same data of influent and effluent coliform concentrations in a wastewater treatment plant. The chart on the left represents the two series in an arithmetic scale in the *Y*-axis, while the graph on the right represents the two series on a logarithmic scale in the *Y*-axis.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

concerned with the interpretation of individual data, but rather to identify their more general temporal behaviour. One way of looking at the series in a clearer way is by **smoothing** its variations. A widely used procedure for data smoothing is to use **moving averages** (also called *rolling* averages).

We will explain how to calculate moving averages in a minute. But, for now, let us take the example shown in Figure 6.8. This example shows the influent flow to a treatment plant, based on daily measurements over a period of four years (taken from the Excel master spreadsheet described in Section 5.1; worksheet *Time Series*). Charts (a) and (b) present the original time series with (a) and

S. 5.1

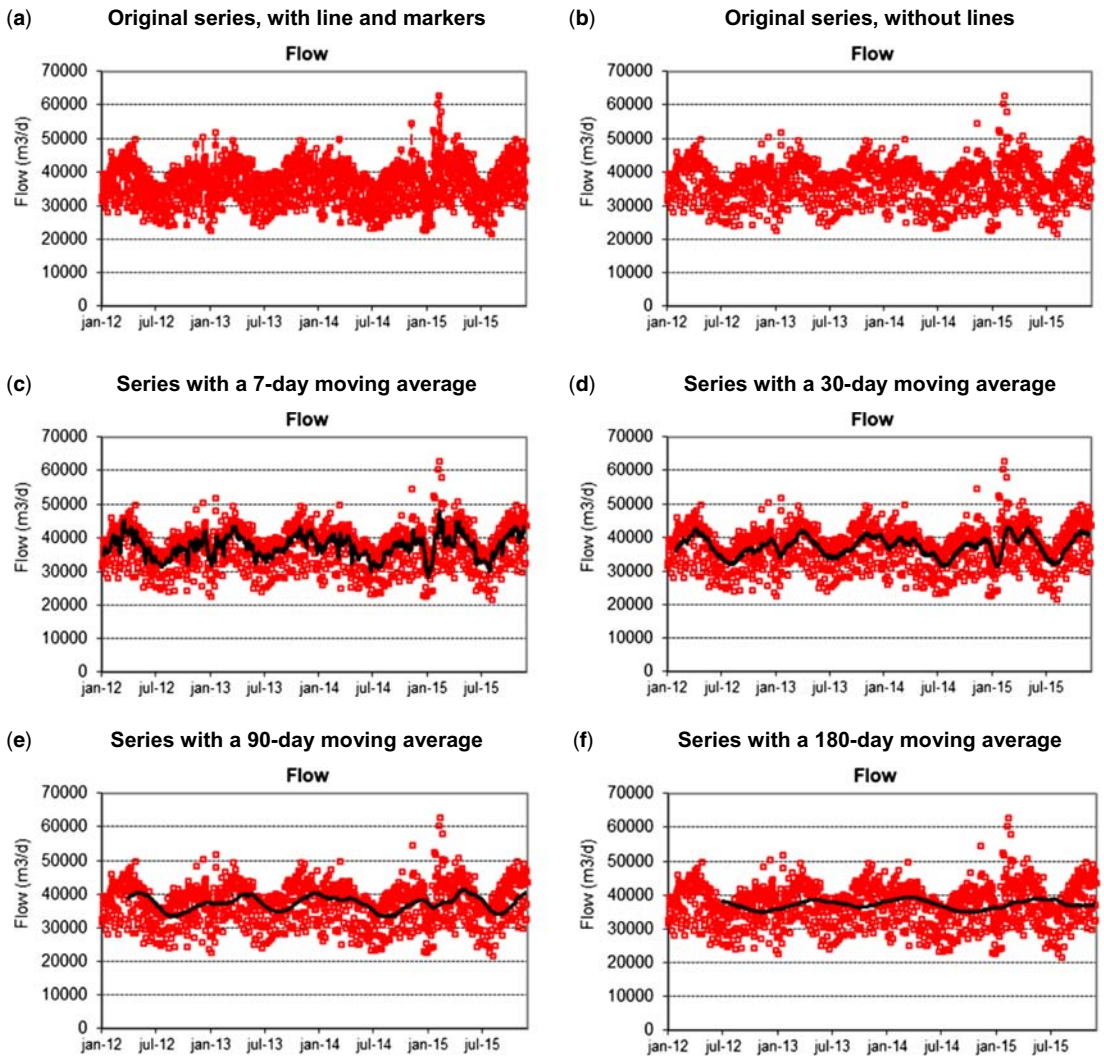


Figure 6.8 Time series of influent flow to a treatment plant, with daily data over a period of four years. Chart (a) shows the original data (markers) connected with lines and chart (b) shows the same data, but without connecting lines (only the markers). Charts (c)–(f) include a moving average line, with different time periods (7, 30, 90, and 180 days).

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

without (b) connecting lines. Because there are so many data with substantial scatter, you can hardly detect any possible trends. Then you decide to plot the moving average of the series, with a period of seven days, resembling what could look like weekly averages (chart c). You can now start to see the line, but there are still ups and downs that are troubling you. Then, you decide to extend the period of the moving average, that is, to smooth even further the series. You try with 30 days (approaching the concept of monthly averages), then 90 days (quarterly averages), and finally, 180 days (half-yearly averages). You can see that **the longer the period, the smoother the series**. With the longer periods, you start to see that the time series shows a seasonal pattern, with more or less well-defined periods of increase and decrease. But you should note the following points: (i) *the longer the period, the larger the quantity of data* ***points you lose*** *at the beginning* in the smoothed series; for instance, for a 30-day moving average, the curve does not show the first 29 days; (ii) *the longer the period, the larger will be the* ***lag*** between seasonal peaks in the data and the peak of your smoothed series (see chart f).

Depending on the frequency of data collection, you can select other periods. For instance, if you have a long time series with monthly values, you could think about using periods of 6 or 12, in order try to represent half-yearly of yearly averages.

It is very easy to incorporate **moving averages** in **Excel** charts. Check the version you have and obtain on-line assistance, but usually it involves right-clicking on your plotted series, adding *Trendline* and selecting the *Period* (default value is 2).

But for you to understand the concept behind it, let us give the example based on the same data set shown in Figure 6.8. The difference is that now, for the sake of simplicity, we will show only the first 20 days, knowing that the sequence will be done in the same manner.

Example

EXAMPLE 6.2 CALCULATION OF MOVING AVERAGES

You have a time series of daily values of inflow to your treatment plant. Calculate the 7-day moving average for the first 20 days in the series.

Excel

Note: This example is also available as an Excel spreadsheet.

Data:

Day	Flow (m^3/d)	Day	Flow (m^3/d)	Day	Flow (m^3/d)	Day	Flow (m^3/d)
1	32,180	6	39,398	11	39,251	16	37,785
2	32,470	7	34,464	12	36,934	17	36,870
3	31,560	8	32,522	13	38,093	18	37,872
4	37,486	9	39,152	14	32,149	19	37,541
5	35,990	10	38,489	15	29,431	20	38,382

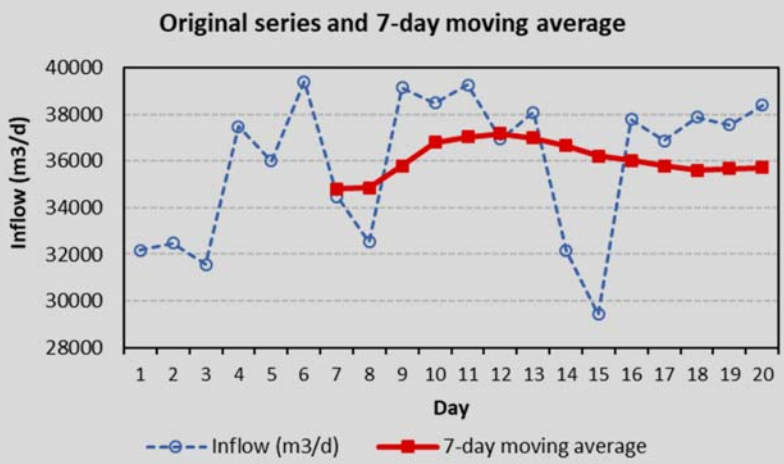
Solution:

Structure the following table. The first two columns will be with your original data. The third column will have the calculation of the moving average, as explained in the fourth column.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Day	Flow (m^3/d)	Seven-day Moving Average (m^3/d)	Comment on the Moving Average
1	32,180		
2	32,470		
3	31,560		
4	37,486		
5	35,990		
6	39,398		
7	34,464	34,793	Average from days 1 to 7
8	32,522	34,841	Average from days 2 to 8
9	39,152	35,796	Average from days 3 to 9
10	38,489	36,786	Average from days 4 to 10
11	39,251	37,038	…
12	36,934	37,173	…
13	38,093	36,986	…
14	32,149	36,656	…
15	29,431	36,214	…
16	37,785	36,019	…
17	36,870	35,788	…
18	37,872	35,591	…
19	37,541	35,677	…
20	38,382	35,719	Average from days 14 to 20

The resulting graph is shown below. Note that the seven-day moving average is smoother than the original series, and that it starts to be plotted only on the seventh day.



Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

6.3 FREQUENCY DISTRIBUTION

6.3.1 Frequency distributions and frequency histograms

Basic

You have already seen graphs such as the ones shown in Figure 6.9. These are **frequency histograms**, widely used in statistics, and also in monitoring treatment plants and water bodies. What we see in a histogram is the number of occurrences of values of some parameter, separated by ranges of values (known as bins). Figure 6.9 shows an example of COD concentrations in the influent (left chart) and the effluent (right chart) of a treatment plant. From the graph, we can see that the frequencies are separated according to ranges of values; for instance, 125 influent data points have values that fall between 515 and 584 mg/L, and 170 effluent data points have values that fall between 44 and 48 mg/L (these are the modes). The limits of the bin ranges seem a bit awkward, but these graphs have been done automatically by Excel using data from the master Excel spreadsheet on treatment plant performance. What is important is their interpretation: we see ranges where most data are concentrated and also the shape of the distributions. Shapes of distributions were briefly discussed in Section 5.6.1, but they will be covered in more detail in Chapter 8.

S. 5.6.1
C. 8

Any general statistical software you use will be able to make frequency histograms, because of their popularity in data analysis. However, in order to understand how histograms are made and how they should be interpreted, we need to analyse the concept of **frequency distributions**.

A frequency distribution is computed as a simple table that shows the ranges of values into which a sample can be divided, and the frequency at which the data are observed within each range. Frequency distributions can be either (a) **absolute** or **relative** and (b) **simple** or **cumulative**. Absolute frequencies refer to the total number of data points that fall within each particular range (or bin). Relative frequencies refer to the percentage of data points that fall within each range. A simple frequency only accounts for the data points within a particular range, while a cumulative frequency also includes the data points falling within all ranges below the range in question.

To construct a frequency distribution table, you need to divide the data range into intervals, which are often called **class intervals** or **bin intervals**. *The intervals should be of equal width* to increase the visual information in the frequency distribution. Some judgment should be used in the selection of the **number of class intervals** (**NCIs**), so that a reasonable graph can be developed. The number of class intervals depends on the **number of observations** (***n***) and the dispersion of the data. In most cases, 5–20 intervals are satisfactory, and the number of intervals should grow with the value of '*n*' (Montgomery & Runge, 2003). If NCI is too small, the visualization of the main characteristics of the histogram will be affected. If NCI is too large (class width too narrow), there will be excessive fluctuations in the class

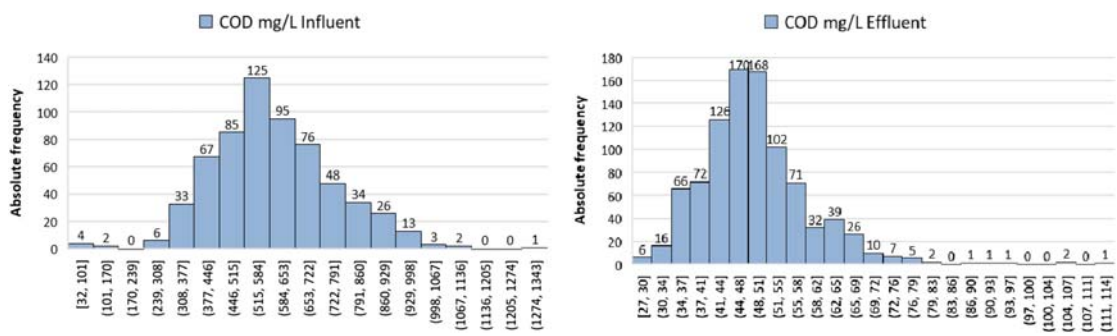


Figure 6.9 Example of a frequency distribution histogram, showing influent (left) and effluent (right) COD concentrations.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

frequencies (Naghettini & Pinto, 2007). Two simple approaches for deciding the number of class intervals NCI are (Oliveira, 2017; Naghettini & Pinto, 2007)

- NCI = integer number closest to $\sqrt{n}$
- $NCI = 1 + 3.3 \log_{10}(n)$
- In practice, NCI should have a minimum of 5 and a maximum of 25, with the additional comment that *histograms are not informative when the sample size is less than 25*

Excel uses an internal criterion to propose automatic NCIs.

Example 6.3 shows you how to build a frequency distribution table and, from it, a frequency histogram chart. As we mentioned, this is automatically done by most statistical software packages, and Excel has a data analysis tool kit that automates the process for you. But if you decide to do it by yourself, as shown in the example, you can use the **Excel** function **FREQUENCY**.

Example

EXAMPLE 6.3 FREQUENCY DISTRIBUTION TABLE AND FREQUENCY HISTOGRAM

You collected samples of a certain constituent at the effluent from your treatment plant (or at the water body you are studying). Structure the frequency distribution table and plot the frequency distribution histograms (absolute and relative; simple and cumulative).

Excel

Note: This example is also available as an Excel spreadsheet.

Data:

2.8	4.2	3.9	3.3	2.8	1.7	1.9	2.5	3.1	3.8	2.7	4.1
4.3	4.8	5.1	5.2	3.9	3.5	2.7	3.1	2.8	3.4	4.9	2.8
2.8	1.8	2.1	2.6	2.3	2.4	2.5	1.3	2.9	2.4	2.1	3.6

Solution:

Based on a simple analysis of your data, you obtain the following information:

Number of data points: $n = 36$
Minimum value: 1.7 mg/L
Maximum value: 5.8 mg/L

Based on this, you decide to specify the width (= max − min) of each class interval to be equal to 1.0 mg/L, in order to have rounded up values in your table and chart. This will lead to six class intervals, if we start at 0 mg/L. This value of NCI = 6 class intervals is supported by the simplified proposals previously presented to you:

- $NCI = \sqrt{n} = \sqrt{36} = 6.0 \rightarrow NCI = 6$
- $NCI = 1 + 3.3 \log_{10}(n) = 1 + 3.3 \log_{10}(36) = 6.1 \rightarrow NCI = 6$

After that you structure a table like the one below and fill in the cells. The table is made up like this:

- *Simple absolute frequency.* There are no values ≤1.0; therefore, the first value in the column is 0. There are four values >1.0 and ≤2.0 (1.7, 1.9, 1.8, and 1.9), so the second value in the column is 4, and so on.
- *Cumulative absolute frequency.* You sum up the simple absolute frequency values in order to obtain the cumulative values in the third column. For instance, the third value, 20, is equal to 0 + 4 + 16. You

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

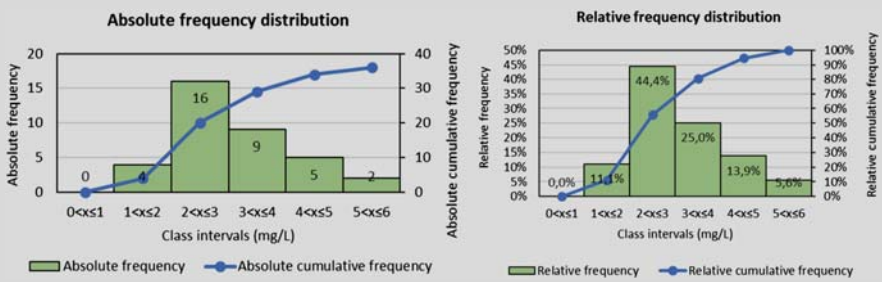
do the same for all other values. You can also use the Excel function FREQUENCY to obtain the number of data that are less than or equal to the upper value of the range of your class interval.

- *Simple relative frequency*. You divide the values in the column 'simple absolute frequency' by the number of data, which, in this case, is $n = 36$. Therefore, the second value in the column is $4/36 = 0.111 = 11.1\%$. You do the same for the other values.
- *Cumulative relative frequency*. You sum up the simple relative frequency values in order to obtain the cumulative values in the last column. For instance, the third value, 55.6%, is equal to $0.0 + 11.1 + 44.4$ (small difference due to rounding errors for reporting with only one decimal case). You do the same for all other values.

Class Intervals (mg/L)	Absolute Frequency		Relative Frequency (%)	
	Simple	Cumulative	Simple	Cumulative
$0.0 < x \leq 1.0$	0	0	0.0	0.0
$1.0 < x \leq 2.0$	4	4	11.1	11.1
$2.0 < x \leq 3.0$	16	20	44.4	55.6
$3.0 < x \leq 4.0$	9	29	25.0	80.6
$4.0 < x \leq 5.0$	5	34	13.9	94.4
$5.0 < x \leq 6.0$	2	36	5.6	100.0
Total	36	–	100	–

The frequency distribution table can also be interpreted in terms of compliance with target values for your effluent concentration. If the legislation specified that the discharge standard for your constituent (or the maximum allowable value in the water body) is 4.0 mg/L, you can see from the cumulative distributions that 29 out of 36 samples (80.6%) were below or equal to this value and thus complied with the standard. Seven samples out of 36 were not in compliance ($7/36 = 19.4\%$; also calculated as $100 - 80.6 = 19.4\%$).

The resulting frequency histograms are presented below. The left chart presents the *absolute frequencies (expressed as numbers)*, while the right chart shows the *relative frequencies (expressed as percentages)*. The histograms simply plot the values from the table. As a matter of fact, the results of the simple frequencies are shown in their respective bars.



In both graphs, the simple frequency (left *Y*-axis) allows the visualization of the distribution of the data. It is observed that this specific data follow, albeit in a more or less rudimentary way, a bell shape with a skew to the right, which is usually associated with a log-normal distribution. Tests on the normality or non-normality of the distribution can be performed through specific statistical procedures, covered in Section 8.2.

S. 8.2

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

The cumulative frequency (right *Y*-axis) allows inference on the percentage of values below a certain concentration. Thus, the above-mentioned statement that 80.6% of the values are below the concentration of 4.0 mg/L can also be derived from this graph (although without the accuracy of the calculated values). In the graph, the concentration value of 4.0 mg/L should be read on the *X*-axis. From this mark, rise with a vertical line, which, when meeting the curve of the cumulative frequency, determines the percentage of values below or equal to 4.0 mg/L, read on the right *Y*-axis.

6.3.2 Frequency polygon

Advanced

Frequency polygons are another type of presentation of the frequency table and are useful as a *diagnosis of the pattern of the distribution of your variable*. The polygon is formed by connecting the mid points at the top of the bars of the histogram. The line should start and finish with the values of zero on the left and right sides of the polygon.

It is more usual to make polygons of *simple relative frequencies* instead of absolute frequencies. In this case, the *Y*-axis plots the frequencies of occurrence, limited between the extreme values of 0 and 1. With the increase in the number of data, and the resulting *decrease in the width of the class intervals*, the relative frequency *polygon becomes a frequency distribution curve*. At the limit case of a *sample of infinite size*, *this curve becomes the probability density function of the population* (Naghettini & Pinto, 2007). Important examples are the normal distribution and the log-normal distribution, which will be covered in detail in Chapter 8.

C. 8

Example

EXAMPLE 6.4 FREQUENCY POLYGON

Plot the frequency polygon using the data from Example 6.3.

Excel

Note: This example is also available as an Excel spreadsheet.

Solution:

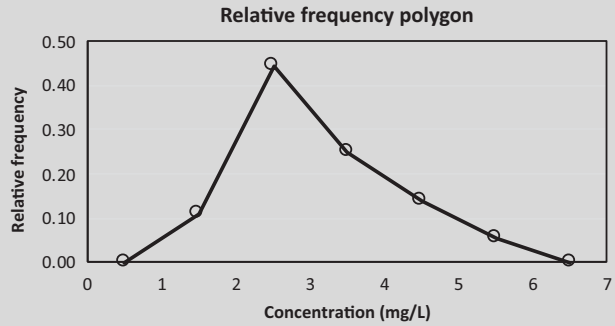
Taking the mid-values of each of the class intervals presented in the table of Example 6.3 and using the values of the simple relative frequency distribution shown in the same table, you have

Class Intervals (mg/L)	Mid-range Values	Relative Frequency
$0.0 < x \leq 1.0$	0.5	0.00
$1.0 < x \leq 2.0$	1.5	0.11
$2.0 < x \leq 3.0$	2.5	0.44
$3.0 < x \leq 4.0$	3.5	0.25
$4.0 < x \leq 5.0$	4.5	0.14
$5.0 < x \leq 6.0$	5.5	0.06
$6.0 < x \leq 7.0$	6.5	0.00

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Note that we inserted the last row with mid-point of 6.5 in order to finalize with a frequency of zero. The first class interval already had a frequency of zero. For polygons, we need to start and finish with zero frequencies.

The values of the second and third columns of the table are then plotted, leading to the frequency polygon chart shown below. Compare it with the relative frequency histogram from Example 6.3 and you will see the similarities.



6.3.3 Percentile graphs

Advanced

Another way of presenting the frequency distribution of your data is by plotting **percentile graphs**. These are basically *relative cumulative frequency distribution* charts, but using *percentiles in one of the axes*. Some people plot, for instance, concentrations in the *X*-axis and percentiles in the *Y*-axis. Here, we will plot percentiles in the *X*-axis and concentrations in the *Y*-axis, because they are easier to make with Excel.

Take the graphs shown in Figure 6.10, based on actual data from the Excel master sheet on the performance of a treatment plant. Effluent COD concentrations are displayed in the *Y*-axis of the left graph and COD removal efficiencies in *Y*-axis of the right graph. In both graphs, the *X*-axis shows *percentiles*, representing the *percentage of values that are less than or equal to the corresponding value in the Y-axis*. The graphs have been made using the Excel function PERCENTILE. We calculated the

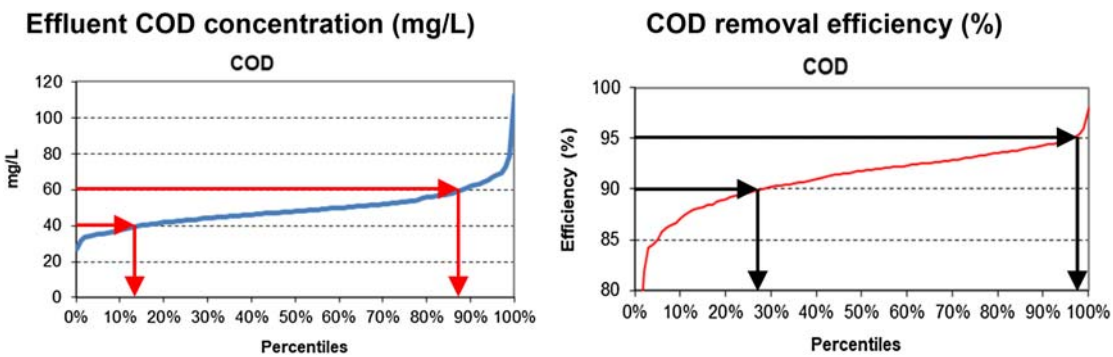


Figure 6.10 Example of percentile graphs for effluent COD concentrations and COD removal efficiencies in a treatment plant.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

values (concentration and efficiency) that corresponded to the percentiles 0%, 1%, 2%, 3%, …, 99%, 100%, and plotted them in the *Y*-axis (see how to make percentile graphs in Example 6.5).

We can use this plot to make inferences about compliance with standards based on concentrations (treatment plant effluent or water body). We will now use the left graph in Figure 6.10, in which we plot percentiles of concentrations**.** If the discharge standard were 40 mg/L, we would start from this value in the *Y*-axis and draw a horizontal line. Where this line crossed with the percentile curve, we would draw a vertical downward line. Where this line crossed the *X*-axis, we would read the value, and would get a value around 15% (the exact value can be seen from the calculations using the PERCENTILE function). This means that only 15% of your samples would comply with the discharge standard of 40 mg/L. However, if the discharge standard were a bit more relaxed, say, 60 mg/L, then we would see that almost 90% of the values would be below it, indicating a conformity around 90%. From the graph, we can also see that most of the effluent concentrations lie in the range between 40 and 60 mg/L. For you to see this, look at the *X*-axis in the range delimited by the two arrows, and you will see that around 75% ($= 90 - 15$) of the data are situated in this range.

Now let us take the case of **removal efficiencies** (right chart). The way we deal with the graph is the same, but the interpretation is the opposite. *For effluent concentrations, the lower the values, the better. The opposite occurs with removal efficiencies: the higher the values, the better*. If our target value was 90% removal, drawing the horizontal and vertical lines, we would see that around 28% of the values are below the target. In other words, $100 - 28 = 72\%$ of the values are above the target and comply with the specified target value. If, on the other hand, we had a more stringent standard removal efficiency of, say, 95%, we would see that around 98% of the values are below the target value, meaning that only $100 - 98 = 2\%$ of the values conform with this more stringent removal efficiency target. Note that in this graph, in order to improve visualization, we made our *Y*-axis scale vary from 80% to 100%, and not 0% to 100%. Also remember that *you can have negative efficiencies in a treatment plant*, in the case that the effluent concentration exceeds the value of the influent concentration (which may happen occasionally). Negative values can appear in your percentile graphs, provided that you allow your *Y*-axis to be automatically scaled by Excel. But also remember that *we cannot (by definition) have removal efficiencies above 100% (you cannot remove more of a constituent than whatever was present to begin with).*

In summary, when interpreting the percentile graphs *in terms of compliance with discharge standards*, keep in mind:

- **Effluent concentrations or concentrations in water bodies**. The *value read in the X-axis gives you the percentage of compliance* with the standard for effluent concentrations or water bodies.
- **Removal efficiency in treatment plant**. *100 minus the value read in the X-axis gives you the percentage of compliance* with the standard for removal efficiencies.

Notice that if your monitoring is undertaken at **fixed time intervals** (e.g., every day, every week, and every month), the *percentage of samples will be equal to the percentage of time*. This is the concept of **permanence curves**, widely used in hydrology to represent frequency distributions of flows in rivers. In the examples above, we will be able to say, for instance, that for only *15% of the time* your treatment plant was complying with the discharge standard of 40 mg/L of COD. Also, you would be able to say that for *72% of the time*, your treatment plant was in conformity with the minimum required COD removal efficiency of 90%.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Example

Excel

EXAMPLE 6.5 PERCENTILE GRAPH

Plot the percentile graph using the data from Example 6.3.

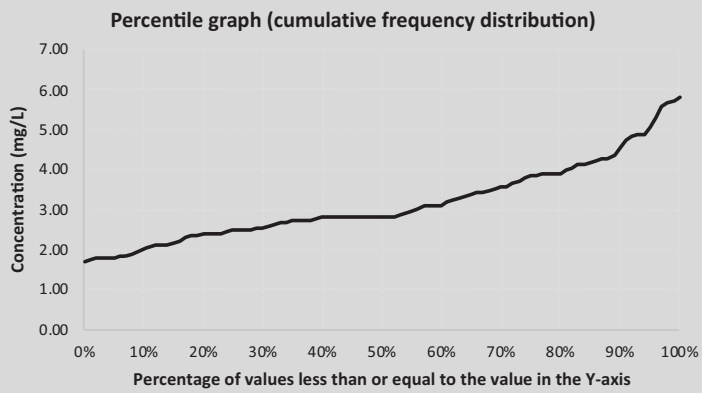
Note: This example is also available as an Excel spreadsheet.

Solution:

Using the Excel PERCENTILE function, you structure the following table, for percentile values varying from 0% to 100%. In order to draw a smooth graph, we will do calculations at every 1 percentile.

Percentiles (%)	Concentration (mg/L)	Comment
0	1.70	Minimum value in the sample
1	1.74	
2	1.77	
3	1.80	
4	1.80	
…	…	
25	2.48	First quartile
…	…	
50	2.80	Second quartile (median)
…	…	
75	3.71	Third quartile
…	…	
98	5.66	
99	5.73	
100	5.80	Maximum value in the sample

The resulting graph is shown below. The interpretation can be done following the structure of the comments made for Figure 6.10.



Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

6.4 BOX-AND-WHISKER GRAPHS (BOX PLOTS)

Basic

Box-and-whisker graphs, or **box plots**, are a very useful way of presenting descriptive statistics for your monitoring data. They are popular in scientific texts, because they give a general view of the *central values and the data dispersion, as well as possible asymmetry, tails, and outliers*.

If you are presenting data from only one sample location (with multiple replicates), possibly a frequency histogram will be more adequate, especially if you are introducing your results to an audience or readership that is not familiar with box plots. However, if you want to *compare two or more samples (each with multiple replicates) in a single chart, box plots will be very useful*. But remember, they are not so well-known by the non-scientific community, and if you use them in an oral presentation for such an audience, you will need to explain their structure.

S. 5.8

The essence of the box plot is the box, which is defined by the first and third quartiles (see Section 5.8):

- **First quartile ($Q1$) = 25 percentile**: the bottom part of the box
- **Third quartile ($Q3$) = 75 percentile**: the top part of the box

Inside the box, we include the median or second quartile:

- **Second quartile ($Q2$) = 50 percentile or median**: marker inside the box

These are *integral parts of all box plots*. You need to check with the statistical software you are using *how it complements the other parts of the graph*, especially the whiskers that extend outside (above and below) the box. The whiskers may include limits for outliers, outliers and minimum and maximum values. Excel also prepares box plots and gives different options to structure the chart.

In this book, we have prepared our own box plots and included the following **additional elements**, in order to have an even better visualization of central tendency, dispersion, and data asymmetry:

- Minimum value
- 10 percentile
- Arithmetic mean
- 90 percentile
- Maximum value

Excel

The Excel spreadsheet box plot allows you to draw these graphs, including the elements listed above. If you prefer, you can substitute the 10 and 90 percentiles by 5 and 95 percentiles, simply by changing the values in the sheet.

Let us use the data from sample 5 of the Excel file box plot. Figure 6.11 presents the corresponding box plot. In this graph, we have included explanatory text to assist you in the identification of all elements comprising the chart. Our Excel file has prepared three options for you: with explanatory text, with labels, and without anything.

From this graph, we can see the following points:

- The box comprised of the percentiles 25% and 75% indicate that 50% ($75 - 25 = 50\%$) of the data are between 2.0 and 3.2 mg/L.
- The percentiles 10% and 90% indicate that 80% ($90 - 10 = 80\%$) of the data are between 1.7 and 3.6 mg/L.
- The minimum and maximum values are not very distant to the 10^{th} percentile and the 90^{th} percentile, indicating that there are no extreme outliers in this data set.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

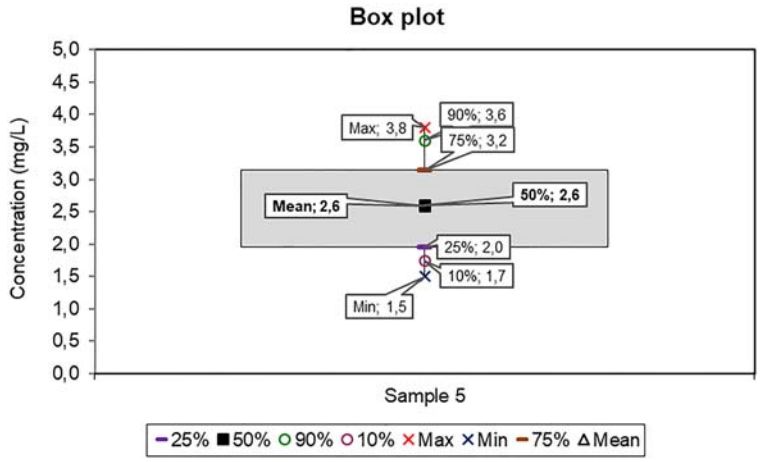


Figure 6.11 Box plot of a data set in Excel file box plot, showing a relatively symmetrical distribution of the data.

- The distribution seems to be relatively symmetrical: the median is situated in the middle of the box, and the points and markers above and below the median seem to be distributed at similar distances.
- The mean and the median are very similar, which is another indication of the approximate symmetry of the data.

Let us now use the data from Example 6.3, which was complemented with Examples 6.4 and 6.5. In these examples, we saw their frequency histograms, frequency polygons, and percentile graphs. The structure is the same, and you now know how to interpret it. Although the *Y*-axis scale in Figure 6.12 is different from Figure 6.11, we can see that the data now are not symmetrically distributed. The upper whisker is longer than the lower whisker, the higher percentiles are more distant from the median than the lower percentiles, and the mean is higher than the median.

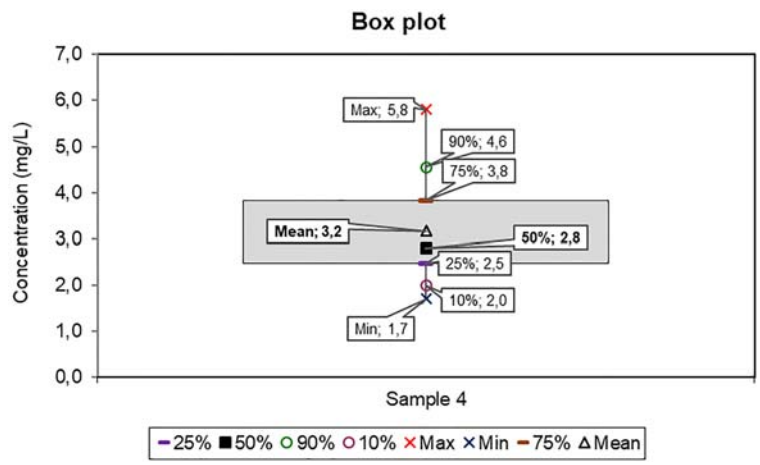


Figure 6.12 Box plot of the data used in Example 6.3, showing a relatively *non-symmetrical* distribution of the data.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

We showed above examples of box plots containing only one data set in each, in order to show you how to interpret its structure. However, box plots are *especially useful when plotting together more than one data set, in order to allow comparisons* among them. Possible examples are shown in Figure 6.13 for effluent concentrations (see also Figure 5.1 for a description of different types of comparisons in treatment plant studies). Similar examples could also be performed using removal efficiencies. The statistical support for hypothesis testing used in the comparison among different data sets is described in Chapter 10.

C. 10

For **water quality monitoring in water bodies**, you could have adaptations in Figure 6.13, such as:

- Upstream/downstream (instead of input/output in figure 'a')
- Sampling points 1, 2, 3, …, *n* (instead of input and output 1, 2, 3, 4 in figure 'b')
- Summer/winter, wet/dry, before/after intervention (in figure 'e')
- Water bodies 1, 2, 3, …, *n* (instead of plant 1, 2, 3, 4, 5 in figure 'f')

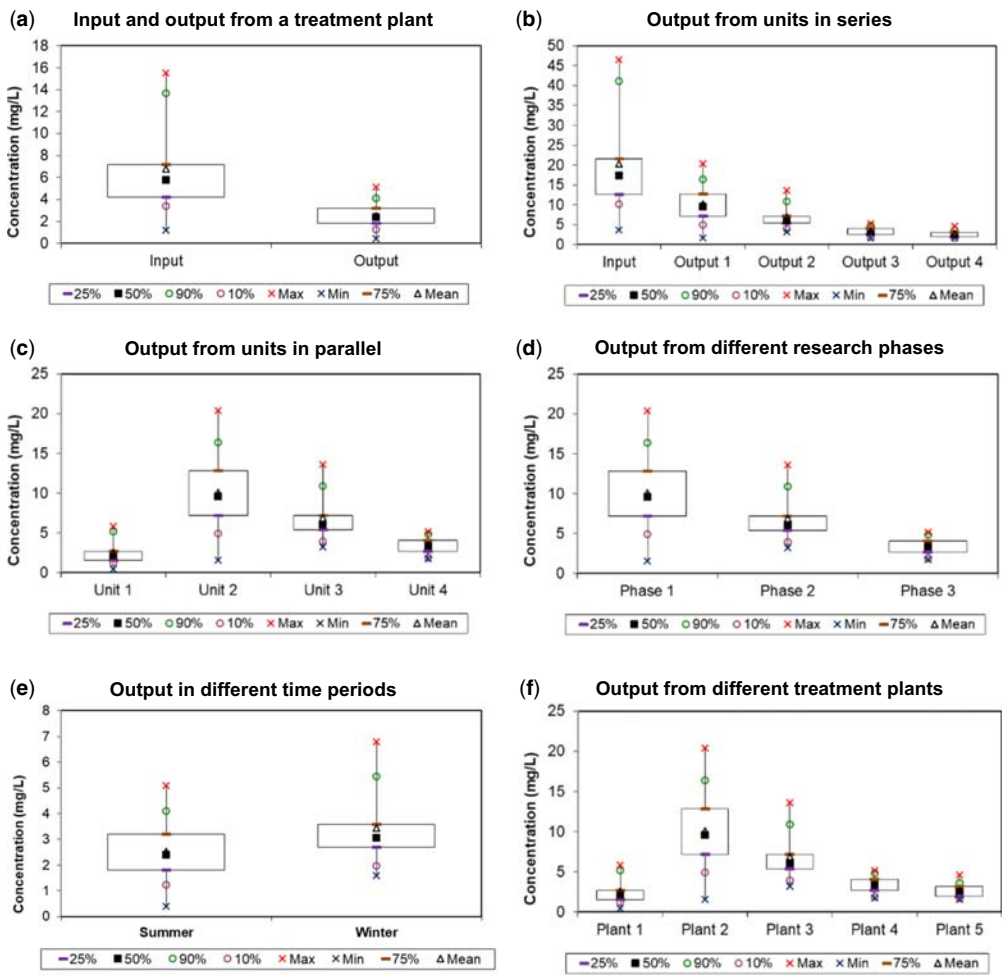


Figure 6.13 Different applications of box plots in treatment plant monitoring.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

6.5 SCATTER PLOTS

Basic

C. 11

Scatter plots are simple graphs that relate the *values of one variable to the values of another variable*. Each variable is shown on one of the axes of the chart, and the values are represented in the chart in the form of points or **markers**. This graph is quite useful for analysing the *relationship between two variables* and is the starting procedure for a correlation or regression analysis (see Chapter 11). Scatter plots are easily made in Excel and in any statistical package you might use.

If you expect some *degree of dependence* between the two variables you are analysing, you should plot on the *X-axis the independent variable* and on the *Y-axis the dependent variable* (the dependent variable is the variable that you judge to be influenced by the independent variable). Note that creating a scatter plot is just a preliminary visual analysis, and *you cannot infer anything on the possible causality in the association* between the variables based on the plot alone. With the scatter plot, you may only suggest a possible correlation between them, which can be further investigated in a regression analysis (see Chapter 11). The dependence between the two variables, even if endorsed by the results of the regression analysis, will ultimately depend on your judgement and your knowledge of the system you are investigating and the role of each variable.

C. 11

Let us analyse the scatter plot displayed in Figure 6.14. The data have been taken from the master spreadsheet entitled Descriptive Statistics Treatment Plant. They represent monthly averages of effluent SS and COD concentrations from a treatment plant. Because you are familiar with the basic principles of wastewater treatment, you know that particulate COD in the effluent is associated with SS, and because of this, you decide to plot SS in the *X*-axis and COD in the *Y*-axis.

From the figure, you observe that increases in effluent COD are associated with higher values of effluent SS. As expected, you find *spread* in the data, since it does not all fall neatly along a single line. This type of spread or variability is not uncommon especially when you are dealing with environmental variables and all the inherent variation associated with them. You consider that maybe the relationship between the two variables could be *linear* but do not discard the possibility that a *non-linear* relationship exists. Finally, you notice what appears to be an *outlier* on the right-hand side (24 mg/L of SS; 71 mg/L of COD). However, the outlier is situated following approximately the same trajectory of the other data points; therefore, it might indeed follow the same trend as the rest of the data points.

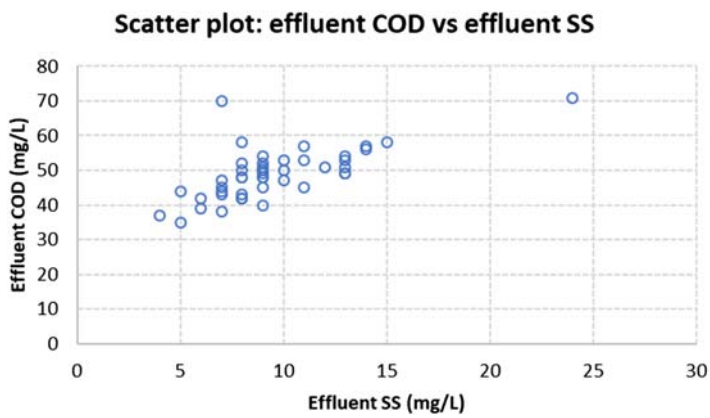


Figure 6.14 Scatter plot between effluent COD and effluent SS from a treatment plant.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

S. 6.2

The first visual impression of your scatter plot will be also influenced by the *scales* you choose for the *X* and *Y* axes (see discussion on axes scales in Section 6.2). Also, it is not infrequent to see a *cloud of points* in scatter plots without showing any definite pattern. You should interpret it and dig more into the data and the other chart types shown in this chapter to see whether there is any other good way of visually representing your data.

In conclusion, the usefulness of the scatter plots will depend very much on your ability to understand the behaviour of the data in your treatment plant or water body. Scatter plots open a window for you to see possible correlations and relationships between variables, and it is your knowledge of the system that will expand their usefulness.

6.6 GRAPHS FOR QUALITATIVE (CATEGORIZED) DATA

Basic

S. 6.1

As mentioned in Section 6.1, **qualitative data** are those that cannot be measured or expressed in a quantitative scale. They represent **categorical values**, which can be characterized by codes, names, letters, or numbers. Descriptive graphical analysis of qualitative data usually uses **bar/column charts** or **pie charts** (Mendenhall & Sincich, 1988):

- **Bar charts or column charts**. Bar charts are usually horizontal, while column charts are vertical, and both represent counts or the absolute or relative frequency corresponding to each data category. The height or length of each bar is proportional to the count or frequency of occurrence within the category.
- **Pie charts**. Pie charts divide a circle (or pie) into sectors, one corresponding to each category, with the angle of each sector proportional to the relative frequency in the category.

Example 6.6, adapted from von Sperling *et al.* (1996), presents applications of these charts. You should decide which one communicates in a clearer way your data and the message you want to convey.

Example

Excel

EXAMPLE 6.6 BAR/COLUMN CHARTS AND PIE CHARTS FOR QUALITATIVE DATA

Make bar/column charts and pie charts for the following data, related to the statistics of sewage treatment in four cities you are studying. The data you obtained were categorized into three different types of treatment processes: stabilization ponds, UASB reactors, and other processes.

Note: This example is also available as an Excel spreadsheet.

Data:

Percentage of the produced sewage that is treated in each of the four cities, according to the three different treatment processes (%).

Process	City A	City B	City C	City D
Ponds	20	15	45	4
UASB	10	23	27	3
Others	3	12	18	1
Total	33	50	90	8

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

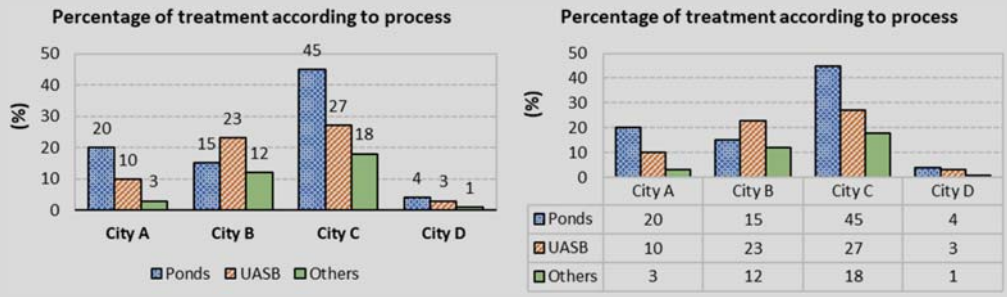
Sewage flow treated in each city, according to three different treatment processes (L/s).

Process	City A	City B	City C	City D	Total
Ponds	10	45	22	12	89
UASB	7	66	14	9	96
Others	6	36	9	5	56
Total	23	147	45	26	241

Solution:

The figures below use column charts and bar charts illustrating the comparison among the four cities in terms of the **percentage** of the **total** sewage flow generated that is treated (coverage of sewage treatment in terms of flow). Both graphs present the same information, and you can decide on them based on your preference and the clarity of the resulting chart.

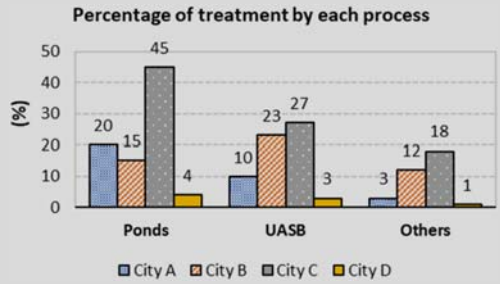
The following column charts present the **percentages** of sewage treatment in each **city**, separated in terms of the three sewage **treatment categories**. The difference between them relates to formatting details. The left graph shows the labels with the values right above the columns, while the right graph includes a table with the data. Again, it is a matter of preference. Take care that the graph or the table is not confusing because of excessive information. Make sure that the colours or fill options for the columns representing each category are clear enough and distinguishable from the others, even if printed in black and white.



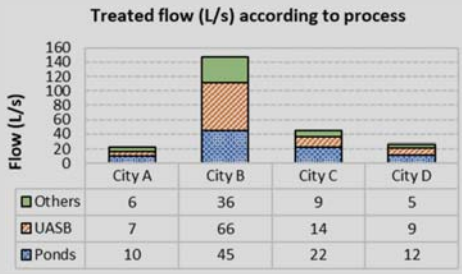
The graph below shows the same data from the left graph above, but they are organized in a different way. Now, the categories (*X*-axis) are organized by the treatment process, and each treatment process

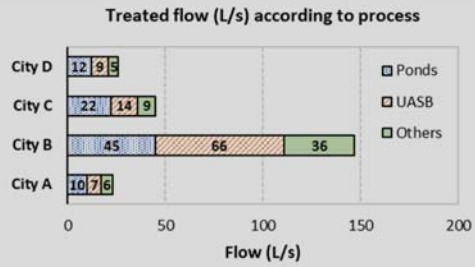
Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

covers data from the four cities. You have to decide upon which element you want to put more emphasis: organization by cities (above) or by the process (below).

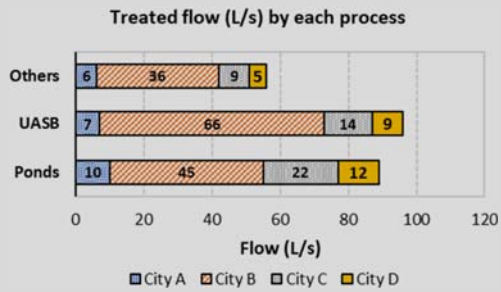


Now, we present graphs related to **treated flow** (L/s) and not percentages of treatment. We want to visualize the information on the flows treated by each process and still have a view on the total flow treated in each city. For this, we use charts with **stacked columns** or **stacked bars**, since we can sum up flows. Although city C had the highest percentage of treated sewage (90%, according to the graphs above), we see that in absolute terms, it treats a smaller flow, compared with city B. The charts below illustrate this comparison in absolute terms, using columns and bars, and different formatting options. Again, you have to decide, in this specific case, which charts better convey your message.





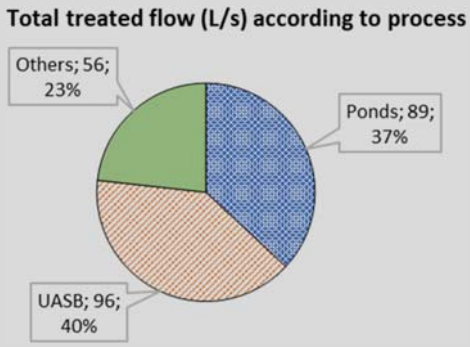
You may want to organize this information in a different way. You change the above right graph by the one presented below, which puts more emphasis on the treatment process and less on the city. From this chart, we directly see that the process treating the highest flow is UASB reactors, followed closely by ponds.



Finally, you can also use pie charts to illustrate your data. In the figure below, we show in a direct way the **total flows** (L/s) and the associated percentages (%) per treatment process. For instance, with the

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

formatting we used, the label shows you the process, flow in L/s, and respective percentage in terms of the total flow. We can conclude again that UASB reactors and stabilizations ponds are the most widely used processes, accounting for 77% (=40 + 37) of the treated flow. Pie charts are easily understandable by a non-technical audience and readership.



Basic

6.7 GENERAL ADVICES ON PRESENTING GRAPHS

Besides all the comments made in the preceding sections, we can highlight the following points:

- Scientific graphics should be clear, without an excess of details. Just because the spreadsheets or statistical software you are using offers many different formatting features does not mean you have to include every one of them in your graph.
- Every chart must be clearly identified by its general title and the titles of its *X* and *Y* axes, including the units (if applicable) of the variables.
- Every chart should be self-sufficient. If necessary, include footnotes and a detailed legend key. In a report or a manuscript, figures have titles (typically shown below the figure), and these titles should be descriptive enough for a reader to be able to interpret the chart without having to read the full text of the report. Thus, if the chart is reproduced in any other publication by another author, its communication and information capacity will be maintained.
- The internal lines of the charts (gridlines) should be included whenever the chart is used for the approximate interpolation of values. However, gridlines can be omitted in order to make the chart simpler and cleaner, especially if there is an appreciable number of curves in the chart, or if it is not going to be used to read off values from the series. If gridlines are used, it is important to make sure that the gridlines are formatted to appear lighter and thinner than the lines used for curves or bars.
- Try not to present a large number of data series on the same chart, as this will make the figure appear heavy and confusing. In this case, the chart should be divided into two or more figures if possible.
- The legend location must be selected in order to allow a balance on the chart, utilizing some unused internal area or defined areas outside the chart.
- If the graph presents series with experimental (observed) and estimated (calculated) values, the observed values should be presented in the form of markers (possibly linked), and the estimates in the form of lines. If the chart contains a series, which is partly composed by observed data and partly by estimated values, the estimated portion should appear as dashed lines.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

- Although graphs in three dimensions (3D or *X*–*Y*–*Z* graphs) or with features of perspective might seem visually elegant at first, usually the gain in elegance is impaired by a loss of clarity and a lack of interpretability for the reader. Reading the values in 3D column and bar charts is especially confusing. As a general rule of thumb: the simpler the better.
- In column and bar charts, the columns and bars should be wider than the spaces between them.
- In stacked column and bar charts, make the data set that is more important or with higher values closer to the base (near the *X*-axis). The strongest hatch or colour should also be in the series closer to the base.
- In pie charts, if there are two or more thin slices, you can combine them into a single slice, which can be called 'other'.
- In the case of colour graphs, you should remember that they may be copied or printed in black and white by a reader. If the graph requires colour for interpretation, the communication might be lost. You should therefore use other formatting features (e.g., different dashed/dotted lines, shapes of the markers, internal hatches or patterns) that allow the interpretation of the graph, even if it is presented in black and white.
- If you use hatches or fill patterns for graphs, check that the different hatch or fill patterns in contiguous bars or columns are not similar to each other, and that they do not create optical illusions or an impression of distortion.
- If you are giving an oral presentation at a conference or a seminar, for any slide containing a graph, you should always state what is represented on each axis and mention if there is a secondary *Y*-axis, if logarithmic scales are used, or if the scale of either axis does not start at zero.

6.8 CHECK-LIST FOR YOUR REPORT

- ✓ Make sure you have tried different types of graphs and selected those that you judge will do the best job at highlighting the important results you obtained and communicating your main idea.
- ✓ Check that the titles of the graph and of the *X* and *Y* axes have been included, including respective units if appropriate (mg/L, m^3/d, %, etc.).
- ✓ Confirm that you have included a legend, in case your graph includes more than one data set, and that the legend is placed in a proper position to clearly identify and distinguish each data set.
- ✓ Check that you have adequately selected the axis scales to best represent your data without confusing or misleading the reader (choose appropriate minimum and maximum values, choose an arithmetic or log scale, whichever is most appropriate).
- ✓ Verify that your data series are well identified by markers and/or lines, and all of them are clearly distinguishable from each other, even if reproduced in black and white.
- ✓ Analyse whether the correct font size is being used and will be readable in your report or presentation.
- ✓ Make sure your axis scale does not show values that have no physical meaning, such as negative concentrations or removal efficiencies greater than 100%.
- ✓ Confirm that the figure containing your chart is self-sufficient and include footnotes, if necessary, so that it can be reproduced and still stand by itself.
- ✓ Verify that you have referenced the figure in the text of your report and have discussed it. Do not make the reader have to interpret it alone, but rather guide the reader through the main elements and take-away messages that you want your graph to convey.
- ✓ Read the suggestions included in Section 6.7 and see whether they can be useful for your report.

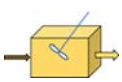
S. 6.7

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Chapter 7

Removal efficiencies

This chapter advances on the topic of descriptive statistics (Chapters 5 and 6), focusing now on removal efficiencies and the specificities associated with their calculation and interpretation. Different ways of presenting removal efficiencies (percentages versus log reduction values) are explained. Specific aspects associated with the determination of removal efficiencies are discussed, such as the influence of water losses, the handling of censored data, and the consideration of minimum and maximum possible values. We provide guidance regarding the interpretation of removal efficiencies, and we emphasize the joint interpretation of removal efficiencies together with effluent concentrations. Different ways of determining measures of central tendency for removal efficiencies are presented and discussed, and typical patterns of the associated frequency distributions are also described.



The contents in this chapter are applicable only to treatment plant monitoring, since the concept of removal efficiencies does not apply to water quality monitoring in water bodies.

CHAPTER CONTENTS

7.1 The Concept of Removal Efficiency **182**
7.2 How to Calculate and Report Removal Efficiencies **182**
7.3 Specific Aspects in the Calculation of Removal Efficiencies **188**
7.4 How to Interpret Values of Removal Efficiency **194**
7.5 The Importance of Analysing Effluent Concentrations and Removal Efficiencies Together **195**
7.6 Measures of Central Tendency for Removal Efficiencies **200**
7.7 Frequency Distribution of Removal Efficiencies **204**
7.8 Check-List for Your Report **206**

© 2020 The Authors. This is an Open Access book chapter distributed under the terms of the Creative Commons Attribution Licence (CC BY-NC-ND 4.0), which permits copying and redistribution for non-commercial purposes with no derivatives, provided the original work is properly cited (https://creativecommons.org/licenses/by-nc-nd/4.0/). This does not affect the rights licensed or assigned from any third party in this book. The chapter is from the book *Assessment of Treatment Plant Performance and Water Quality Data: A Guide for Students, Researchers and Practitioners*, Marcos von Sperling, Matthew E. Verbyla and Sílvia M. A. C. Oliveira (Authors).
doi: 10.2166/9781780409320_0181

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

7.1 THE CONCEPT OF REMOVAL EFFICIENCY

Basic

Throughout this book, we emphasize the importance of using data from influent and effluent sampling points to calculate the removal efficiency of pollutants to help you in your assessment of treatment plant performance. In this chapter, we provide details about how to *calculate*, *report*, and *interpret* removal efficiencies using data from influent and effluent sampling points. The descriptive statistics (numerical and graphical methods) for removal efficiencies will be the same as those described in Chapters 5 and 6 for flows, concentrations, and loads. However, in this chapter we bring some specificities in the calculation and interpretation of removal efficiencies.

C. 5
C. 6

As you read through this chapter, consider the true meaning of the word **removal**. We use 'removal' and 'removal efficiency' most frequently in this book to avoid confusion, because their use is *widespread in other textbooks and in the literature*. However, you should consider that other words may reflect more clearly what actually takes place in a treatment plant. *Most water quality constituents can be produced and removed* in a treatment plant, and sometimes we do not know the individual contribution of each of these factors because we only have measured values of the influent and effluent concentrations. In this case, we can only infer the overall reduction in the concentration of the constituent. We cannot infer anything about the balance between production and removal.

Furthermore, *removal may not be the true mechanism explaining the change in concentrations.* Consider the example of a wastewater treatment plant. If we only sample the liquid influent and the **liquid** effluent of the system (but not the **sludge**), we do not necessarily know if the constituent was removed from the liquid and transferred to the sludge or if it was simply transformed or destroyed. Take nitrogen as an example – without an understanding about the concentrations of nitrogen species in wastewater sludge, we do not necessarily know if the nitrogen was biochemically transformed to N_2 (and released to the atmosphere) or transformed to organic nitrogen and incorporated into the biomass of the sludge. Another example is with pathogens. If the concentration is lower in the liquid effluent than it is in the liquid influent, we do not necessarily know if the pathogens were inactivated (i.e., killed) or if they were simply settled out in the solids.

S. 7.3.4

See Section 7.3.4 for a more detailed discussion on these aspects, taking as examples organic matter, nitrogen, and pathogens.

Rather than using the word removal, a more accurate term to use might be **reduction**, because it reflects better the balance between measured values of input and output, as well as the fact that we do not necessarily know if the constituent was removed to the solids portion, destroyed, or transformed (and released as a gas, for example).

7.2 HOW TO CALCULATE AND REPORT REMOVAL EFFICIENCIES

7.2.1 Expressing removal efficiencies as relative values or percentages

Basic

The removal efficiency of a certain constituent in a treatment plant or treatment unit is frequently reported as the **percent (%) removal efficiency**. The removal efficiency is given by the well-known concept described in Equation 7.1.

$$E = \frac{\text{Input concentration} - \text{Output concentration}}{\text{Input concentration}}$$
$$= \frac{\text{Removed concentration}}{\text{Input concentration}} = 1 - \frac{\text{Output concentration}}{\text{Input concentration}} \quad (7.1)$$
$$= 1 - \text{Remaining fraction}$$

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

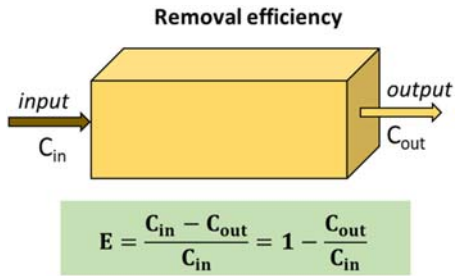


Figure 7.1 Removal efficiencies expressed as relative values.

Converting into our usual nomenclature of input (C_{in}) and output (C_{out}) concentrations, we obtain the removal efficiency, expressed in relative terms (Equation 7.2 and Figure 7.1):

$$E = \frac{C_{in} - C_{out}}{C_{in}} = 1 - \frac{C_{out}}{C_{in}} \tag{7.2}$$

If we want to express removal efficiencies as percentages, we simply multiply the value from Equation 7.2 by 100:

$$E(\%) = \frac{C_{in} - C_{out}}{C_{in}} \times 100 = \left(1 - \frac{C_{out}}{C_{in}}\right) \times 100 \tag{7.3}$$

The units of concentration must be the same for C_{in} and C_{out}, and are those traditional for reporting concentrations (mg/L, g/m^3, μg/L, etc.).

For instance, if you have an influent concentration of 300 mg/L and an effluent concentration of 30 mg/L, the removal efficiency will be $(300 - 30)/300 = 0.90 = 90\%$ or $(1 - 30/300) = 1 - 0.10 = 0.90 = 90\%$.

We also have the concept of the **remaining fraction**, which is given by Equation 7.4:

$$\begin{aligned} \text{Remaining fraction} &= 1 - \text{Removed fraction} \\ &= 1 - E = \frac{C_{out}}{C_{in}} \end{aligned} \tag{7.4}$$

For instance, in the example in the paragraph above, $C_{out}/C_{in} = 30/300 = 0.10 = 10\%$. Therefore, the plant removes 90% ($E = 0.90$) of the constituent, and the remaining percentage is 10% (remaining fraction = 0.10, according to Equation 7.4), meaning that 10% of the constituent has not been removed.

Basic

7.2.2 Expressing removal efficiencies as logarithmic units removed

For some constituents, we express concentrations in terms of their **order of magnitude**, which is a way of comparing their relative size. In engineering, water quality and treatment systems, base 10 logarithmic comparisons are usually always implied, and this is the type of logarithmic comparison used throughout this book. However, you should note that in other natural science fields such as ecology or biology, natural logarithms may be more commonly used. For example, if one number is 10 ($=10^1$) times greater

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

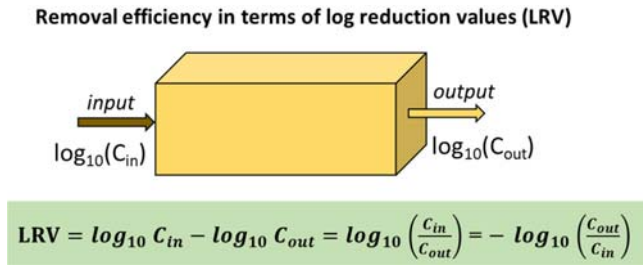


Figure 7.2 Removal efficiencies expressed as $\log_{10}$ reduction values (LRV).

than another number, it is one order of magnitude greater (using the base 10 logarithmic scale). If a number is 1000 ($=10^3$) times greater than another number, it is said to be 3 orders of magnitude greater.

In the case of **pathogens and indicators of faecal contamination** (see von Sperling *et al.,* 2018, from which part of this section is based), the concentrations can be very high and are often log-normally distributed in the system. Thus, *more attention is given to the order of magnitude of the concentrations instead of the absolute values themselves*. For instance, a concentration of 183,098,765 MPN/100 mL is usually expressed as 1.83×10^8 MPN/100 mL, giving more emphasis on the order of magnitude of 10^8 and recognizing that there is not much accuracy on the digits that come after 183.

Given these high numbers, another way of expressing this concentration is by taking the $\log_{10}$ of the original value (this is known as the **$\log_{10}$-transformed concentration**). For instance, a concentration of 1.00×10^8 MPN/100 mL has a log-transformed value of 8.00 (i.e., $\log_{10}(1.00 \times 10^8) = 8.00$). Likewise, for instance, 1.46×10^8 MPN/100 mL has a log-transformed value of 8.16 (i.e., $\log_{10}(1.46 \times 10^8) = 8.16$).

An alternative to expressing reductions as a percentage is to use the **$\log_{10}$ reduction value (LRV)**, which is defined as the difference between the log-transformed concentrations of the influent and effluent across a particular treatment unit or across the whole system. Log reduction values are seldom used for chemical constituents, but the reduction of pathogens and faecal indicators such as *E. coli* and coliforms should almost always be expressed this way (see Equation 7.5 and Figure 7.2):

$$\text{LRV} = \log_{10} C_{\text{in}} - \log_{10} C_{\text{out}} = \log_{10}\left(\frac{C_{\text{in}}}{C_{\text{out}}}\right) = -\log_{10}\left(\frac{C_{\text{out}}}{C_{\text{in}}}\right) \tag{7.5}$$

Advanced

7.2.3 Relationship between removal efficiencies as percentages and log reduction values

The LRV is related to percent reduction, and one can be calculated from the other as shown in Equations 7.6 and 7.7:

$$\mathbf{LRV} = -\log_{10}(1 - E) = -\log_{10}\left(1 - \frac{\boldsymbol{E}(\%)}{100}\right) = \log_{10}\left(\frac{100}{100 - \boldsymbol{E}(\%)}\right) \tag{7.6}$$

$$\boldsymbol{E}(\%) = 100 \times \left(1 - 10^{-\boldsymbol{LRV}}\right) \tag{7.7}$$

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

In Equation 7.6, note that the term $(1 - E)$ corresponds to the remaining fraction (see Equation 7.4). Therefore, *LRV is directly associated with the $\log_{10}$ value of the remaining fraction.*

Example

EXAMPLE 7.1 EXPRESSING REDUCTION EFFICIENCIES AS PERCENTAGE VALUES AND LOG REDUCTION VALUES

In the treatment plant you are studying, you obtained the following E. coli values: influent = 1.00×10^8 MPN/100 mL; effluent concentration = 1.00×10^5 MPN/100 mL. Calculate the reduction efficiencies as percentage and $\log_{10}$ reduction values (LRV).

Solution:

From Equation 7.2, you obtain the reduction efficiency in relative and in percentage values:

$$E = \frac{C_{in} - C_{out}}{C_{in}} = \frac{1.00 \times 10^8 - 1.00 \times 10^5}{1.00 \times 10^8} = \mathbf{0.999} = \mathbf{99.9}\%$$

In order to express the reduction efficiency in terms of LRV, using Equation 7.5, you have three different options:

$\text{LRV} = \log_{10}(1.00 \times 10^8) - \log_{10}(1.00 \times 10^5) = 8.0 - 5.0 = \mathbf{3.0}$

$\text{LRV} = \log_{10}[(1.00 \times 10^8)/(1.00 \times 10^5)] = \log_{10}(1.00 \times 10^3) = \mathbf{3.0}$

$\text{LRV} = -\log_{10}[(1.00 \times 10^5)/(1.00 \times 10^8)] = -\log_{10}(1.00 \times 10^{-3}) = -(-3.0) = \mathbf{3.0}$.

Equations 7.6 and 7.7 can be used to convert $E(\%)$ into LRV and vice versa:

$$\mathbf{LRV} = -\log_{10}\left(1 - \frac{\boldsymbol{E}(\%)}{100}\right) = -\log_{10}\left(1 - \frac{99.9}{100}\right) = -(-3.0) = \mathbf{3.0}$$

$$\boldsymbol{E}(\%) = 100 \times (1 - 10^{-\mathbf{LRV}}) = 100 \times (1 - 10^{-3}) = 100 \times 0.999 = \mathbf{99.9}\%$$

Using Equations 7.6 and 7.7, you can see that *an efficiency of 90% corresponds to an LRV of 1 $\log_{10}$ unit; 99% → 2 $\log_{10}$ units; 99.9% → 3 $\log_{10}$ units; 99.99% → 4 $\log_{10}$ units; 99.999% → 5 $\log_{10}$ units, and so on.* This relationship between percent reduction efficiency (E) and $\log_{10}$ reduction value (LRV) is shown in Table 7.1 and Figure 7.3. You may find strange a reduction efficiency of LRV of 12 log units, but in California, USA, a 12-log reduction of viruses is required for potable wastewater reuse! From the figure, you can see that it is difficult to visualize removal efficiencies in a graph if the values of the removal efficiency are above 99% (2 $\log_{10}$ units). Again, this is one of the reasons why we adopt the concept of LRV when dealing with very high removal efficiencies, as is often the case with pathogens and faecal indicators.

In this book and in most of the literature, *pathogen and coliform reduction efficiencies are generally expressed as LRVs, using $\log_{10}$ units.* This is because in some cases, pathogens and coliforms in wastewater must be reduced by six or more orders of magnitude for the treated effluent or sludge to be safely reused, for example in unrestricted agriculture (WHO, 2006). As another example, for indirect potable water reuse systems in California, the overall LRV for pathogens needs to be as high as 10–12 log units, which is equivalent to 99.99999999–99.9999999999% removal! Another reason for using LRV instead of percent reduction in this context is because it is cumbersome to refer to reduction as 99.9999%; it is much easier to say '6-log' reduction. Note that the term 'log' here implies a *base 10*

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Table 7.1 Relationship between equivalent percent reduction efficiency (%) and $\log_{10}$ reduction values (LRV).

Percent Reduction Efficiency (%)	Equivalent LRV	Percent Reduction Efficiency (%)	Equivalent LRV
1	0.004	99.99	4
10	0.05	99.999	5
25	0.1	99.9999	6
50	0.3	99.99999	7
75	0.6	99.999999	8
90	1.0	99.9999999	9
95	1.3	99.99999999	10
99	2	99.999999999	11
99.9	3	99.9999999999	12

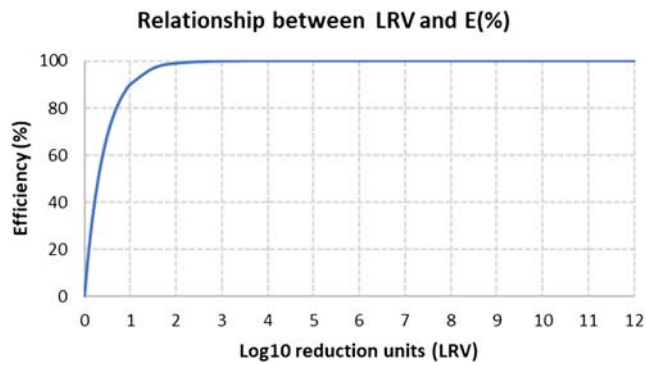


Figure 7.3 Relationship between LRV and efficiency (%).

logarithm (*log_{10}), even if the subscript 10 does not necessarily appear after* '*log*'. Pathogen reduction is almost never described in terms of natural logarithms, but if the natural logarithm is used in this context, it is denoted by the notation 'LN' (von Sperling *et al.*, 2018).

It is *possible to have an LRV that is greater than the order of magnitude of the pathogen concentration in the influent*. For instance, a pathogen that has an influent concentration of 1.00×10^5 CFU/100 mL can be subjected to a treatment with an LRV of, say, 7. Rearranging Equation 7.5, you can see that this will lead to $\log_{10}(10^5) - 7 = 5 - 7 = -2$. The effluent concentration will then be $10^{(-2)} = 0.01$ CFU/100 mL $= 1$ CFU per mL. Of course, in this case, you must check whether this value is below the detection limits of the lab method used to enumerate the pathogen in question (considering the dilutions made to the sample prior to analysis and whether or not the sample was concentrated from a larger volume).

Advanced

7.2.4 Removal efficiencies (% and LRV) for units in series

For treatment units placed in series, the ***overall percentage removal efficiency*** *of the combined treatment units is based on the multiplication of the* ***remaining fractions*** *of the constituent in each unit.*

$$E_{\text{overall}} = 1 - [(1 - E_1) \times (1 - E_2) \times \cdots \times (1 - E_n)] \tag{7.8}$$

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

or, expressing removal as percentages:

$$E_{\text{overall}}(\%) = 100 \times \left(1 - \left[\left(1 - \frac{E_1(\%)}{100}\right) \times \left(1 - \frac{E_2(\%)}{100}\right) \times \cdots \times \left(1 - \frac{E_n(\%)}{100}\right)\right]\right) \tag{7.9}$$

For instance, in a complete treatment system there may be three process units placed in series, with the following reduction efficiencies: Unit 1 = 90%, Unit 2 = 99.9%, and Unit 3 = 99%. In this situation, the overall reduction efficiency will be, according to Equations 7.8 and 7.9:

$$E_{\text{overall}} = 1 - [(1 - 0.90) \times (1 - 0.999) \times (1 - 0.99)] = \mathbf{0.999999}$$

or

$$E_{\text{overall}}(\%) = 100 \times \left(1 - \left[\left(1 - \frac{90}{100}\right) \times \left(1 - \frac{99.9}{100}\right) \times \left(1 - \frac{99}{100}\right)\right]\right) = \mathbf{99.9999}\%$$

If we use Equation 7.6, we can see that 99.9999% corresponds to 6 log-units reduction (LRV = 6).

However, we can also express the reduction efficiencies in terms of LRV. In this case, the relationship between the units in series is **additive** *in terms of their individual LRV values*, and much easier to calculate:

$$\boxed{\text{LRV}_{\text{overall}} = \text{LRV}_1 + \text{LRV}_2 + \cdots + \text{LRV}_n} \tag{7.10}$$

In the example above, the LRV values in each unit are Unit 1 = 1 log, Unit 2 = 3 log, and Unit 3 = 2 log. Therefore, the overall efficiency expressed in LRV can be simply calculated as the sum of each individual LRV:

$$\text{LRV}_{\text{overall}} = 1 + 3 + 2 = 6 \log_{10} \text{unit reduction}$$

From Equation 7.7, we know that 6 log-units reduction is equivalent to 99.9999%.

Figure 7.4 summarizes the equations for calculating removal efficiencies in units in series.

Please note that *we are not advocating here that removal efficiencies remain the same, regardless of the treatment that takes place upstream of the unit*. For instance, if we are dealing with organic matter in a biological treatment system, the removal efficiency of the biodegradable fraction of organic matter is affected by the possible existence of treatment ahead, which will reduce the biodegradability of organic

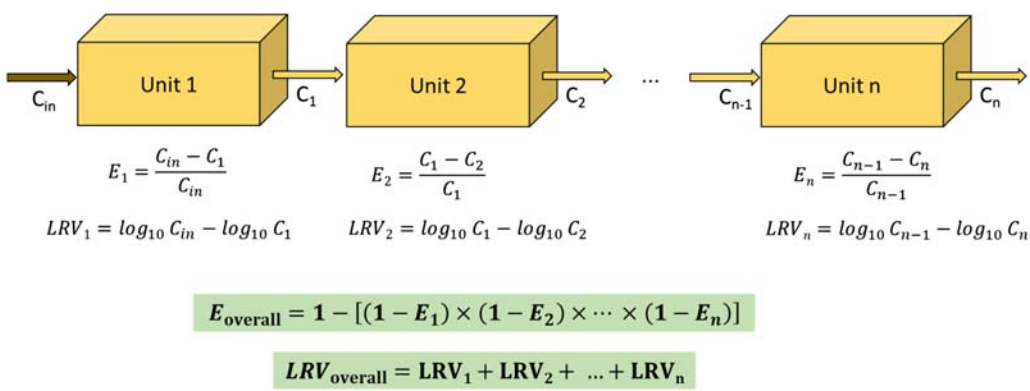


Figure 7.4 Calculations involved in the determination of reduction efficiencies in units in series. Note: LRV = log reduction value.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

matter. For instance, an aeration tank of an activated sludge system that receives raw sewage will receive a more biodegradable organic matter compared with one aeration tank that receives UASB reactor effluent, and the activated sludge stage will probably have a higher removal efficiency compared with the latter option (for the same loading conditions). Therefore, please understand that *our objective here is just to present the math for how to calculate the overall removal efficiency of a system if you have removal efficiencies from each of the individual units*.

7.3 SPECIFIC ASPECTS IN THE CALCULATION OF REMOVAL EFFICIENCIES

Advanced

7.3.1 The influence of water losses on the calculation of removal efficiencies

S. 10.2

In Section 10.2 on water and mass balances, Example 10.3 illustrates what should be done when you have a situation in which there are **substantial water losses in your tank or reactor**, due to evaporation and evapotranspiration. This may occur in treatment units such as *constructed wetlands, overland flow systems, and maturation ponds*. If water is simply lost to the atmosphere, this lost water is pure (void of the constituent), and the constituents in the effluent thus become more concentrated by simple water loss. This will result in an apparent reduction in the actual removal that took place in the unit.

In this case, we need to compute **removal efficiencies in terms of removed loads** (flow × concentration), and *not of removed concentrations*, as shown before in this chapter. In the previous calculations we have done here we assumed that the flow did not change in the system, that is, the outflow was equal to the inflow. However, if this is not the case, we need to calculate removal efficiency by

$$E = \frac{\text{Load}_{in} - \text{Load}_{out}}{\text{Load}_{in}} = \frac{Q_{in} \cdot C_{in} - Q_{out} \cdot C_{out}}{Q_{in} \cdot C_{in}} \tag{7.11}$$

Example 11.3, in the chapter on water and mass balances, illustrates these computations.

A similar reasoning applies to a **sludge** subject to water removal in the stages of *thickening* and *dewatering*. In these steps of sludge treatment, water is removed, but the mass of pollutants may remain incorporated in the solids. If we interpret values in terms of the usual concentrations reported as mass or count per unit volume (e.g., g/m^3, MPN/100 mL), we will be led to think that the pollutant concentration increased, while the only thing that may have happened is that the sludge volume diminished. Therefore, when we deal with sludges, it is more convenient to report values in terms of **grams of total solids (TS)**, which are understood as the mass of dry solids, because these will not be affected by the removal of water and the reduction of volume during thickening and dewatering processes. Example of such units may be g of pollutant per g of TS, MPN of coliforms per g TS, etc.

Advanced

7.3.2 The influence of censored data on the calculation of removal efficiencies

S. 5.4

In Section 5.4, we presented the concept of censored data, and analysed different ways of handling this situation. You should review this topic again to understand what we will cover in the section below.

In Section 5.4, we saw how to calculate summary statistics for a censored data set, and we will now analyse the influence of censored data on the calculation of removal efficiencies. As we saw in the previous sections, removal efficiencies are calculated from the influent and effluent data sets, both of which can potentially be censored. *Left-censored data* (*values below the detection limit*) *are likely to occur more in effluent concentrations* (because they have lower values), while *right-censored data*

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

(*values above the upper detection limit*), if they occur, will probably be at a *higher frequency in the influent concentrations* (because they have higher values).

S. 5.4

In Section 5.4, we saw the following *substitution techniques* for left-censored data:

- Eliminating non-detects from the data set (not recommended)
- Substituting the non-detects with a value of zero
- Substituting the non-detects with a value equal to the detection limit
- Substituting the non-detects with a value equal to a fraction of the detection limit
- Using the maximum likelihood estimation (MLE) method to estimate the mean and standard deviation of a censored data set

In Example 5.2, we illustrated these five procedures with a focus on the concentration of the constituent at a single sample point. Now, we will revisit this example, analysing the impact of the different substitution techniques on the **calculation of removal efficiencies** using the same effluent concentrations as in the previous example, and the following characteristic influent concentrations: (a) low influent concentrations, close to the lower detection limit and (b) high influent concentrations, much greater than the lower detection limit.

For *left-censored data* (values below the lower detection limit) of *effluent concentrations*, their impact on the calculated values of removal efficiencies will be different:

- If ***influent concentrations are low*** (e.g., only slightly above the detection limit), the impact of censored effluent concentration data on the calculated removal efficiency is likely to be **higher**.
- If ***influent concentrations are high*** (e.g., much higher than the detection limit), the impact of censored effluent concentration data on the calculated removal efficiency is likely to be **lower**.

Of course, other factors may influence the above results, including the percentage of non-detects in the data set and the values of the non-censored data points.

Example

EXAMPLE 7.2 WORKING WITH CENSORED DATA IN THE CALCULATION OF REMOVAL EFFICIENCIES

You analysed censored data relative to the concentration of a certain constituent in the effluent from a treatment plant in Example 5.2. In this example, the value of the method detection limit (MDL) was 0.10 mg/L. Using different techniques for handling censored data, in Example 5.2 you obtained the following values for the arithmetic mean of the effluent concentration (C_{out}):

Technique for Handling Left-censored Data	Resulting Mean Effluent Concentration C_{out} (mg/L)
Exclusion of <MDL values (mg/L)	0.16
<MDL values substituted by zero (mg/L)	0.11
<MDL values substituted by MDL (mg/L)	0.14
<MDL values substituted by MDL/2 (mg/L)	0.12
Maximum Likelihood Estimation (MLE) Method	0.13

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Now calculate the removal efficiency of the constituent based on the mean concentrations, assuming two different scenarios:

- *Low mean influent concentration (0.20 mg/L), only slightly above the MDL of 0.10 mg/L*
- *High mean influent concentration (5.0 mg/L), much greater than the MDL of 0.10 mg/L*

Solution:

(a) First scenario: low influent concentration (mean of 0.20 mg/L)

We set up the following computational table. The removal efficiency is calculated based on the mean influent concentration (0.20 mg/L) and the mean effluent concentration (calculated using the five different techniques for handling censored data), using Equation 7.3.

Technique for Left-censored Data	Mean Influent Concentration C_{in}	Mean Effluent Concentration C_{out}	Mean Removal Efficiency (%)
Exclusion of below detection limit (BDL) values (mg/L)	0.20	0.16	**21**
BDL values substituted by zero (mg/L)	0.20	0.11	**48**
BDL values substituted by MDL (mg/L)	0.20	0.14	**31**
BDL values substituted by MDL/2 (mg/L)	0.20	0.12	**39**
MLE method	0.20	0.13	**35**

We see that the value of mean removal efficiency varies widely among the five methods, from 21% to 48%. Therefore, the procedure for handling censored data is influential in this case that C_{in} is close to the detection limit.

(b) Second scenario: high influent concentrations (mean of 5.0 mg/L)

We set up the same type of computational table, with the difference that now the removal efficiency is calculated based on the mean influent concentration of 5.0 mg/L.

Technique for Left-censored Data	Mean Influent Concentration C_{in}	Mean Effluent Concentration C_{out}	Mean Removal Efficiency (%)
Exclusion of BDL values (mg/L)	5.0	0.16	**97**
BDL values substituted by zero (mg/L)	5.0	0.11	**98**
BDL values substituted by MDL (mg/L)	5.0	0.14	**97**
BDL values substituted by MDL/2 (mg/L)	5.0	0.12	**98**
MLE method	5.0	0.13	**97**

We can now see that the value of mean removal is very similar for the five methods, from 97% to 98%. Therefore, the procedure for handling censored data is not as influential in this case when C_{in} is much above the detection limit.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Advanced

7.3.3 Minimum and maximum possible values of removal efficiencies

Consider our basic equation for calculating removal efficiencies, expressed as percentages (Equation 7.3):

$$E(\%) = \frac{C_{in} - C_{out}}{C_{in}} \times 100$$

From this equation, we have the following situations for **removal efficiencies expressed as %**:

- If $C_{in} = 0 \rightarrow E(\%) =$ error (we cannot have zero in the denominator of the equation)
- If $C_{out} = 0 \rightarrow E(\%) = 100$
- If $C_{out} < C_{in} \rightarrow E(\%) > 0$
- If $C_{out} = C_{in} \rightarrow E(\%) = 0$
- If $C_{out} > C_{in} \rightarrow E(\%) < 0$

A similar analysis can be done for the basic equation for calculating LRV (Equation 7.5):

$$LRV = \log_{10} C_{in} - \log_{10} C_{out}$$

From this equation, we have the following situations for **removal efficiencies expressed as LRV**:

- If $C_{in} = 0 \rightarrow$ LRV = error (we cannot calculate $\log_{10}$ of zero)
- If $C_{out} = 0 \rightarrow$ LRV = error (we cannot calculate $\log_{10}$ of zero)
- If $C_{out} < C_{in} \rightarrow$ LRV > 0
- If $C_{out} = C_{in} \rightarrow$ LRV $= 0$
- If $C_{out} > C_{in} \rightarrow$ LRV < 0

In conclusion, *mathematically speaking*, we can say that:

- It is possible to have negative removal ($E\% < 0$ and LRV < 0); this is the case when you have an increase in concentration due to growth or production
- There is no minimum limiting value for $E\%$ and LRV
- The maximum limiting value for $E\%$ is 100% (when C_{out} is equal to zero)
- There is no maximum limiting value for LRV (but note that C_{out} cannot be zero)

If you think that *negative removal efficiencies are a strange concept*, they indeed occur in treatment plants. *You do not expect that you would have **mean** negative values for removal efficiencies* for the majority of constituents, because you would come to the frustrating conclusion that your treatment plant is doing a worse job than a non-existing plant. But if you give a detailed analysis on the *individual monitoring data* in your records, you may find days in which, for instance, the effluent concentration was higher than the influent concentration. Maybe there were episodes of solids loss in the final clarifier, which could have caused an increase in effluent SS and particulate COD. Another example would be for some nitrogen species, such as nitrate in a wastewater treatment plant. You might actually expect negative removal (i.e., an increase) for the concentration of nitrate in an aeration basin, as illustrated in Section 7.3.4. Other similar situations, with different explanations, can take place for other constituents.

S. 7.3.4

Advanced

7.3.4 Differences between removal and reduction

S. 7.3.1

As mentioned in Section 7.3.1, we may use both expressions (removal and reduction) in this book, but we would like to point out that there are fundamental underlying differences in the two terms. When we obtain values of influent and effluent concentrations, we are able to make inferences only on the overall **reduction**

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

in the concentration between the inlet and the outlet. However, unless we undertake specific studies, we cannot say if there were simultaneous **removal** and **production** of the constituent in the treatment plant, and whether the reduction we calculated is simply a result of the combined effect of the factors in the mass balance (production: positive term; removal: negative term). In a broad mass balance of constituents in a treatment plant, we might think about the concept of **conversion**: some constituents may be converted into others, and, while they are removed (or consumed), they may be converted into another constituent which would be produced.

We will illustrate these points with three examples: organic matter, nitrogen, and pathogens. Other examples could have been cited, and *it is up to you, based on your knowledge of the treatment processes, to decide for which constituents it is appropriate to calculate removal efficiency and for which constituents this type of calculation is not appropriate.*

(a) Organic matter

In a biological wastewater treatment plant, we have **conversion** of organic matter (BOD or COD) into water, gases, and new cell material (biological cells). The organic matter from the influent that has been converted is considered to be **removed**. However, the new biological cells that are **produced** in the reactor are, themselves, organic matter. In order to analyse this balance between consumption and production, we would need to resort to using mathematical models of the treatment process (which is outside of the scope of this book).

Although organic matter may have been converted efficiently in the biological reactor, if we have solids loss in the effluent from the subsequent solids-separation unit (secondary clarifiers), we will increase the organic matter concentration again (in the form of particulate BOD and COD). Usually, the major episodes of deterioration of effluent quality in a treatment plant are associated with suspended solids loss, which can mask the possible good conversion that may have taken place in the biological reactor. In other words, *organic matter conversion may have been very high, but the introduction of particulate matter in the effluent from the secondary sedimentation tank will decrease the calculated value of removal efficiency*.

Because of this, for this particular case, it is useful to calculate what is known as the **biological removal efficiency** in addition to calculating the removal efficiency in the traditional way (i.e., using Equations 7.2 and 7.3). This calculation of biological removal efficiency applies to the actual conversion that takes place in the biological reactor:

$$E_{\text{biological}} = \frac{\text{Influent } \mathbf{total} \text{ COD} - \text{Effluent } \mathbf{soluble} \text{ COD}}{\text{Influent } \mathbf{total} \text{ COD}} \tag{7.12}$$

where:

Influent total COD = usual (total) COD measured in the influent (mg/L)

Effluent soluble COD = soluble or filtered COD, reflecting the remaining fraction of influent COD, and excluding the COD associated with suspended solids (particulate COD) (mg/L)

Equation 7.12 can be used for COD or BOD; the concept is the same.

(b) Nitrogen

Nitrogen undergoes conversions according to its biogeochemical cycle, and part of this cycle takes place in treatment plants.

Organic nitrogen is converted into **ammonia** by the process of *ammonification*, which takes place under normal operating conditions in wastewater treatment plants and causes ammonia concentrations to initially increase. Organic nitrogen is thus 'removed' at the expense of its

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

conversion into ammonia. If we were to compute ammonia removal in treatment units in which only ammonification is taking place, the effluent concentration will be higher than the influent one, which leads to a 'negative' removal efficiency for ammonia.

Ammonia may then be converted into **nitrite**, and nitrite may be converted into **nitrate**, in the process of *nitrification*, which takes place in treatment plants that are capable of supporting it. Thus, ammonia may simultaneously be produced (via ammonification) and consumed (via nitrification). The calculation of the 'removal efficiency' for ammonia will be affected by this, and if we simply take into account the concentrations of influent ammonia and effluent ammonia, we will not be able to say what has effectively been 'removed' and what has been 'produced'. A similar comment can be made for nitrite: it is both produced (ammonia → nitrite) and consumed (nitrite → nitrate), and therefore the expression 'removal' does not seem appropriate. Finally, nitrate may also undergo a similar fate: it may be produced via nitrification (nitrite→ nitrate) and it may also be consumed (nitrate → nitrogen gas) via a process called *denitrification*. Again, the term 'removal' may not necessarily be the most suitable for this situation.

Let us analyse part of the N cycle to expand our comments. Nitrate is usually not present in the influent to a wastewater treatment plant but, in the process of nitrification, it can be formed in the biological reactor. Therefore, nitrate may change from a negligible concentration in the influent into a higher value in the effluent (let us not consider denitrification here). If we apply Equation 7.3, we will obtain negative values of the removal efficiencies, since $C_{out} > C_{in}$.

Ammonification and nitrification do not indicate any problem with the treatment process, but rather they are desired processes that should be taking place in the treatment system. Because we know this, *we should not say that the removal efficiency is negative*. Indeed, we can conclude that there is *no interest in doing this calculation* for these constituents. In this case, it is better that we *use the expression* ***conversion*** *instead of* ***removal***: conversion of organic nitrogen into ammonia, conversion of ammonia into nitrite, conversion of nitrite into nitrate, etc.

S. 4.6

S. 5.4

Also, we could employ the term **production** (e.g., production of nitrate) because it more accurately describes what is taking place. Note that we cannot use Equation 7.3 if we have a zero concentration in the influent, because we will get an 'error' message. Also, given your knowledge about method detection limits and censored data sets (Sections 4.6 and 5.4), you should *not be reporting concentrations of zero*. Instead you should report that the concentration is below the method detection limit (and of course, you should also report the value of the method detection limit).

Instead of mentioning the efficiency of removal or conversion, we can also specify the **efficiency of the processes** involved (in the nitrogen cycle, *nitrification* and *denitrification*, if we exclude other processes):

$$\text{Efficiency nitrification} = \frac{(\text{TKN}_{in} - \text{TKN}_{out})}{\text{TKN}_{in}} = 1 - \frac{\text{TKN}_{out}}{\text{TKN}_{in}} \tag{7.13}$$

$$\text{Efficiency denitrification} = 1 - \frac{\text{NOx}_{out}}{\text{NOx}_{produced}} = 1 - \frac{(\text{Nitrite}_{out} + \text{Nitrate}_{out})}{(\text{TKN}_{in} - \text{TKN}_{out})} \tag{7.14}$$

where:

TKN = total Kjeldahl nitrogen (=organic N + ammonia N) (mg/L)
NOx = oxidized forms of nitrogen (=nitrite N + nitrate N) (mg/L)

In Equation 7.13, we use TKN instead of ammonia, because we know that most of the organic nitrogen will eventually become ammonia. In Equation 7.14, we do not know how much NOx

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

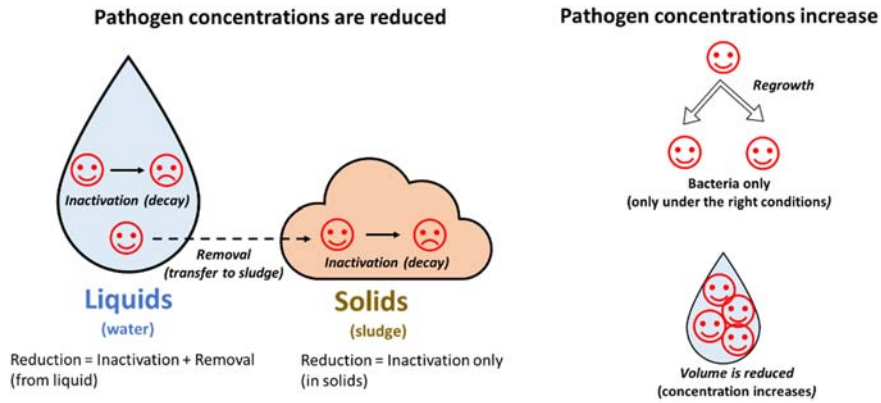


Figure 7.5 Dynamics associated with the increase or reduction of pathogen concentrations in water and solids.

will be produced by simply measuring influent and effluent concentrations, because nitrate can be produced (nitrification) and also removed (denitrification). Therefore, we estimate that NOx production will be equal to TKN removal (nitrification). This calculation assumes that denitrification is the main process associated with nitrogen removal.

(c) Pathogens and coliforms

There are several important terms that are used to describe changes in the measured concentrations of pathogens and coliforms along with a treatment plant (Figure 7.5). **Removal** refers to the physical elimination of pathogens from water or wastewater. Often, pathogens removed are simply transferred to sludge or sediments, where they may still remain viable. **Inactivation** or decay refers to the physical destruction of pathogens resulting in a loss of viability – this can happen to pathogens in water, wastewater, or in sludge. **Regrowth** refers to the replication of pathogens in the treatment system. Some opportunistic, zoonotic, and bacterial pathogens may be capable of regrowth within treatment systems (Jjemba *et al.*, 2010), but parasites and enteric viruses require a human host to replicate and cannot regrow within treatment systems (von Sperling *et al.*, 2018). Like any constituent, pathogen concentrations may also increase as a result of the reduction of volume.

In the case of pathogens and coliforms, the term **reduction** seems more appropriate, because it refers to the combined *removal* and *inactivation* in water and wastewater systems.

7.4 HOW TO INTERPRET VALUES OF REMOVAL EFFICIENCY

Basic

In the preceding sections of this chapter we have presented and discussed the concept of removal efficiency. Now we would like to expand this interpretation with the following polemical statement you may have heard from a colleague:

> *In my system of facultative and maturation ponds, I have a **good** BOD removal of **80%** and a **poor** coliform removal of **99%**.*

Suppose other colleagues present at the meeting laughed at this apparent counter sense (good = 80%; poor = 99%?), but you, based on your knowledge of wastewater treatment systems nodded in agreement. The reason for your posture was that you know that sewage treatment with ponds has BOD removal efficiencies ranging from 75% to 85% (von Sperling, 2005), and thus the treatment plant of your colleague

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

was performing as expected, as far as BOD is concerned. But you also know that systems with maturation ponds are expected to reduce more than 99.9% (3 log-units) of coliform concentrations, and therefore this treatment plant, with only 99% (2 log-units) was underperforming. Based on this, you told your colleagues:

> The interpretation of what is a **good** or **poor** removal efficiency depends on the expectations based on the *capacity of your treatment plant in terms of the processes* it utilizes.

After this, during a discussion on compliance with discharge standards, another colleague stated that:

In my system, I have sufficient nitrogen removal of 60% but insufficient COD removal of 85%.

Your colleagues looked at you, because they knew you would have a reasonable explanation for this (sufficient = 60%; insufficient = 85%?). You then stated:

> The interpretation of what is a **sufficient** or an **insufficient** removal efficiency depends on *specified legal standards or desired target values established for your treatment plant*.

In this case, the local legislation specified a minimum total nitrogen removal efficiency of 50%, which is lower than that plant was providing. On the other hand, the minimum required removal efficiency for COD was 90%, but this treatment plant was not in compliance with that level.

The moral of this story is that presenting the removal efficiency alone, without any context, does not tell the full story of the treatment plant performance. In your report, you should also mention what the expected removal efficiency is for the type of system you are studying, as well as any regulatory guidelines or performance standards that are relevant for the system in question. Furthermore, removal efficiencies should be presented together with effluent concentrations in order to gain a more complete understanding about the performance of the system (see Section 7.5).

S. 7.5

7.5 THE IMPORTANCE OF ANALYSING EFFLUENT CONCENTRATIONS AND REMOVAL EFFICIENCIES TOGETHER

Basic

When evaluating the performance of your treatment plant, we strongly recommend that you analyse the **removal efficiencies together with the effluent concentrations**. These two pieces of information complement each other and together they tell a more complete story about the performance of the system being studied.

Another good reason for analysing removal efficiencies together with effluent concentrations is that some countries and regions specify wastewater **discharge standards** in terms of both the *effluent concentration* (maximum allowable value) and the *removal efficiency* (minimum allowable value).

(a) Example in the comparison of different treatment plants

We have emphasized the importance of the joint analysis of effluent concentrations and removal efficiencies in other parts of this book, but we will try to make it clearer with Example 7.3. In this example, we have created two different scenarios: (*a*) *a plant with high influent concentrations, high effluent concentrations, and high removal efficiencies* and (*b*) *a plant with low influent concentrations, low effluent concentrations, and low removal efficiencies*. We will then interpret on what we judge to be a good performance: a **low effluent concentration** or a **high removal**

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

efficiency? Apart from the more direct interpretation of legal requirements for the effluent concentrations and removal efficiencies (if they exist), there will be no single answer for this question. Ultimately, it requires your judgement, together with your knowledge of the treatment process, to decide whether the performance of the system can be considered 'good' or 'bad'.

Example

EXAMPLE 7.3 INTERPRETING TOGETHER EFFLUENT CONCENTRATIONS AND REMOVAL EFFICIENCIES

You are studying the performance of two wastewater treatment plants and would like to decide if their performance is 'good' or 'bad'. The two plants have different characteristics. Plant A has high influent concentrations, high effluent concentrations, and high removal efficiencies. Plant B has low influent concentrations, low effluent concentrations, and low removal efficiencies. Based on monitoring data that you collected over 10 days in each plant, make your judgement.

Excel

Note: This example is also available as an Excel spreadsheet.

Data:

	Plant A				Plant B		
Day	**C_{in} (mg/L)**	**C_{out} (mg/L)**	**Efficiency (%)**	**Day**	**C_{in} (mg/L)**	**C_{out} (mg/L)**	**Efficiency (%)**
1	1000	100	90.0	1	100	40	60.0
2	980	85	91.3	2	98	34	65.3
3	1120	88	92.1	3	112	35	68.8
4	1090	79	92.8	4	109	32	70.6
5	1030	83	91.9	5	103	33	68.0
6	970	87	91.0	6	97	35	63.9
7	1010	88	91.3	7	101	35	65.3
8	1050	92	91.2	8	105	37	64.8
9	950	86	90.9	9	95	34	64.2
10	930	91	90.2	10	93	36	61.3

Solution:

If we take the mean of the influent and effluent concentrations and removal efficiency values from both plants, we will obtain the following results:

Mean values of concentrations and removal efficiencies

Plant A			Plant B		
C_{in} (mg/L)	**C_{out} (mg/L)**	**Efficiency (%)**	**C_{in} (mg/L)**	**C_{out} (mg/L)**	**Efficiency (%)**
1013	88	91.3%	101	35	65.2%
	Bad?	***Good?***		***Good?***	***Bad?***

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Note the questions we have put in the last line of the table. These will be discussed here:

- Plant A has a higher effluent concentration (88 mg/L) compared with Plant B (35 mg/L). Can we say that the effluent concentration from Plant A is **bad** and from Plant B is **good**?
- Plant A has a higher removal efficiency (91.3%) compared with Plant B (65.2%). Can we say that the removal efficiency from Plant A is **good** and from Plant B is **bad**?

Apparently, we have contradictory results, but you have to look at the broader picture, and analyse the *influent concentrations* as well. In Plant A, these are much higher than in Plant B. Plant A is doing a good job, with a high removal efficiency but, because the influent concentrations are very high, the effluent concentrations are still somewhat high. With Plant B we have the opposite. The influent concentrations are low and even with low removal efficiencies, the effluent concentrations are still lower than those of Plant A.

It is up to you, based on your knowledge of the system, to interpret this with respect to the treatment objectives of these two plants and the possible requirements of the legislation in terms of effluent concentrations and removal efficiencies. There is no single correct answer to this example problem, based on the information given. If these plants were operating in a jurisdiction that specified maximum allowable effluent concentrations and minimum required removal efficiencies, then we could make a more definitive assessment about whether or not the two systems are in compliance.

(b) Example in the comparison of different operational periods

We can make a similar analysis for your treatment plant, comparing the results between different operational periods, for instance, between **summer** and **winter**. You know that higher temperatures favour more efficient chemical conversion processes, and therefore *you expect removal efficiencies to be higher in the summer than in the winter*. However, you also know that rainfall in your region is concentrated in the summer, and the intrusion of rain water in the combined sewerage system dilutes the influent concentrations. This is the background of Example 7.4, which illustrates the comparison of the plant performance during different periods of the year, based on an interpretation of its removal efficiencies, effluent concentrations, and removed loads.

Example

EXAMPLE 7.4 INTERPRETING TOGETHER EFFLUENT CONCENTRATIONS, REMOVAL EFFICIENCIES, AND REMOVED LOADS IN DIFFERENT OPERATIONAL PERIODS

You obtained the following *mean* values of influent and effluent concentrations of the constituent you are analysing in your treatment plant: (a) *summer*: $C_{in} = 25$ mg/L, $C_{out} = 10$ mg/L and (b) *winter*: $C_{in} = 45$ mg/L, $C_{out} = 15$ mg/L. Assume the following mean flow rates: *mean* influent flow during the *summer* (*rainy period*) $= 1000$ m^3/d and *mean* influent flow during the *winter* (*dry period*) $= 400$ m^3/d. Interpret the results.

Solution:

Based on the mean influent and effluent concentrations, you calculate the mean removal efficiencies:

- *Summer*: $E = (C_{in} - C_{out})/C_{in} = (25 - 10)/25 = 0.60 = \mathbf{60\%}$;
- *Winter*: $E = (C_{in} - C_{out})/C_{in} = (45 - 15)/45 = 0.67 = \mathbf{67\%}$.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Initially, a less-experienced researcher might hastily conclude that the treatment plant is doing the opposite of what could be expected, with poor removal efficiency occurring during periods of higher temperatures. But you know that you cannot judge based on removal efficiencies alone, that you should also consider effluent concentrations: *in the summer, the mean effluent concentration* (*10 mg/L*) *was lower than it was in the winter* (*15 mg/L*), as expected. Therefore, the lower removal efficiency in summer did not necessarily affect the effluent concentration but may rather be due to the fact that the influent concentration during that period was lower, leading to a lower removal efficiency.

S. 7.3.1

You could also expand the analysis of the paragraph above by studying the **removal in terms of loads** instead of concentrations, as shown in Section 7.3.1. For this, you obtain the following calculated mean loads and removal efficiencies:

Summer:

- Influent load $= Q_{in} \times C_{in} = (1000\ m^3/d) \times (25\ g/m^3) = 25{,}000\ g/d$
- Effluent load $= Q_{out} \times C_{out} = (1000\ m^3/d) \times (10\ g/m^3) = 10{,}000\ g/d$
- Efficiency $= (Load_{in} - Load_{out})/Load_{in} = (25{,}000 - 10{,}000)/25{,}000 = 0.60 =$ **60%**;
- Removed load $= Load_{in} - Load_{out} = 25{,}000 - 10{,}000 =$ **15,000 g/d**

Winter:

- Influent load $= Q_{in} \times C_{in} = (400\ m^3/d) \times (45\ g/m^3) = 18{,}000\ g/d$
- Effluent load $= Q_{out} \times C_{out} = (400\ m^3/d) \times (15\ g/m^3) = 6000\ g/d$
- Efficiency $= (Load_{in} - Load_{out})/Load_{in} = (18{,}000 - 6000)/18{,}000 = 0.67 =$ **67%**
- Removed load $= Load_{in} - Load_{out} = 18{,}000 - 6000 =$ **12,000 g/d**

You now see that even though the calculation of *removal efficiencies* using loads in this example led to the same values as with the concentrations (because there were no water losses inside the plant), the *load removed was higher in the summer*, as you would have expected (15,000 g/d in summer, compared with 12,000 g/d in winter).

Now you can see that by digging deeper into the analysis of your plant's performance, you are able to see that your results make more sense.

(c) Example of variations in influent concentrations

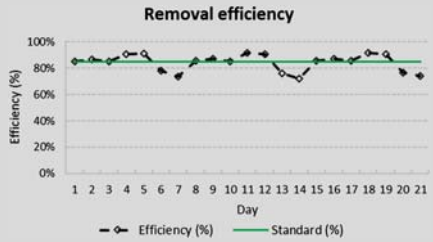
Now let us analyse another example, in which the *influent concentration of a certain constituent varies widely*, while the *effluent concentration remains stable*. This could be the situation, for instance, of the treatment plant of an industry that operates only some activities on weekends (low influent concentrations to the treatment plant) and has heavy processing of materials on some days of the week (high influent concentrations to the treatment plant). *Although the effluent concentrations remain low and stable, the variation of the influent concentrations affects the results of removal efficiency*. See Example 7.5.

Example

EXAMPLE 7.5 INTERPRETING REMOVAL EFFICIENCY INFLUENCED BY VARYING INFLUENT CONCENTRATIONS

You receive data from the last three weeks for an industrial wastewater treatment plant that you are supervising. The data comprise a time series graph of the removal efficiencies of a certain constituent (*see below*).

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest



You know that the legislation specifies a standard of minimum removal efficiency of 85% (*shown as a horizontal line in the chart*) *and you were concerned, because you noticed that 6 out of the 21 values of removal efficiency* (*29%*) *were not in compliance with the standard. You decide to investigate the situation and make the necessary clarifications to the environmental agency.*

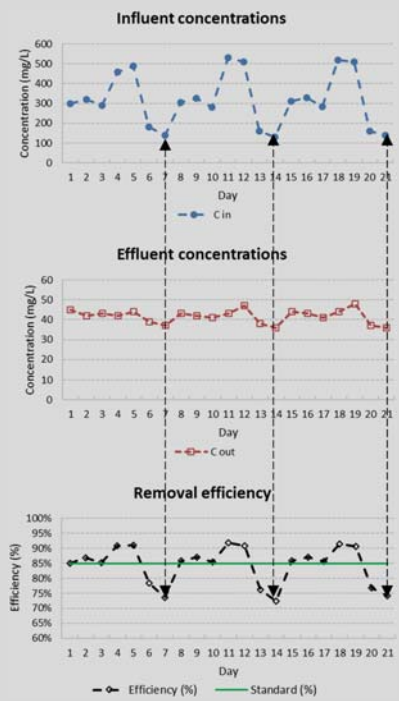
Excel

Note: this example is also available as an Excel spreadsheet.

Solution:

You first ask for the full data set containing influent and effluent concentrations for the period in question. You know that your industry has five days per week with full production, and therefore discharges higher concentrations on those days. Also, you know that during weekends the activities change and production decreases, causing a decrease in the influent concentrations.

You plot the time series of influent and effluent concentrations together with the chart of removal efficiencies (see below; note that *Y*-axis scales have been changed, for clarity) and observe the following points.



Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

The periods with low removal efficiency and non-conformity with the standard for minimum removal efficiency (85%) are not associated with a deterioration in the effluent quality. As a matter of fact, the **effluent** concentrations in this period are also reduced, indicating an improvement in the effluent quality. You notice that *the decrease in removal efficiency is only associated with a decrease in the **influent** concentrations*, which are much lower during these weekend days (see the arrows in the charts).

Conversely, you notice that the periods with higher removal efficiencies are simply higher due to an increase in the influent concentration, giving the erroneous impression that the effluent quality might be better on these days. Indeed, they are not, and are slightly increased during the peak days.

But, overall, you observe that your treatment plant is very robust to fluctuations of the influent, producing very stable effluent concentrations. This is endorsed by the values of the coefficients of variation CV (=standard deviation/mean): $C_{in} = 0.43$; $C_{out} = 0.08$ (calculations done on the attached spreadsheet). Although the influent concentrations vary widely, the variations in the effluent concentrations are very small, indicating stability.

The mean value of the effluent concentration is 42 mg/L. You consult the discharge standards and see that the maximum allowable value is 60 mg/L. You check your data, and see that all values are complying with this requirement.

You then prepare a good report and submit it to the environmental agency with all the above clarifications.

7.6 MEASURES OF CENTRAL TENDENCY FOR REMOVAL EFFICIENCIES

7.6.1 Two different ways of calculating central tendency of removal efficiencies

Advanced

In your reports, you will have to present the measure of central tendency for removal efficiencies. We have thoroughly discussed measures of central tendency in Section 5.6 (mean, median, and geometric mean), and you should use the concepts presented there.

S. 5.6

We will expand the possibility of calculating the **mean** value of removal efficiencies based on the mean values of influent and effluent concentrations. The same comment applies to **median** or **geometric mean**, but, for the sake of simplicity at this moment, we will present these two alternatives based on arithmetic means:

- **Mean of efficiencies**: mean value of the time series of calculated values of removal efficiencies. *The resulting value is more influenced by fluctuations in the data*. This approach could be seen as conceptually closer to that of a paired two-sample test (each pair is made up of simultaneous values of influent and effluent concentrations – see Chapter 10).
- **Mean efficiency**: calculated using the mean values of influent and effluent concentrations = (Mean C_{in}–Mean C_{out})/Mean C_{in}. *The resulting value is less influenced by fluctuations in the data*. This approach could be seen as conceptually closer to that of an independent non-paired two-sample test.

C. 10

We will see from Example 7.6 that **both calculations lead to different values**. The *frequency distribution of the data set of removal efficiency values is usually not symmetrical*, which affects the calculation of measures of central tendency. This is an important point, and because of this, frequency distributions of removal efficiencies are further discussed in Section 7.7.

S. 7.7

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

7.6.2 The case of missing data

Another challenge frequently associated with treatment plant monitoring data sets relates to **missing data**, and this is worth discussing here. When you are monitoring, you may lose some samples on certain days, and have missing values for the influent and effluent concentrations. For instance, you might have five days of monitoring data, but you obtained *influent concentrations* only on three days ($n = 3$): days 1, 2, and 3 (suppose you missed days 4 and 5). For *effluent concentrations*, for some reason, you obtained values only on four days ($n = 4$): days 2, 3, 4, and 5 (suppose you missed day 1). Therefore, you can calculate the *removal efficiency* for only two days ($n = 2$): days 2 and 3, in which you have simultaneous data on influent and effluent concentrations. Your **mean of efficiencies** will be based on only two values (two days). However, if you compute the mean value of the influent concentration based on three data points ($n = 3$) and the mean value of the effluent concentration based on four data points ($n = 4$), then you can calculate the **mean efficiency** [(Mean C_{in}–Mean C_{out})/Mean C_{in}] using the mean values of the influent and effluent concentrations. The mean values of removal efficiencies will be different for the two different methods.

7.6.3 The case of outliers

Remember also that we can have **negative values for removal efficiencies**, in the case where $C_{out} > C_{in}$ (which is not uncommon during specific periods in your treatment plant, as discussed in Sections 7.3.3 and 7.3.4). When searching for measures of central tendency, mean values may be strongly affected by these negative values, while median values are likely to be more robust. Let us imagine that your treatment plant generally performs well but suffered poor performance during one episode where it experienced strange conditions. Suppose you have five values of removal efficiencies, and four were very good, but one reflected this abnormal system malfunction: 91%, 88%, −165%, 87%, and 92%. The mean of these values is 39%, while the median is 88%. It is up to you to investigate the causes and consequences of this irregular episode of low efficiency, and you might consider it an outlier in your data set. However, you must also make a balanced interpretation of the measures of central tendency to decide what you consider to be the 'typical' behaviour of your plant. *In this case, the median efficiency provides a better indication of central tendency compared with the mean of efficiencies. The mean efficiency would probably also give you a better indication of central tendency.*

S. 7.3.3
S. 7.3.4

7.6.4 Mean efficiency versus mean of efficiencies: impact on results

Example 7.6 is supported by an Excel spreadsheet. The idea is that you will change the values as per the instructions in the sheet and interpret the results. We will present here only one of the possibilities, which is the one directly available in the sheet.

Example

EXAMPLE 7.6 CALCULATION OF MEASURES OF CENTRAL TENDENCY OF REMOVAL EFFICIENCIES

You received data from a treatment plant for eight samples that were collected on different days. The data you obtained are for the influent and effluent concentrations of a certain constituent. For each of the eight days, you calculate the associated removal efficiency [$(C_{in}-C_{out})/C_{in}$] *and include it in your summary table.*

Calculate the measure of central tendency for the removal efficiency using the following two concepts: ***mean of efficiencies*** *and* ***mean efficiency****. Also calculate the median and geometric means using the Excel functions MEDIAN and GEOMEAN* (*see* Section 5.6 *on their calculation*).

S. 5.6

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Data:

Sample	Concentrations (mg/L)		Efficiency $= (C_{in} - C_{out})/C_{in}$ (%)
	Influent C_{in}	Effluent C_{out}	
1	100	50	50
2	60	25	58
3	140	30	79
4	75	35	53
5	90	60	33
6	130	40	69
7	85	25	71
8	120	20	83

Excel

Note: this example is also available as an Excel spreadsheet.

Solution:

Initially, we present in the summary table below the measures of central tendency based on the calculations involving the eight values in the data set. For instance, the mean is the **mean of efficiencies**, and is simply the mean of the eight values of efficiency.

Measure of Central Tendency	Concentrations (mg/L)		Efficiency (%)
	Influent	Effluent	
Mean	100.0	35.6	**62.1**
Median	95.0	32.5	**63.8**
Geometric mean	96.5	33.5	**59.9**

Note: these statistics should have been presented without the extra digit to the right of the decimal point, because the original values do not have this many significant digits. However, we included one extra digit so that you can check on the accuracy of the calculations and compare the different ways of calculating the measures of central tendency.

Now we will present another frequently used way of expressing means, based on the *mean values of influent concentrations and mean values of effluent concentrations*. We are calling this the **mean efficiency**. A similar procedure can be done for *medians* and *geometric means*.

Mean efficiency	**64.4%** $= (Mean\ C_{in} - Mean\ C_{out})/Mean\ C_{in}$ $= (100.0 - 35.6)/100.0$
Median efficiency	**65.8%** $= (Median\ C_{in} - Median\ C_{out})/Median\ C_{in}$ $= (95.0 - 32.5)/95.0$
Geometric mean efficiency	**65.3%** $= (Geom\ mean\ C_{in} - Geom\ mean\ C_{out})/Geom\ mean\ C_{in}$ $= (96.5 - 33.5)/96.5$

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

We can see that the **mean of the efficiencies** is **62.1%** and the **mean efficiency** is **64.4%.** Usually, the *values of the mean efficiency are higher than those of the mean of efficiencies*, because the latter are influenced by occasional values of low efficiency that affect the interpretation of arithmetic means as measures of central tendency. Note that this comment is similar to those made repeatedly in Section 5.6, when we mentioned that few high values of concentration can push the mean concentration to reach high values. Now, with efficiencies, the situation is opposite: a few low values can push the mean to reach low values.

S. 5.6

Now you could make the comparisons in terms of medians and geometric means and make your conclusions.

In the Excel spreadsheet, we propose that you do several different exercises: (a) keep all influent concentrations the same; (b) keep all effluent concentrations the same; and (c) introduce empty cells (missing data) for the influent and effluent concentrations. We will not carry out these exercises here, but you can complete them on your own using the linked Excel spreadsheet. The conclusions we have included in the sheet regarding the comparison of **mean of efficiencies** versus **mean efficiency** are:

- If the influent concentrations are all the same in their own series (and effluent concentrations are the same or different in their own series), the mean efficiency and mean of efficiencies will both be the same;
- If the values of effluent concentrations are all the same in their own series, but influent concentrations are different, the mean efficiency and mean of efficiencies will produce different results;
- If there are any missing data (influent or effluent), the mean efficiency and mean of efficiencies will yield different values (even if influent and effluent values are the same in their own series);
- If the influent concentrations are different in their own series, the mean efficiency and mean of efficiencies will yield different values (even if the effluent concentrations are the same) (equal to comment 2 but stated in a different way).

7.6.5 Mean efficiency versus mean of efficiencies: which one to use?

There are some situations where it might be more appropriate to use the mean of efficiencies rather than the mean efficiency, despite the fact that the resulting value for this measure of central tendency is more influenced by fluctuations in the data.

In short, *if your influent and effluent samples are **independent samples**, then you should use the **mean efficiency*** (or the geometric mean or median equivalent). Alternatively, *if your influent and effluent samples are **paired samples**, then you should use the **mean of efficiencies*** (or its geometric mean or median equivalent). The difference between independent and paired samples is briefly discussed here, but we provide a more detailed discussion on this topic in Chapter 10.

C. 10

Samples are considered to be **independent** if there is 'no natural structure in the order of observations across groups' (Helsel and Hirsch, 2002). In other words, the influent samples are not directly linked to the effluent samples. This is the case for many flow-through treatment systems, which naturally have some degree of mixing within the reactors. As such, the water sampled at the inlet is not the same plug of water sampled at the outlet. In this case, it could seem more natural that you use the concept of **mean efficiency**.

Alternatively, in the reasoning of **mean of efficiencies**, samples are considered to be **paired** (**dependent**) when there is a logical union between the two groups. In other words, the influent samples are directly linked to the effluent samples. One example of a type of treatment system that satisfies this requirement is a sequencing batch reactor. In a batch system, the plug of water entering the reactor is typically the same

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

plug of water that is discharged after the treatment period. In studies of natural water bodies, upstream/downstream sampling techniques could be argued to be paired samples (e.g., one sample directly upstream of a contamination source, such as a discharge pipe, and another sample directly downstream of the contamination source). This is especially true if the replicate samples are collected on different days throughout the year. The reason is that the ambient concentration of the constituent in the water body might be subject to large fluctuations throughout the year, but the source of contamination is hypothesized to be a relatively steady source of the pollutant into the water body. In this case, you would use the same approach to calculate the percent difference or log difference between the upstream and downstream concentrations, but the desired value would be a percent or log increase rather than a percent or log removal efficiency.

S. 7.7

S. 5.6

Another point of view on the decision of which measure of central tendency to adopt for removal efficiencies is the fact that their *distribution is likely to be skewed* (see following Section 7.7). In this case, we could have the same argument when we questioned arithmetic mean to be a good representative of central tendency when data are skewed (see Section 5.6), and mentioned about the appropriateness of other measures (medians or geometric means). Therefore, in our case here, we could consider that *mean of efficiencies is subject to the influence of the skewness of the distribution*.

7.7 FREQUENCY DISTRIBUTION OF REMOVAL EFFICIENCIES

Advanced

S. 5.7

S. 6.3

S. 6.3.1

In this chapter, we have already seen how to calculate and interpret removal efficiencies, and also how to determine measures of *central tendency* in different ways. *Measures of variation* will be calculated following the same procedures that we previously presented in Section 5.7. However, now we will analyse **frequency distributions of removal efficiency data**. Frequency distributions, in general, have already been independently explained in Section 6.3.

Let us start by analysing Figure 7.6. This figure shows **frequency histograms** (see Section 6.3.1 for a detailed explanation of frequency histograms) for the *influent concentration* (mg/L), the *effluent concentration* (mg/L), the *removal efficiency* (%), and the *remaining fraction* [$=$ *100–efficiency* (%)]. Assume for this example that the constituent of interest is total phosphorus (total P). The data shown here are based on actual measurements of a wastewater treatment plant collected over a period of four years (see the master spreadsheet Descriptive Statistics Treatment Plant). For now, do not worry about the actual values in the graphs, but focus instead only on the shape of their distributions. In this example, the number of class intervals in the four graphs is the same (equal to 20).

C. 8

These graphs from Figure 7.6 reflect usual patterns, although the degree of non-symmetry or skewness may vary from plant to plant and from constituent to constituent. We see that the *influent concentration* has a slight skew to the right and the *effluent concentration* has a strong skew to the right. The influent concentration could probably fit to a normal or a log-normal distribution (see Chapter 8), while the effluent concentration would not be a good fit for a normal distribution, and most likely would be a candidate for fitting to a log-normal distribution.

C. 8

Now, if we look at the third panel of Figure 7.6, which shows **removal efficiency**, we see a *skew to the **left***, which is a *typical pattern for frequency distributions of removal efficiencies*. Note that this has the opposite shape of a log-normal distribution (see Chapter 8). If we now analyse the pattern of the frequency distribution of the **remaining fractions**, we see an almost mirror image, with a *skew to the **right***, and a strong candidate for a log-normal distribution fitting. Remembering, the remaining fraction is the fraction that has not been removed (see Equation 7.4). For instance, if the value of removal efficiency is 90%, the remaining fraction is $100 - 90 = 10\%$.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

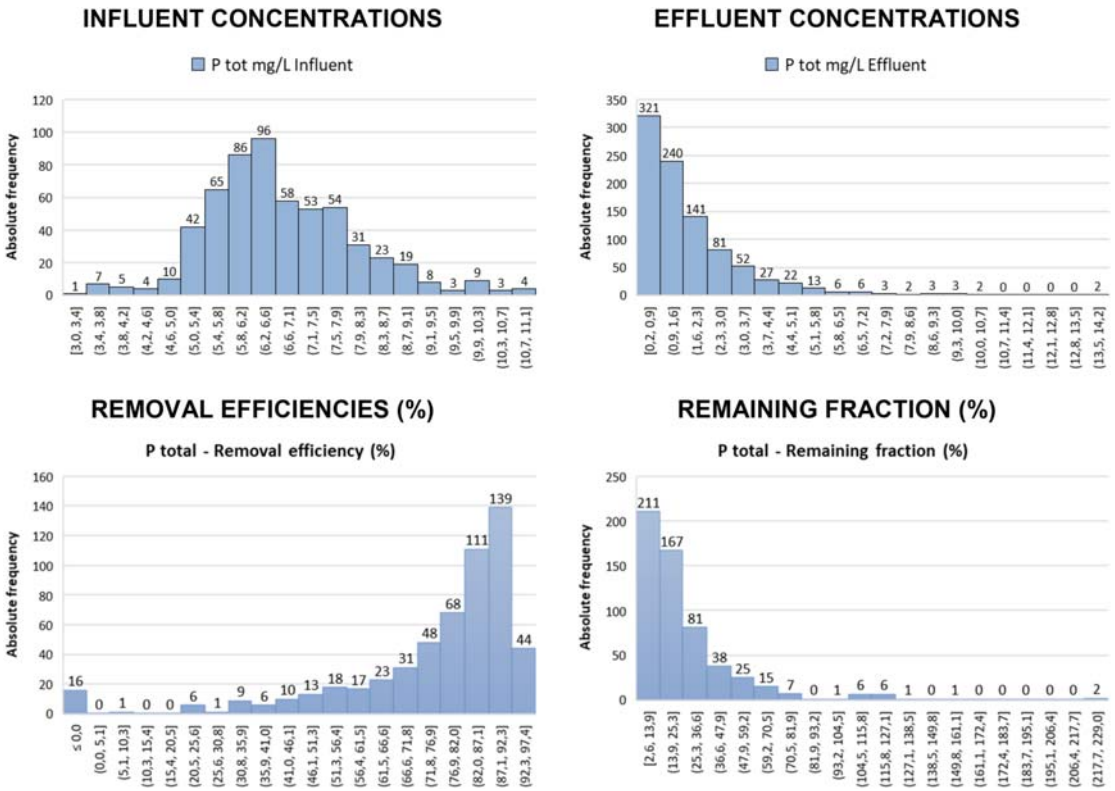


Figure 7.6 Frequency histograms of influent concentrations (mg/L), effluent concentrations (mg/L), removal efficiencies (%), and remaining fractions (%). The constituent is total phosphorus (total P), and the histograms are based on four years of monitoring in a wastewater treatment plant.

S. 5.6.1

Figure 7.7 shows *'typical' frequency distributions for removal efficiency and the remaining fraction* in **wastewater treatment** plants (this figure is an adaptation of Figure 5.8 from Section 5.6.1). Note the emphasis on 'typical': there is no theoretical guarantee that the patterns will be like those shown here; it is just our experience, based on a wide survey of wastewater treatment plants (Oliveira *et al.*, 2012), and on the fact that removal efficiencies have a maximum value (upper bound) of 100% and remaining fractions a minimum value (lower bound) of 0%.

C. 5

As mentioned before, if your removal efficiency data display these characteristic skews in their frequency distributions, then keep in mind that the arithmetic mean may not be the most representative measure of central tendency (see Chapter 5). Instead or in addition, you may want to report the medians or geometric means of your calculated removal efficiencies and remaining fractions. Also note that if your distribution is skewed this way, if you convert the percent removal efficiency values to log reduction values (LRVs), then the shape of the frequency distribution might look more like a bell-curve (i.e., normal distribution).

The implications of these types of characteristic frequency distributions will be discussed in the next chapter, with a detailed view on theoretical aspects of normal and log-normal distributions, together with other distributions of importance in monitoring data from treatment plants.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

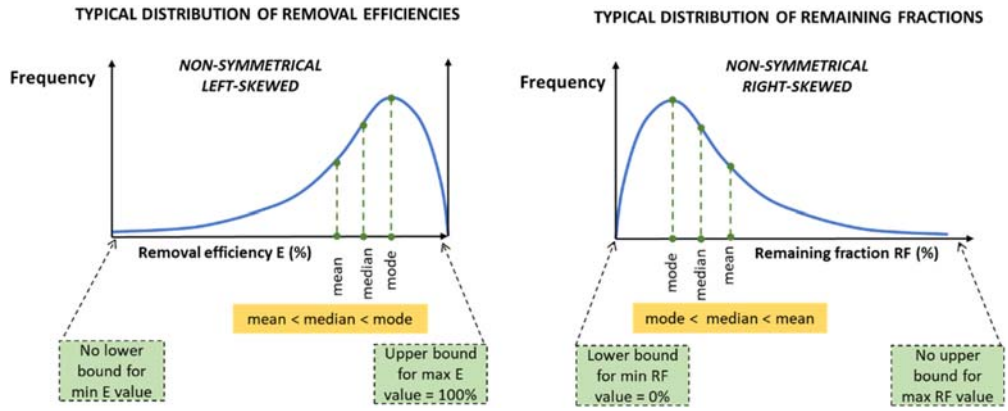


Figure 7.7 'Typical' frequency distributions for removal efficiencies (E) and remaining fractions (RF) for treatment plants.

7.8 CHECK-LIST FOR YOUR REPORT

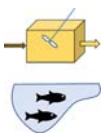
- ✓ Verify that you are adopting a correct terminology for what you are trying to describe: removal or reduction. Also, check that you are not reporting removal of a constituent that should not be removed in the plant, but rather that you expect would be produced (e.g., via chemical conversions).
- ✓ Check that you are using efficiencies expressed as percentages or LRVs (log reduction values) in the correct way. Percentage removal is the most frequent way for reporting removals for most constituents, while LRV is the best way to report reduction efficiencies for microbial constituents such as coliforms and pathogens.
- ✓ If the treatment unit you are representing has substantial water losses, calculate the removal efficiencies in terms of loads, not concentrations, and state this clearly in your report.
- ✓ If you dealt with censored data using any specific substitution technique, clearly state which approach you used to handle the censored data and consider the impact this might have on your estimated removal efficiency.
- ✓ Check that you have investigated possible episodes of very low or even negative removal efficiencies and have provided a reasonable explanation for them.
- ✓ Make sure that you interpret whether the removal efficiency was 'good' or 'poor' in light of the expected capacity of your treatment plant for removing that specific constituent. Also, verify that you characterize removal efficiency as being 'sufficient' or 'insufficient' as specified by discharge standards in your legislative district, or by target values established by your company (e.g., in the case of an industrial treatment plant).
- ✓ Confirm that you have interpreted together *removal efficiencies* and *effluent concentrations*.
- ✓ Ensure that you have clearly described how you calculated the measures of central tendency for the efficiencies: did you use the mean of efficiencies or the mean efficiency (or the median or geometric mean equivalents)? Determine if your samples should be considered as independent or paired. If your samples are independent, you should use the mean efficiency; if they are paired, you should use the mean of efficiencies.
- ✓ If necessary, interpret the pattern of the frequency distribution of your effluent concentrations and removal efficiencies together.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Chapter 8

Symmetry and asymmetry in monitoring data. Normal and log-normal distributions

This chapter presents the foundations of two of the most important frequency distributions in environmental monitoring: normal and log-normal distributions. The main characteristics, properties, and parameters are presented, with indications on how to define and interpret them. Normal distributions are widely used and are associated with a symmetric interpretation of data distribution; however, this may not be the case with environmental monitoring data. Therefore, we provide conceptual background to think about data in a non-symmetrical way, taking the log-normal distribution to represent this typical asymmetrical behaviour of environmental data. Chapters 9 and 10 will make use of the concepts presented here.



The contents in this chapter are applicable to both treatment plant monitoring and water quality monitoring.

CHAPTER CONTENTS

8.1 Frequency Distributions of Monitoring Data **208**
8.2 Normal Distribution **211**
8.3 Log-normal Distribution **222**
8.4 Moment Matching to Use Other Distributions **235**
8.5 Check-List for Your Report **239**

© 2020 The Authors. This is an Open Access book chapter distributed under the terms of the Creative Commons Attribution Licence (CC BY-NC-ND 4.0), which permits copying and redistribution for non-commercial purposes with no derivatives, provided the original work is properly cited (https://creativecommons.org/licenses/by-nc-nd/4.0/). This does not affect the rights licensed or assigned from any third party in this book. The chapter is from the book Assessment of Treatment Plant Performance and Water Quality Data: A Guide for Students, Researchers and Practitioners, Marcos von Sperling, Matthew E. Verbyla and Sílvia M. A. C. Oliveira (Authors).
doi: 10.2166/9781780409320_0207

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

8.1 FREQUENCY DISTRIBUTIONS OF MONITORING DATA

Basic

The subject of frequency distributions and **probability distribution functions** is widely covered in statistical textbooks, and comprises an impressive amount of accumulated knowledge in the literature. Our objective here is to draw from this theoretical background the most important concepts you need to know in order to interpret your monitoring data. As such, the coverage here will avoid many theoretical considerations, instead stressing the most important aspects for you to understand probability distribution functions. If you decide to go into more detail on these topics, you will definitely need to consult your reference statistics textbooks. Some good examples that demonstrate applicability to environmental data are found in Sokal and Rohlf (2011) and Barnett (2004).

As mentioned before, we have two types of **quantitative data** and their associated variables:

- **Discrete variables:** represented only by *integer numbers.* Examples: variables that can be counted, such as the number of samples per year complying with discharge standards or the number of treatment plants using activated sludge.
- **Continuous variables:** expressed as numbers that can be measured or represented on a *numerical scale.* Examples: the majority of monitoring data we cover in this book, such as flows, concentrations, loads, and removal efficiencies.

Each of these variables may have a distribution that is represented by different **theoretical functions**:

- **Discrete variables.** The main models of discrete random variables, which find a wide range of environmental applications, can be grouped into three major categories. The first is related to variations of the so-called Bernoulli processes and includes the *binomial*, *geometric*, and *negative binomial* distributions. The second refers to the Poisson processes, which include the *Poisson* distribution. The third includes the *hypergeometric* and *multinomial* distributions. We will not cover distributions for discrete variables in this book, however there are plenty of other texts that adequately cover this topic (e.g., Biometry).
- **Continuous variables.** There are several models of probability distributions pertaining to continuous random variables. Among them, some more well-known distributions include the *uniform*, *exponential*, *normal*, and *log-normal* distributions. The **normal** distribution will be presented first in this chapter, since it has traditionally served as a theoretical basis for the development of confidence intervals, statistical hypothesis tests, as well as the regression and correlation theory. We will also strongly emphasize the importance of **log-normal** distributions for the representation of treatment plant and water quality monitoring data.

C. 6
C. 7
S. 8.2
S. 8.3

Expanding upon our interest in **continuous variables**, we remind you that we have already seen in this book (Chapters 6 and 7) frequency histograms and frequency polygons; we have mentioned that they are interlinked; and we have stressed their importance in the evaluation of monitoring data. Let us show again here these two graphs (see Figure 8.1), taken from Examples 6.3 and 6.4. Our aim in Sections 8.2 and 8.3 of this chapter will be to provide the essential theoretical background for us to represent this frequency distribution by a **probability density function (pdf)** using the **normal** and **log-normal functions.**

Several statistical tests shown in our book are based on the *assumption that the data are* ***normally distributed***, or that they are, at least, *symmetrical around the mean*. If this assumption is fulfilled, the so-called **parametric tests** can be applied. However, if your data are ***not normally distributed***, or are *not symmetrical around the mean*, and you see that it can be represented by a *log-normal distribution*, you should not use directly parametric tests, and need to perform transformations on the data set to make

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

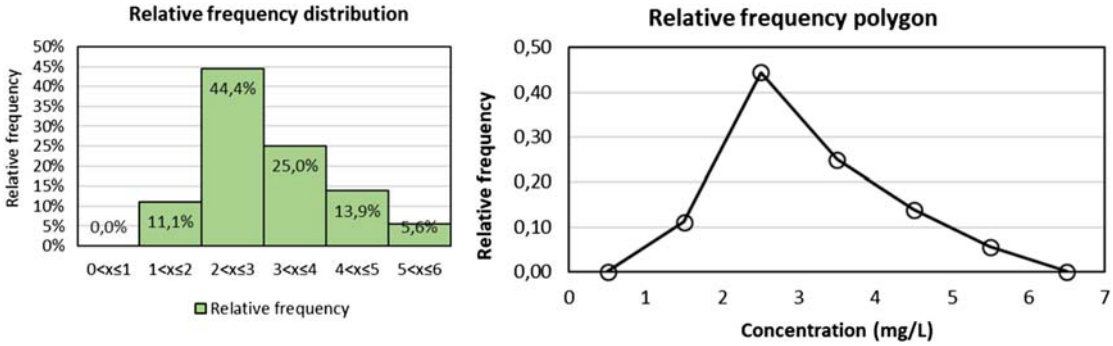


Figure 8.1 Examples of relative frequency histograms and relative frequency polygons.

them normally distributed (for example, if you have log-normally distributed data points, if you take the log of each of them, the resulting log-transformed values will be normally distributed). Otherwise, you must resort to using **non-parametric tests**, which work with *ranked data*, and so *do not depend on the distribution of your original data*. These two categories of tests will be explained in Chapter 10, with due *emphasis to non-parametric tests*, given the fact that most environmental data are not symmetrically nor normally distributed (Oliveira *et al.*, 2012; Limpert *et al.*, 2001). In this chapter, we will show you how to assess whether your data approach a normal or a log-normal distribution, so that you can take the following decisions:

C. 10

- If your data follow a **normal distribution**, you can use **parametric tests.**
- If your data follow a **log-normal distribution**, you can convert the original values to their **$\log_{10}$ values** and use **parametric tests** with the log-transformed data set.
- If you are **not sure about the distribution of your data**, do not want to make transformations in your original data, or simply cannot or do not want to decide on this regard, you can apply **non-parametric tests.**

In this chapter, we will focus mainly on the description of **normal** and **log-normal distributions** because of their importance in treatment plant and water quality monitoring. Sometimes you may find that too much detail is being presented, but here we want to lay the foundations for other statistical tests, described later in the book (Chapters 9 and 10), which depend on your prior understanding of these two distributions. If you want to study other continuous variable distributions, you should consult statistical textbooks. Books on hydrology are also a good source on frequency distributions.

C. 9
C. 10

Another aim we have here is to give you an incentive to open your mind to **thinking on a non-symmetrical basis**. We are very used to think with a symmetrical and even-balanced approach, having the mean as the centre of the balance. **Symmetry** will be highlighted in the **normal** distribution. However, our presentation of the concepts behind the **log-normal** distribution, which is considered to prevail in many environmental data, will hopefully open your mind to also incorporate **non-symmetrical** thinking in the interpretation of your data (see Figure 8.2).

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

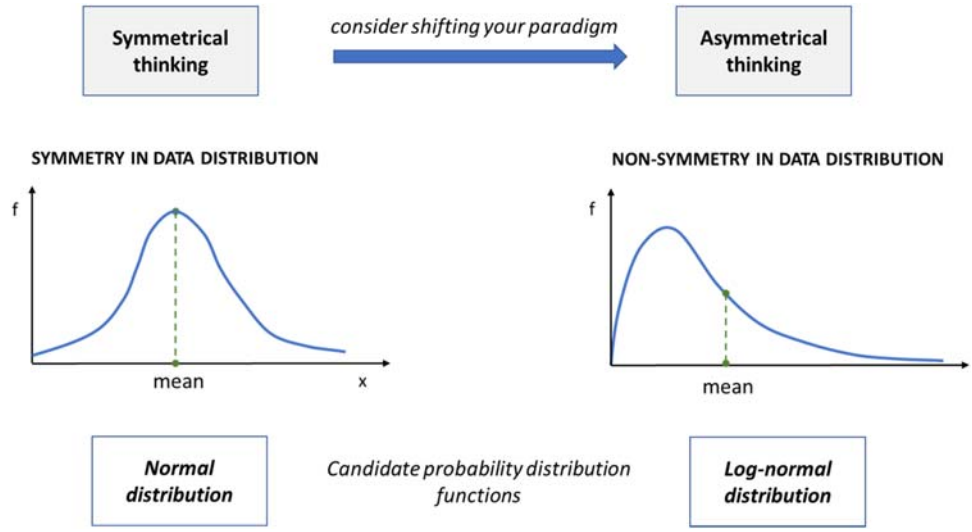


Figure 8.2 Symmetry and asymmetry in distributions of environmental data, and influence on the approach we need to adopt for their interpretation.

Recall that when we look at **probability distribution** plots such as the ones shown in Figure 8.2, the *X*-axis shows the value of the variable and the *Y*-axis shows the probability of encountering that value in the population. So, for instance, if our 'population' is the concentration of total dissolved solids (TDS) in all of the raw water brought into the influent of a water treatment plant, then the *X*-axis would represent the concentration of TDS in a sample of that water. If we picked a concentration at random and found the corresponding *y* value associated with that *x* value, the *y* value would be equal to the probability of encountering that TDS concentration in the population. Because of this, *the area under a pdf curve is by definition equal to* ***1*** *or* ***100%***.

It is important to note that when we deal with continuous variables common to monitoring water systems such as concentrations and loadings, the **physical limits** of these variables extend from **0 to** $+\infty$ (they cannot be negative). In other words, they are generally non-negative continuous variables. In theory, the ***normal distribution*** *extends from* $-\infty$ *to* $+\infty$. Therefore, if we use the normal distribution to describe concentrations of constituents in environmental samples, there are certain regions of the distribution (everything <0) that are impossible in reality. In other words, the normal distribution is not the perfect distribution for these variables, although in practice, the normal distribution may generally meet our needs when evaluating certain aspects of treatment plant performance and natural water systems.

For certain modelling applications, the *negative regions of the normal distribution may present computational problems* (e.g., since negative concentrations are impossible in reality). One way around this is to use another distribution such as the *gamma distribution* or the **lognormal distribution**, since both have a *range from 0 to* $+\infty$ (this is one of many reasons why the lognormal distribution is often a more appropriate distribution to use for certain environmental data sets). Another way around the problems associated with negative regions of the normal distribution is to use a variation called the truncated normal distribution. This distribution looks just like the normal distribution, but it stops at zero. These more advanced distributions are outside the scope of this book, but if you have a need for

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

these approaches, you should consult a statistics textbook. This issue is also discussed in more detail in Section 8.2.3.

S. 8.2.3

8.2 NORMAL DISTRIBUTION

8.2.1 Basic concepts about the normal distribution

Basic

You already know about the bell-shaped normal (or Gaussian) distribution, and we have discussed it in other parts of this book. This distribution is vastly important in statistics, and forms the conceptual basis *for the development of confidence intervals, statistical hypothesis tests, as well as the regression and correlation theory*. All these topics are covered in this book, in the context of treatment plant and water quality monitoring data.

The theoretical normal distribution has the following characteristics:

- The curve is symmetrical around the mean (it has a characteristic 'bell' shape).
- The average (mean), median and mode are the same.
- The curve has two inflection points, which correspond to x values located, respectively, at the distance of **one standard deviation (σ)** above and below the **mean (μ)**.
- The area below the curve totals 1 or 100% (this is true of all distributions).
- The mean splits equally the total area into 50% to the left and 50% to the right.

Some of these characteristics are illustrated in Figure 8.3.

The probability density function (pdf) of the normal distribution is defined by the following *two parameters*:

- **Mean μ**. *Location parameter*. The central value around which the variable is dispersed. It does not influence the shape of the distribution.
- **Standard deviation σ**. *Scale parameter*. The value that indicates the degree of dispersion around the central value. It influences the shape of the distribution.

In this section, we use the following representation:

- *Theoretical normal probability distribution function of a population*: **mean** and **standard deviation** are represented by $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$, respectively.
- *Data sample*: **mean** and **standard deviation** are represented by $\bar{\boldsymbol{x}}$ and **s**, respectively.

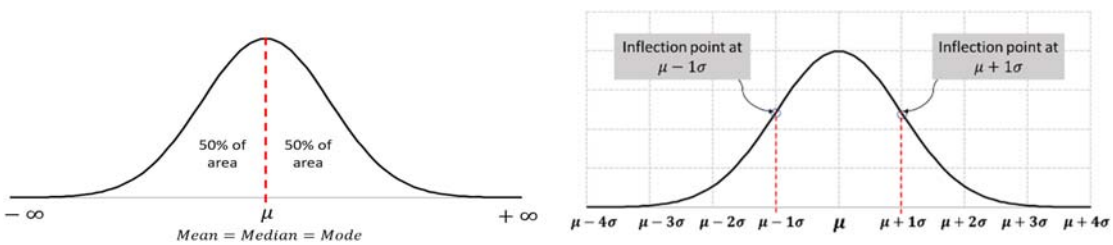


Figure 8.3 Important properties of the normal distribution. μ and σ are the mean and standard deviation of the normal random variable *Y*.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

8.2.2 Influence of the mean and standard deviation on the normal distribution

Advanced

Excel

In order to understand the influence of mean and standard deviation on the normal distribution, let us use the *Excel spreadsheet made for generating a normal probability distribution*. In this spreadsheet, you are invited to vary the values of the mean and standard deviation and see the impact in the resulting graphs of distribution functions, together with other useful information that will allow you to have a good understanding of the normal distribution.

A first example can be with *varying values of the mean* **(location)** and *fixed values of the standard deviation* **(scale)**. Let us compare the resulting normal distributions with the following input data: (i) *mean* = 100, *standard deviation* = 20; (ii) *mean* = 200, *standard deviation* = 20; and (iii) *mean* = 300, *standard deviation* = 20. These values could represent, for instance, concentrations, in mg/L. The resulting values of the coefficient of variation (CV = standard deviation ÷ mean) are 0.20; 0.10; and 0.07. The three **probability density functions** are shown together in Figure 8.4. You can see that the shape of the distributions is the same (because the standard deviations are the same), and only their position is different (because the means are different). Given the importance of **box-plots** in the context of this book, we also present them in Figure 8.4. In order to facilitate the understanding of a comment we make in Section 8.2.5 regarding the concept of the mean ± standard deviation, in this specific box-plot we include, besides the traditional quartiles, also the percentiles 2.275%, 15.866%, 84.135%, and 97.725%. You can also note that the three distributions have the same shape, and the only variation is in their relative position. A clearer view of both graphs can be seen in the Excel spreadsheet.

S. 8.2.5

A second example can be with *fixed values of the mean* **(location)** and *varying values of the standard deviation* **(scale)**. Let us now compare the resulting normal distributions with the following input data: (i) *mean* = 100, *standard deviation* = 20; (ii) *mean* = 100, *standard deviation* = 40; and (iii) *mean* = 100, *standard deviation* = 60. The resulting values of the coefficient of variation (CV = standard deviation ÷ mean) are 0.20; 0.40; and 0.60. The three **probability density functions** are shown together in Figure 8.5. You can see that the shapes of the distributions are now different (because the standard deviations are different), and only their position is the same (because the means are the same). The distribution with the larger CV spreads more than the distribution with the lower CV. Also note that negative values are obtained, which is a matter that deserves attention in the context of treatment plant and water quality monitoring, and that will be discussed in Section 8.2.3. The same conclusions can be obtained in the adjoining **box-plots**.

S. 8.2.3

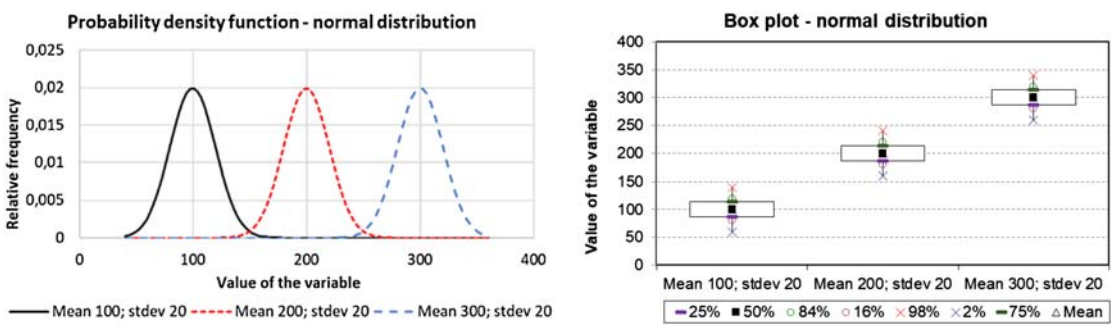


Figure 8.4 Plot of three probability density functions of the normal distribution, together with their corresponding box-plots, for different values of the location parameter (mean) and the same values of the scale parameter (standard deviation).

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

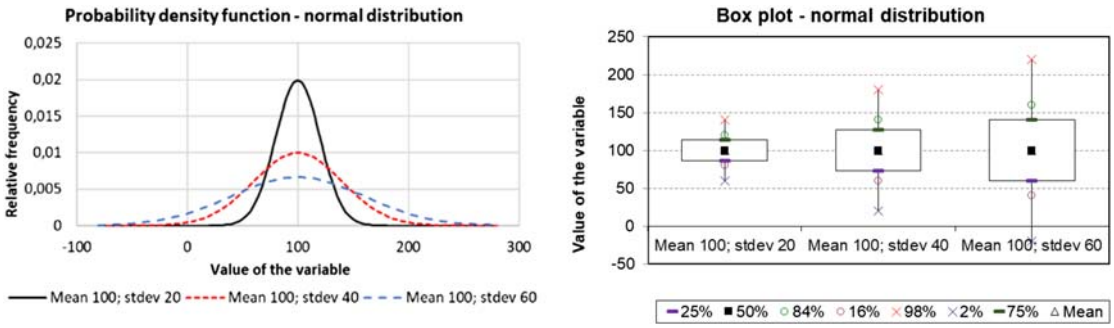


Figure 8.5 Plot of three probability density functions of the normal distribution, together with their corresponding box-plots, for the same values of the location parameter (mean) and different values of the scale parameter (standard deviation).

8.2.3 Negative values for concentrations and values above 100% for removal efficiencies in normal distributions

Advanced

S. 7.7

When analysing Figure 8.5, we highlighted the fact that negative values had been obtained. If we are representing concentrations, negative values have no physical meaning, and should not be considered. In Section 7.7 we had anticipated this, and discussed two situations that have no conceptual support: *concentration values lower than zero* and *removal efficiencies greater than 100%*. Let us analyse this matter in more detail, with respect to the normal distribution.

S. 8.2.2

We will use the same Excel spreadsheet already used in Section 8.2.2. Our first simulation will represent a **concentration** that has a mean value of 10 mg/L. The following values of input data for the normal distribution are used: (i) *mean* = 10, *standard deviation* = 10; (ii) *mean* = 10, *standard deviation* = 20; and (iii) *mean* = 10, *standard deviation* = 30. The resulting values of the coefficient of variation (CV = standard deviation ÷ mean) are high, but still within the range of what can be found in treatment plant and water quality monitoring: 1.00; 2.00; and 3.00. In Figure 8.6, we plot the probability density functions and the box-plots of the resulting normal distributions. Note that the scale of the concentration axis is forced to have a minimum value of 0 mg/L, not allowing **negative values**. We can clearly see

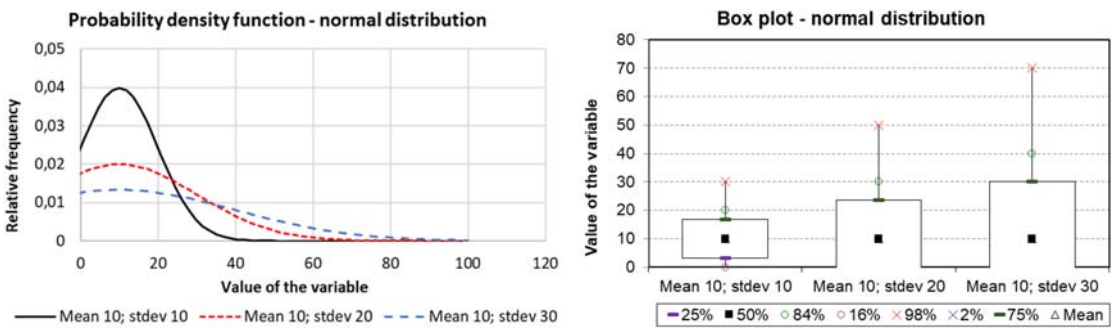


Figure 8.6 Plot of three probability density functions of the normal distribution, together with their corresponding box-plots, for the same values of the location parameter (mean) and different values of the scale parameter (standard deviation). Scale of the variable has been forced to have a minimum of zero.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

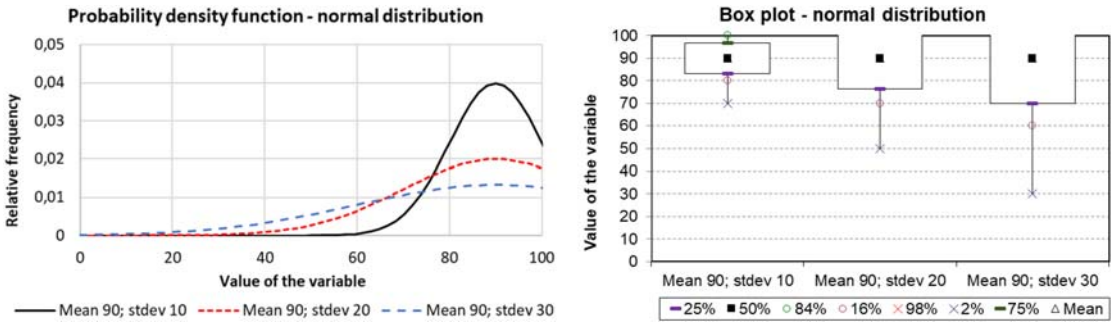


Figure 8.7 Plot of three probability density functions of the normal distribution, together with their corresponding box-plots, for the same values of the location parameter (mean) and different values of the scale parameter (standard deviation). Scale of the variable has been forced to have a maximum of 100.

that the *normal distribution cannot be applied in this case, especially for higher CV values*, indicating a limitation of this distribution for representing this type of data.

Now let us simulate a situation that can take place with **removal efficiencies (%)** at treatment plants, with a mean value of, say, 90%. The following values of input data for the normal distribution are used: (i) *mean* = 90, *standard deviation* = 10; (ii) *mean* = 90, *standard deviation* = 20; and (iii) *mean* = 90, *standard deviation* = 30. The resulting values of the coefficient of variation are 0.11; 0.22; and 0.33. In Figure 8.7, we plot the probability density functions and the box-plots of the resulting normal distributions. Note that the scale of the removal efficiency axis is forced to have a maximum value of 100%, not allowing **values greater than 100**. Again, we can see that the *normal distribution cannot be applied in this case,* indicating a limitation in this distribution for representing this type of data.

As a general conclusion from this section, we see that the **normal distribution** has to be used with **caution** when *representing concentrations and removal efficiencies*, due to the fact that it can produce values that are outside the boundaries of conceptual acceptance. This is one of the motivations for our coverage of the log-normal distribution (see Section 8.3), widely applicable to environmental data.

S. 8.3

Theoretically speaking, the normal distribution can only be used to represent random variables that range from $-\infty$ to $+\infty$. However, in practice, the normal distribution can sometimes be useful for representing concentrations in treatment systems and natural water bodies. However, it is generally *not an appropriate distribution to represent per cent removal efficiency*. An alternative distribution to use in this case would be the **beta distribution**, which represents a continuous random variable that can take on values between 0 and 1. This approach is also not perfect as it prevents the inclusion of negative removal efficiencies, which, as we demonstrated in Chapter 7, are possible in some systems (especially for certain constituents) and problematic operating conditions. Nevertheless, you may find the beta distribution to be useful to represent removal efficiencies that are higher and closer to 100%. We will not go into depth on the theoretical considerations for the beta distribution in this book. However, in Section 8.4, we will show you a brief example of *how to apply your knowledge of the mean and the standard deviation to represent removal efficiencies with a beta distribution, using a technique called moment matching*.

C. 7

S. 8.4

8.2.4 Generation of values for the normal distribution

Advanced

The equations for returning the values of the probability density function of the normal distribution are presented in most **statistical textbooks**, frequently coupled with look-up tables with values. Our option

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

here is to present only Excel functions, because the formulae are complex to handle. Of course, *you should consult the textbooks for a complete view on this matter*.

Note that there are **different variants of Excel functions** for normal distribution, depending on whether you want the normal or standard normal and cumulative or non-cumulative values. They also vary with the version of Excel, and you should consult the manual of your version to select the correct function for your application. Here, we will basically introduce these two functions:

- **NORM.DIST function.** *Returns the normal distribution* (*cumulative or non-cumulative*). You provide the value of the variable for which you want to calculate the frequency, plus the mean and standard deviation of your data, and specify whether you want a cumulative or non-cumulative distribution. You obtain the corresponding value of the relative frequency. For instance, for mean = 100, standard deviation = 20, for a value of your variable equal to 121, the relative frequency according to the normal distribution is 0.011 (specify FALSE in the syntax, meaning that you do not want a cumulative distribution). For a cumulative frequency, you obtain the value of 0.853, meaning that 85.3% of your data have a value ≤121 (the cumulative probability ranges from 0 to 1).
- **NORM.INV function.** *Returns the inverse of the normal cumulative distribution*. You provide the cumulative value of probability (a value between 0 and 1), the mean, and standard deviation of your data. You obtain the corresponding value of the variable. For instance, suppose that for a mean = 100, a standard deviation = 20, and a value of cumulative relative frequency equal to 0.75 (75%), you obtain the value of your variable equal to 113. This would mean that 75% of the distribution is ≤113.

Excel

S. 8.2.2

If you have a statistical software, it will probably have similar functions to manipulate values of the normal distribution. If you do not have one, you can use the Excel spreadsheet already mentioned in Section 8.2.2.

8.2.5 Standard normal variable (*Z*)

Basic

The normal distribution with mean $\mu = \mathbf{0}$ and standard deviation $\sigma = \mathbf{1}$ is called the **standard normal distribution**, and its variable, commonly denoted by Z, is called the **standard normal variable** and is given by

$$Z = \frac{X - \mu}{\sigma} \tag{8.1}$$

The value of Z informs the distance that the variable X has from the mean, measured in terms of the number of standard deviations. For instance, for a mean = 100 mg/L and a standard deviation = 20 mg/L, the variable X with a value of 80 mg/L will be at a negative distance of one standard deviation from the mean [$Z = (80-100)/20 = -1$]. If the variable X has a value of 150, it will have a distance of 2.5 standard deviations from the mean [$Z = (150-100)/20 = 2.5$].

Figure 8.8 presents the **correspondence between mean ± standard deviation and *Z***. It also shows the **percentage of the population that falls between each range**. In this figure, μ is the *true mean of the population* and σ is the *true standard deviation of the population*. Recall that we never actually know the true mean and standard deviation. Instead, we collect samples and estimate the mean and standard deviation from those samples. For a **sample**, instead of using μ and σ, we use $\bar{x}$ for the **sample mean** and **s** for the **sample standard deviation**. Table 8.1 shows that if you calculate the sample mean ($\bar{x}$) and standard deviation (s) of your sample, you can use them to estimate the percentage of future data points that will fall within each range.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

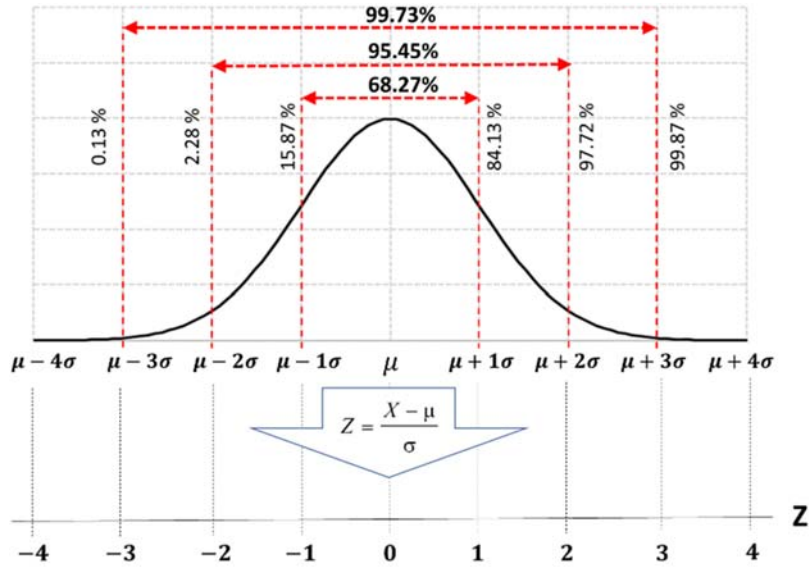


Figure 8.8 Plot of the normal distribution for a mean μ and a standard deviation σ, showing the correspondence between different intervals of $\mu \pm \sigma$ and Z. The graph also shows the percentage of the area under the curve that is contained within each interval.

Table 8.1 Values of the normal cumulative function estimated for different intervals using $\bar{x} \pm s$ and the Z value.

Range	$\bar{x} \pm s$	Z value	Value of Normal Cumulative Function (Percentile) %	Approximate Percentage of Future Data Points Inside Range %
$\bar{x} \pm 1\,s$	$\bar{x} - 1\,s$	-1	15.87	~68
	$\bar{x} + 1\,s$	$+1$	84.13	
$\bar{x} \pm 2\,s$	$\bar{x} - 2\,s$	-2	2.28	~95
	$\bar{x} + 2\,s$	$+2$	97.72	
$\bar{x} \pm 3s$	$\bar{x} - 3\,s$	-3	0.13	~99.7
	$\bar{x} + 3\,s$	$+3$	99.87	

For instance, if you sample a water source for a particular constituent and you measure a mean concentration of 100 mg/L with a standard deviation of 20 mg/L, assuming this was a normally distributed variable, approximately 68% of your future measurements of this water source will have concentrations between the interval of 80 and 120 mg/L, since $100 - 1 \times 20 = 80$ and $100 + 1 \times 20 = 120$ (between $Z = -1$ and $Z = +1$). Likewise, approximately 95% of future samples will have concentrations between 60 and 140 mg/L, since $100 - 2 \times 20 = 60$ and $100 + 2 \times 20 = 140$ (between $Z = -2$ and $Z = +2$). Nearly all future samples should have concentrations between 40 and 160 mg/L, since $100 - 3 \times 20 = 40$ and $100 + 3 \times 20 = 160$. It is important to note that your predictions for $\bar{x}$ and s from your data set are not the true values of the underlying distribution – they are simply predictions. This means that there is no guarantee that 68%, 95%, and ~100% of future data points will fall within

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

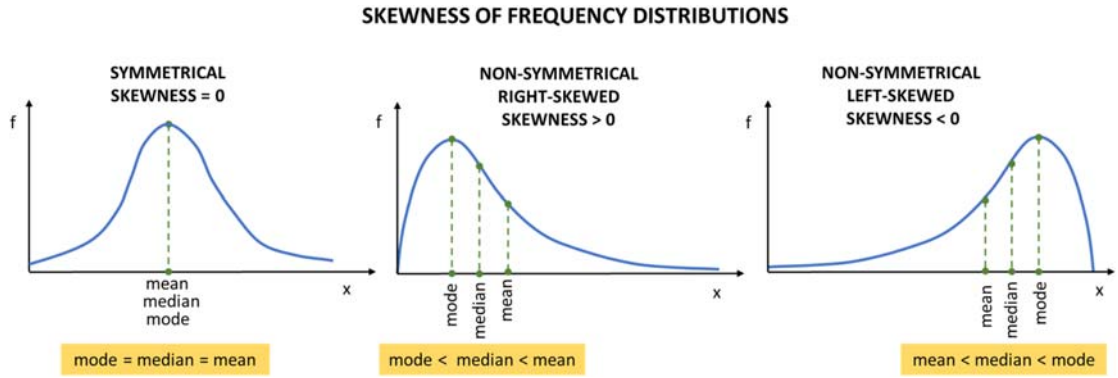


Figure 8.9 Different types of skewness of frequency distributions and influence on the relative values of mean, median, and mode.

the ranges defined by 1, 2, and 3 standard deviations away from the mean, respectively. You should consider the uncertainty associated with your estimate of the mean (e.g., using its confidence interval).

Advanced

8.2.6 Skewness of a distribution

We can also characterize the distribution of the variable we are studying by its **skewness**. This parameter can be calculated for your variable and compared with the theoretical value for a perfect normal distribution. Many statistics textbooks describe skewness in great detail and give the formula for the calculation of a skewness coefficient. However, we will take a simpler approach here and highlight only the most important points of the **skewness coefficient** (Metcalf & Eddy, 2003; Naghettini & Pinto, 2007):

- It characterizes the *degree of asymmetry of a distribution around its mean*.
- The skewness coefficient is **dimensionless** and its values are negative, zero, or positive.
- For a perfect **normal distribution**, it has a value of **zero**. For a **right-skewed** distribution, the value is **positive**, and for a **left-skewed** distribution, the value is **negative** (see Figure 8.9).
- In treatment plant and water quality monitoring, the distribution of **concentrations** is frequently **positively skewed**, while the distribution of **removal efficiencies** is frequently **negatively skewed**.
- The **Excel** function for calculating the skewness of a data set is **SKEW**.

Advanced

8.2.7 Fitting a normal distribution to your data

You have your monitoring data, you have plotted a frequency histogram and frequency polygons, and now you want to try to fit a normal distribution to your data. For you to understand how this procedure is done, consider these three simple steps:

- Calculate the mean and standard deviation of your data set.
- Create a column for the values of your variable X, for which you want the normal frequency distribution to be calculated. For instance, if your data set ranges from 10 mg/L (minimum value) to 85 mg/L (maximum value), you may make calculations for increments of 1 mg/L, starting with 0 mg/L and going up to 100 mg/L, that is, having 100 intervals with a class interval width of 1 mg/L.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

- Use the Excel function **NORM.DIST** in a second column, picking up the value of X in the cell in the first column, and the fixed values of mean and standard deviation. Specify FALSE, meaning that you do not want a *cumulative* frequency distribution.

Example 8.1 illustrates the use of these steps for fitting a normal distribution of a data set (this example builds on Examples 6.3 and 6.4, which showed you how to make a frequency distribution table and plot a frequency histogram and polygon). However, this example is simply to show you the principles of distribution fitting – if you are planning to fit distributions to your data sets on a routine basis, you would probably benefit from learning to use a more advanced statistical software. It is probable that other statistical software programs will provide you with better graphs than the one shown here (Example 8.1). For the sake of simplicity, we do not calculate the goodness-of-fit for the normal distribution to the data set in this example. However, assessing goodness-of-fit is an important part of model fitting and is something that you should also learn how to do.

Example

EXAMPLE 8.1 FITTING A NORMAL DISTRIBUTION TO YOUR FREQUENCY POLYGON

In Example 6.3 you calculated the frequency distribution and plotted the frequency histogram of the data below. In Example 6.4 you plotted the frequency polygon. Now, fit a normal distribution to your frequency polygon, plot both distributions and make a visual interpretation.

Excel

Note: This example is also available as an Excel spreadsheet.

Data:

2.8	4.2	3.9	3.3	2.8	1.7	1.9	2.5	3.1	3.8	2.7	4.1
4.3	4.8	5.1	5.2	3.9	3.5	2.7	3.1	2.8	3.4	4.9	2.8
2.8	1.8	2.1	2.6	2.3	2.4	2.5	1.3	2.9	2.4	2.1	3.6

Solution:

The calculated values of the mean and standard deviation of your data are as follows (note that when you report the mean and standard deviation in your report, you should use only two significant digits since your original data points are represented by only two significant digits; however, since we are using these values to calculate Z values for the normal distribution, it is okay to use some additional digits in the calculation to increase the precision):

- Mean = 3.16 mg/L (reported as 3.2 mg/L).
- Standard deviation = 1.04 mg/L (reported as 1.0 mg/L).

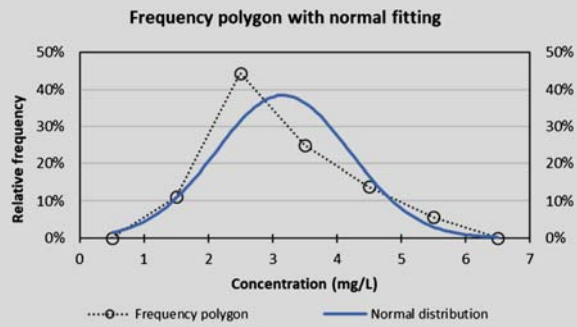
Set up a computational table with 100 class intervals (100 rows), starting from 0.5 mg/L (which is the lowest mid-point value of your frequency polygon table from Example 6.4) and going up to 6.5 mg/L (which is the highest mid-point value from the same table). Since there are 100 intervals, the width of each class interval or the increment in the values of the variable in the computational table will be $(6.5 - 0.5)/100 = 0.06$ mg/L.

For the last column, we will use the NORM.DIST Excel function in the following way:
NORM.DIST (X_i value in the cell to the left; mean; standard deviation; FALSE)

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Interval Number	X_i Value of the Variable (Values in *X*-axis)	Syntax	Value of the Normal Distribution
1	0.500	NORM.DIST(0.500; 3.16; 1.04; FALSE)	0.01445
2	0.560	NORM.DIST(0.560; 3.16; 1.04; FALSE)	0.01672
3	0.620	NORM.DIST(0.620; 3.16; 1.04; FALSE)	0.01930
4	0.680	NORM.DIST(0.680; 3.16; 1.04; FALSE)	0.02219
…	…	…	…
99	6.440	NORM.DIST(6.440; 3.16; 1.04; FALSE)	0.00259
100	6.500	NORM.DIST(6.500; 3.16; 1.04; FALSE)	0.00215

The normal distribution plot, superimposed over the frequency polygon made in Example 6.4, results in the following plot. You analyse it and consider that, at least visually speaking, in this case the normal distribution appears to approximately follow the main trends of the data set but does not reproduce the peak values very well. Again, please note that to definitively determine whether or not the normal distribution is a good fit for this data set, you should use goodness-of-fit tests.



8.2.8 Tests for normality and goodness-of-fit tests for a normal distribution

Advanced

Checking to see whether the distribution of your data **follows a normal distribution** may be an important step when you want to decide whether to use **parametric** or **non-parametric** tests. If your data can be *reasonably well represented by a normal distribution*, you can use a family of *parametric tests* for statistical hypothesis testing, regression, and correlation theory.

These goodness-of-fit tests are widely covered in statistical textbooks and are part of most statistical software packages. Because of their complexity, we will not show how to undertake their calculations here. If you decide to go further into this, you should consult relevant references. For testing whether the normal distribution fits into your data, you can employ simultaneously the following techniques (Oliveira *et al.*, 2012; Action Stat manual, 2019):

- Graphical analysis
  - Normal probability plots
  - Q–Q plots (quantile–quantile plots)

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

- Interpretation of the skewness coefficient
- Statistical tests for normality or goodness-of fit tests
  - Shapiro–Wilk W test
  - Anderson–Darling
  - Lilliefors test
  - Ryan–Joiner
  - Chi-squared test
  - Kolmogorov–Smirnov test (although this test seems to present poor power properties for normality testing)

(a) Normal probability plot and quantile–quantile plot

Typically, when you use a statistical software, you will obtain **normal probability plots** and **Q–Q plots**, and you will analyse *adherence to a straight line*. The graphs have similar concepts, but are presented in different ways:

- The Q–Q plots present only the quantiles Z on both axes (theoretical quantiles in the X axis and quantiles of the measured data in the Y axis)
- The normal probability plot presents the theoretical quantiles on the X axis and the values of the measured variable or its probability of occurrence on the Y axis. Some people prefer to invert the positions of the X and Y axes.

Excel

An example of both plots is shown in Figure 8.10 (using the data from Example 8.1). This calculation is supported by an Excel spreadsheet.

From both graphs presented in Figure 8.10, we can see some adherence to a straight line, but also some departures, especially on higher values of $\pm Z$. Let us analyse typical plots from different types of distributions and how to interpret them (Figure 8.11).

(b) Skewness coefficient

The **skewness coefficient** will assist you in analysing asymmetry of the data (the skewness coefficient for a normal distribution is zero). For a **right-skewed** distribution, the value is **positive**, and for a **left-skewed** distribution, the value is **negative** (see Section 8.2.6).

S. 8.2.6

(c) Tests for normality of the data

In many situations, you may not be very interested in applying a goodness-of-fit test to your data. *You may be only concerned in knowing whether your monitoring data approximately follow a* ***normal distribution***, so that you can apply **parametric statistical tests** (confidence intervals, hypothesis testing, regression, and correlation theory), because a basic underlying assumption

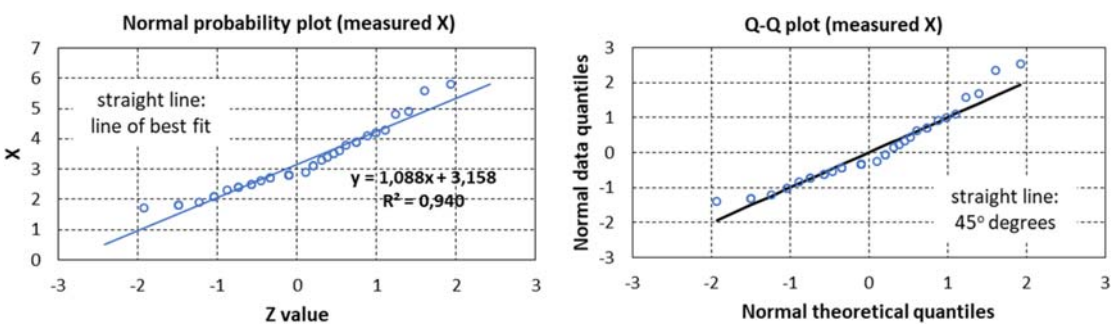


Figure 8.10 Example of a probability plot and a Q–Q plot using the data from Example 8.1. The plots are constructed in the accompanying Excel spreadsheet.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

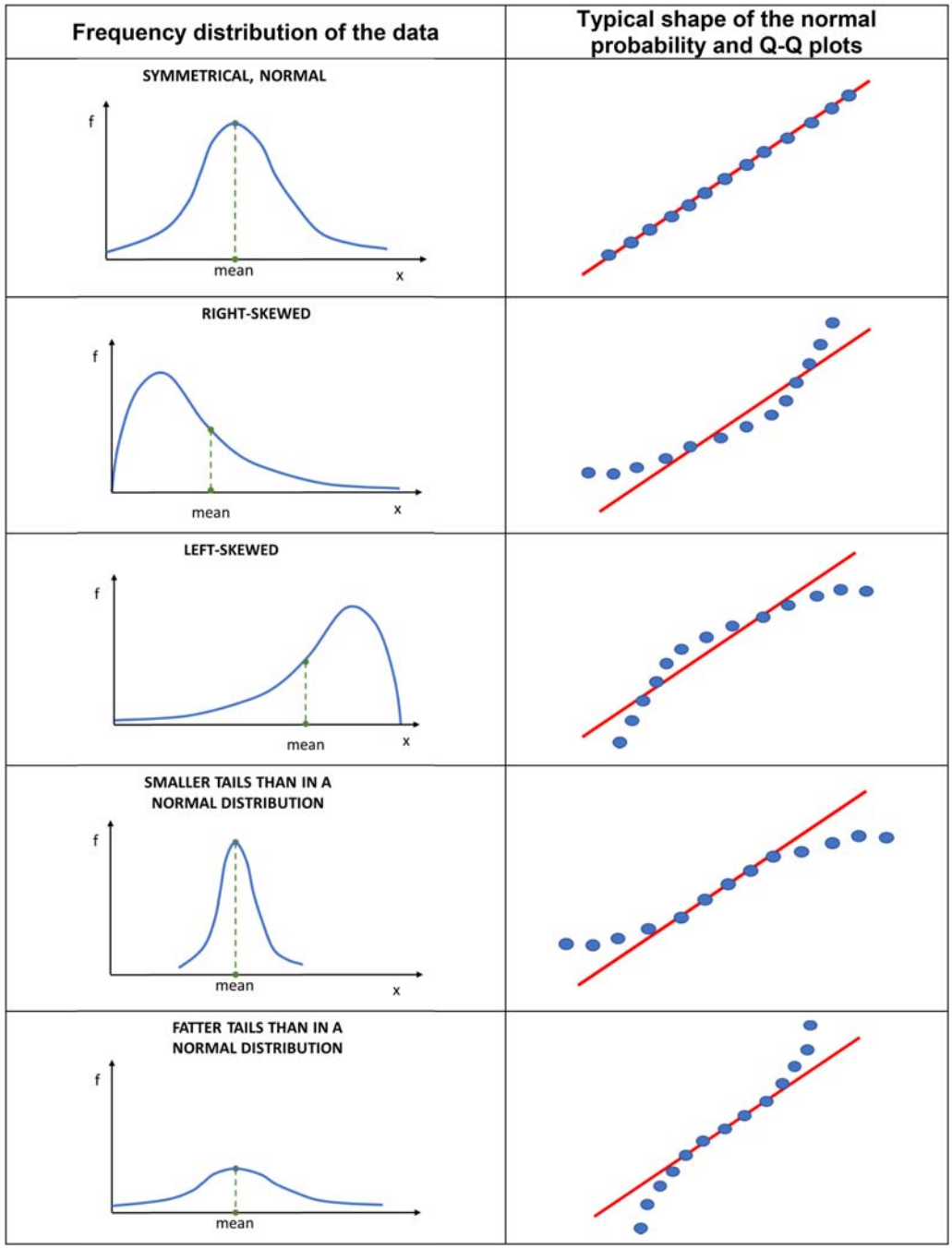


Figure 8.11 Interpretation of normal probability and Q–Q plots for different shapes of frequency distributions. *Source*: Right-hand column adapted from Skymark (2019).

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

about these tests is that the data are normally distributed (or at least symmetrical). If your data **do not follow a normal distribution**, you may need to use **non-parametric tests**.

Testing for normality can be done using the **Shapiro–Wilk** test. Because this test is more complex to perform, it will not be described here, but you can use a statistical software. The main information we look for is the ***p*-value**. The *p*-value is the final result of the test, and should be interpreted in comparison with a certain specified **significance level (α)**. Usually, a significance level of $\alpha = 0.05$ (5%) is used, implying a confidence level of 0.95 (95%). The interpretation of the *p*-value from a Shapiro–Wilk test is as follows:

- If ***p*-value** $<$**0.05**: the distribution of your data is *significantly different from a normal distribution.*
- If ***p*-value** $\geq$**0.05**: the distribution of your data is *not significantly different from a normal distribution.*
- **Higher *p*-values** mean that you have less evidence that the distribution is *significantly different from a normal distribution.*

For instance, if we use a statistical software and perform the **Shapiro–Wilk** test for normality, using the data from Examples 6.3 and 8.1, we obtain a *p*-value of 0.0305 (using Action Stat software). Since this value is lower than 0.05, we conclude that, at the 5% significance level (95% confidence level), *the distribution of our data is significantly different from a normal distribution*. The major conclusion is that, if you want to do statistical analyses that involve confidence intervals, hypothesis testing, regression, and correlation theory, you should not use parametric tests, because your data do not follow a normal distribution. However, if you had considered a more rigorous significance level, say, 1% (99% confidence level), your conclusion would be different, since 0.0305 is greater than 0.01. Still, if you had used other tests, besides the Shapiro–Wilk, you could have arrived at different *p*-values, and the interpretation would be dependent on your judgement. In this case, also include in your analyses the graphical interpretation of the normal probability and Q–Q plots and the skewness coefficient.

8.3 LOG-NORMAL DISTRIBUTION

8.3.1 Basic concepts about the log-normal distribution

Advanced

We have already emphasized several times in this book that it is common to see a lack of symmetry in the frequency distributions of environmental data and, in our case, treatment plant and water quality monitoring data. We have mentioned that these distributions are commonly *skewed to the right* (positive skewness – see Figure 8.12). A theoretical probability distribution that has been reported to fit well to this type of data is the **log-normal distribution**.

Before we delve into the log-normal distribution, let us make a **warning** about not giving attention to the shape of the distribution, and assuming, without questioning, the *usual normal distribution* (as is frequently done by many researchers). If the *assumption of normality does not hold*, significance levels may be distorted, the methods may suffer loss of statistical power, and the estimates obtained based on this assumption may be severely inaccurate. Users frequently disregard the assumptions required for parametric tests, and their results are likely to be correspondingly incorrect or unreliable (Potvin & Roff, 1993; Modarres *et al.*, 2005).

Theoretically, **positive skewness** (distribution towards higher values) is to *be expected with monitoring data* since there is usually a lower bound on concentration data (lower limit equal to zero or the detection limit, with no negative values), but there are no upper limits. In fact, many wastewater treatment research

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

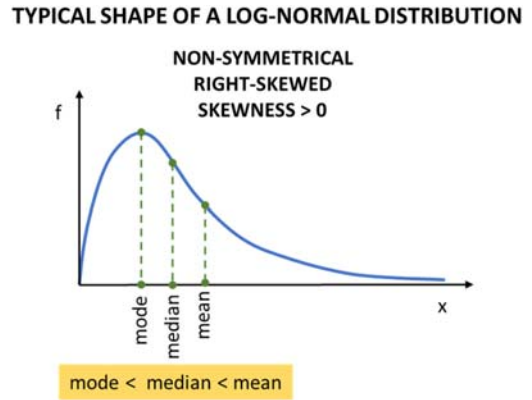


Figure 8.12 Typical shape of relative frequency of log-normally distributed data.

studies relate that, in most cases, the log-normal distribution provided a reasonable description of the effluent BOD (biochemical oxygen demand) and TSS (total suspended solids) concentration data (Dean & Forsythe, 1976a, b; Niku *et al.*, 1979, 1981 and 1982; Berthouex & Hunter, 1981, 1983; Metcalf & Eddy, 2003; Charles *et al.*, 2005). Oliveira *et al.* (2012) expanded this statement to other wastewater constituents, both in the influent and in the effluent of treatment plants.

The **implications of non-normality or lack of symmetry** in water quality and treatment performance evaluation are (for further details, see Oliveira *et al.*, 2011):

S. 5.6
S. 5.7
S. 9.6
S. 9.7
S. 9.8

- Implications for measurements of **central tendency** (widely discussed in this chapter, and also in Section 5.6).
- Implications for measurements of **data dispersion** (also discussed here and in Section 5.7).
- Implications for the assessment of the compliance with **water quality** and **effluent quality standards** (see Section 9.6).
- Implications for **reliability assessment** (see Section 9.7).
- Implications for **quality control charts** (see Section 9.8).

The probability density function of the **log-normal distribution** is defined by the following *two parameters* (Limpert *et al.*, 2001):

- **Geometric mean (μ_g).** *Scale parameter.* Equal to the median. A change in μ_g affects the scaling in horizontal and vertical directions but does not affect σ_g.
- **Geometric standard deviation (σ_g)** or multiplicative standard deviation**.** *Shape parameter.* Increasing values of σ_g imply increased skewness, the mode approaches zero, but the median value does not change.

In this section, we use the following representation:

- *Theoretical log-normal probability distribution of a population*: **geometric mean** and **geometric standard deviation** are represented by $\boldsymbol{\mu_g}$ and $\boldsymbol{\sigma_g}$, respectively.
- *Data sample*: **geometric mean** and **geometric standard deviation** are represented by $\boldsymbol{M_g}$ and $\boldsymbol{s_g}$, respectively.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

S. 5.6
S. 5.7

The concept of **geometric mean** (M_g) of a **data set** was explained and exemplified in Section 5.6, while the concept of **geometric standard deviation** (s_g) was presented in Section 5.7. In these sections, we saw that we need to take the *log_{10} values of the original data*, and calculate, from these log-transformed data, their arithmetic mean and standard deviation. After that, we calculate the geometric mean and geometric standard deviation by

$$\textbf{Geometric mean } M_g = 10^{(\textbf{mean of the } \log_{10} \textbf{transformed data})} \tag{8.2}$$

$$\textbf{Geometric standard deviation } s_g = 10^{(\textbf{standard deviation of the } \log_{10} \textbf{transformed data})} \tag{8.3}$$

The **geometric mean** has values that are **greater than 0** and the **geometric standard deviation** has values that are **greater than 1**.

Note that we have made here the log-transformation using **$\log_{10}$**, and then we reconvert it to the original base for calculating M_g by using **power 10** (as in Equation 8.2). We could use the natural log transformation using **LN** instead of $\log_{10}$ and do the back transformation using **EXP** (number 'e', equal to EXP(1) = 2.718) instead of 10. A similar comment applies to s_g. We need to be coherent on the log base that is used.

Advanced

8.3.2 Influence of geometric mean and geometric standard deviation on the log-normal distribution

Excel

In order to understand the influence of geometric mean ($\boldsymbol{\mu_g}$) and geometric standard deviation ($\boldsymbol{\sigma_g}$) on the log-normal distribution, let us use the *Excel spreadsheet made for generating a log-normal probability distribution*. Use this spreadsheet by varying the values of $\boldsymbol{\mu_g}$ and $\boldsymbol{\sigma_g}$ and analyse the resulting graphs of the log-normal distribution, together with other useful information.

A first example can be with *varying values of the geometric mean* $\boldsymbol{\mu_g}$ (assuming that it is equal to the median in the log-normal distribution), and *fixed values of the geometric standard deviation* $\boldsymbol{\sigma_g}$. Let us compare the resulting log-normal distributions with the following input data: (i) ***geometric mean*** = 100, ***geometric standard deviation*** = 1.5; (ii) *geometric mean* = 200, *geometric standard deviation* = 1.5; and (iii) *geometric mean* = 300, *geometric standard deviation* = 1.5. The three **probability density functions** are shown together in Figure 8.13. You can see that changing μ_g affects the scaling in horizontal and vertical directions.

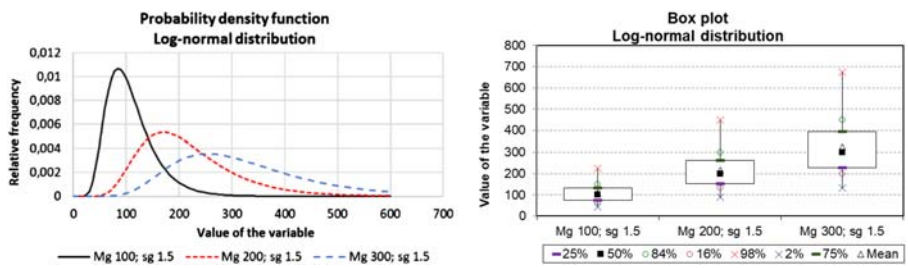


Figure 8.13 Plot of three probability density functions of the log-normal distribution, together with their corresponding box-plots, for different values of the geometric mean $\boldsymbol{\mu_g}$ and the same values of the geometric standard deviation $\boldsymbol{\sigma_g}$.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

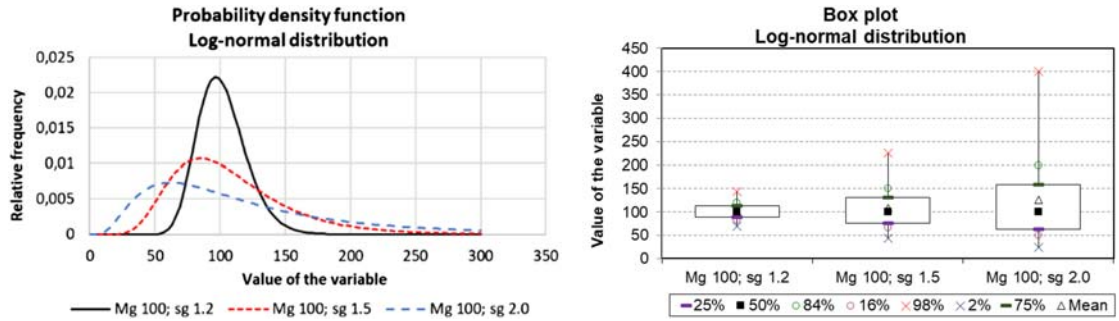


Figure 8.14 Plot of three probability density functions of the log-normal distribution, together with their corresponding box-plots, for the same values of the geometric mean $\boldsymbol{\mu_g}$ and different values of the geometric standard deviation $\boldsymbol{\sigma_g}$.

A second example can be with *fixed values of the geometric mean* $\boldsymbol{\mu_g}$ and *varying values of the geometric standard deviation* $\boldsymbol{\sigma_g}$. Let us now compare the resulting log-normal distributions with the following input data: (i) ***geometric mean*** $= 100$; ***geometric standard deviation*** $= 1.2$; (ii) *geometric mean* $= 100$; *geometric standard deviation* $= 1.5$; and (iii) *geometric mean* $= 100$; *geometric standard deviation* $= 2.0$. The three **probability density functions** are shown together in Figure 8.14. You can see that the shapes of the distributions are different (because the geometric standard deviations are different), and their median is the same (see in the box-plots). The distribution with the larger $\boldsymbol{\sigma_g}$ spreads more than the distribution with the lower $\boldsymbol{\sigma_g}$. Also note that the distribution becomes more asymmetrical with larger values of $\boldsymbol{\sigma_g}$ (in the box-plots, the upper values are more distant to the median than the lower values). Also note, in the box-plots, that the arithmetic means become higher than the median with increasing $\boldsymbol{\sigma_g}$.

Using the box plots, and analysing the lack of symmetry around the mean, we may add the following comment, regarding *graphs based purely on mean* $\pm$ *standard deviation bars* (*e.g.*, a bar plot where the bars show the mean and equal vertical bars show the standard deviation, symmetrically situated around the mean). As we are repeatedly seeing here, we should not force representations that imply symmetry if our data are non-symmetrical around the mean. Therefore, you are much better off **using box-plots instead of mean** $\pm$ **standard deviation plots**, because the box plots can provide a much better visual representation of any asymmetry in your data.

S. 5.2

Likewise, we can repeat here the comment made in Section 5.2 about the **inconvenience of reporting values of mean and standard deviation as mean** $\pm$ **standard deviation**. Frequently, summary tables report mean ($\bar{x}$) and standard deviation (s) in the form of $\bar{x} \pm s$. The comment here applies to the symbol $\pm$. When it is placed like this, some readers might misinterpret this syntax to mean that the data are distributed symmetrically around the mean, and that the standard deviation is an indicator of the variability in the data in a symmetrical way, below and above the mean. This of course is not true, since you cannot use the standard deviation to infer anything about the symmetry of the data – however it could be used to indicate the uncertainty you have in your estimate of the *true mean*, which is symmetrical (i.e., the sampling distribution of means is normal, in accordance with the central limit theorem). Still, we do not want to suggest that the variability in the *population* will be symmetrical around the mean, as would occur with a normal distribution. If you have to report mean and standard deviation in the same cell of your table, consider using the notation $\bar{x}(s)$, that is, with the *standard deviation inside parentheses*.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

8.3.3 Generation of values for the log-normal distribution

Advanced

The equations for returning the values of the probability density function of the log-normal distribution are presented in some statistical textbooks. We will present only Excel functions, because the mathematical formulae are too complex for the purposes of this book. As mentioned before, you should consult statistical textbooks for a complete view on this matter.

The first thing you need to do is to specify the values of the geometric mean $\boldsymbol{\mu_g}$ and geometric standard deviation $\boldsymbol{\sigma_g}$. To obtain these statistics from your data set (in this case, M_g and s_g), see Equations 8.2 and 8.3. After this, calculate the log-transformation of these values by taking their **natural logarithms**: **LN(M_g)** and **LN(s_g)**. For instance, for $\boldsymbol{M_g = 100}$ and $\boldsymbol{s_g = 2.0}$, we have LN(M_g) = 4.61 and LN(s_g) = 0.69. You can then use these values directly in the Excel functions for the log-normal distribution.

Note that there are **different variants of Excel functions** for the log-normal distribution, and you should consult the manual of your version to select the correct function for your application. These two functions are relevant for us.

- **LOGNORM.DIST function**. *Returns the log-normal cumulative distribution*. You provide the value of the variable for which you want to calculate the frequency, plus the values of LN(M_g) and LN(s_g) of your data, and specify whether you want a cumulative or non-cumulative distribution, then you obtain the corresponding value of the relative frequency. For example, suppose LN(M_g) = 4.61 and LN (s_g) = 0.69. For a value of your variable equal to 120, the relative frequency according to the log-normal distribution is 0.0046 (make sure to specify FALSE in the Excel function syntax, meaning that you do not want the cumulative distribution). For a cumulative frequency, you obtain the value of 0.604, meaning that 60.4% of your data have a value ≤120 (the cumulative probability ranges from 0 to 1).
- **LOGNORM.INV function**. *Returns the inverse of the log-normal cumulative distribution*. You provide the cumulative probability (a value between 0 and 1) and the values of LN(M_g) and LN(s_g) from your data set, then you obtain the corresponding value of the variable. For instance, suppose again that LN(M_g) = 4.61 and LN(s_g) = 0.69. For a cumulative frequency of 0.75 (75%), you will obtain the value of 160 for your variable, meaning that approximately 75% of future data points will have values ≤160. It is important to note that your predictions for LN(M_g) and LN(s_g) from your data set are not the true values of the underlying distribution – they are simply predictions. This means that it is not a guarantee that 75% of future data points will fall in this range ≤160. You should consider the uncertainty associated with your estimate of the mean (e.g., using its confidence interval).

Excel

If you have a good statistical software, it will probably have a function that can be used to generate and manipulate log-normal distributions. If you do not have one, you can use the Excel spreadsheet already mentioned in Section 8.3.2:

S. 8.3.2

8.3.4 Fitting a log-normal distribution to your data

Advanced

If you want to **fit a log-normal distribution to your data**, you can follow similar steps to those employed for the normal distribution (see Section 8.2.7 and Example 8.1). The difference is that you have to calculate the $\log_{10}$ values of your data and work with them. You then calculate the mean and standard deviation of your log-transformed data. With these values, and using Equations 8.2 and 8.3, you calculate the geometric mean and geometric standard deviation. With this information in hand, you use the Excel function LOGNORM.DIST with the values of your variable X, for which you want the log-normal frequency distribution to be calculated. You may use the same Excel spreadsheet used in Example 8.1, because it is also set for fitting a log-normal distribution to these data.

S. 8.2.7

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Example

EXAMPLE 8.2 FITTING A LOG-NORMAL DISTRIBUTION TO YOUR FREQUENCY POLYGON

In Example 6.3 you calculated the frequency distribution and plotted the frequency histogram of the data below. In Example 6.4 you plotted the frequency polygon, and in Example 8.1 you fitted a normal distribution to your data. Now, fit a log-normal distribution to your frequency polygon, plot both distributions and make a visual interpretation.

Excel

Note: this example is also available as an Excel spreadsheet.

Data:

2.8	4.2	3.9	3.3	2.8	1.7	1.9	2.5	3.1	3.8	2.7	4.1
4.3	4.8	5.1	5.2	3.9	3.5	2.7	3.1	2.8	3.4	4.9	2.8
2.8	1.8	2.1	2.6	2.3	2.4	2.5	1.3	2.9	2.4	2.1	3.6

Solution:

You calculate the $\log_{10}$ values of your original data, and obtain the following table:

0.447	0.623	0.591	0.519	0.447	0.230	0.279	0.398	0.491	0.580	0.431	0.613
0.633	0.681	0.748	0.763	0.591	0.544	0.431	0.491	0.447	0.531	0.690	0.447
0.447	0.255	0.322	0.415	0.362	0.380	0.398	0.255	0.462	0.380	0.322	0.556

The calculated values of the mean and standard deviation of the log-transformed data are:

- Mean of log-transformed data = **0.478**.
- Standard deviation of log-transformed data = **0.138**.

The geometric mean and geometric standard deviation are:

- Geometric mean $= 10^{(\text{mean of log-transformed data})} = 10^{0.478} =$ **3.01 mg/L**.
- Geometric standard deviation $= 10^{(\text{standard deviation of log-transformed data})} = 10^{0.138} =$ **1.37 mg/L**.

You then calculate the natural logarithm of the geometric mean and geometric standard deviation:

- LN(geometric mean) = LN(3.01) = **1.100** (same as 0.478 × LN(10)).
- LN(geometric standard deviation) = LN(1.37) = **0.317** (same as 0.138 × LN(10)).

Note that we could have used only LN in this example but wanted to show you the conversions from one log base to another.

Set up a computational table with 100 class intervals (100 rows), starting from 0.5 mg/L (which is the lowest mid-point value of the frequency polygon table you structured in Example 6.4) and going up to 6.5 mg/L (which is the highest mid-point value in the same table). Since there are 100 intervals, the width of each class interval or the increment in the values of the variable in the computational table will be $(6.5 - 0.5)/100 = 0.06$ mg/L.

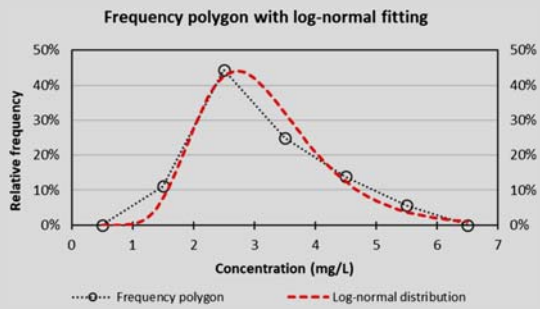
The last column will use the LOGNORM.DIST Excel function in the following way:

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

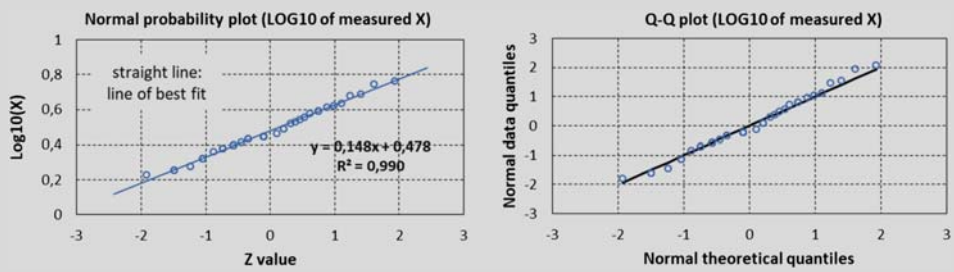
LOGNORM.DIST (X_i value in the cell to the left; LN(geometric mean); LN(geometric standard deviation); FALSE)

Interval Number	X_i Value of the Variable (values in *X*-axis)	Syntax	Value of the Log-normal Distribution
1	0.500	LOGNORM.DIST (0.500; 1.100;0.317; FALSE)	2.79011×10^{-7}
2	0.560	LOGNORM.DIST (0.560; 1.100;0.317; FALSE)	1.76836×10^{-6}
3	0.620	LOGNORM.DIST (0.620; 1.100;0.317; FALSE)	8.32684×10^{-6}
4	0.680	LOGNORM.DIST (0.680; 1.100;0.317; FALSE)	3.10747×10^{-5}
…	…	…	…
99	6.440	LOGNORM.DIST (6.440; 1.100;0.317; FALSE)	0.010853984
100	6.500	LOGNORM.DIST (6.500; 1.100;0.317; FALSE)	0.010018868

The log-normal distribution plot, superimposed over the frequency polygon made in Example 6.4, results in the following plot. You analyse it and consider that, at least visually speaking, the fitting of the log-normal distribution seems to be good. In fact, it appears to be better than that of the normal distribution (Example 8.1).



If you take the $\log_{10}$ of your original data, you can make a normal probability plot or a Q–Q plot of your log-transformed data. For instance, the resulting plots (structured in the spreadsheet associated with Example 8.1) are plotted below. If you compare them with the plots for the original, non-transformed data, shown in Figure 8.10, you will see that the log transformation led to a better adjustment to the straight-line, suggesting that the log-normal distribution likely provides a better representation of your data compared with the normal distribution.



Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

8.3.5 Measures of central tendency and variation in the log-normal distribution

Advanced

The measures of central tendency and variation for a **log-normal distribution** are given by the following equations (Naghettini & Pinto, 2007; www.bertolo.pro.br/FinEst/Estatistica/Planilhas/distribs.htm, accessed 2019).

The **notation** for the theoretical log-normal probability function is

μ = arithmetic mean
σ = arithmetic standard deviation
CV = arithmetic coefficient of variation ($\sigma \div \mu$)
μ_g = geometric mean
σ_g = geometric standard deviation

Measures of central tendency in a log-normal distribution:

- *Equality between geometric mean (μ_g) and median:*

$$\mu_g = \text{median} \tag{8.4}$$

- *Arithmetic mean (μ) for given values of the geometric mean (μ_g) and geometric standard deviation (σ_g):*

$$\mu = \text{EXP}\left(\text{LN}(\mu_g) + \frac{[\text{LN}(\sigma_g)]^2}{2}\right) \tag{8.5}$$

- *Arithmetic mean (μ) for given values of geometric mean (μ_g) and CV:*

$$\mu = \mu_g \times \sqrt{\text{CV}^2 + 1} \tag{8.6}$$

- *Geometric mean (μ_g) for given values of the arithmetic mean (μ) and CV:*

$$\mu_g = \frac{\mu}{\sqrt{\text{CV}^2 + 1}} \tag{8.7}$$

- *Mode for given values of the geometric mean (μ_g) and geometric standard deviation (σ_g):*

$$\text{Mode} = \text{EXP}[\text{LN}(\mu_g) - [\text{LN}(\sigma_g)]^2] \tag{8.8}$$

Measures of variation in a log-normal distribution:

- *Arithmetic standard deviation (σ) for given values of μ_g and σ_g:*

$$\sigma = \sqrt{\text{EXP}[2 \times \text{LN}(\mu_g)] \times \{\text{EXP}[2 \times [\text{LN}(\sigma_g)]^2] - \text{EXP}[\text{LN}(\sigma_g)]^2\}} \tag{8.9}$$

- *Coefficient of variation CV ($= \sigma \div \mu$) for a given value of σ_g:*

$$\text{CV} = \sqrt{\text{EXP}\{[\text{LN}(\sigma_g)]^2 - 1\}} \tag{8.10}$$

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

- *Geometric standard deviation (σ_g) for a given value of CV ($= \sigma \div \mu$):*

$$\sigma_g = \mathrm{EXP}\left(\sqrt{\mathrm{LN}(\mathrm{CV}^2 + 1)}\right) \tag{8.11}$$

- *Geometric standard deviation (σ_g) for given values of the cumulative distribution function of 0.1587 and 0.8413 (percentiles 15.87% and 84.13%, which correspond to $\bar{\mu} - 1\sigma$ and $\bar{\mu} + 1\sigma$ in the normal distribution):*

$$\sigma_g = \frac{\mu_g}{P_{15.87}} = \frac{P_{84.13}}{\mu_g} \tag{8.12}$$

- *Skewness coefficient for a given value of CV:*

$$\mathrm{Skewness} = 3 \times \mathrm{CV} + \mathrm{CV}^3 \tag{8.13}$$

Note the **influence of the parameter CV** on the parameters of the log-normal distribution (geometric mean and geometric standard deviation). Ott (1995) states that for CV values less than 1/6 (=0.167), the probability density function of the lognormal distribution shows a very similar behaviour to the normal distribution. Indeed, if we calculate $\mathrm{CV}^2 + 1$, that is part of several of the above equations, for CV = 0.167, we obtain 1.028, which is very close to 1.0, typical of the normal distribution.

The question is then: **what are typical values of the coefficient of variation CV**? It is not difficult to obtain this information based on simple monitoring of treatment plants and water quality. Oliveira and von Sperling (2008), investigating 166 full-scale **wastewater treatment plants**, obtained mean CV values ranging from 0.3 to 1.0 for effluent concentrations of BOD, COD, TSS, TN, and TP, for six different treatment technologies. However, for coliforms, as expected, CV values were higher, ranging from around 1.0 up to 3.0. For all parameters investigated, *log-normal distribution provided a better fit than the normal distribution* (Oliveira *et al.*, 2012). As an additional comment, if we use Equation 8.11 to convert these values of CV into geometric standard deviation s_g, we obtain the following results: for CVs of 0.3, 1.0, and 3.0, we get s_g values of 1.3, 2.3, and 4.6. Melo (2019) investigated 45 **water treatment plants** in Brazil and obtained CV values between **0.2 and 0.8** for effluent turbidity. These CV values correspond to geometric standard deviations of 1.2 to 2.0.

Example

EXAMPLE 8.3 PARAMETERS OF THE LOG-NORMAL DISTRIBUTION

For given values of the arithmetic mean ($\mu = 127.15$) and arithmetic standard deviation ($\sigma = 99.86$), calculate the corresponding parameters of the log-normal distribution using Equations 8.4–8.12. We use such specific values in this example because they will lead to round figures of geometric mean and standard deviation, as you will see below.

Excel

S. 8.3.2

Note: You can use the spreadsheet for the log-normal distribution mentioned in Section 8.3.2.

Solution:

The coefficient of variation (CV) is

$$\mathrm{CV} = \frac{\text{standard deviation}}{\text{mean}} = \frac{\sigma}{\mu} = \frac{99.86}{127.15} = \mathbf{0.785}$$

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

The geometric mean is (Equation 8.7)

$$\mu_g = \frac{\mu}{\sqrt{CV^2 + 1}} = \frac{127.15}{\sqrt{0.785^2 + 1}} = \mathbf{100.00}$$

The geometric standard deviation is (Equation 8.11)

$$\sigma_g = EXP\left(\sqrt{LN(CV^2 + 1)}\right) = EXP\left(\sqrt{LN(0.785^2 + 1)}\right) = EXP\left(\sqrt{0.480}\right) = \mathbf{2.00}$$

If we use the Excel function LOGNORM.INV to obtain the values of the cumulative density function for the percentiles 15.87 and 84.13, we obtain the values of 50.00 and 200.00, respectively. Using Equation 8.12, with $\mu_g = 100.00$ (calculated above), we obtain:

$$\sigma_g = \frac{\mu_g}{P_{15.87}} = \frac{P_{84.13}}{\mu_g} = \frac{100.00}{50.00} = \frac{200.00}{100.00} = \mathbf{2.00}$$

The mode is (Equation 8.8):

$$\text{Mode} = EXP[LN(\mu_g) - [LN(\sigma_g)]^2] = EXP[LN(100.00) - [LN(2.00)]^2] = EXP(4.12)$$
$$= 61.85$$

Advanced

S. 8.2.5

For the *normal distribution* (Section 8.2.5), we saw that the dispersion of the data around the mean μ for different quantities of standard deviation σ (standard normal variable Z) depended on an additive relationship: $\mu \pm \sigma$. For the **log-normal distribution**, the relations are multiplicative: $\boldsymbol{\mu_g \times / \div \sigma_g}$. Simply stated, we can say that:

- whatever is 'addition' in normal distribution, is 'multiplication' in log-normal distribution;
- whatever is 'subtraction' in normal distribution, is 'division' in log-normal distribution;
- whatever is 'multiplication' in normal distribution, is 'raising to a power' in log-normal distribution

Having said that, we present in Table 8.2 the percentage of the data that is included inside different ranges of dispersion around the geometric mean, expressed as $\mu_g \times / \div \sigma_g$. Table 8.3 shows that if you have a data set and you estimate the values of M_g and s_g, you can use those values to estimate approximately what percentage of future data points will fall within each range. Note that the last two columns have the same values as in Table 8.1 for normal distribution. Figure 8.15 shows the example of a log-normal

Table 8.2 Values of the log-normal cumulative function for different intervals expressed as $\mu_g \times / \div \sigma_g$.

Range	$\mu_g \times / \div \sigma_g$	Value of Normal Cumulative Function (percentile)	Percentage of Area (data points) Inside Range
$\mu_g \times / \div (\sigma_g)^1$	$\mu_g \div (\sigma_g)^1$	15.87	68.27
	$\mu_g \times (\sigma_g)^1$	84.13	
$\mu_g \times / \div (\sigma_g)^2$	$\mu_g \div (\sigma_g)^2$	2.28	95.45
	$\mu_g \times (\sigma_g)^2$	97.72	
$\mu_g \times / \div (\sigma_g)^3$	$\mu_g \div (\sigma_g)^3$	0.13	99.73
	$\mu_g \times (\sigma_g)^3$	99.87	

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Table 8.3 Values of the log-normal cumulative function estimated for different intervals using M_g and s_g.

Range	$\mu_g \times/\div \sigma_g$	Value of normal cumulative function (percentile)	Approximate percentage of future data points inside range
$M_g \times/\div (s_g)^1$	$M_g \div (s_g)^1$	15.87	~68
	$M_g \times (s_g)^1$	84.13	
$M_g \times/\div (s_g)^2$	$M_g \div (s_g)^2$	2.28	~95
	$M_g \times (s_g)^2$	97.72	
$M_g \times/\div (s_g)^3$	$M_g \div (s_g)^3$	0.13	~99.7
	$M_g \times (s_g)^3$	99.87	

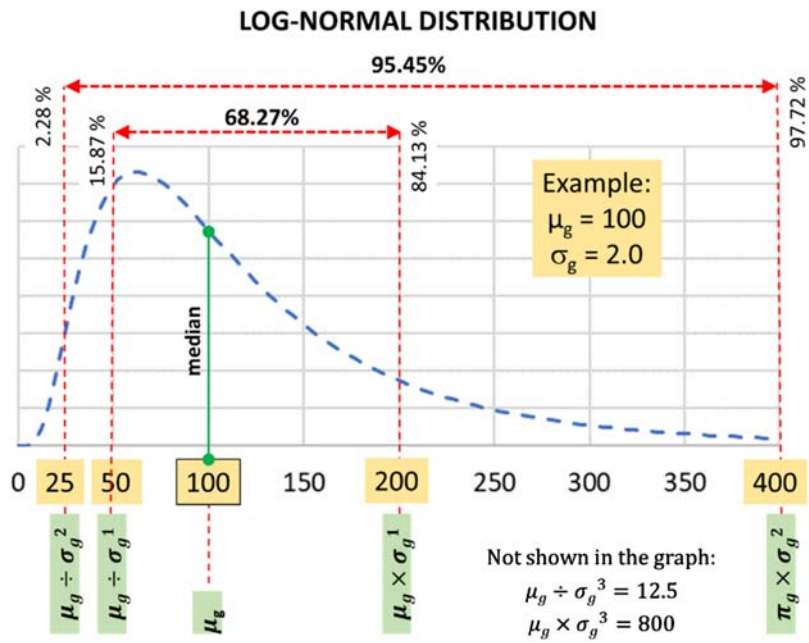


Figure 8.15 Plot of the log-normal distribution for a geometric mean $\mu_g = 100$ and a geometric standard deviation $\sigma_g = 2.0$, showing the correspondence between different intervals of $\mu_g \times/\div \sigma_g$. The graph also shows the percentage of the area under the curve that is contained within each interval.

distribution with geometric mean $\mu_g = 100$ and geometric standard deviation $\sigma_g = 2.0$ to help you in understanding these relationships.

The interpretation of Figure 8.15 is as follows. If you have a population with a geometric mean of $\boldsymbol{\mu_g} = \mathbf{100}$ and a geometric standard deviation of $\boldsymbol{\sigma_g} = \mathbf{2.0}$ of a log-normally distributed variable, **68.27%** of the data will be inside the interval of 50 and 200, since $100/(2.0)^1 = 50$ and $100 \times (2.0)^1 = 200$. In the same way, **95.45%** of the data will be inside the interval of 25 and 400, since $100/(2.0)^2 = 25$ and $100 \times (2.0)^2 = 400$. Finally, **99.73%** of the data will be inside the interval of 12.5 and 800, since $100/(2.0)^3 = 12.5$ and $100 \times (2.0)^3 = 800$. Note that the dispersion is not symmetrical around the geometric mean (σ_g = median). The lower values are much closer to the median, while the upper values are far away, situated in the tail of the distribution.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Let us make a *correspondence between log-normal distribution and normal distribution.* If we take the **natural logarithm (LN)** of the geometric mean μ_g and geometric standard deviation σ_g, we obtain LN (100) = 4.605 and LN(2.0) = 0.693. By taking the LN of the original values of the log-normal distribution, we have transformed it into a normal distribution. We can now use the additive expressions from Table 8.1 (normal distribution) and calculate the percentage of values that will be in the range of $\mu \pm 1\sigma$, or between 3.912 and 5.298 ($4.605 - 1 \times 0.693 = 3.912$ and $4.605 + 1 \times 0.693 = 5.298$). Since $Z = 1$ (one standard deviation below and above the mean), **68.27%** of the data will be in this range between 3.912 and 5.298 (same value as the one calculated above for the log-normal distribution and one geometric standard deviation). Similar calculations can be done for $Z = 2$ and $Z = 3$. You can use the spreadsheet for the normal distribution to check this.

Table 8.4 presents a summary of the main parameters and values of interest for a log-normal probability distribution having a mean of one ($\mu = 1$) and different values of the coefficient of variation CV.

Excel

Figure 8.16 contains a plot of the **main parameters of the log-normal distribution** shown in Table 8.4, and it shows how these parameters change with respect to different CV values. These values and the graphs can be accessed by the following Excel spreadsheet, in which CV values are calculated in increments of 0.1. The graphs are general and can be used for any mean value. For instance, if the arithmetic mean of your series is 10 mg/L, you need to multiply all values by 10 (or use the spreadsheet, with the mean value desired). You can clearly see that, as the CV increases, the difference between the arithmetic mean (in

Table 8.4 Summary of the main parameters and values of interest for a log-normal probability distribution having mean of one ($\mu = 1.00$) and different values of the coefficient of variation CV (ranging from 0.00 to 4.00).

Mean μ	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Coefficient of variation CV	0.00	0.50	1.00	1.50	2.00	3.00	4.00
Measures of central tendency							
Geometric mean μ_g	1.00	0.89	0.71	0.55	0.45	0.32	0.24
Median	1.00	0.89	0.71	0.55	0.45	0.32	0.24
Mode	1.00	0.72	0.35	0.17	0.09	0.03	0.01
Ratio Mean/Median	*1.00*	*1.12*	*1.41*	*1.80*	*2.24*	*3.16*	*4.12*
Measures of dispersion							
Standard deviation σ	0.00	0.50	1.00	1.50	2.00	3.00	4.00
Geometric standard deviation σ_g	1.00	1.60	2.30	2.96	3.56	4.56	5.38
Values for different cumulative frequencies							
0.13%	1.00	0.22	0.06	0.02	0.01	0.00	0.00
2.28%	1.00	0.35	0.13	0.06	0.04	0.02	0.01
15.87%	1.00	0.56	0.31	0.19	0.13	0.07	0.05
50.00%	1.00	0.89	0.71	0.55	0.45	0.32	0.24
84.13%	1.00	1.43	1.63	1.64	1.59	1.44	1.31
97.72%	1.00	2.30	3.74	4.86	5.66	6.58	7.03
99.87%	1.00	3.69	8.59	14.41	20.11	29.99	37.83

Note: for different values of mean (μ), keeping these same values of CV, all values in the table should be multiplied by μ. The exceptions are 'Ratio Mean/Median' and 'Geometric standard deviation σ_g' which remain the same, since they are a function of only CV.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

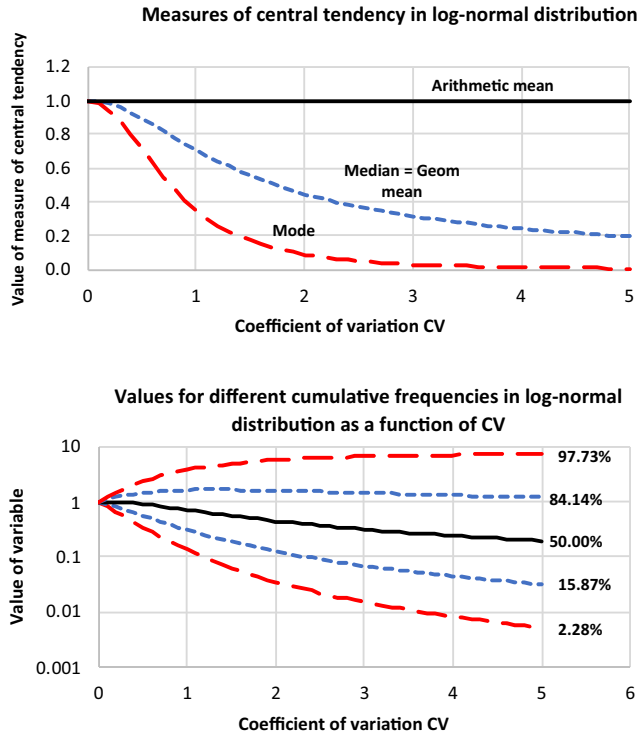


Figure 8.16 Visualization of the main parameters of central tendency and dispersion in a log-normal distribution, for a fixed value of the arithmetic mean (1.0) and varying values of CV (from 0.0 to 5.0). For different values of the mean (μ), simply multiply the values in the *Y*-axis by μ.

Table 8.5 Summary comparison between normal and log-normal distributions

Property	**Distribution**	
	Normal	**Log-normal**
Effects	Additive	Multiplicative
Shape of distribution	Symmetrical	Positively skewed
Characterization:		
Mean	$\bar{x}$ (arithmetic)	M_g (geometric)
Median	$= \bar{x}$	$= M_g$
Standard deviation	s (additive)	s_g (multiplicative)
Measure of dispersion	$CV = s/\bar{x}$	$CV = \sqrt{EXP\{[LN(\sigma_g)]^2 - 1\}}$
Approximate prediction interval		
68%	$\bar{x} \pm 1s$	$M_g \times/\div s_g$
95%	$\bar{x} \pm 2s$	$M_g \times/\div (s_g)^2$
99.7%	$\bar{x} \pm 3s$	$M_g \times/\div (s_g)^3$

Notes: CV = coefficient of variation; $\pm$ plus/minus; $\times/\div$ times/divide.
Source: Adapted from Limpert *et al.* (2001).

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

this example, fixed at a value of 1.0) and the median, geometric mean, and mode increase. In the log-normal distribution, the arithmetic mean is always greater than the median (geometric mean), and the mode is always the lowest measure of central tendency, approaching zero as CV increases (as the peak of the distribution gets close to zero). Also, with increasing CV, the ranges that define $\times/\div\ 1\sigma$ (15.87% and 84.14%) and $\times/\div\ 2\sigma$ (2.28% and 97.73%) increase sharply (the graph is on a log-scale).

8.3.6 Comparison between normal and log-normal distributions

Advanced

Table 8.5 presents a synthesis of the comparison between the normal and log-normal distributions.

8.4 MOMENT MATCHING TO USE OTHER DISTRIBUTIONS

Advanced

In addition to the normal and log-normal distributions, there are many other distributions that may be useful for environmental, water treatment, and water quality monitoring data sets. For example, the **gamma distribution** *does not include negative values* and may be an appropriate distribution to use for low concentrations with large standard deviations. The **beta distribution** is a continuous probability distribution defined between the **interval of 0 and 1** (i.e., *0% and 100%*). As such, it can be useful for modelling the distribution of variables that fall predominantly within these limits. Although removal efficiencies can be negative for some constituents that are produced at the treatment plant (and, as such, should not be analysed in terms of removal) and in treatment plants during extremes cases of malfunctioning, in general they will fall within the limits of 0% and 100%. Therefore, in the present case we will illustrate the *application of beta distribution for removal efficiencies*.

Now that you have some basic knowledge about the mean and standard deviation, we will teach you a technique called **moment matching**, which can be used *to manipulate other distributions to determine parameters of fit for your data set*. To demonstrate this, we will take the example of the **beta distribution**. The beta distribution is defined by **two shape parameters**, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, which must both be *greater than zero*. So, if you have a data set of removal efficiencies, and you want to evaluate it with respect to the beta distribution, you must determine the best fit values for the parameters α and β.

C. 5

S. 8.2

You have previously learned about descriptive statistics such as the mean and the variance (Chapter 5), and we have also discussed the topics of symmetry and skewness (Section 8.2). These important features of a distribution are something known in statistics as moments (specifically, the **mean** is the *first moment*, the **variance** is the *second central moment*, and the **skewness** is the *third normalised moment*).

The mean (μ) and the variance (σ^2) of the beta distribution are given by the following equations:

$$\mu = \frac{\alpha}{\alpha + \beta} \tag{8.14}$$

$$\sigma^2 = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)} \tag{8.15}$$

Solving Equation 8.14 for β gives us the following:

$$\beta = \frac{\alpha(1 - \mu)}{\mu} \tag{8.16}$$

If we then substitute β from Equation 8.16 into Equation 8.15, we can simplify to find that

$$\boxed{\alpha = \frac{\mu^2 - \mu^3 - \mu\sigma^2}{\sigma^2}} \tag{8.17}$$

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Finally, substituting α from Equation 8.17 back into Equation 8.16 gives us the following parameterization of β:

$$\beta = \frac{\mu - 2\mu^2 + \mu^3 - \sigma^2 + \mu\sigma^2}{\sigma^2} \tag{8.18}$$

Therefore, if we have a *sample mean* ($\bar{x}$) *and sample standard deviation* (s)*, we can use those values as estimates for* μ *and* σ, and we can use Equations 8.17 and 8.18 to estimate parameters for the beta distribution that best represent our data set. Beta distribution is covered by Excel function BETA.DIST (). See Example 8.4 that illustrates the application of beta distribution.

Example

EXAMPLE 8.4 FINDING BEST FIT PARAMETERS FOR BETA DISTRIBUTION USING MOMENT MATCHING

Consider the following wastewater treatment plant for which you have influent and effluent concentrations and the resulting removal efficiency of a constituent for a period of 30 consecutive days. Use the techniques you learned in Example 6.3 to construct a relative frequency table for the data and plot a relative frequency polygon. Then, use the techniques you learned in Example 8.1 to fit and plot a normal distribution to this frequency polygon. Finally, use the moment matching technique to fit and plot a beta distribution to the frequency polygon. Overlay the two distributions and decide which one appears to provide a better representation of the data. Use Equations 8.17 and 8.18 to find the best values for the parameters α *and* β *and use the Excel function BETA.DIST the same way you used the function NORM.DIST.*

Excel

Note: This example is also available as an Excel spreadsheet.

Data:

Day	C_{in} (mg/L)	C_{out} (mg/L)	Efficiency (%)
1	1112	70	93.7
2	1201	78	93.5
3	952	136	85.7
4	1034	82	92.1
5	998	81	91.8
6	971	65	93.3
7	780	158	79.7
8	1009	98	90.3
9	1014	112	89.0
10	978	101	89.7
11	1022	18	98.2
12	904	17	98.2
13	1060	62	94.1
14	905	156	82.8

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Day	C_{in} (mg/L)	C_{out} (mg/L)	Efficiency (%)
15	851	35	95.9
16	894	155	82.6
17	1007	96	90.4
18	1131	88	92.2
19	909	120	86.7
20	760	54	92.9
21	957	10	99.0
22	969	77	92.1
23	957	24	97.5
24	1098	83	92.5
25	988	125	87.4
26	1102	55	95.0
27	830	14	98.3
28	1055	115	89.1
29	1012	137	86.5
30	1120	107	90.4

Solution:

Compute the absolute and relative frequency of the data as you have done already in Example 6.3. This time use the following class intervals: $76\% < x \leq 80\%$; $80\% < x \leq 84\%$; $84\% < x \leq 88\%$; $88\% < x \leq 92\%$; $92\% < x \leq 96\%$; and $96\% < x \leq 100\%$.

Class Intervals (mg/L)	Mid-Range Values (%)	Absolute	Relative	Absolute Frequency	Relative Frequency (%)
		Cumulative Frequency	Cumulative Frequency (%)		
$76\% < x \leq 80\%$	78	1	3	1	3
$80\% < x \leq 84\%$	82	3	10	2	7
$84\% < x \leq 88\%$	86	7	23	5	17
$88\% < x \leq 92\%$	90	14	47	9	30
$92\% < x \leq 96\%$	94	25	83	16	53
$96\% < x \leq 100\%$	98	30	100	14	47

Next, if you calculate the mean and standard deviation of the removal efficiency, you should get a value of **91.36% for the mean** and **4.88% for the standard deviation**. Using Equations 8.17 and 8.18, we can estimate a value of 29.35 for α and a value of 2.78 for β.

$$\alpha = \frac{0.9126^2 - 0.9126^3 - 0.9126 \times 0.0488^2}{0.0488^2} = \mathbf{29.35}$$

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

$$\beta = \frac{0.9126 - 2 \times 0.9126^2 + 0.9126^3 - 0.0488^2 + 0.9126 \times 0.0488^2}{0.0488^2} = \mathbf{2.78}$$

Now, set up a computational table with 100 class intervals (100 rows). In the first column, enter the numbers starting from 0.01 (1%) and going up by an increment of 0.01 (1%) in each row until you reach a maximum value of 1 (100%). The width of each class interval or the increment in the values of the variable in the computational table is 0.01 (1%).

For the last column, use the NORM.DIST Excel function as you did in Example 8.1:

NORM.DIST (X_i value in the cell to the left; mean = 0.9126; standard deviation = 0.0488; FALSE)

Interval Number	X_i Value of the Variable (values in *X*-axis)	Syntax	Value of the Normal Distribution
1	0.01	NORM.DIST(0.01; 0.9126; 0.0488; FALSE)	3.62164×10^{-74}
2	0.02	NORM.DIST(0.02; 0.9126; 0.0488; FALSE)	1.5689×10^{-72}
3	0.03	NORM.DIST(0.03; 0.9126; 0.0488; FALSE)	6.51731×10^{-71}
4	0.04	NORM.DIST(0.04; 0.9126; 0.0488; FALSE)	2.59614×10^{-69}
…	…	…	…
99	0.99	NORM.DIST(0.99; 0.9126; 0.0488; FALSE)	2.3989
100	1.00	NORM.DIST(1.00; 0.9126; 0.0488; FALSE)	1.7047

Then, do the same in the next table for the beta distribution, using the BETA.DIST Excel function:

BETA.DIST (X_i value; alpha = 29.35; beta = 2.78; FALSE)

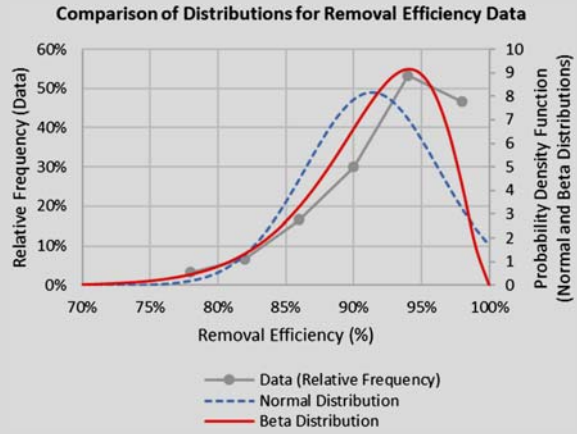
Interval Number	X_i Value of the Variable (values in *X*-axis)	Syntax	Value of the Beta Distribution
1	0.01	BETA.DIST(0.01; 29.35; 2.78; FALSE)	1.57332×10^{-53}
2	0.02	BETA.DIST(0.02; 29.35; 2.78; FALSE)	5.26854×10^{-45}
3	0.03	BETA.DIST(0.03; 29.35; 2.78; FALSE)	5.07095×10^{-40}
4	0.04	BETA.DIST(0.04; 29.35; 2.78; FALSE)	1.73175×10^{-36}
…	…	…	…
99	0.99	BETA.DIST(0.99; 29.35; 2.78; FALSE)	1.6495
100	1.00	BETA.DIST(1.00; 29.35; 2.78; FALSE)	0

Note: The values in the tables have been calculated using Excel, and there may be small rounding differences if you calculate the distribution parameters in a calculator and rounds the values.

The plots with the normal and beta distributions, superimposed over the frequency polygon, result in the following plot shown below. You analyse it and consider that, at least visually speaking, in this case the beta distribution follows the main trends of the data set and reproduces the peak value *better than the normal distribution*. Furthermore, based on your knowledge of the normal distribution, you know that the full distribution extends beyond the maximum possible value of 100%, while the beta distribution comes down to zero at a value of

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

1. Again, please note that to definitively determine whether or not the beta distribution is a good fit for this data set, you should use goodness-of-fit tests.



8.5 CHECK-LIST FOR YOUR REPORT

- ✓ Check whether the distribution of your data follows the basic assumptions for the normal distribution (e.g., symmetry), so that you know whether certain statistical tests that depend on data normality or symmetry can be used. If the data do follow a normal distribution, include in your report whether you have done any data transformations and include the results of a goodness-of-fit test for normality.
- ✓ Check that any frequency distributions of concentrations, flows, and loads have a minimum value of zero (no negative values) and that any frequency distributions of removal efficiencies (%) have a maximum value of 100% (no values $>$100%). Consider using the beta distribution for removal efficiencies that are very close to 100%.
- ✓ Consider not representing mean and standard deviation in summary tables as $\bar{\boldsymbol{x}} \pm \mathbf{s}$., because some readers may misinterpret this notation as an implication that your data are symmetrical around the mean (which may not be true). We recommend using an alternative way to show the same information, using the notation $\bar{\boldsymbol{x}}(\mathbf{s})$.
- ✓ Likewise, we recommend that you give preference to box-plots over simple graphs that show mean $\pm$ standard deviation error bars. Box plots represent well the asymmetry of the data, whereas mean $\pm$ standard deviation graphs may be misinterpreted to imply symmetry in the data (which may not exist).
- ✓ If you are using a log-normal distribution, make it clear whether you are reporting values of arithmetic mean and arithmetic standard deviation or geometric mean and geometric standard deviation.
- ✓ If you have to present geometric means and geometric standard deviations in an oral presentation, make it in such a way that they will be understandable to an audience that may be not familiar with them.

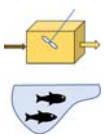
Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Chapter 9

Compliance with targets and regulatory standards for effluents and water bodies

This chapter presents a wide range of material to help you determine how to assess conformity with targets established by managers or with standards specified by regulatory agencies for the quality of water bodies, drinking water, or discharges from wastewater treatment plants. You will learn to use a suite of statistical methods that will provide a complete view of compliance assessment. One-sample parametric and non-parametric hypotheses tests are covered, together with an advanced view on frequency analysis, reliability analysis, and control charts, using the assumptions of both the normal and the log-normal distributions. The concepts here are useful for students, researchers, plant managers, and water, sanitation, and environmental quality regulators.



Most of the contents in this chapter are applicable to both treatment plant monitoring and water quality monitoring.

CHAPTER CONTENTS

9.1 Regulatory Standards and Targets for Treatment Plant Effluents and the Quality of Drinking Water and Ambient Water Bodies **242**
9.2 Graphical Methods for Comparing Monitored Data with Quality Standards **243**
9.3 Evaluation of Compliance Based on Average Values. **247**
9.4 Evaluation of Compliance Based on the Proportion of Non-conformity with Standard Using *Z*-test For Proportions. **257**
9.5 Probabilities of Conformity or Non-conformity Obtained Directly from the Monitoring Data **260**
9.6 Estimation of Compliance with the Standard Based on Frequency Analysis Using Normal and Log-normal Distributions **263**
9.7 Reliability Analysis **270**
9.8 Control Charts. **281**
9.9 Check-List for Your Report **315**

© 2020 The Authors. This is an Open Access book chapter distributed under the terms of the Creative Commons Attribution Licence (CC BY-NC-ND 4.0), which permits copying and redistribution for non-commercial purposes with no derivatives, provided the original work is properly cited (https://creativecommons.org/licenses/by-nc-nd/4.0/). This does not affect the rights licensed or assigned from any third party in this book. The chapter is from the book Assessment of Treatment Plant Performance and Water Quality Data: A Guide for Students, Researchers and Practitioners, Marcos von Sperling, Matthew E. Verbyla and Sílvia M. A. C. Oliveira (Authors).
doi: 10.2166/9781780409320_0241

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

9.1 REGULATORY STANDARDS AND TARGETS FOR TREATMENT PLANT EFFLUENTS AND THE QUALITY OF DRINKING WATER AND AMBIENT WATER BODIES

Basic

If you are going to monitor treatment plant performance or water quality in full-scale systems, you almost certainly will need to assess *compliance with quality standards or regulations*. In this chapter, we cover the analysis of compliance with the following types of standards or specifications (see Figure 9.1):

- Quality of **water** from a water supply or drinking water treatment plant
- Quality of **treated effluent** from a wastewater treatment plant (to be discharged or reused)
- Quality of ambient water in a **water body** (e.g., for recreational use)

These specifications may be in the form of **standards** or **target** values:

- **Quality standard**: They are usually specifications dictated by **legislation** at the *municipal, state, regional, or country levels* and enforced by health organizations, environmental agencies, water management agencies, regulators, or other official entities. Being a legal or official determination, conformity with these regulations is considered **compulsory**.
- **Quality target**: They are usually quality specifications determined by the *water and sanitation service provider, by the industry using the water or producing the wastewater, or other company or organization*, reflecting their view on what should be the **desired quality** (also called 'quality *goals*'), but without the character of an official regulation. Another example could be *guidelines* established by an entity that are not implemented as a legal standard. As such, they are not compulsory, but may be controlled by the entity directly involved in different ways, even including incentives for meeting the targets.

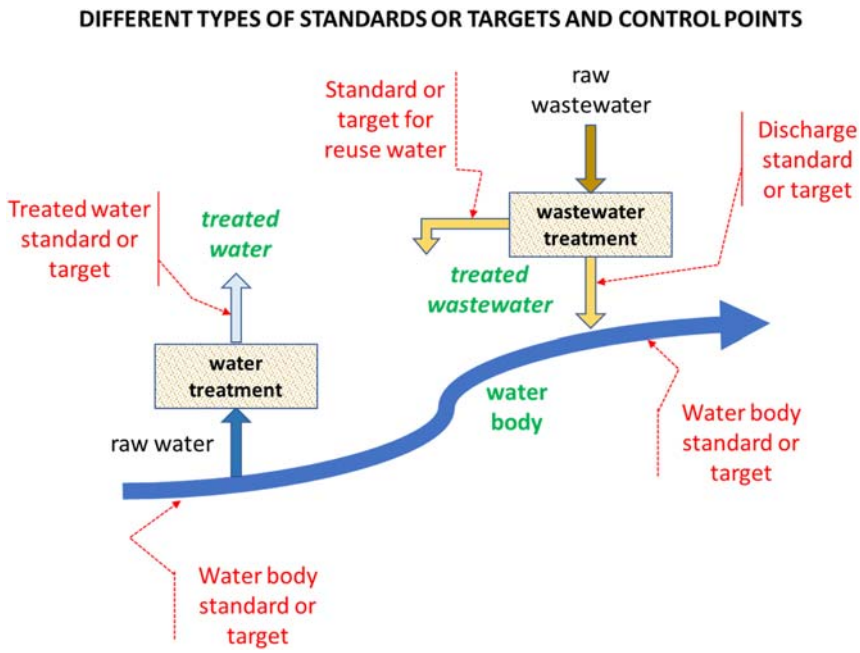


Figure 9.1 Different types of quality standards or targets and related control points.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

The specifications may cover the following variables:

- **Concentrations** of constituents (compounds or pollutants) in treated water, treated wastewater, and water bodies
- **Removal efficiencies** (usually for wastewater treatment plants)

Standards or targets may use the following specifications for the concentrations or removal efficiencies:

- *Maximum or minimum **permissible** values*:
  - maximum permissible values for concentrations of pollutants
  - minimum permissible values for concentrations of desired compounds (e.g., dissolved oxygen in a water body)
  - acceptable ranges (minimum and maximum values) for concentrations (e.g., pH)
  - minimum permissible values for removal efficiencies
- *Maximum or minimum **average** values*:
  - maximum average values for concentrations of pollutants
  - minimum average values for concentrations of desired compounds
  - ranges of minimum and maximum average values for concentrations
  - minimum average values for removal efficiencies
- *Minimum **percentage of samples** that comply* (or maximum percentage of samples that do not comply) with the standard or the target:
  - for concentrations
  - for removal efficiencies

The topics covered in this chapter are

- **Graphical methods** for comparing monitoring data with standards/targets
- Evaluation of compliance based on **average values** using hypothesis testing
- Evaluation of compliance based on the **proportion of samples** in conformity with the standard/target using hypothesis testing
- Determination of **probabilities** based on the monitoring data
- Estimation of compliance with the standard/target based on **frequency analysis** using normal and log-normal distributions
- Estimation of compliance with the standard/target based on **reliability analysis** using normal and log-normal distributions
- Evaluation of **quality control charts** based on your monitoring data and control limits using normal and log-normal distributions

Although there are subtle *differences in the concepts of **compliance** and **conformity*** (or non-compliance and non-conformity), we will use them interchangeably in this chapter. Also, although we have characterized the differences between *standard* and *target*, for the sake of simplicity in the text, considering that the mathematical treatment of both is the same, we will primarily use the term '**standard**'.

9.2 GRAPHICAL METHODS FOR COMPARING MONITORED DATA WITH QUALITY STANDARDS

In Chapter 6, you learned how to present **monitoring data** using different types of charts. Most of these charts can also be adapted to also show the applicable **quality standard**. Usually, this is accomplished by including a horizontal line representing the value of the standard. Typical examples

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

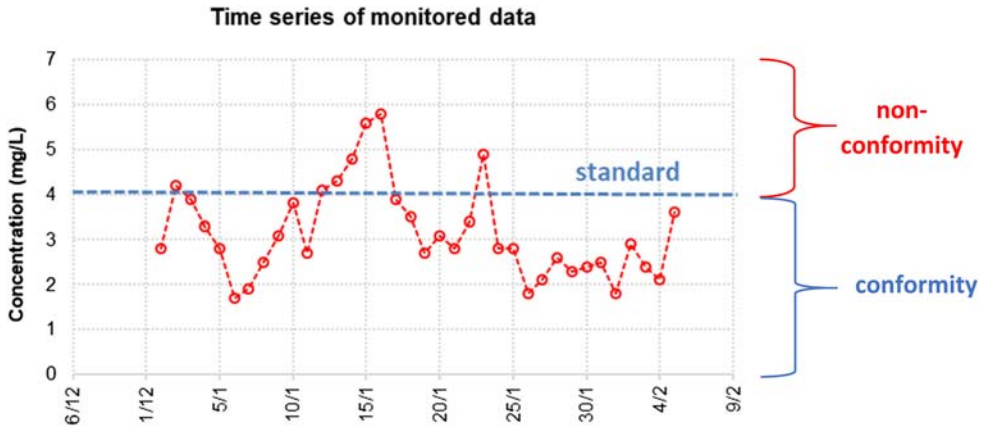


Figure 9.2 Time series of data from Example 6.3, showing the value of the quality standard and the regions of conformity and non-conformity. *Note*: For constituents that are not pollutants, but desired compounds to be preserved, and also for removal efficiencies, the interpretation is the opposite: the region of conformity is above the standard value (the higher, the better), and non-conformity is below the standard.

C. 6

are time series graphs, box plots, and percentile graphs (Figures 9.2–9.4). In these three examples, which use the same data from Example 6.3 and other subsequent examples, the adoption of a standard value of, say, 4.0 mg/L is considered and applied for treatment plant effluents and the quality of an ambient water body.

(a) Time series graphs

S. 6.2

Time series graphs were explained in Section 6.2. This is a **useful** chart – it is **simple to understand** because it plots the *original monitored data* without further processing (see Figure 9.2). We have placed a horizontal line to show the standard which is for the concentration to be less than 4.0 mg/L. In total, there are 36 data points, and you can see that 7 of them ($7/36 = 19.4\%$) are above 4.0 mg/L, that is, are not in conformity with the standard. Conversely, 29 points ($29/36 = 80.6\%$) are in conformity. Of course, if you have a long time series, with many data points, you will not be able to do this type of visual analysis. However, you will still be able to identify trends of deterioration or improvement, and periods with and without conformity with the standards.

(b) Box plots

S. 6.4

Box-and-whiskers plots were explained in Section 6.4. They are highlighted in this book as very **practical** and important graphs, especially when you want to **compare different samples**. Figure 9.3 shows the comparison between the effluent concentrations from five treatment plants (you could also make a plot like this for five different water bodies). The value of the standard (4.0 mg/L) is the horizontal line plotted over all the data sets, allowing a quick comparison of the degree of compliance of all samples. In this figure, plant 4 corresponds to the data from Example 6.3. We can see that plant 1 has the highest concentrations, with the 25th percentile in the non-compliance region (above the line of the standard). This means that at least 75% of the samples from plant 1 were not in conformity with the standard of 4.0 mg/L. In plant 5, all values are below the standard. For this type of plot, it is important to show the number of samples.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

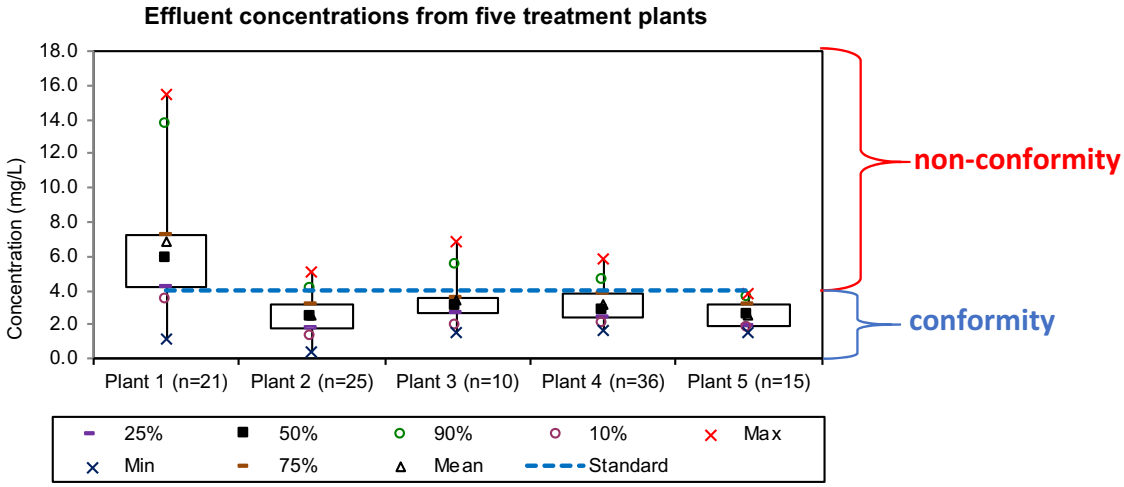


Figure 9.3 Box plots comparing the effluent concentrations from five treatment plants (or five water bodies), including the line with the value of the quality standard. Data from plant 4 are the data from Example 6.3. *Note*: For constituents that are not pollutants, but desired compounds to be preserved, and also for removal efficiencies, the interpretation is the opposite: the region of conformity is above the standard value (the higher, the better), and non-conformity is below the standard.

Excel

The Excel file *Box Plot* gives you different types of box plots, including this option of including the line with the standard value.

(c) Percentile graphs

S. 6.3.3

Percentile graphs were explained in Section 6.3.3. When plotted together with the value of the standard, they show directly **the percentage of your data that are in compliance with the standard**. Read Section 6.3.3 on how to build and interpret this graph, both for concentrations and removal efficiencies. Consider the example in Figure 9.4 (data from Example 6.3). Suppose the standard is 4.0 mg/L. Draw a horizontal line starting from this value on the *Y*-axis, and when it reaches the percentile curve, draw a descending vertical line. The point where this line crosses the *X*-axis gives you the percentage of samples in your data set with concentrations less than or equal to the standard value. In this example, you visually see that around *80% of the values in your data set are in conformity with the standard (the other 20% are not)*.

If this were a desired constituent that we want to preserve (such as dissolved oxygen in a water body), the interpretation would be the opposite. We would say that around *20% of the data points are in conformity with the standard*, and *80% of the values are not*.

S. 6.3.3

Please read again Section 6.3.3 for additional comments regarding the interpretation of percentile graphs used for representing **removal efficiencies**.

Besides the percentile graph, you may also use a percentile table, using the values that are the basis for the percentile graph. Following the instructions given in Section 6.3.3, using the Excel function PERCENTILE, we can structure a table like the one presented in Example 6.5 (using the data from Example 6.3 and representing the graph in Figure 9.4). This table is reproduced here, with all the percentiles, from 0% to 100%, and with the specific indication of the percentile associated with the standard of 4.0 mg/L (see Table 9.1).

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

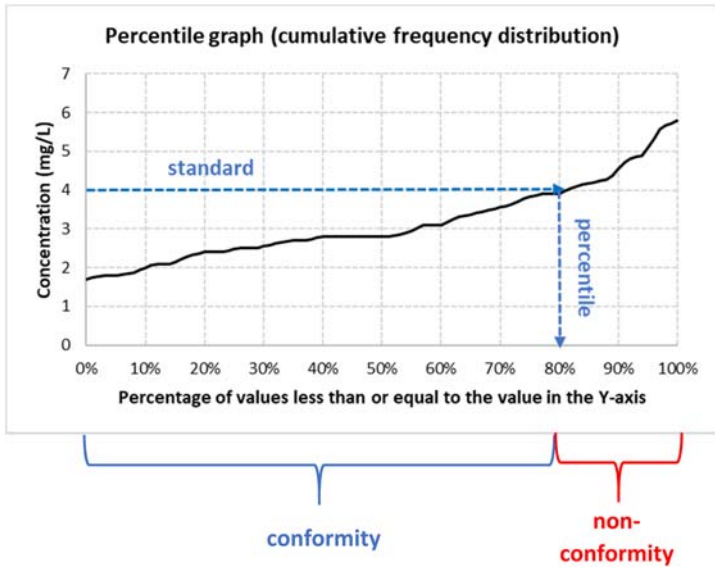


Figure 9.4 Percentile graphs using data from Example 6.3, showing a horizontal line with the value of the quality standard and a vertical line leading to the probability of obtaining a value less than or equal to the value of the standard. Also shown are the regions of conformity and non-conformity. *Note*: For constituents that are not pollutants, but desired compounds to be preserved, and also for removal efficiencies, the interpretation is the opposite: the region of conformity is above the standard value (the higher, the better), and non-conformity is below the standard.

Table 9.1 Example of a table with percentile values, calculated using the Excel function PERCENTILE. Data taken from Example 6.3 and corresponding to the graph in Figure 9.4.

Percentile (%)	Conc (mg/L)	Percentile (%)	Conc (mg/L)	Percentile (%)	Conc (mg/L)	Percentile (%)	Conc (mg/L)	Percentile (%)	Conc (mg/L)
0%	1.70	20%	2.40	40%	2.80	60%	3.10	80%	3.90
1%	1.74	21%	2.40	41%	2.80	61%	3.17	81%	3.97
2%	1.77	22%	2.40	42%	2.80	62%	3.24	82%	4.04
3%	1.80	23%	2.41	43%	2.80	63%	3.31	83%	4.11
4%	1.80	24%	2.44	44%	2.80	64%	3.34	84%	4.14
5%	1.80	25%	2.48	45%	2.80	65%	3.38	85%	4.18
6%	1.81	26%	2.50	46%	2.80	66%	3.41	86%	4.21
7%	1.85	27%	2.50	47%	2.80	67%	3.45	87%	4.25
8%	1.88	28%	2.50	48%	2.80	68%	3.48	88%	4.28
9%	1.93	29%	2.52	49%	2.80	69%	3.52	89%	4.38
10%	2.00	30%	2.55	50%	2.80	70%	3.55	90%	4.55
11%	2.07	31%	2.59	51%	2.80	71%	3.59	91%	4.73
12%	2.10	32%	2.62	52%	2.82	72%	3.64	92%	4.82
13%	2.10	33%	2.66	53%	2.86	73%	3.71	93%	4.86
14%	2.10	34%	2.69	54%	2.89	74%	3.78	94%	4.89
15%	2.15	35%	2.70	55%	2.95	75%	3.83	95%	5.08
16%	2.22	36%	2.70	56%	3.02	76%	3.86	96%	5.32
17%	2.29	37%	2.70	57%	3.09	77%	3.90	97%	5.57
18%	2.33	38%	2.73	58%	3.10	78%	3.90	98%	5.66
19%	2.37	39%	2.77	59%	3.10	79%	3.90	99%	5.73
								100%	5.80

Standard = 4.0 mg/L (between percentiles 81 and 82%)

Note: The value of the standard (4.0 mg/L) lies between percentiles 81% and 82% (see shaded area). The exact value is percentile 81.4%. This means that 81.4% of the data are lower than the value of the standard (4.0 mg/L). This value can also be calculated directly in Excel using the PERCENTRANK function (see the Excel file for Example 6.5).

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

9.3 EVALUATION OF COMPLIANCE BASED ON AVERAGE VALUES

In this section, we start by assessing compliance based on the **average value** from our data set, since this method is fairly straightforward. Specifically, we will introduce the concept of one-sample hypothesis testing. However, *we are not advocating that this is the best approach to assess compliance*, since putting our focus only on the average may be short-sighted, potentially concealing other important aspects of our data set, such as its variability and its distribution. Later in this chapter, we provide several other approaches that do not simply rely on the average value to help you develop a much more complete suite of evaluation methods to assess conformity with standards or regulations.

9.3.1 Introductory concepts

Basic

In many situations, specifications for compliance with the standard are explicitly based on **average values of the monitored data**. As a matter of fact, even if this is not explicitly stated in the regulatory text, many people frequently tend to compare the average concentration from their sample with the value of the standard. If the average is lower than the standard value, some people are tempted to conclude that the system is in conformity with the legislation. However, we need to *exercise caution when making this type of comparison*.

C. 5

C. 6

C. 8

First of all, from Chapters 5, 6, and 8, we already know that small average values can be obtained if your data set contains many small values. However, suppose we are dealing with the quality of a water body, there may be cases where even though most values in the data set are very low and below the specified standard, the few data points in the set that are higher may still cause water quality or toxicity problems in the water body.

Consider again the data from Example 6.3. The **mean** of the data is **3.2 mg/L**. If the **standard** is **4.0 mg/L**, it might appear that the system is complying with the standard. However, this apparent compliance would be justifiable only *if the regulation explicitly stated that the standard should simply be compared with the average value*. In the absence of such language, it is not safe to assume that the system is in compliance simply because the average value of the data set is below the standard.

S. 3.2.3

We have already seen in Section 3.2.3 and in other parts of this book that when we calculate the mean value of a data set (e.g., for concentrations), it is simply our *estimate of the true mean concentration*. We do not ever actually know what the true mean concentration is, because we only collect a limited number of samples, plus our methods for measuring the constituents of concern in the field or in the laboratory are subject to natural variability. Thus, if we are assessing compliance based solely on the average value, we must ask ourselves: **is the average concentration *significantly* lower than the standard value?**

9.3.2 Fundamentals of hypothesis testing

C. 10

To determine if the difference between the standard and the average values from our data set is statistically significant, we need to use **hypothesis testing**. The topic of hypothesis testing in general and the concepts of tests involving one sample and more than one sample are covered in much more detail in Chapter 10. Here, in Section 9.3, we cover the concept of hypothesis testing briefly in the context of our example of assessing compliance based on average values – i.e., determining if the average value from a data set is significantly below the regulatory standard. For this, we use what is called a **one-sample hypothesis test**. But before we start learning how to apply this type of hypothesis test, let us first cover a few basics.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Basic

To conduct a hypothesis test, we must start by defining a **null hypothesis H_0** and an **alternative hypothesis H_a**. Think of them like this:

- The *null hypothesis* is typically *the one that you do not believe to be true, the situation that you believe you can invalidate with your study*.
- The *alternative hypothesis is the one that you believe to be true or that you want to try to validate*.

We want to determine if there is a significant difference between the average value of our data and the fixed standard. Therefore, the null hypothesis is that the true concentration is equal to the standard value, and the alternative hypothesis is that the true concentration is not equal to the standard value (i.e., it is either greater than or less than the standard).

The hypothesis test will result in one of the following two conclusions: either

- we **reject the null hypothesis** (*in favour of our alternative hypothesis*) or
- we **fail to reject the null hypothesis** (*which does not mean that the null hypothesis is true*, by the way!)

For our example, if we **reject the null hypothesis**, this means that there is enough evidence to say that there is a *significant difference* between our average value and the standard and

- If the **average value is below the standard**, then we can say that it is *significantly* less than the standard and is in compliance or conformity.
- If the **average is higher than the standard**, then we can say it is *significantly* higher and that it is non-complying or in non-conformity.

If we **fail to reject the null hypothesis**, this means that *we do not have enough confidence to say whether the true average value is above or below the standard*. It could be either. We cannot say that the null hypothesis is true, but we cannot say it is false either. We also cannot draw any conclusions about the alternative hypothesis in this case! It also often means that we may need to collect more data. The more data we collect, the more likely we are to be able to reject the null hypothesis.

Do not worry if it takes you a while to understand these concepts. The logic of hypothesis testing is not so straightforward and can be difficult to comprehend at first. One analogy that may help is the 'presumption of innocence' principle used in law, which is where a person is considered innocent until there is enough evidence to prove that they are guilty. With statistical hypothesis testing, we assume the null hypothesis until we have enough evidence to 'prove' the alternative hypothesis.

A practical way of expressing results from hypothesis tests is by presenting the ***p*-value** (also called the *probability level* or *observed significance level*). The *p*-value is the *probability of incorrectly rejecting the null hypothesis when it is actually true* (i.e., finding misleading results by chance). The *p*-value is interpreted in comparison with a prespecified **significance level** (also known as the *α level* or the *type I error*):

- **If the *p*-value is less than the significance level α, then we reject the null hypothesis.**
- **If the *p*-value is greater than or equal to α, then we fail to reject the null hypothesis.**

Usually a significance level of $\alpha = 0.05$ **(5%)** is used, implying a *confidence level of 0.95 (95%)*. However, if you want to be even more rigorous on the test, you may use lower values for α, which will increase the confidence associated with the test. For example, an α level of 0.01 (1%) would imply a confidence level of 0.99 (99%). Chapter 10 presents more information about the meaning of the α level.

C. 10

More detailed information about the fundamentals and theory behind hypothesis tests is presented in the following chapters, but here, we simply want to highlight the **importance of conducting a statistical test** to

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

draw conclusions about compliance in terms of average values of our data set, rather than simply comparing the calculated average value to the standard.

In summary, here is the most important information you need to know about hypothesis testing in order to complete the applied examples presented in this chapter:

- Hypothesis tests are needed to determine if there is a **significant** difference between our average value and the standard.
- When doing hypothesis tests, we need to **obtain strong conclusions**, and our ability to do so will depend on how we formulate our hypotheses.
- The **significance level (α) directly affects the confidence level** of the hypothesis test; typically, we use a value of $\alpha = 0.05$, but if you use a lower value, it will make for a more rigorous hypothesis test.
- The hypothesis test produces a *p*-value, which allows us to draw a conclusion about the test: **if *p*-value $< \alpha$, we reject H_0; if *p*-value $\geq \alpha$, we fail to reject H_0.**
- The ***p*-value is the probability of incorrectly rejecting the null hypothesis when it is actually true** (i.e., finding misleading results by chance).
- **Rejecting the null hypothesis H_0 is a strong conclusion**.
- **Failing to reject the null hypothesis H_0 is generally a weak conclusion** (it usually suggests that we need to collect more data to draw a stronger conclusion).
- **Failing to reject the null hypothesis *does not mean that we can accept the null hypothesis***; we can only say that *the null hypothesis cannot be rejected*.
- The **alternative hypothesis H_a is usually the theory we want to support**; we typically do not believe the null hypothesis to be true, and we are attempting to provide evidence against it (in favour of our alternative hypothesis).

9.3.3 Different types of hypothesis tests

Basic

In the context of this chapter, hypothesis tests can be used to compare a single sample to a quality standard value. This is called a **single-sample test**. Remember, here we are talking about one statistical sample, which is a single data set comprising values obtained from many water samples collected at the same point. In this chapter, you will learn to use single-sample hypothesis tests, but hypothesis testing can also be used to compare two or more samples to each other, and these tests are presented in Chapter 10.

C. 10

When using a one-sample hypothesis test to assess compliance with a standard, there are three approaches you can use: **two-tailed**, **left-tailed**, and **right-tailed**. We recommend the two-tailed test, but you may find that *other people may use a left-tailed or right-tailed test in some cases*. Chapter 10 goes into detail about the meaning of one-tailed and two-tailed tests and when it is appropriate to use each one. In the box below there is a description of the *three approaches in the context of assessing compliance based on the average value compared to the quality standard*.

One-tailed and two-tailed hypothesis tests

- **Two-tailed test.** The first approach is to assume a *null hypothesis* H_0: mean = standard and an *alternative hypothesis* H_a: mean ≠ standard. This would lead to a 'two-tailed' test (both right- and left-tailed). If you use a two-tailed *t*-test, then you may have one of the following three outcomes:
  - **The average concentration is *less than* the standard (and the *p*-value is less than α, so the difference is significant).** This result would be good news for the manager of a treatment system who wants to prove that the system is working well and *is complying with the quality standards.*

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

  - **The average concentration is *greater than* the standard (and the *p*-value is less than α, so the difference is significant).** This result might be helpful for a member of an environmental organization who is trying to prove that a nearby industry *is not complying with the quality standards*.
  - **The average concentration may be less than or greater than the standard (but the *p*-value is greater than α, so the difference is *not significant*).** This means that, based on the data, you do not have enough confidence to know if the system is complying or not. This result may indicate that you need to collect more samples.
- **Left-tailed test.** If you assume a *null hypothesis* H_0: mean $\geq$ standard and an *alternative hypothesis* H_a: mean $<$ standard, it would lead to a one-sided, 'left-tailed' or 'inferior' test. Some people might use this approach to try to prove that a system is in compliance or conformity with the standard.
- **Right-tailed test.** Likewise, if you assume a *null hypothesis* H_0: mean $\leq$ standard and an *alternative hypothesis* H_a: mean $>$ standard, it would lead to a one-sided, 'right-tailed' or 'superior' test. Some people might use this approach to try to prove that a system is out of compliance or in non-conformity with the standard.

In your **report**, it is a good strategy to state **which type of test you used and present the resulting *p*-values** of your statistics. By doing so, the readers will be able to have an idea on the confidence level of the decision and also compare your observed *p*-value with another significance level (α), different from the traditional 0.05 (maybe a more rigorous value of 0.01).

Also, depending on the distribution of your data, you should use one of the following two approaches for conducting this type of hypothesis test:

- Use a **parametric test**, in case the distribution of your data *does not depart substantially from a normal distribution*. Typically, you compare the **mean** value of your sample with the value of the standard.
  - One-sample, one-tailed *t*-test.
- Use a **non-parametric test**, in case the distribution of your data *departs substantially from a normal distribution* (the non-parametric tests do not depend on the distribution of your data). Typically, you compare the **median** value of your sample with the value of the standard.
  - Sign test using the binomial distribution.
  - Sign test using the standard normal *Z* statistic.
  - Wilcoxon signed-rank test.

C. 8
C. 10

Review Chapter 8 for more insight about how to determine if your data are distributed normally or if their distribution departs substantially from a normal distribution. Refer to Chapter 10 for more detailed information about how to implement parametric and non-parametric hypothesis tests.

9.3.4 Parametric one-sample test (*t*-test)

Basic

The most widely used parametric test is the **one-tailed *t*-test for a single-sample mean**. Please see Chapter 10 for a more complete explanation on the relevant equations of the *t*-test and how to conduct it and its applicability. We will present here only a brief summary of the relevant aspects of the *t*-test used when you are comparing your sample mean with the value of the quality standard.

C. 10

Any general *statistical software* will perform a *t*-test. *Excel* has several built-in functions associated with the *t*-test and the *t*-distribution. Example 9.1 presents, in a direct way, the calculations of the *p*-value for the *t*-test and other tests (non-parametric).

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

The *t*-test requires calculation of a ***t*-statistic**, which is calculated as follows:

$$t = \frac{\bar{x} - M_0}{s/\sqrt{n}} \tag{9.1}$$

where

$\bar{x}$ = mean value of your statistical sample
M_0 = mean value for the null hypothesis (in this case, the value of the quality standard)
s = standard deviation of your statistical sample
n = number of data points in your statistical sample.

Note that the denominator of Equation 9.1 ($s/\sqrt{n}$) is called the **standard error**.

Finally, to complete the *t*-test, you can calculate the ***p*-value** using the Excel function T.DIST.2T([x], [degrees_freedom]). For [x], you will enter the value of the *t*-statistic calculated using Equation 9.1, and the degrees of freedom are the sample size minus one ($n - 1$).

A summary of the ***t*-test** is presented below (extracted from comments by Levine *et al.*, 1998):

***t*-test (one sample)**

- **Description**: Compares the **mean** value of the sample with a specified value (in our case, the value of the quality or regulatory **standard**).
- **Type**: *Parametric* test.
- **Input data required**: Number of data points, mean and standard deviation, value of the standard, plus the specification of the desired significance level for the test.
- **Output data produced**: *t*-statistic, *p*-value.
- **Assumptions**: The data have been obtained independently and represent a random sample from a population, which is *normally distributed*.
- **Comment**: If the sample size is small ($n < 30$) and if we cannot ascertain that the population from which the sample was obtained is normally distributed, non-parametric tests may be needed. If the sample size is large, the *t*-test is relatively robust to small departures from normality.

9.3.5 Non-parametric one-sample test (sign test)

Advanced

S. 9.3.6

If the distribution of your data is *not symmetrical* and *departs substantially from the shape of a normal distribution*, you should consider using a test that does not depend on the distribution, that is, a **non-parametric test**. In this section, we cover the simple sign test, and in Section 9.3.6, we describe the Wilcoxon test.

(a) Sign test using the binomial distribution for the *S*-statistic

For our application, the **sign test** is the simplest non-parametric technique. It is specifically designed for testing hypotheses about the **median** of any continuous population. The sign test uses the test statistic *S*, where *S is the number of values that exceed the value M_0* you specified in the null hypothesis (in our case here, M_0 is the value of the quality standard). You should notice that the *sign test depends only on the* ***sign*** (positive or negative) of the difference between each sample value and the value of the null hypothesis (quality standard). *S* has a **binomial distribution**, and you can use this distribution to calculate the *p*-value (Mendenhall & Sincich, 1988).

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

With the number of values that exceed the value of the null hypothesis (positive sign in the difference between the value and the null hypothesis M_0 value) and the total number of values in your sample, you can use the Excel function BINOM.DIST with the following syntax:

- *Null hypothesis* H_0: **Sample median** $= \boldsymbol{M_0}$ (*two-tailed*):
  p-value $= 2 \times (1 -$ BINOM.DIST(maximum value between the number of data points that exceed (S) and the number of data points that do not exceed M_0; number of data points in the sample; 0.5; TRUE)).
- *Null hypothesis* H_0: **Sample median** $\geq \boldsymbol{M_0}$ (*left-tailed*):
  p-value $=$ BINOM.DIST(number of values that exceed M_0 (S); number of data points in the sample; 0.5; TRUE).
- *Null hypothesis* H_0: **Sample median** $\leq \boldsymbol{M_0}$ (*right-tailed*):
  p-value $= 1 -$ BINOM.DIST(number of values that exceed M_0 (S); number of data points in the sample; 0.5; TRUE).

The value of 0.5 in the BINOM.DIST function comes from the fact that if H_0 states that your *median is equal to* M_0, you have a probability of 50% of having values higher than M_0 and 50% of having values lower than M_0 (this is, indeed, the concept of a median). Therefore, $p = 0.5$ represents this probability of 50%.

Note that unlike the ***t*-test** (in which you needed the number of data n, the mean, and standard deviation), for the **sign test** you need the number of data and the number of data that exceed the standard. You can use the Excel function COUNTIF (counts the number of cells within a range that meet given criteria):

COUNTIF(array with your data range; criterion).

The criterion is '>the value of M_0' (quality standard). If M_0 is in, say, cell E26, then in the field for entering the criterion, you enter: ">"&E6.

A summary of the sign test is presented below (extracted from comments by Mendenhall & Sincich, 1988; Hines *et al.*, 2003).

Sign test with binomial distribution of *S* (one sample)

- **Description**: Compares the **median** value of the sample with a specified value (in our case, the value M_0 of the **standard**).
- **Type**: *Non-parametric* test.
- **Input data required**: Number of data, number of data with values above M_0, plus the specification of the desired significance level for the test.
- **Output data produced**: p-value.
- **Assumptions**: No assumptions have to be made about the shape of the probability distribution.
- **Comments**:
  - This version of the test, which uses the binomial distribution for S, can be done with small and large samples.
  - Comparing the parametric t-test with the sign test: usually the t-test may lead to better results compared with the sign test, unless the distribution of the data has substantial departures from normality.
  - Comparing the sign test with the Wilcoxon signed-rank test (Section 9.3.6): the Wilcoxon test is usually a better choice.

S. 9.3.6

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

(b) Sign test using the standard normal *Z* statistic (for $n \geq 10$)

Advanced

If our sample size is $n \geq 10$, we can implement the sign test using the familiar standard normal Z statistic. This comes from the fact that, for $p = 0.5$, the normal approximation to the binomial distribution performs reasonably well, even for sample sizes as small as 10 (Mendenhall & Sincich, 1988).

The Z_0 statistic for the normal approximation to the sign test is (Hines *et al.*, 2003; Mendenhall & Sincich, 1988)

$$Z_0 = \frac{S - 0.5n}{0.5\sqrt{n}} \tag{9.2}$$

where

S = number of data points with values greater than M_0 (where M_0, value of the standard). If you use the number of data with values lower than M_0, you will get the same result, but with an opposite sign. This will not influence the calculations we use here, because in the determination of the p-value, we will use the absolute value of Z_0.

n = number of data of your sample.

Once you find the value for Z_0, you can then calculate the p-value using the Excel function NORM.S.DIST([absolute value of Z], TRUE for cumulative distribution):

- *Null hypothesis* H_0: **sample median** $= \boldsymbol{M_0}$ (*two-tailed*):
  p-value $= 2 \times (1 -$ NORM.S.DIST(ABS(Z_0); TRUE))
- *Null hypothesis* H_0: **sample median** $\geq \boldsymbol{M_0}$ (*left-tailed*):
  p-value $= (1 -$ NORM.S.DIST (ABS(Z_0); TRUE))
- *Null hypothesis* H_0: **sample median** $\leq \boldsymbol{M_0}$ (*right-tailed*):
  p-value $=$ (NORM.S.DIST (ABS(Z_0); TRUE))

As we saw previously for hypothesis tests, if the p-value is less than your significance level α, then the median of your data set is significantly differently (or below or above, for one-tailed tests) the standard value. If the p-value is greater than α, then the results are not significant.

A summary of the sign test using the standard normal approximation is presented below (extracted from comments by Mendenhall & Sincich, 1988):

Sign test using approximation by the standard normal *Z* statistic (for a large sample, $n \geq 10$) (one sample)

- **Description**: Compares the **median** value of the sample with a specified value (in our case, the value M_0 of the **standard**).
- **Type**: *non-parametric* test.
- **Input data required**: Number of data, number of data with values above M_0, plus the specification of the desired significance level for the test.
- **Output data produced**: p-value.
- **Assumptions**: No assumptions have to be made about the shape of the probability distribution.
- **Comment**:
  - This version of the test that uses the normal approximation to the binomial distribution for S requires a sample size with $\boldsymbol{n \geq 10}$.
  - See the previous box for a comparison of the sign test with the *t*-test and the Wilcoxon test.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

9.3.6 Non-parametric one-sample test (Wilcoxon signed-rank test)

Advanced

The Wilcoxon signed-rank test is a very good alternative to the *t*-test, and a preferable choice over the Sign test. As a non-parametric test, it does not rely on the values of data but on their ranking, from smallest to largest.

The data need to be processed, and a good description of the procedure can be found on the Excel spreadsheet associated with Example 9.1. Basically, it involves calculating the difference between each value and M_0 (in our case, the value of the quality standard). Some differences will be positive (when the value is greater than M_0) and negative (when the value is lower than M_0). All differences are ranked, and the sum of the ranks of the positive differences (R^+) and the sum of the ranks of the negative differences (R^-) are calculated. The smallest of the two values (R^+ and R^-) is called R and is used for the calculation of the test statistic.

If the **sample is small**, you need to use look-up tables to consult the critical values of the statistic of the Wilcoxon test. However, if the sample is relatively large ($\boldsymbol{n \geq 20}$), the distribution of R is approximately normal and we can use the Z statistic, according to the following equation (Hines *et al.*, 2003):

$$Z_0 = \frac{R - n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}} \tag{9.3}$$

where

R = smallest value between R^+ (sum of the ranks of the positive differences) and R^- (sum of the ranks of the negative differences)
n = number of data of your sample.

As with the sign test, once you find the value for Z_0, you can then calculate the p-value using the Excel function NORM.S.DIST:

- *Null hypothesis H_0:* **sample median = M_0** (*two-tailed*):
  p-value = 2 × (1 − NORM.S.DIST(ABS(Z_0); TRUE))
- *Null hypothesis H_0:* **sample median ≥ M_0** (*left-tailed*):
  p-value = (NORM.S.DIST (ABS(Z_0); TRUE))
- *Null hypothesis H_0:* **sample median ≤ M_0** (*right-tailed*):
  p-value = 1 − (NORM.S.DIST (ABS(Z_0); TRUE))

A summary of the Wilcoxon signed-rank test is presented below (extracted from comments by Hines *et al.*, 2003):

Wilcoxon signed-rank test using a normal approximation for the test statistic (for a large sample, $n \geq 20$) (one sample)

- **Description**: Compares the sum of the ranks of the positive differences (R^+) and the sum of the ranks of the negative differences (R^-), where the differences are between the values of the sample and a specified value (in our case, the value M_0 of the **standard**).
- **Type**: *non-parametric* test.
- **Input data required**: Data from your sample, which will be further processed to calculate the ranks, plus the specification of the desired significance level for the test.
- **Output data produced**: *p*-value.
- **Assumptions**: No assumptions have to be made about the shape of the probability distribution.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

- **Comments**:
  - This version of the test that uses the normal approximation for the test statistic is to be used for large samples ($n \geq 20$).
  - Comparing the parametric *t*-test with the Wilcoxon signed-rank, in general, the Wilcoxon test will perform almost as well as the parametric *t*-test, and will even be superior, when the data distribution departs from normality.

9.3.7 Application of one-sample hypothesis tests to assess compliance

Advanced

Example 9.1 presents the application of the one-sample parametric (*t*-test) and non-parametric (sign test and Wilcoxon signed-rank test) tests for comparing your sample means and medians with the value of the quality standard. You can use the spreadsheet associated with the example or your own statistical software. The example presents the *p*-values from the tests using Excel functions and includes the interpretation of the results.

Example

EXAMPLE 9.1 COMPARING THE MEAN AND MEDIAN VALUE OF YOUR SAMPLE WITH THE VALUE OF THE STANDARD USING PARAMETRIC AND NON-PARAMETRIC TESTS

You monitored a certain constituent in the effluent from a treatment plant (or in a water body). Analyse whether the mean or median of your sample is in significant compliance with the standard. Use the same data from Example 6.3. Consider that the value of the regulatory standard is 4.0 mg/L.

Data (values are in mg/L and are the same as in Example 6.3):

2.8	4.2	3.9	3.3	2.8	1.7	1.9	2.5	3.1	3.8	2.7	4.1
4.3	4.8	5.1	5.2	3.9	3.5	2.7	3.1	2.8	3.4	4.9	2.8
2.8	1.8	2.1	2.6	2.3	2.4	2.5	1.3	2.9	2.4	2.1	3.6

Excel

Note: This example is also available as an Excel spreadsheet.

Solution:

From these data, we have the following basic descriptive statistics:

- Number of data: $\boldsymbol{n = 36}$
- Mean = **3.16 mg/L**
- Standard deviation = **1.04 mg/L**
- Median = **2.80 mg/L**

We also obtain the following information on the number of samples complying with the standard:

- Number of data with value ≤ standard: **29**
- Number of data with value > standard: **7**

These statistics are calculated in the spreadsheet.
For performing the tests, we specify the following information:

- Significance level for the test (α) = 0.05 (confidence level of 0.95 or 95%)

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

We will use the Excel spreadsheet and will not show the calculations here. Using the spreadsheet, we can make three analyses, taking into account whether our sample mean or median should be '=', '≤', or '≥' than the standard value:

- Two-tailed test (mean value = standard?)
- One-tailed test (left-tailed) (mean value ≥ standard?)
- One-tailed test (right-tailed) (mean value ≤ standard?)

S. 9.3.3

From the three choices, the first one (**two-tailed test**) will be adopted here (see discussion in Section 9.3.3).

Since our data set is composed of $n = 36$ data, we consider it to be a large sample and we are able to implement the four tests described in this section:

- One sample *t*-test
- Wilcoxon signed-rank test (for samples with $n \geq 20$)
- Sign test with the binomial distribution
- Sign test with approximation by the standard normal Z statistic (for samples with $n \geq 10$)

(a) One sample parametric and non-parametric tests

- Null hypothesis H_0: mean (or median) = standard
- Alternative hypothesis H_a: mean (or median) ≠ standard

Based on the *p*-value obtained and comparing with the established significance level of 0.05, we can draw the following conclusions:

- If *p*-value < 0.05: reject null hypothesis that mean = standard
- If *p*-value ≥ 0.05: do not reject null hypothesis that mean = standard

From the calculations, we obtain the following *p*-values:

- *t*-test: *p*-value = 2.39879×10^{-5}
- Wilcoxon signed-rank test: *p*-value = 0.000157843
- Sign test (binomial distribution): *p*-value = 6.96017×10^{-5}
- Sign test (standard normal Z approximation): *p*-value = 0.000245733

(we are leaving the values with all these significant figures for you to be able to check your calculations; naturally, in your report you do not need all of them).

From the parametric *t*-test, since *p*-value < 0.05, we **reject** the hypothesis that the sample **mean** (3.16 mg/L) = standard (4.00 mg/L), accepting instead the alternative hypothesis that the mean concentration is *significantly different* from the standard. Since the mean (3.16 mg/L) is lower than the standard (4.00 mg/L), we can say that the mean **is significantly below the standard**.

From the non-parametric tests, since in all of them the *p*-value < 0.05, we **reject** the hypothesis that the sample **median** (2.80 mg/L) = standard (4.00 mg/L), accepting instead the alternative hypothesis that the median concentration is *significantly different* from the standard. Since the median (2.80 mg/L) is lower than the standard (4.00 mg/L), we can say that the median **is significantly below the standard**.

All four tests have given consistent results. All *p*-values (observed significance levels) were very low (much lower than 0.05, giving us a high confidence in the test results.

(b) Analysis with different values for the tests

Use the spreadsheet and substitute some of the input data and interpret the results.

For instance, if we had a very small **sample size** ($n = 6$). Applying the *t*-test (without checking its adequacy in terms of the sample distribution), we would get a *p*-value > 0.05, leading us to the

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

opposite conclusion that we should not reject the hypothesis that mean = standard. Sample size does influence the test results:

- *The **smaller the sample size**, the wider the non-rejection region (reducing the rejection region) → **more difficult to reject the null hypothesis***
- *The **higher the sample size**, the wider the rejection region (decreasing the non-rejection region) → **easier to reject the null hypothesis***

When we have very large samples, we have to be careful, because even small differences between the hypothetical value and the sampled value may be detected by the test of hypothesis (Hines *et al.*, 2003).

Now, coming back to the original value of $n = 36$, keeping the mean, try changing the value of the **standard deviation**. Put, for instance, a very high value of the standard deviation (say, 5.0 mg/L). The test outcome changes again, increasing the *p*-value up to a point in which we could conclude that we should not reject the hypothesis that mean = standard. For this type of test, we could say that:

- *The **smaller the standard deviation**, the wider the non-rejection region (reducing the rejection region) → **more difficult to reject the null hypothesis***
- *The **higher the standard deviation**, the wider the **rejection** region (decreasing the non-rejection region) → **easier to reject the null hypothesis***

Now, with the same original conditions, try different values of the **quality standard** and interpret the results. For instance, if you specify a standard of 3.0 mg/L, you would get a *p*-value ≥ 0.05, thus leading to the conclusion that we should not reject the hypothesis that mean = standard (even if we had a mean of 3.16 mg/L, higher than the standard of 3.00 mg/L).

9.4 EVALUATION OF COMPLIANCE BASED ON THE PROPORTION OF NON-CONFORMITY WITH STANDARD USING *Z*-TEST FOR PROPORTIONS

Basic

S. 9.3

In Section 9.3, we saw one approach for evaluating compliance, based on the average values in comparison with some standard value. However, another way to assess conformity or compliance with a standard is by assessing the proportion of samples that do not conform (fail) or conform with the particular standard. To assess compliance using this approach, you can use the *Z*-test for proportions.

In many cases, a regulatory agency may specify a minimum **percentage of samples that must be in compliance** with a standard limit. For instance, the legislation may specify that at least 90% of your data set must have a concentration below the standard limit. In other words, a maximum of $100 - 90 = 10\%$ of samples may be **out of conformity** (failure to meet the standard limit).

You can simply *count the number of non-conforming data points in your sample and divide it by the total number of data points in your sample*. For instance, in Example 9.1, we saw that out of a total sample size of 36 data points, 7 data points were exceeding the standard value (non-conformity). Therefore, the proportion of **failure** (non-conformity) would be $7/36 = 0.194 = 19.4\%$, and the proportion of **conformity** would be $1 - 0.194 = 0.806 = 80.6\%$. This proportion of conformity of 80.6% is lower than the minimum required by the legislation (90%), and thus it would appear that you are not complying with the regulation. Suppose the minimum compliance required by the regulators was 80%, then it would appear that you *are* conforming

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

with the legislation. But what about the **data variability** and the conclusions based on a **confidence level**? Is this measured proportion really representing the true probability of conformance for your data population? To answer this question, we must use hypothesis testing.

Advanced

S. 9.3

The hypothesis test for proportions is similar to the one-sample hypothesis test introduced in Section 9.3, with some slight differences. For example, we often assume that concentration data follow a normal or log-normal distribution. Proportion data, on the other hand, follow a *binomial distribution*. Nevertheless, if you have a large enough sample, you can assume a *normal approximation to the binomial distribution*, and thus use the ***Z*-test for proportions**.

Like the hypothesis tests described in Section 9.3, the *Z*-test for proportions produces a *p*-value, which must be evaluated against the chosen significance level (typically 0.05). To use the *Z*-test, we simply need the number of data in non-conformity and the total number of data, and we specify the maximum allowable proportion of non-conformity. The Z_0 statistics can be calculated as (Levine *et al.*, 1998)

$$Z_0 \cong \frac{X - n \cdot p_0}{\sqrt{n \cdot p_0 \cdot (1 - p_0)}} \tag{9.4}$$

where

X = number of data failing with the standard
p_0 = proportion of data specified in the standard (value for the null hypothesis H_0)
n = number of data.

The ***p*-values** can be simply found using the Excel function NORM.S.DIST as used in the other examples.

To use this *Z*-test, we need to verify whether our sample is large enough for us to use the normal approximation to the binomial distribution. Hines *et al.* (2003) suggest that the *proportion* of failures (p) should not be very close to 0 or 1. Furthermore, Mendenhall and Sincich (1988) report the rule of thumb described below.

Sample size (n) is considered sufficient if

$$P - 2\sqrt{\frac{P \cdot (1 - P)}{n}} > 0 \tag{9.5}$$

and

$$P + 2\sqrt{\frac{P \cdot (1 - P)}{n}} < 1 \tag{9.6}$$

A summary of the ***Z*-test** for proportions is presented below (extracted from comments by Mendenhall & Sincich, 1988; Levine *et al.*, 1998):

***Z*-test for proportions (one sample)**

- **Description**: Compares the **proportion** of failures in our sample with a specified proportion (in our case, the maximum proportion allowed in **standard**).
- **Type**: *Parametric* test.
- **Input data required**: Number of data, proportion of data failing with the standard, maximum proportion of failure allowed in the standard, plus the specification of the desired significance level for the test.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

- **Output data produced**: *p*-value.
- **Assumptions**: The size of the data is large enough, as specified in the criteria shown in Equations 9.5 and 9.6, such that we can use the normal approximation to the binomial distribution.

Example

EXAMPLE 9.2 COMPARISON OF THE PROPORTION OF FAILURE OF YOUR SAMPLE WITH THE MAXIMUM ALLOWABLE PROPORTION OF FAILURE, AS SPECIFIED IN THE STANDARD, USING THE Z-TEST FOR PROPORTIONS (LARGE SAMPLES)

Using the same data from Example 9.1, we observed that seven samples were not conforming with the standard. The total sample size was 36. Analyse the compliance with the regulations, taking into account that it requires a minimum proportion of compliance of 90%.

Excel

Note: This example is also available as an Excel spreadsheet.

Solution:

The number of failures and sample size are
Number of samples failing: $X = 7$
Total number of samples: $n = 36$

The **proportion of failure** (P) is

$$P = \frac{\text{number of samples failing}}{\text{total number of samples}} = \frac{X}{n} = \frac{7}{36} = 0.194 = 19.4\%$$

Therefore, the **proportion of compliance** with the standard is: proportion of compliance $= 1 - P = 1 - 0.194 = 0.806 = 80.6\%$. This value is lower than $\mathbf{1 - p_0}$, the minimum required percentage of 90% ($p_0 = 0.10$ or 10%).

The Z statistic is

$$Z_0 \cong \frac{X - n \cdot p_0}{\sqrt{n \cdot p_0 \cdot (1 - p_0)}} = \frac{7 - 36 \times 0.10}{\sqrt{36 \times 0.10 \times (1 - 0.10)}} = 1.888$$

S. 9.3.3

Following the same reasoning of Example 9.1 and the discussion in Section 9.3.3, we will use a **two-tailed test**:

- Null hypothesis H_0: proportion of failure = maximum allowable proportion of failure; $P = p_0$ (0.10)
- Alternative hypothesis H_a: proportion of failure ≠ maximum allowable proportion of failure; $P \neq p_0$ (0.10)

The *p*-value is obtained from the normal distribution, using the Excel function NORM.S.DIST and the value of the Z_0 statistic. The result is ***p*-value = 0.0589**. Since this value is greater than our significance level of 0.05, our conclusion is '**Do not reject the null hypothesis that proportion of failure in the sample *P* (0.194) = maximum allowable proportion of failure p_0 (0.10)**'. Therefore, there are not sufficient evidences that the system is not complying with the regulation in terms of the minimum required proportion of compliance.

However, note that, in this example, the resulting *p*-value is only marginally higher than the 0.05 significance level, and that the conclusion was based on this comparison. It is up to you to interpret the results and possibly request a larger number of samples or do additional investigations to be able to draw conclusions with more confidence.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

We can check whether **our sample size is sufficient** for undertaking this analysis. For this, we will use the rule of thumb described by Mendenhall and Sincich (1988) and stated in Equations 9.5 and 9.6:

$$P - 2\sqrt{\frac{P \cdot (1-P)}{n}} = 0.194 - 2\sqrt{\frac{0.194 \times (1-0.194)}{36}} = \mathbf{0.063} > \mathbf{0} \rightarrow \mathbf{ok}$$

and

$$P + 2\sqrt{\frac{P \cdot (1-P)}{n}} = 0.194 + 2\sqrt{\frac{0.194 \times (1-0.194)}{36}} = \mathbf{0.326} < \mathbf{1} \rightarrow \mathbf{ok}$$

Both conditions have been simultaneously satisfied, what indicates that **our sample size ($n = 36$) is sufficient.**

You can test different specifications for the required proportions of compliance with the legislation. Here, we have used the minimum proportion of 90% conformity. You may test other values, such as 95% or 80%, and see the test outcome.

9.5 PROBABILITIES OF CONFORMITY OR NON-CONFORMITY OBTAINED DIRECTLY FROM THE MONITORING DATA

Basic

We can also assess conformity with standards by calculating probabilities directly from the monitoring data. In Sections 6.3 and 9.2, we presented percentile graphs, which associated values of the variable you are analysing with percentages of occurrence. The graphs and table have been constructed using **calculated** values based on the Excel function PERCENTILE.

S. 6.3

S. 9.2

Now, let us do the same type of analysis but calculating the **probability associated with each measured value**. In order to do this, we need to **sort** the measured data in ascending or descending order (in the example provided here, we will use ascending order). When the data are in sequence, we assign the order (or **rank**) of each value (1, 2, 3, …, n). The **number of data** in our data set is n and the **order** of each value is called m. Based on the order of each data, we can calculate the **probability** of occurring values '<' or '>' the respective value of the variable, using the following equations:

- *Probability of occurrence of values 'lower than' the value ($P_<$):*

$$P_< = \frac{m}{n+1} \tag{9.7}$$

- *Probability of occurrence of values 'greater than' the value (exceedance probability) ($P_>$):*

$$P_> = 1 - P_< = 1 - \frac{m}{n+1} = \frac{n+1-m}{n+1} \tag{9.8}$$

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

where

$P_<$ = probability of occurrence of values lower than the value
$P_>$ = probability of occurrence of values greater than the value
n = number of data
m = order or rank of the data in the data sequence.

For instance, if 2.3 mg/L is the 7th value in a set of 36 values, $m = 7$ and $n = 36$. Therefore, applying Equation 9.7, the probability that there will be a value less than 2.3 mg/L is $7/(36 + 1) = 0.189 = 18.9\%$. Conversely, the probability of having a value greater than 2.3 mg/L is, according to Equation 9.8, $1 - 0.189 = 0.811 = 81.1\%$ or $(36 + 1 - 7)/(36 + 1) = 0.811 = 81.1\%$.

Instead of having '$n + 1$' in the equations above, we could have 'n' or even other possibilities, as discussed by Chow *et al.* (1988) on the concept of **plotting position**. Actually, the equation using 'n' gives us the direct probability but may bring difficulties from the fact that when $m = n$, the probability is 100%, what may not be easily plotted on probability scale in graphs. Although there are other possibilities for establishing the plotting positions, *we will use the common approach of adopting '$n + 1$'*.

For all the values in your data set, prepare a computational table. From the table, extract the probability value that is closest to the value of the standard. Example 9.3 illustrates the construction of the table and associated graphs.

Example

EXAMPLE 9.3 DIRECT DETERMINATION OF PROBABILITIES BASED ON THE MONITORING DATA

Using the same data from Examples 9.1 and 9.2, prepare a table and graphs for the probability of occurrence of values lower than or greater than any of the monitored data.

Data (values are in mg/L and are the same as in Example 6.3):

2.8	4.2	3.9	3.3	2.8	1.7	1.9	2.5	3.1	3.8	2.7	4.1
4.3	4.8	5.1	5.2	3.9	3.5	2.7	3.1	2.8	3.4	4.9	2.8
2.8	1.8	2.1	2.6	2.3	2.4	2.5	1.3	2.9	2.4	2.1	3.6

Excel

Note: This example is also available as an Excel spreadsheet.

Solution:

Sort your data in ascending order, rank the values, and estimate the probabilities using Equations 9.7 and 9.8. In this example, the number of data is $n = 36$.

Note that in the Excel spreadsheet associated with this example, we have used the Excel function RANK.AVG for ranking the data, which has the attribute that if more than one value has the same rank (data with the same values), the average rank is returned. Other people prefer to list the data in sequential order, regardless of repetitions. As a matter of fact, in the Excel spreadsheet provided, by using this function, we would not need to sort the data, and all calculations would be done automatically if you just entered your data in the same sequence as it had been obtained. However, the computational table becomes more organized and easier to interpret if the data are sorted.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Monitoring data (sorted in ascending order)	Rank (m)	Probability of occurrence of value 'less than' (%) $P_< = m/(n+1)$	Probability of occurrence of value 'greater than' (%) $P_> = 100 - P_<$
1.7	1	2.70%	97.30%
1.8	2.5	6.76%	93.24%
1.8	2.5	6.76%	93.24%
1.9	4	10.81%	89.19%
2.1	5.5	14.86%	85.14%
2.1	5.5	14.86%	85.14%
2.3	7	18.92%	81.08%
2.4	8.5	22.97%	77.03%
2.4	8.5	22.97%	77.03%
2.5	10.5	28.38%	71.62%
2.5	10.5	28.38%	71.62%
2.6	12	32.43%	67.57%
2.7	13.5	36.49%	63.51%
2.7	13.5	36.49%	63.51%
2.8	17	45.95%	54.05%
2.8	17	45.95%	54.05%
2.8	17	45.95%	54.05%
2.8	17	45.95%	54.05%
2.8	17	45.95%	54.05%
2.9	20	54.05%	45.95%
3.1	21.5	58.11%	41.89%
3.1	21.5	58.11%	41.89%
3.3	23	62.16%	37.84%
3.4	24	64.86%	35.14%
3.5	25	67.57%	32.43%
3.6	26	70.27%	29.73%
3.8	27	72.97%	27.03%
3.9	28.5	77.03%	22.97%
3.9	28.5	77.03%	22.97%
4.1	30	81.08%	18.92%
4.2	31	83.78%	16.22%
4.3	32	86.49%	13.51%
4.8	33	89.19%	10.81%
4.9	34	91.89%	8.11%
5.6	35	94.59%	5.41%
5.8	36	97.30%	2.70%

Equal values

(2+3)/2 = 2.5 or 2 and 3

Standard = 4.0 mg/L

Probability is between these two values

In the table, we may look for the value of the quality standard (4.0 mg/L, the same value adopted in the previous examples). There is no monitoring data with this value, and the closest ones are 3.9 mg/L (position 28.5 and probability of 77.03% of having values less than 3.9 mg/L and probability of 22.97% of having values greater than 3.9 mg/L) and 4.1 mg/L (position 30 and probability of 81.08% of having values less than 4.1 mg/L and probability of 18.92% of having values greater than 4.1 mg/L). From this table, we can conclude that the probability of having values less than the standard of 4.0 mg/L lies **between 77.03% and 81.08%**. These values match with the other ones already presented in this chapter.

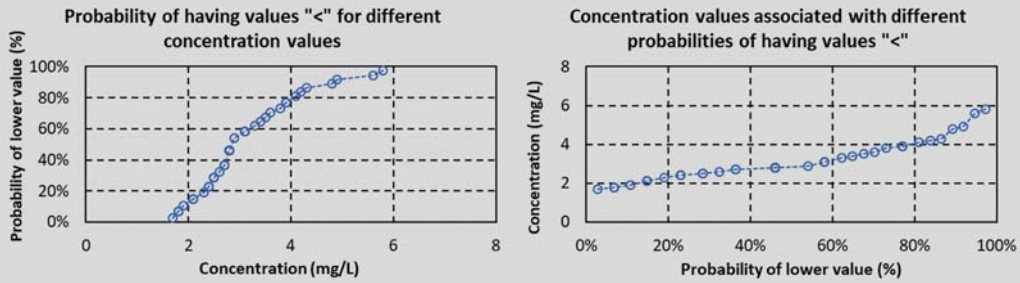
Note that the exact value can be computed directly using the Excel function PERCENTRANK.EXC([array],[x]), specifying the array of data points for [array] and the regulatory limit value of 4.0 for [x]. This produces a result of 79.7%, which is the probability of having values less than the standard of 4.0 mg/L.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

We can plot the values from this table (monitoring data and probability of having values 'less than') in two different ways, as shown below. In the **left-hand side graph**, the probability is in the *Y*-axis. For you to find the probability of having a value lower than the standard of 4.0 mg/L, you move upwards with a vertical line originating from 4.0 mg/L. Where this line crosses the curve, you draw a horizontal line towards the *Y*-axis. In this example, the line reaches a value **around 80%** (probability of having a value less than 4.0 mg/L). The **right-hand side graph** is similar to Figure 9.4 and the percentile graphs presented in Section 6.3.3. The difference is that the chart below was made with the 36 data points from the monitoring, while the percentile graphs have been made with 100 percentile values (from 1 to 100 percentile), calculated using the Excel function PERCENTILE.

S. 6.3.3

In the graphs below, you will find less than 36 markers for the data points because we have data with equal values, and they are plotted exactly on top of each other.



9.6 ESTIMATION OF COMPLIANCE WITH THE STANDARD BASED ON FREQUENCY ANALYSIS USING NORMAL AND LOG-NORMAL DISTRIBUTIONS

Advanced

Frequency analysis is another method that can be used to estimate compliance with a standard. This approach is a very common procedure in the field of hydrology for analysing probabilities of occurrence of flow values in rivers, especially extreme events such as minimum values during droughts or peak flows during storm events. These methods are widely described in hydrology textbooks (e.g., Chow, 1988) and also covered in the context of the analysis of influent flow rate to a treatment plant in Metcalf and Eddy (2003, 2014). We can use the same method here for undertaking a *frequency analysis of our monitoring data*.

S. 9.5

The difference from the approach adopted in Section 9.5 is that, there, the *probabilities have been calculated **directly** based on monitoring data*. With the frequency analysis approach described in this section, we complete the calculations based on **probability models** that allow us to *estimate the probability associated with any value of the constituent, not just the measured values*.

A vast array of frequency distributions may be used (and are indeed used in hydrological studies) to fit our data, but here we will confine ourselves to the two traditional distributions we have been applying throughout the book: **normal** and **log-normal** distributions.

To undertake a frequency analysis, our data are assumed to be **independent**. If extreme events are selected to form subsamples (which is not the case here), they should be identically distributed (Chow *et al.*, 1988). *The requirement for having independent data is a difficulty we face for all methods described in this book*: monitoring data from treatment plants may not always be independent. In some cases, they may be serially correlated. For instance, in some cases, today's value of the concentration may be affected by the value that occurred yesterday. This is not always the case, but it is important to

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

note that in a situation where data show a strong serial correlation, this may be a violation of some of the underlying assumptions associated with the methods presented here.

Before we do our calculations, we will introduce here the useful concept of **return period** so that we can analyse future failure times based on a probabilistic approach. Suppose that a failure event occurs if the concentration X is greater than or equal to the value of the regulatory standard. *Then, the return period T of a failure event is the expected average time until the next failure event will happen* (adapted from Chow *et al.*, 1988).

For instance, in the examples we have been using here in this chapter, we have 36 observations and 7 of them were above the standard value, that is, they failed to comply with the regulations. Thus, we may expect a failure every $36/7 = 5.1$ samples on average. If the samples have been collected at **fixed time intervals** (e.g., every week), we could assume that a failure event would occur every 5.1 weeks. If our monitoring frequency is one measurement per day, we could expect a failure to occur every 5.1 days on average. Note that *there is no guarantee whatsoever that the failure events will occur at regular time intervals* of 5.1 days: the failure event could occur, say, on days 3 and 4, consecutively, and then remain for 15 days before the next failure. However, we would expect that, over a longer time period, for instance, 100 days, failure events would occur approximately $100/5.1 = 19.6 \approx 20$ times. Let us emphasize again that the *correspondence between number of measurements and time presupposes fixed monitoring intervals* (hourly, daily, weekly, monthly, etc.). The number of data (36) could be associated with 36 h, 36 days, 36 weeks, 36 months, and so on if monitoring was kept at a regular fixed frequency. The larger the number of measurements, the higher our confidence in the calculations.

The **return period is the reciprocal of the probability of occurrence of the event** so that we have:

- *Return period ($T_<$) associated with the probability ($P_<$) of occurrence of values 'lower than' the value of the standard:*

$$T_< = \frac{1}{P_<} = \frac{1}{1 - P_>} \tag{9.9}$$

- *Return period ($T_>$) associated with the probability ($P_>$) of occurrence of values 'greater than' the value of the standard (exceedance probability):*

$$T_> = \frac{1}{P_>} = \frac{1}{1 - P_<} \tag{9.10}$$

where

$P_<$ = probability of occurrence of values lower than the value of the standard (calculated in Equation 9.7)

$P_>$ = probability of occurrence of values greater than the value of the standard (calculated in Equation 9.8).

S. 9.5

Now let us shift our focus to **the fitting of normal and log-normal distributions**. For this, we will build on the previous calculations highlighted in Section 9.5 and Example 9.3. The full calculation is included in the Excel spreadsheet for Example 9.4. However, here, for the sake of simplicity, we will present only the probability of occurrence of values 'less than' ($P_<$) and the value of the return period for values 'greater than' ($T_>$). The calculation of the probabilities of exceedance ($P_>$) and the return period ($T_<$) are included only in the Excel spreadsheet.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

From the probabilities of occurrence of values 'less than' the standard ($P_<$) in the monitoring data, we calculate the corresponding standard normal variable Z, using the Excel function NORM.S.INV($P_<$).

To use the log-normal distribution, we need to calculate the **$\log_{10}$** values of the measured data. After that we calculate the **mean** and **standard deviation** of the *original data series* and of the *$\log_{10}$-transformed series*. With the values of the mean, standard deviation, and Z, we calculate the **estimated** values of our variable (X_{est}) for the different probabilities ($P_<$) according to the normal and log-normal distributions using the following equations:

- **Normal distribution:**

$$X_{est} = \text{mean of data} + Z \cdot (\text{standard deviation of data}) \tag{9.11}$$

- **Log-normal distribution:**

$$\log_{10}(X_{est}) = \text{mean of } \log_{10}(\text{data}) + Z \cdot (\text{standard deviation of } \log_{10}(\text{data})) \tag{9.12}$$

The values of mean and standard deviation are fixed, and what varies, from data to data, are the Z values (which are directly associated with the previously calculated probabilities $P_<$).

After we have fitted the normal and log-normal distributions to our data, we can estimate the **probability of occurrence of any value of our variable** using the Excel functions NORM.S.DIST (for the normal distribution) and LOGNORM.DIST (for the log-normal distribution):

- **Normal distribution**: NORM.S.DIST(variable value; mean of measured data; standard deviation of measured data; TRUE for cumulative)
- **Log-normal distribution**: LOGNORM.DIST(variable value; mean of $\log_{10}$-transformed data × LN(10); standard deviation of $\log_{10}$-transformed data × LN(10); TRUE for cumulative)

We can also calculate the **value of the variable associated with a certain cumulative probability** using the Excel functions NORM.S.INV (for normal distribution) and LOGNORM.INV (for log-normal distribution):

- **Normal distribution**: NORM.INV(cumulative probability of having a value 'less than'; mean of measured data; standard deviation of measured data; TRUE for cumulative)
- **Log-normal distribution**: LOGNORM.INV(cumulative probability of having a value 'less than'; mean of $\log_{10}$-transformed data × LN(10); standard deviation of $\log_{10}$-transformed data × LN(10); TRUE for cumulative)

Example

EXAMPLE 9.4 ESTIMATION OF PROBABILITIES USING FREQUENCY ANALYSIS EMPLOYING NORMAL AND LOG-NORMAL DISTRIBUTIONS TO FIT THE DATA

Using the same data from Examples 9.1–9.3, undertake a frequency analysis, estimating (a) the probabilities associated with different values of the variable and (b) values of the variable corresponding to different probability values. Make a special evaluation of the probability of complying with the quality standard of 4.0 mg/L.

The data are in mg/L and are the same as in Example 6.3 and Examples 9.1–9.3. We will build up from the probability analysis already undertaken in Example 9.3 using the monitored data.

Excel

Note: This example is also available as an Excel spreadsheet.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Solution:

Prepare the following computational table, which was already started in Example 9.3 (columns 1–3).

Measured data					Normal	Log-normal		
(1) X	(2) m	(3) $P_<$ (%)	(4) $T_>$	(5) Z	(6) X_{est} (normal)	(7) $\log_{10}(X)$	(8) $\log_{10}(X_{est})$	(9) X_{est} (log-normal)
1.7	1	2.70	1.03	−1.926	**1.2**	0.230	0.213	**1.6**
1.8	2.5	6.76	1.07	−1.494	**1.6**	0.255	0.272	**1.9**
1.8	2.5	6.76	1.07	−1.494	**1.6**	0.255	0.272	**1.9**
1.9	4	10.81	1.12	−1.237	**1.9**	0.279	0.308	**2.0**
2.1	5.5	14.86	1.17	−1.042	**2.1**	0.322	0.334	**2.2**
2.1	5.5	14.86	1.17	−1.042	**2.1**	0.322	0.334	**2.2**
2.3	7	18.92	1.23	−0.881	**2.2**	0.362	0.357	**2.3**
2.4	8.5	22.97	1.30	−0.740	**2.4**	0.380	0.376	**2.4**
2.4	8.5	22.97	1.30	−0.740	**2.4**	0.380	0.376	**2.4**
2.5	10.5	28.38	1.40	−0.572	**2.6**	0.398	0.399	**2.5**
2.5	10.5	28.38	1.40	−0.572	**2.6**	0.398	0.399	**2.5**
2.6	12	32.43	1.48	−0.456	**2.7**	0.415	0.415	**2.6**
2.7	13.5	36.49	1.57	−0.345	**2.8**	0.431	0.430	**2.7**
2.7	13.5	36.49	1.57	−0.345	**2.8**	0.431	0.430	**2.7**
2.8	17	45.95	1.85	−0.102	**3.1**	0.447	0.464	**2.9**
2.8	17	45.95	1.85	−0.102	**3.1**	0.447	0.464	**2.9**
2.8	17	45.95	1.85	−0.102	**3.1**	0.447	0.464	**2.9**
2.8	17	45.95	1.85	−0.102	**3.1**	0.447	0.464	**2.9**
2.8	17	45.95	1.85	−0.102	**3.1**	0.447	0.464	**2.9**
2.9	20	54.05	2.18	0.102	**3.3**	0.462	0.492	**3.1**
3.1	21.5	58.11	2.39	0.205	**3.4**	0.491	0.506	**3.2**
3.1	21.5	58.11	2.39	0.205	**3.4**	0.491	0.506	**3.2**
3.3	23	62.16	2.64	0.310	**3.5**	0.519	0.521	**3.3**
3.4	24	64.86	2.85	0.382	**3.6**	0.531	0.530	**3.4**
3.5	25	67.57	3.08	0.456	**3.6**	0.544	0.541	**3.5**
3.6	26	70.27	3.36	0.532	**3.7**	0.556	0.551	**3.6**
3.8	27	72.97	3.70	0.612	**3.8**	0.580	0.562	**3.6**
3.9	28.5	77.03	4.35	0.740	**3.9**	0.591	0.580	**3.8**
3.9	28.5	77.03	4.35	0.740	**3.9**	0.591	0.580	**3.8**
4.1	30	81.08	5.29	0.881	**4.1**	0.613	0.599	**4.0**
4.2	31	83.78	6.17	0.986	**4.2**	0.623	0.614	**4.1**
4.3	32	86.49	7.40	1.102	**4.3**	0.633	0.630	**4.3**
4.8	33	89.19	9.25	1.237	**4.4**	0.681	0.648	**4.4**
4.9	34	91.89	12.33	1.398	**4.6**	0.690	0.670	**4.7**
5.6	35	94.59	18.50	1.607	**4.8**	0.748	0.699	**5.0**
5.8	36	97.30	37.00	1.926	**5.2**	0.763	0.743	**5.5**

- Column (1): Measured data (X)
- Column (2): Rank (m)
- Column (3): Probability of $<$ value (%); $P_< = m/(n+1)$
- Column (4): Return period for value $>$; $T_> = 1/(1-P_>)$
- Column (5): Z value from a normal distribution; Excel function NORM.S.INV(probability $P_<$)
- Column (6): $X_{est\ normal}$ = mean of measured data + $Z \times$ (standard deviation of measured data) = $3.158 + Z \times 1.038$
- Column (7): $\log_{10}$ of measured data; $\log_{10}(X)$
- Column (8): $\log_{10}(X_{est\ log\text{-}normal})$ = mean of $\log_{10}(X) + Z \times$ (standard deviation of $\log_{10}(X)$) = $0.478 + Z \times 0.138$
- Column (9): $X_{est\ log\text{-}normal} = 10\text{^}[\log_{10}(X_{est\ log\text{-}normal})]$

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

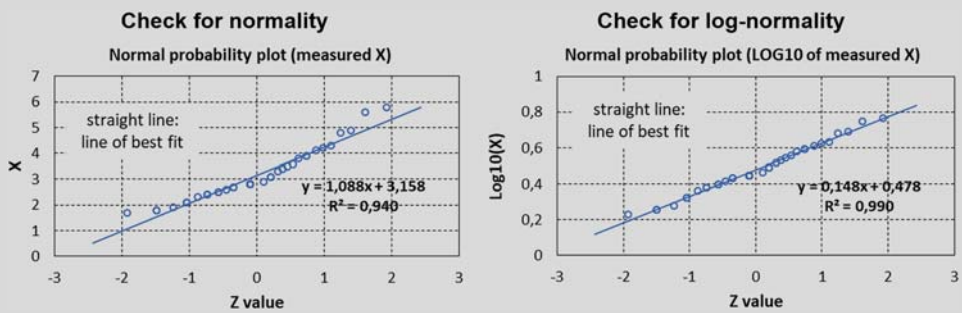
From columns (1) and (7) from the table, we draw the following descriptive statistical data which are used in columns (6), (8), and (9):

- $n = 36$
- mean of measured data $(X) = 3.158$
- standard deviation of measured data $(X) = 1.038$
- mean of log-transformed measured data $(\log_{10}(X)) = 0.478$
- standard deviation of log-transformed measured data $(\log_{10}(X)) = 0.138$

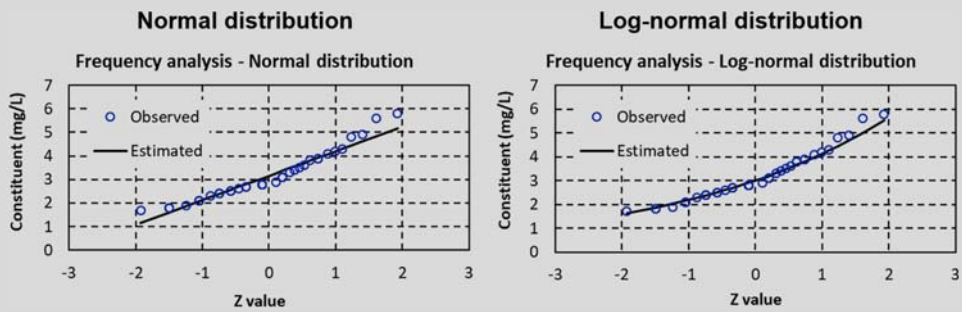
Note that columns (6) and (9) present the estimated values of our variable, as calculated using the normal and log-normal distributions, respectively.

With the calculated Z values and the measured data (X), we can prepare the normal probability plot for the observed data to see whether the measured data appear to follow a normal distribution (in case the plotted points fall reasonably well on a straight line). Similarly, with the calculated Z values and the log-transformed measured data $(\log_{10}(X))$, we can draw the normal probability plot for the log-transformed observed data to see whether the data appear to follow a log-normal distribution. Both plots are shown in the figure below. We can see that the original data show some departures from normality (left-hand-side graph) and are better represented by a log-normal distribution (right-hand-side graph, in which the log-transformed data follow a straight line). See Section 8.2.8 for a description of tests for normality and goodness-of-fit tests for a normal distribution.

S. 8.2.8



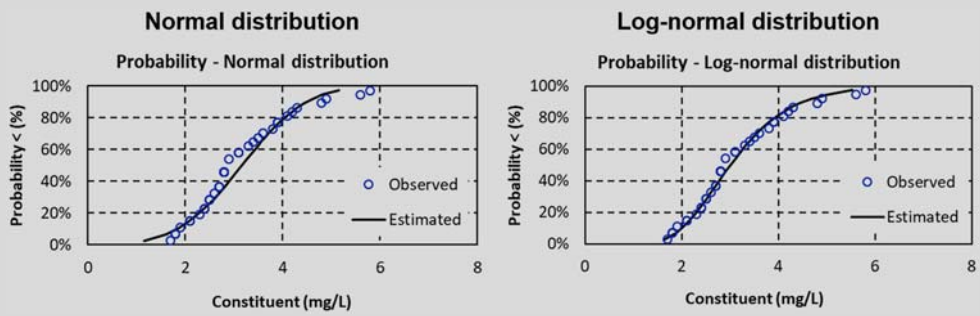
Now, based on the Z values calculated in column (5), the measured values shown in column (1) and the calculated values for the normal distribution (column 6) and log-normal distribution (column 9), we can draw the frequency analysis graphs, showing the fitting of the distributions to the measured data.



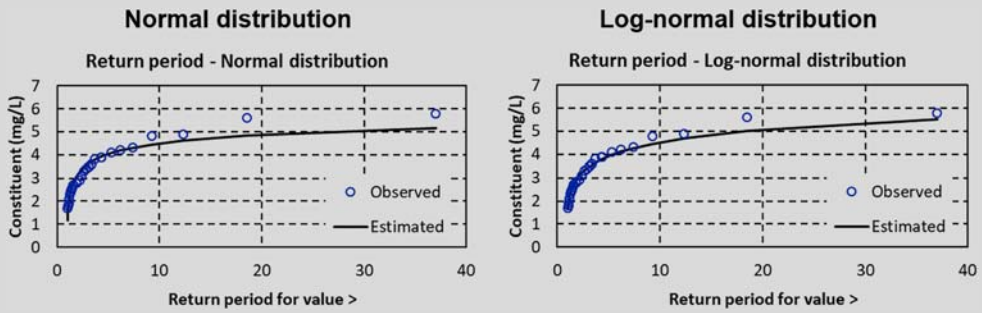
Similar frequency analysis graphs can be plotted, using the probabilities of having values 'lower than' $(P_<)$, as calculated in column (3). We can see that the fitting of the log-normal distribution was superior, given the fact that the data were somewhat skewed to the right (see Section 6.3).

S. 6.3

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest



Graphs showing the return period for exceedance values ($T_<$), calculated in column (3), are shown below. We can see that the higher the value of the constituent, the higher the return period, indicating that it would take longer time periods to reach them. The return period, to be expressed in time units, needs to have a fixed frequency of monitoring associated with our data. The graphs simply show, in the *X*-axis, the number of sampling events. If our data are collected on a daily basis, the values in the *X*-axis are expressed in days. If the data are obtained every week, the values in the *X*-axis are expressed in weeks, and so on.



Now, inside the Excel spreadsheet, we move to the worksheet in the tab 'Estimation of probabilities'. There, we use the Excel functions to calculate the **probability of compliance with the quality standard** of 4.0 mg/L.

- **Normal distribution:** NORM.S.DIST(variable value; mean of measured data; standard deviation of measured data; TRUE for cumulative) = NORM.S.DIST(4.0; 3.158; 1.038; TRUE) = **79.1%**
- **Log-normal distribution:** LOGNORM.DIST(variable value; mean of $\log_{10}$-transformed data $\times$ LN (10); standard deviation of $\log_{10}$-transformed data $\times$ LN(10); TRUE for cumulative) = LOGNORM. DIST(4.0; 0.478 $\times$ LN(10); 0.138 $\times$ LN(10); TRUE) = **81.6%**

The interpretation of these calculations is that if we assume a **normal distribution** of the data, we obtain a probability of **79.1%** that our monitoring data will be lower than the standard of 4.0 mg/L (compliance of 79.1%). If, on the other hand, we use the **log-normal distribution**, which showed to fit better to the data, we obtain a probability of **81.6%** of compliance. Both values are similar and roughly indicate compliance around 80%.

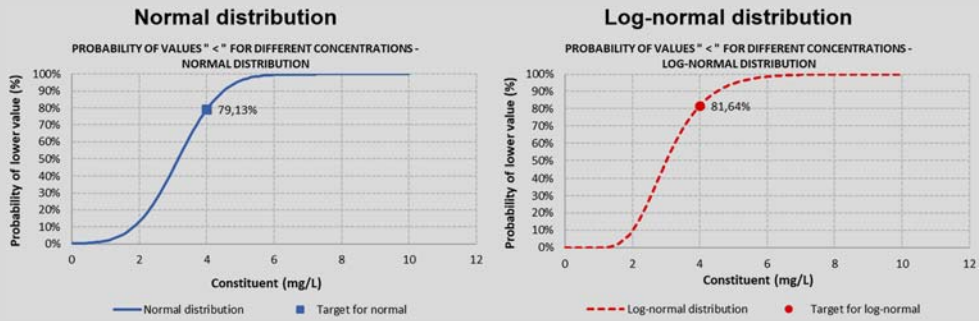
The **return periods** associated with the occurrence of values greater than the standard of 4.0 mg/L are calculated using the probabilities estimated above for the normal and log-normal distributions:

- Normal distribution: $T = 1/(1 - P_<) = 1/(1 - 0.791) =$ **4.8**
- Log-normal distribution: $T = 1/(1 - P_<) = 1/(1 - 0.816) =$ **5.4**

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Again, the interpretation is that if the data are collected systematically, say, on a weekly basis, there will be, on average, one event of exceedance of the standard every 5.4 weeks (for the calculation using log-normal fitting). Over one year (52 weeks), it may be expected that there will be $52/5.4 = 9.6$ events of non-conformity. The association with the probability for failure is that $P_{>} = 1/5.4 = 9.6/52 = 18.4\%$. The probability of conformity ($P_{<}$) has been already calculated and is equal to $100 - 18.4 = 81.6\%$ (for log-normal distribution).

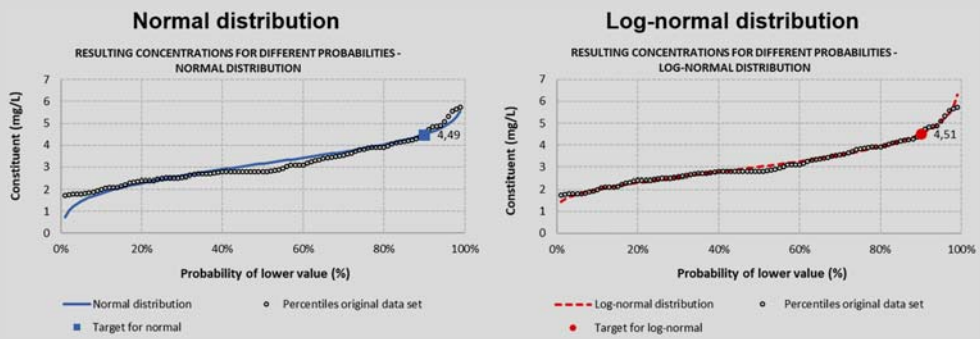
The graphs of the frequency analysis using the calculated values from the normal and log-normal distributions are shown below, with a special indication of the value of the standard and the associated probability of compliance.



Finally, if the regulations instead specify that 90% (=0.90) of the data must be below a certain value, we may calculate the resulting value using Excel functions:

- **Normal distribution**: NORM.INV(cumulative probability of having a value 'less than'; mean of measured data; standard deviation of measured data; TRUE for cumulative) = NORM.INV(0.90; 3.158; 1.038; TRUE) = **4.49 mg/L**
- **Log-normal distribution**: LOGNORM.INV(cumulative probability of having a value 'less than'; mean of $\log_{10}$-transformed data $\times$ LN(10); standard deviation of $\log_{10}$-transformed data $\times$ LN (10); TRUE for cumulative) = LOGNORM.INV(0.90; 0.478 $\times$ LN(10); 0.138 $\times$ LN(10); TRUE) = **4.51 mg/L**

The graphs of concentration as a function of the probability are shown for the calculated values using the normal and log-normal distributions, plus the percentile values calculated using the Excel function PERCENTILE. The concentration values associated with the probability of 90% are also shown.



Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

9.7 RELIABILITY ANALYSIS

9.7.1 Reliability and stability

Advanced

A **treatment plant** should aim to perform in a stable and reliable manner. This is frequently a challenge because the input variables keep changing due to their dynamic nature, and also other influencing factors vary over time such as environmental conditions, the status of mechanical equipment, and other elements that may impact performance, such as conditions set by the operator to control performance, such as recycle flows, sludge wastage flows, aeration levels, chemical dosing levels, etc.

A **water body** may experience varying water quality on a seasonal basis due to climatic conditions, especially rainfall and temperature dynamics, and also because of interference arising from land use and activities in the catchment area, including the discharge of treated or untreated wastewater effluents. The stability and reliability of water quality in a water body will depend on these factors. Apart from very specific actions practiced in only a limited number of water bodies (e.g., localized aeration, *in situ* chemical dosing), there are generally no manipulated variables in the water bodies to control their quality. All corrective actions must be done at the watershed level.

Stability is evaluated based on the variation of the water quality, as inferred from the standard deviation or the coefficient of variation ($CV = standard\ deviation/mean$). **Reliability** is associated with the percentage of compliance with an established standard or target for the water quality. These concepts apply for both treatment plants and water bodies.

Figure 9.5 illustrates possible combinations of stability and reliability, with a plot of a time series of a water quality constituent, a line for the mean value of the monitored variable, and a line for the specified

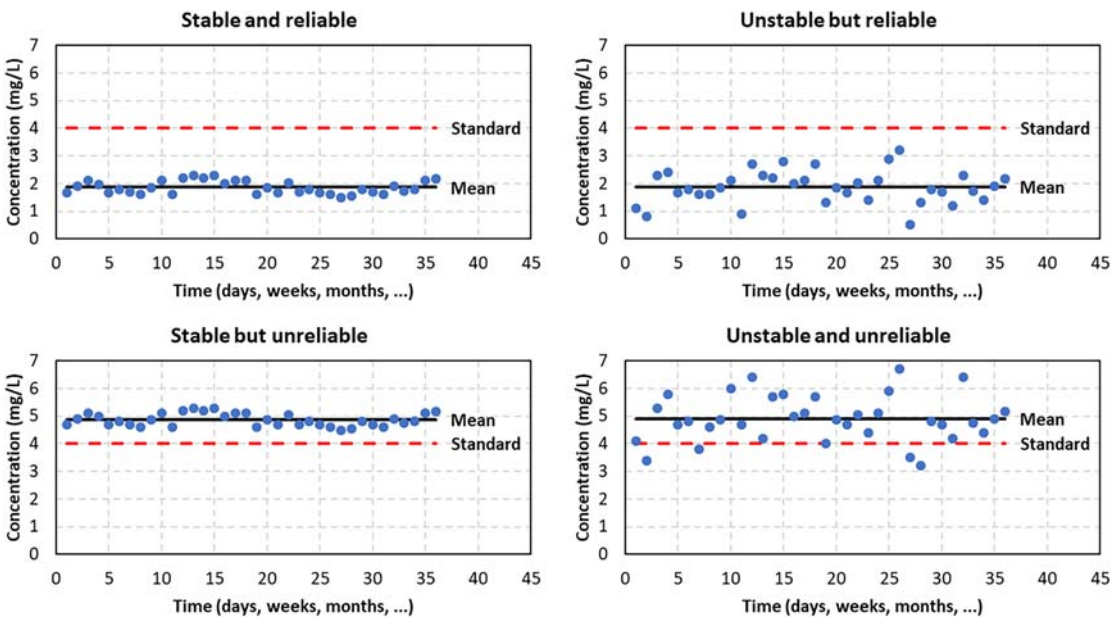


Figure 9.5 Possible combinations in terms of stability and reliability of the performance of a treatment plant or the conditions in a water body.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

quality standard (maximum allowable value). The same concept can also be used for analysing removal efficiencies in treatment plants. Four situations are depicted:

- *Stable and reliable performance*. Variability is small, since the values are close to their mean. The monitored values and the resulting mean are well below the maximum allowable value (standard), indicating reliability.
- *Unstable but reliable performance*. The values are widely variable around the mean, indicating instability. However, performance is reliable, because all values are in conformity with the quality standard.
- *Stable but unreliable*. Stability is high, since the values are close to their mean (low variability). However, the values and also their mean are above the stipulated quality standard, indicating non-conformity with the regulatory limit.
- *Unstable and unreliable*. Data variability is high, indicating low stability. Also, most of the values, including their mean, are above the maximum allowable standard value, indicating low reliability.

Therefore, we can conclude that the analysis of the values of the **mean** and **standard deviation** of your monitored data can provide useful information for making inference about the performance of treatment plants or prevailing conditions in a water body. This information can be combined into useful equations that comprise what is known as a reliability analysis, which is the main topic of this section.

Because reliability analysis is so easily done, we encourage you to utilize this method. You will only need the arithmetic **mean** and **standard deviation** of the constituent you are analysing, based on your monitoring data, plus the value of the **quality standard** or target applicable to your treatment plant or water body. After that you should decide whether you will use the simple equations associated with either the **normal** or **log-normal** distribution, knowing that, in most cases, the log-normal distribution will be the most applicable one.

Advanced

9.7.2 Background concepts about reliability analysis

Part of the text in this subsection has been based and adapted from Oliveira and von Sperling (2008). Most of the comments here apply to treatment plants, but the underlying concepts are also applicable to water bodies.

The reliability of a system can be defined as the *probability of achieving adequate performance for a specified period of time under specified conditions*. In terms of performance of a treatment plant or the conditions of a water body, the reliability can be understood as the **percentage of time at which the expected concentrations comply with specified standards or targets** (Dean & Forsythe, 1976a; Metcalf & Eddy, 2003; Niku *et al.*, 1979; Niku *et al.*, 1981; Oliveira & von Sperling, 2008 for *wastewater treatment plants*; Melo *et al.*, 2015; Melo, 2019 for *water treatment plants*).

A treatment plant will be *completely reliable if the process performance response has no failure*, that is, if the limits established by the regulators are not violated. The treatment process will fail when the required effluent standards or targets are exceeded.

Because of numerous **uncertainties** underlying the design and operation of treatment plants, a *failure risk is always unavoidable*, and the treatment plant should be designed and operated based on an acceptable risk or violation level. The minimal reliability requirements must be determined in order to establish the failure probability that can be accepted. The investment and operational costs of the treatment process will be affected by the desired reliability. High expectations with respect to effluent quality may bring the need for physical expansion of units, additional steps in the treatment line, more sophisticated operation, or installation of better control systems.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

The *probability of failure is very sensitive to the distribution of the effluent concentration*. After this distribution is known, an expression may be found to define the fraction of time that a given concentration has been exceeded in the past and, consequently, the future performance of the plant can be predicted, provided that process variables and other conditions remain the same.

Because of variations in performance, a treatment plant should be designed to **produce an average concentration that remains below the specified regulatory standard or limit** at a certain **reliability level**. The **coefficient of reliability (COR)** is a measure that relates the mean values of the constituent (i.e., design or operational mean value) to the required value established by the standards that must be achieved on a probability basis (Niku *et al.*, 1979). This method can be applied to **wastewater** and **water** treatment plants as well as water quality monitoring.

A detailed description of the reliability analysis can be found in Oliveira and von Sperling (2008). In this publication, the COR was used to assess the performance of several different wastewater treatment technologies in the removal of various constituents in a large number of treatment plants.

S. 9.6

We should note that *the concepts shown here, based on reliability analysis, converge with the concepts shown in Section 9.6, related to the application of frequency distribution analysis, using both the normal and log-normal distributions*. The results obtained are the same, and we make inferences on compliance with the standards based on a fitted distribution to our experimental data.

Advanced

9.7.3 The Coefficient of Reliability (COR)

The **Coefficient of Reliability (COR)** relates the values of the mean design concentrations to the standard to be achieved on a probability basis:

$$m_x = (\text{COR}) \cdot X_s \tag{9.13}$$

where

m_x = target mean concentration to be maintained in the effluent of the treatment plant (design or operational values) (mg/L)
X_s = concentration specified by the regulatory standard (mg/L)
COR = coefficient of reliability.

C. 8

For undertaking the reliability analysis, you need to decide whether you will assume that your monitoring data follow a **normal** or a **log-normal** distribution. You can review Chapter 8, which shows the properties of both distributions and demonstrates how to assess their fitting to experimental data. In Section 9.6, when we studied frequency distributions and compliance with quality standards, we also analysed fitting to both distributions (see Example 9.4).

S. 9.6

For the **normal distribution**, using the concept of the standard normal Z statistic, we can assess the value of the standard X_s as the mean plus the product of Z and the standard deviation (see Section 8.2.5 and also Equation 9.11):

S. 8.2.5

$$X_s = m_x + Z_{1-\alpha} \times \text{standard deviation} \tag{9.14}$$

Knowing that the standard deviation is equal to the mean (m_x) times the CV, we can express Equation 9.14 as

$$X_s = m_x + Z_{1-\alpha} \times (m_x \cdot \text{CV}) \tag{9.15}$$

Because COR is equal to m_x/X_s (Equation 9.13), for the **normal distribution**, we obtain the value of the COR as a simple function of CV (for different values of $Z_{1-\alpha}$):

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Normal distribution:

$$\mathrm{COR} = \frac{1}{1 + (Z_{1-\alpha}) \cdot \mathrm{CV}} \tag{9.16}$$

For the **log-normal distribution**, the COR is calculated as proposed by Niku *et al.* (1979).

Log-normal distribution:

$$\mathrm{COR} = \sqrt{\mathrm{CV}^2 + 1} \times \exp\left\{-(Z_{1-\alpha}) \cdot \sqrt{\ln(\mathrm{CV}^2 + 1)}\right\} \tag{9.17}$$

where

CV = coefficient of variation (arithmetic standard deviation divided by arithmetic mean)

α = probability of failure to meet the standard (probability of having values that exceed the standard; called $P_>$ in Section 9.6)

$1 - \alpha$ = **reliability level** or probability of compliance with the standards (probability of having values below the standard; called $P_<$ in Section 9.6), exp is the number

exp = (2.718) raised to a power

$Z_{1-\alpha}$ = standardized normal variate (obtained from the standard normal variate tables or using the Excel function NORM.S.INV($1 - \alpha$).

S. 9.6

For instance, if we accept a probability of failure of 10% ($\alpha = 0.10$), the reliability level is $1 - 0.10 = 0.90$, meaning that we aim at having a conformity of 90% with the standards. Some selected values of the probability $1 - \alpha$ (**reliability level**) and the associated percentiles $Z_{1-\alpha}$, which are used in Equations 9.16 and 9.17, are shown in Table 9.2.

Note that *COR is expressed based on the properties of the original data and not on the logarithm of the data*.

From the coefficients of reliability **COR** obtained, it is possible to determine the **mean design or operating effluent concentrations** that would be required to achieve the specified standards at a certain reliability level by simply using Equation 9.13, with the values of the standard (X_s) and COR.

In order to assist in the interpretation of the concept of the COR, Table 9.3 and Figure 9.6 have been prepared for selected reliability levels and for a wide range of CV values, for the normal and log-normal distributions. The determination and interpretation of the COR values in Table 9.3 are as described below.

Table 9.2 Values of the standard normal variate Z as a function of the reliability level.

Cumulative Probability $(1 - \alpha)$ = Reliability Level	$Z_{1-\alpha}$
99.9	3.090
99	2.326
95	1.645
90	1.282
80	0.842
50	0.000

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Excel

Table 9.3 Coefficient of reliability (COR) as a function of CV and reliability level (50%, 80%, 90%, 95%, 99%, 99.9%) for the normal and log-normal distributions.

Reliability level (%)	CV (standard deviation/mean)														
	0.001	**0.20**	**0.40**	**0.60**	**0.80**	**1.00**	**1.20**	**1.40**	**1.60**	**1.80**	**2.00**	**2.50**	**3.00**	**3.50**	**4.00**
Normal distribution															
50	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
80	1.00	0.86	0.75	0.66	0.60	0.54	0.50	0.46	0.43	0.40	0.37	0.32	0.28	0.25	0.23
90	1.00	0.80	0.66	0.57	0.49	0.44	0.39	0.36	0.33	0.30	0.28	0.24	0.21	0.18	0.16
95	1.00	0.75	0.60	0.50	0.43	0.38	0.34	0.30	0.28	0.25	0.23	0.20	0.17	0.15	0.13
99	1.00	0.68	0.52	0.42	0.35	0.30	0.26	0.23	0.21	0.19	0.18	0.15	0.13	0.11	0.10
99.9	0.99	0.62	0.45	0.35	0.29	0.24	0.21	0.19	0.17	0.15	0.14	0.11	0.10	0.08	0.07
Log-normal distribution															
50	1.00	1.02	1.08	1.17	1.28	1.41	1.56	1.72	1.89	2.06	2.24	2.69	3.16	3.64	4.12
80	1.00	0.86	0.78	0.73	0.71	0.70	0.71	0.72	0.73	0.75	0.77	0.82	0.88	0.94	1.00
90	1.00	0.79	0.66	0.57	0.52	0.49	0.47	0.45	0.45	0.44	0.44	0.44	0.45	0.46	0.48
95	1.00	0.74	0.57	0.47	0.40	0.36	0.33	0.31	0.30	0.29	0.28	0.27	0.26	0.26	0.26
99	1.00	0.64	0.44	0.32	0.25	0.20	0.17	0.15	0.14	0.13	0.12	0.10	0.09	0.09	0.08
99.9	1.00	0.55	0.33	0.21	0.15	0.11	0.08	0.07	0.06	0.05	0.04	0.03	0.03	0.03	0.02

For instance, suppose you monitored the effluent concentrations from your treatment plant and obtained a mean CV value of 0.80. If you aim at a probability of 90% compliance with the standards, you adopt a reliability level of 90% ($\alpha = 0.10$). The percentile $Z_{1-\alpha}$ obtained from the Excel function NORM.S.INV $(1-\alpha)$ or from standard normal variate tables (see Table 9.2 for selected values of $Z_{1-\alpha}$) is equal to $Z_{1-\alpha} = 1.282$ for $1-\alpha = 0.90$. The resulting values for COR, calculated using Equation 9.16 for the normal distribution and Equation 9.17 for the log-normal distribution, are 0.49 and 0.52, respectively, for the normal and log-normal distributions. These values can also be obtained directly from Table 9.3, with CV = 0.80 and reliability level of 90%.

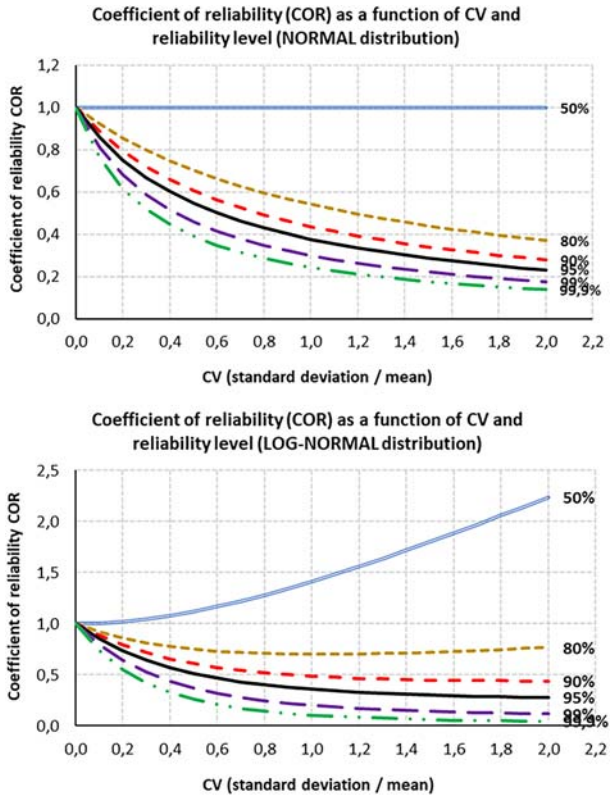
This signifies that, in order to comply with the standards 90% of the time, the *mean* effluent concentration should be, according to these calculations and using Equation 9.13: $m_x = \text{COR} \cdot X_s = 0.49X_s$ (assuming the normal distribution) or $m_x = \text{COR} \cdot X_s = 0.52X_s$ (assuming the log-normal distribution).

If your quality standard is, for instance, 10 mg/L, this indicates that the design or operating effluent concentration (assumed equal to the mean value) should be $m_x = 0.49 \times 10 = 4.9$ mg/L (for the normal distribution) and $m_x = 0.52 \times 10 = 5.2$ mg/L (for the log-normal distribution).

For a lower reliability level of 80% and the same CV of 0.80, from Table 9.3 one obtains COR = 0.60 (normal distribution) and COR = 0.71 (log-normal distribution). Therefore, in order to comply with the standards 80% of the time, a mean effluent concentration of $0.60 \times 10 = 6.0$ mg/L (normal distribution) or $0.71 \times 10 = 7.1$ mg/L (log-normal distribution) should be obtained. These mean concentration values are naturally higher than those calculated for a reliability level of 90%, because now, with a reliability level of 80%, we are less stringent in terms of percentage of compliance.

However, assuming once more a reliability level of 90%, if the CV value of the data were higher, say, CV = 2.0, COR would be 0.28 (normal distribution) or 0.44 (log-normal distribution), indicating that the mean effluent concentration should be $0.28 \times 10 = 2.8$ mg/L (normal distribution)

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest



Excel

Figure 9.6 Coefficients of reliability (COR) as a function of the CV (from 0.0 to 2.0) and the reliability level (50%, 80%, 90%, 95%, 99%, and 99.9%). Top: normal distribution; bottom: log-normal distribution.

or $0.44 \times 10 = 4.4$ mg/L (log-normal distribution). Similar calculations can be done for different CV values and reliability levels. The same inferences can also be drawn from Figure 9.6.

Now, we can see that in order to use reliability analysis, we need to know what the **typical CV values are for water and wastewater treatment plants**. For your treatment plant, you should calculate CV directly from the mean and standard deviation of your monitoring data. To put your measured CV values in context, Oliveira and von Sperling (2008) found that, using monitoring data from 166 **wastewater treatment plants** in Brazil, CV values for the effluent concentrations of biochemical oxygen demand (BOD), chemical oxygen demand (COD), total suspended solids (TSS), total nitrogen (TN), and total phosphorus (TP) mostly ranged between **0.3 and 1.0**. The CV values for thermotolerant coliforms (TTC) were higher, mainly ranging between **1.0 and 3.0**. Melo (2019) investigated 45 **water treatment plants** in Brazil, obtaining CV values between **0.2 and 0.8** for effluent turbidity.

Next, we will *interpret the shapes of the curves of the COR* as a function of the CV and reliability level, shown in Figure 9.6 and Table 9.3. For the **normal distribution**, we can make the following comments:

Normal distribution:

- For the **reliability level of 50%,** COR is equal to 1.0 for all CV values. This is already expected, since the theoretical normal distribution is symmetrical around the mean, and the mean is equal to the median.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Therefore, regardless of the spread of the monitoring data around the mean (different CV values), the median will always be equal to the mean, indicating that, in order to have compliance of 50% (50% = median value), the target mean should also be equal to the value of the standard (COR = 1.0).

- For **all other reliability levels greater than 50%**, COR decreases with increasing CV values and reliability levels. This should also be expected, because, in order to achieve higher probabilities of compliance, lower mean values are necessary as the variation of the data increases and the percentage of required compliance becomes more rigorous.

The interpretation for the **log-normal distribution** requires more detailed comments, since the COR curves in Figure 9.6 decrease and then start to increase again, as CV values become higher. The following comments can be made (Oliveira & von Sperling, 2008):

Log-normal distribution:

- For the **reliability levels frequently** adopted (80% or higher) and CV values frequently found in practice (lower than 1.0), there is a general trend of *decreasing COR with the increase of* CV *and the reliability level*. The interpretation is that the higher the desired reliability level and the effluent variability, the lower the COR.
- **COR values present first a decreasing and then an increasing pattern with respect to the CV**. Looking closely at Figure 9.6 and Table 9.3, you can see that the COR curves start to show an increasing trend after the CV reaches a certain value. This pattern appears for all reliability levels, but at different CV values. Because the distribution is not symmetrical (it is skewed to the right, i.e., the data are clustered more to the left of the mean, with most of the extreme values to the right), the arithmetic mean will be shifted to the right, since it is very influenced by the few large values obtained. As a result, a treatment plant with a large CV value for a particular constituent, even if having a high arithmetic mean value (close to the standard), may have most of the values well below the standard, thus possibly characterizing a high reliability level. See Oliveira and von Sperling (2008) for further discussions on this topic.
- **Some COR values greater than 1.0**. For the *normal distribution*, we saw that the maximum value of COR was 1.0, for reliability levels greater than or equal to 50%. With the *log-normal distribution*, we can find COR values greater than 1.0. But let us analyse the behaviour of the 50% reliability level, for which all COR values are greater that 1.0 (see Table 9.3). This means that, in order to comply with the standards during 50% of the time, the *mean* effluent concentration can be equal to or greater than the standard, according to Equation 9.13. This particular case arises from the fact that, for the log-normal distribution, as mentioned above, the arithmetic mean is greater than the median. If the median is taken to be the value of the discharge standard, thus being a direct interpretation of the 50% reliability level, then the arithmetic mean will be higher that the discharge standard, which justifies a COR that is greater than 1.0. COR values greater than 1.0 can be found for other reliability levels, and the interpretation remains that a certain desired percentage of compliance with the standards may be achieved even though the arithmetic mean is greater than the discharge standard. Once again, this is associated with the non-symmetrical pattern of the log-normal distribution.

Advanced

9.7.4 Expected probability of compliance with the standards

We can also use reliability analysis to estimate the **percentage probability of compliance with the standard**. This is known as the *reliability level*. Here, we will demonstrate a calculation for the reliability level for the normal distribution and the log-normal distribution, using algebraic manipulations of Equations 9.16 and 9.17 (Niku *et al.*, 1979; Oliveira & von Sperling, 2008).

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Below we present the ratio m'_x/X_s. Although this ratio is mathematically the same as the COR (see Equation 9.13), now we use m'_x to represent the actual **arithmetic mean of the monitored data**, instead of using m_x, which represented the target mean value to obtain for the design or operation of the treatment plant (Equation 9.13).

To make it clearer, we show the calculations in two steps. In the first step, we calculate the value of the standard normal variate $Z_{1-\alpha}$, remembering that α is the probability of failure and $1-\alpha$ is the probability of compliance with the regulation (or, if you wish, the reliability level).

Normal distribution:

$$Z_{1-\alpha} = \frac{\left(\frac{1}{m'_x/X_s} - 1\right)}{\text{CV}} \tag{9.18}$$

Log-normal distribution:

$$Z_{1-\alpha} = -\frac{\ln\left[\frac{m'_x}{X_s} \cdot \frac{1}{\sqrt{\text{CV}^2+1}}\right]}{\sqrt{\ln(\text{CV}^2+1)}} \tag{9.19}$$

where

m'_x = arithmetic mean of the monitored data (mg/L)
CV = coefficient of variation (standard deviation/mean) of the monitored data
X_s = concentration as specified by the quality standard (mg/L)
α = probability of failure of meeting the standards (probability of having values > standard; called $P_>$ in Section 9.6)
$1-\alpha$ = **reliability level**, or probability of compliance with the standards (probability of having values < standard; called $P_<$ in Section 9.6)
$Z_{1-\alpha}$ = standardized normal variate.

S. 9.6

After calculation of the value of $Z_{1-\alpha}$, we move into the second step to obtain the values corresponding to the cumulative probability of the standardized normal distribution. These values may be calculated by means of the function NORM.S.DIST in Excel, although they are also easily found in statistics books, corresponding to the cumulative area composed of the standardized normal curve.

$$\%\text{ of compliance} = 100 \times \text{NORM.S.DIST}(Z_{1-\alpha};\ \text{TRUE}) \tag{9.20}$$

Table 9.4 and Figure 9.7 show the values of the **expected probability of compliance** (*reliability level*) as a function of the ratio of the measured mean values and the quality standard (m'_x/X_s) and the CV (standard deviation/mean) for the normal and log-normal distributions.

You should note that Table 9.4 and Figure 9.7 are **very important and simple to use**. You only need the mean concentration (m'_x) from your monitored data, the CV from your monitored data, and the specified value of the regulatory or quality standard (X_s). With this information in hand, you can directly estimate the probability of compliance with the regulation, assuming either a normal or log-normal distribution (usually with a preference for the latter).

For instance, if you have an arithmetic mean value of $m'_x = 7.5$ mg/L, a CV = 0.80, and if the specified quality standard is $X_s = 10.0$ mg/L, the ratio is $m'_x/X_s = 7.5/10.0 = 0.75$. From the table, with $m'_x/X_s = 0.75$ and CV = 0.80, we obtain a percentage of compliance equal to 66% (normal distribution)

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Excel

Table 9.4 Expected probability of compliance with quality standard (%) as a function of the ratio mean/standard (m'_x/X_s) and CV (standard deviation/mean) for the normal and log-normal distributions.

Mean/standard (m'_x/X_s)	CV (standard deviation/mean)														
	0.001	**0.20**	**0.40**	**0.60**	**0.80**	**1.00**	**1.20**	**1.40**	**1.60**	**1.80**	**2.00**	**2.50**	**3.00**	**3.50**	**4.00**
Normal distribution															
0.01	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
0.25	100	100	100	100	100	100	99	98	97	95	93	88	84	80	77
0.50	100	100	99	95	89	84	80	76	73	71	69	66	63	61	60
0.75	100	95	80	71	66	63	61	59	58	57	57	55	54	54	53
1.00	50	50	50	50	50	50	50	50	50	50	50	50	50	50	50
Log-normal distribution															
0.01	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
0.25	100	100	100	100	99	98	97	97	96	96	96	95	95	95	95
0.50	100	100	98	94	91	89	89	88	88	88	88	88	89	89	89
0.75	100	94	83	79	78	78	78	79	79	80	81	82	83	84	84
1.00	50	54	58	61	64	66	68	70	71	73	74	76	78	79	80

Note: Mean (m'_x) and CV based on monitored data.

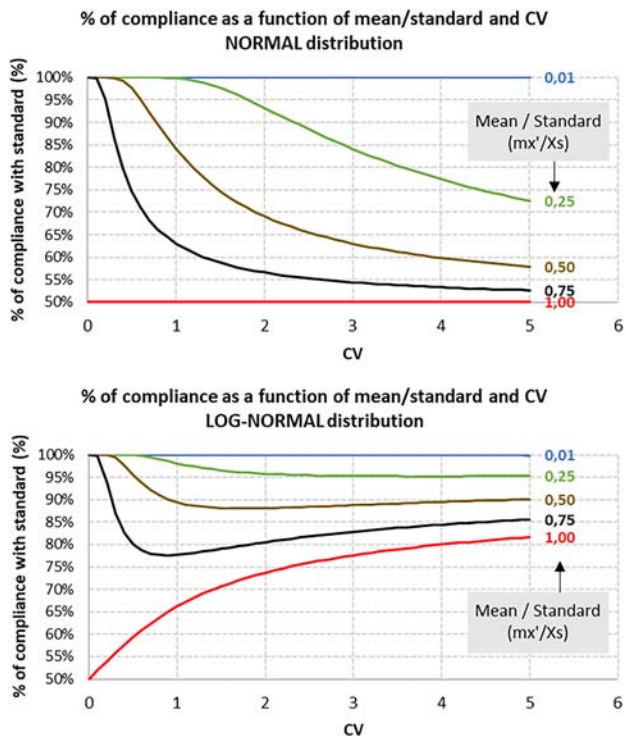


Figure 9.7 Expected probability of compliance with quality standard (%) as a function of the ratio mean/standard (m'_x/X_s) and CV (standard deviation/mean). Top: normal distribution; bottom: log-normal distribution. Mean and CV should be based on monitored data.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

or 78% (log-normal distribution). Notice that log-normal distribution leads to higher percentages of compliance for the same input data, because the mean is higher than the median (and maybe even greater than other higher percentiles, as discussed in Section 9.7.3, as shown in Figure 9.6).

Similar conclusions could be obtained from Figure 9.7, which is a *very powerful and simple representation of the capability of treatment plants or water bodies in complying with quality standards*. The shapes of the curves for the normal and log-normal distributions are very distinctive, and you should try to interpret them with consideration of the properties of the distributions already discussed in Chapter 8 and here in this chapter.

C. 8

So far, all of our examples shown here have been for data on concentrations of some constituent. However, you can also do a similar analysis for **removal efficiencies (E)** in treatment plants; just remember that the distribution of removal efficiencies is skewed to the left, and usually does not follow a log-normal distribution. However, the **remaining fraction ($1 - E$)** usually does follow a **log-normal distribution**, thus the theory above can be applied directly. See Section 7.7 for a discussion on the distribution of removal efficiencies and remaining fractions.

S. 7.7

Example

EXAMPLE 9.5 COMPLETE A RELIABILITY ANALYSIS USING THE NORMAL AND LOG-NORMAL DISTRIBUTIONS

Using the same data from Examples 6.3 and 9.1–9.4, complete a reliability analysis, estimating (a) the required mean concentrations to be maintained in order to comply with different reliability levels and (b) the expected percentages of compliance with the standard.

Input data, already presented in the previous Examples 9.1–9.4:

- *Arithmetic mean of the monitored data:* $m'_x = 3.16\ \text{mg/L}$
- *Arithmetic standard deviation:* $s = 1.04\ \text{mg/L}$
- *Coefficient of variation:* $\text{CV} = 1.04/3.16 = 0.33$
- *Quality standard:* $X_s = 4.0\ \text{mg/L}$
- *Measured mean/quality standard:* $m'_x/X_s = 3.16/4.00 = 0.79 = 79\%$

Excel

Note: This example is also available as an Excel spreadsheet.

Solution:

(a) Estimation of the required mean concentrations to be maintained in order to comply with different reliability levels

We will analyse the following reliability levels (probability of compliance with the regulatory standards): 50%, 80%, 90%, 95%, and 99%.

With these values and the input data presented above, we set up the following computational table:

Reliability level (%)	Z value	COR		Mean concentration to be maintained (m_x)	
		Normal distribution	Log-normal distribution	Normal distribution	Log-normal distribution
50	0.000	1.00	1.05	**4.00**	**4.21**
80	0.842	0.78	0.80	**3.13**	**3.22**
90	1.282	0.70	0.70	**2.81**	**2.79**
95	1.645	0.65	0.62	**2.60**	**2.49**
99	2.326	0.57	0.50	**2.27**	**2.00**

Note: The values above may differ slightly from those calculated below, because the table was built using values calculated in the Excel spreadsheet, without rounding up of numbers.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

The calculations were done using the following formulae:

- *Z* value: Excel function NORM.S.INV(reliability level). The reliability level should be entered as normal numbers (e.g., 0.90) and not as percentages (e.g., 90%).
  For instance, for the reliability level of 90% → NORM.S.INV(0.90) = 1.282.

The following calculations have been done using, for instance, reliability level of 90%, $Z = 1.282$, and $CV = 0.33$ (input data). The calculations for the other reliability levels follow the same structure.

- COR for the *normal distribution* (Equation 9.16):

$$COR = \frac{1}{1 + Z \cdot CV} = \frac{1}{1 + 1.282 \times 0.33} = 0.704$$

- COR for the *log-normal distribution* (Equation 9.17):

$$COR = \sqrt{CV^2 + 1} \times \exp\left\{-Z \cdot \sqrt{\ln(CV^2 + 1)}\right\}$$

$$= \sqrt{0.33^2 + 1} \times \exp\left\{-1.282 \cdot \sqrt{\ln(0.33^2 + 1)}\right\} = 0.698$$

- Mean concentration to be maintained (Equation 9.13) – *normal distribution*:

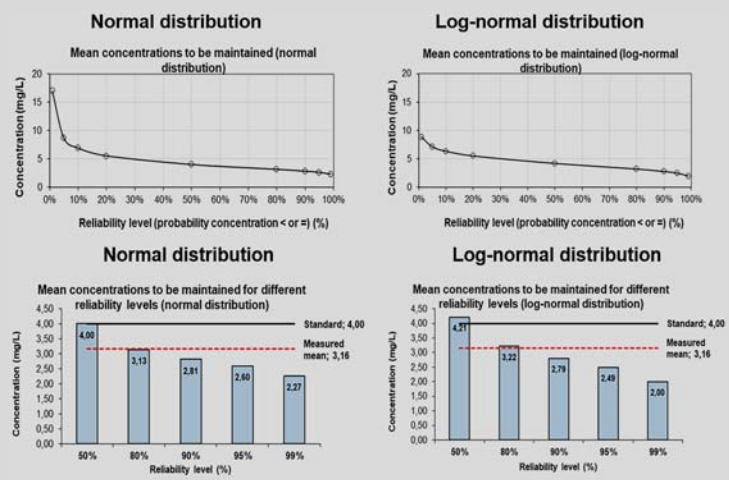
$$m_x = (COR) \cdot X_s = 0.704 \times 4.0 = 2.81 \text{ mg/L}$$

- Mean concentration to be maintained (Equation 9.13) – *log-normal distribution*:

$$m_x = (COR) \cdot X_s = 0.698 \times 4.0 = 2.79 \text{ mg/L}$$

Therefore, we should aim at maintaining the mean concentrations shown in the last two columns of the table, in order to comply with the reliability levels specified in the first column of the table.

The graphs below expand the results from the table. In the Excel spreadsheet, other reliability levels have been included. The graphs present the mean concentrations to be kept, comparing with the actual mean of the data (3.16 mg/L) and the regulatory standard (4.00 mg/L).



Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

(b) Expected probabilities of compliance with the standard

Given the mean value of the monitored data ($m'_x = 3.16$ mg/L), the CV of the monitored data (CV = 0.33), and the regulatory quality standard ($X_s = 4.00$ mg/L), we can estimate the expected probability of compliance with the standard using the normal and log-normal distributions.

- Z value for the *normal distribution* (Equation 9.18):

$$Z = \frac{\left(\frac{1}{m'_x/X_s} - 1\right)}{\text{CV}} = \frac{\left(\frac{1}{3.16/4.00} - 1\right)}{0.33} = 0.811$$

- Z value for the *log-normal distribution* (Equation 9.19):

$$Z = -\frac{\ln\left[\frac{m'_x}{X_s} \cdot \frac{1}{\sqrt{\text{CV}^2 + 1}}\right]}{\sqrt{\ln(\text{CV}^2 + 1)}} = -\frac{\ln\left[\frac{3.16}{4.00} \cdot \frac{1}{\sqrt{0.33^2 + 1}}\right]}{\sqrt{\ln(0.33^2 + 1)}} = 0.898$$

- Expected probability of compliance with standard – **normal distribution** (Equation 9.22):

$$\%\text{ of compliance} = 100 \times \text{NORM.S.DIST}(Z_{1_a}; \text{TRUE})$$
$$= 100 \times \text{NORM.S.DIST}(0.811; \text{TRUE}) = \mathbf{79.1\%}$$

- Expected probability of compliance with standard – **log-normal distribution** (Equation 9.22):

$$\%\text{ of compliance} = 100 \times \text{NORM.S.DIST}(Z_{1_a}; \text{TRUE})$$
$$= 100 \times \text{NORM.S.DIST}(0.898; \text{TRUE}) = \mathbf{81.5\%}$$

S. 9.6

Now, let us compare these values with the probabilities of compliance calculated using the **frequency analysis** approach, explained in Section 9.6 and covered in Example 9.4, using the same input data. In Example 9.4, the calculated probabilities of compliance with the standard were, as expected, the same: **79.1%** (*normal distribution*) and **81.6%** (*log-normal distribution*; difference of one decimal case only).

For the sake of an additional comparison, the **probability of compliance calculated directly using the monitored data** indicates that from the 36 data obtained, 29 complied with the standard, giving a probability of compliance of $29/36 = 0.806 =$ **80.6%** (see Example 9.3). Note that with this sample data set, the three methods produce very similar results. However, with other data sets, the results obtained from these three different methods could be very different.

9.8 CONTROL CHARTS

Advanced

9.8.1 Introductory concepts on statistical process control

Statistical process control is part of a set of statistical methods for the control of a system and the improvement of quality. It is also useful for indicating actions that may lead to better process stability and the reduction of variability. Basically, there are seven main tools used for **statistical process control** (Montgomery, 2009): a histogram or a stem-and-leaf plot, a check sheet, a Pareto chart, a cause-and-effect diagram, a defect concentration diagram, a scatter diagram, and a control chart.

All of these tools are important, but the **control chart** is the most sophisticated process control technique and has some important potential applications for monitoring quality control in treatment plants. Despite being widely used for industrial processes, the use of control charts has been less

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

common for water and wastewater treatment plants, perhaps since the characteristics of operational and environmental factors can sometimes strongly influence performance. Nevertheless, we feel that, as long as the uncertainty and variability are appropriately characterized, control charts can indeed be useful tools for monitoring the quality of treatment plants and ambient water bodies. In particular, we will show you how they can be useful tools for you to identify trends, peaks, disturbances, or unusual sources of variability in the data, and to obtain information that can be used to improve the process operating conditions.

As mentioned, the quality of environmental systems and performance of treatment plants are subject to high variability due to operational and environmental factors. Therefore, we need to learn how to characterize sources of variability in our data. In the context of treatment plants, when *unusual sources of variability* in the data are observed, it may be due to one of the two possible conditions:

- **Non-assignable causes** (also called *random*, *common*, or *chance* causes). They are inherent to the process, such as variations of the influent flows, concentrations, and characteristics, besides environmental conditions. Their occurrence is typical of a *process operating under statistical control*.
- **Assignable causes** (also called *special* causes). The assignable causes are those that arise in a sudden or abnormal way and can be identified and potentially eliminated from the system, such as mechanical failures and operational problems. In general, only assignable causes are susceptible to intervention. Their occurrence is typically associated with a *process operating out of normal control*.

Within the domain of the control charts that comprise statistical process control, there are several variants, such as

- Control chart for means ($\bar{x}$ chart)
- Control chart for process variation (R chart)
- Control chart for proportion of failure (p-chart)
- Control chart for the number of defects per item (c chart)

In this book, we will devote special attention to the **control chart for means ($\bar{x}$ chart)**, given its more widespread use, and the suitable application for treatment plant effluents. Since we have already analysed the percentage of conformity or non-conformity in this chapter, we will also cover the **control chart for proportion of failure (p-chart)**.

In this section, to keep things simple, we will present **only the main concepts associated with control charts**. If you find them useful, you should consult statistical textbooks that provide more details about this topic: many of them have dedicated chapters for statistical process control. Also, there are textbooks entirely devoted to statistical process control, for instance, Burr (1976) and Montgomery (2009).

9.8.2 Concepts behind a control chart for means

Advanced

(a) Structure of a control chart for means

Control charts have been developed and used mainly for industrial processes. In industry, the product is expected to comply with certain conditions, and generally it is expected that the characteristics of the end-product will be as close as possible to a prespecified mean value, with a variability that remains inside **prespecified acceptable boundaries**. The mean value and the acceptable variability boundaries for the product define the main elements of a control chart:

S. 4.5.3

- ***Centre line***: This is defined as the expected mean value.
- ***Upper control limit*** **(UCL)**: This can be defined either as an upper confidence limit or an upper prediction limit (see Sections 4.5.3 and 4.5.4 for more detailed information about confidence and prediction intervals).

S. 4.5.4

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

- ***Lower control limit*** **(LCL)**: This can be defined either as a lower confidence limit or a lower prediction limit (see Sections 4.5.3 and 4.5.4 for more detailed information about confidence and prediction intervals).

With these concepts in mind, a **typical control chart for means** would look like the one presented in Figure 9.8. The graph shows the effluent characteristics, in this case represented by the **average values from multiple samples** collected each day, for example (it could also be the average value for multiple samples collected each week or each month). If multiple independent samples are collected each day, then the average from each day constitutes the average of the group and is plotted in the graph. In this graph, we see nine groups, and each group represents one day with, say, eight samples or measurements made per day. Therefore, we can say that we have a sequence of nine days with a sample size of $n = 8$ per day. But what is more important now is to understand the structure of the control chart. Besides the points representing the effluent characteristics, we see the three control lines. The **centre line** aims to represent the long-term mean of the effluent characteristics, *obtained when the system is under control* (therefore, operating with only non-assignable or random variations). The lines for the **upper control limit** (**UCL**) and **lower control limit** (**LCL**) represent the boundaries of what is considered a system under control.

Note that since our plotted points are *average values computed from, say, $n = 8$ replicate measurements made each day*, then **the upper and lower control limits will be defined as upper and lower confidence limits.**

If the process is under control, a high percentage (this percentage is defined by our chosen confidence level, e.g., 95%) of the plotted daily sample averages will be inside the region defined by UCL and LCL. Therefore, you must choose an α value; see Equations 9.22 and 9.23 below; remember that an α value of 0.05 is equivalent to a confidence level of $1 - 0.05 = 0.95$ or 95%. The variation in the daily sample averages is expected to be due to non-assignable causes, and the spread of the points in the graph should follow a *random pattern*, with only 1 out of every 20 points falling above the UCL or below the LCL (assuming you choose an α value of 5%; e.g., $1/20 = 0.05 = 5\%$). This is the case with the sample points shown in Figure 9.8 – they display a random pattern, and thus the process is assumed to be in control, and no actions are needed.

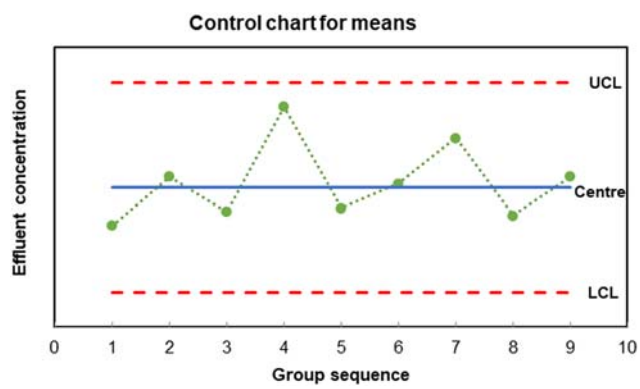


Figure 9.8 Example of the basic components of a control chart for means, with the centre line, the upper control limit (UCL), the lower control limit (LCL), and values of the effluent concentrations (represented by the average of samples collected in each group, e.g., each day).

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

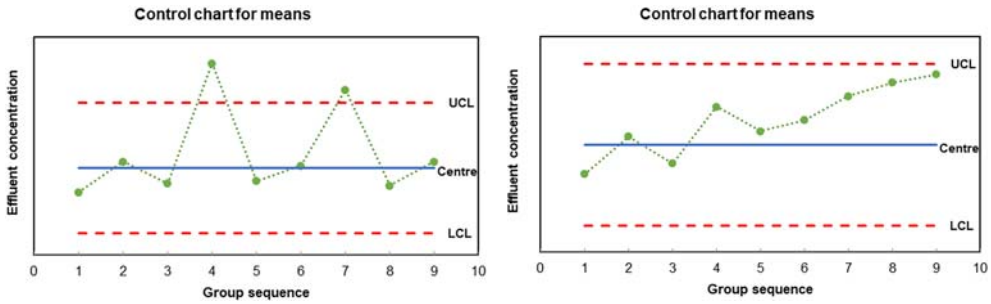


Figure 9.9 Example of control charts for means indicating that the system is not under control: (a) there are sample points above the upper control limit UCL and (b) the sample points are not randomly distributed and present an increasing pattern.

Now, let us analyse **possible departures** from the situation of a system under control. In Figure 9.9a, we see that for two out of nine days ($2/9 = 22\%$), our **sample averages were above the upper control limit**. Therefore, our interpretation is that the process may be out of control (because $22\% > 5\%$). The concentrations obtained on these two days are probably associated with **assignable causes**, and action may be required to bring the system back into control. If we look at the right-hand-side graph (Figure 9.9b), we see that all points are within the control limits. However, they are **not randomly distributed**, and an upward trend is evident from the data sequence. We still do not know for sure, but it is possible that the increasing trend will continue in the next few days to come, and the next data points may soon cross the upper boundary defined by UCL. Therefore, we cannot consider the system to be under control, and actions are probably required to bring the system back into control.

(b) Control limits and associated probabilities for periodic sample means ($\bar{x}$)

When determining the LCL and the UCL, we need to use *long-term averages and standard deviations*, in order to characterize the expected mean and standard deviation *as precisely as possible*. Normally, control charts operate using the population mean (μ) and the population standard deviation (σ) rather than a sample average ($\bar{x}$) and a sample standard deviation (s) (see Chapter 5). These values are, of course, unknown to us, though over a long period of time and many samples collected and analysed under normal operating conditions, we will gain considerable insight about them.

C. 5

In this section, we will use the notation μ and σ for the *long-term average and standard deviation*, even though these measurements will come from samples (which are normally denoted $\bar{x}$ and s). By using this notation, we are distinguishing the long-term average (presumed to be the population average) from averages calculated from smaller periodic samples collected hourly, daily, weekly, etc. Likewise, we distinguish the long-term standard deviation (presumed to be the population standard deviation) from a standard deviation calculated from the smaller periodic samples. We will use the notation $\bar{x}$ and s for the *average and standard deviations calculated from these periodic samples*. As such, it is important to note that your values for μ and σ should be based on many samples collected over a long period of time under normal operating conditions.

Now, imagine that our treatment process produces an effluent with a long-term average concentration of 10.0 mg/L and a long-term standard deviation of 2.0 mg/L. In order to assess adherence to the specifications using a control chart, assume four independent samples are collected each day ($n = 4$), and the average value ($\bar{x}$) of these four samples is plotted in the

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

control chart. Note that the control chart plots the **averages of the samples** collected each day, and that is why it is called control charts for *means*. Of course, other time intervals (hours, weeks, months, etc.) may be used in the control chart, depending on sampling frequency. In these cases, our groups can be constituted by average values of independent samples collected every hour (n samples per hour), every day (n samples per day), every week (n samples per week), and so on.

Figure 9.11 shows a control chart for our example where the long-term mean concentration is 10.0 mg/L and the long-term standard deviation of the concentration is 2.0 mg/L, and suppose that the number of samples per group (i.e., per day in this case) is $n = 4$.

We will now introduce a new concept called the **standard error of the mean** (i.e., the long-term process average), which is calculated as follows (Montgomery, 2009):

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \tag{9.21}$$

where

$\sigma_{\bar{x}}$ = standard error of the long-term (process) average
σ = long-term (process) standard deviation
n = number of samples per group.

In our example, the *standard error of the long-term (process) average* ($\sigma_{\bar{x}}$) can be computed as follows:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{2.0}{\sqrt{4}} = 1.0 \text{ mg/L}$$

Therefore, if the treatment process is in control with a long-term mean value of 10.0 mg/L, the standard error ($\sigma_{\bar{x}}$) for our sample size of $n = 4$ is 1.0 mg/L, and we assume that the data are normally distributed, then using the *central limit theorem*, we would expect $100(1 - \alpha)\%$ of the

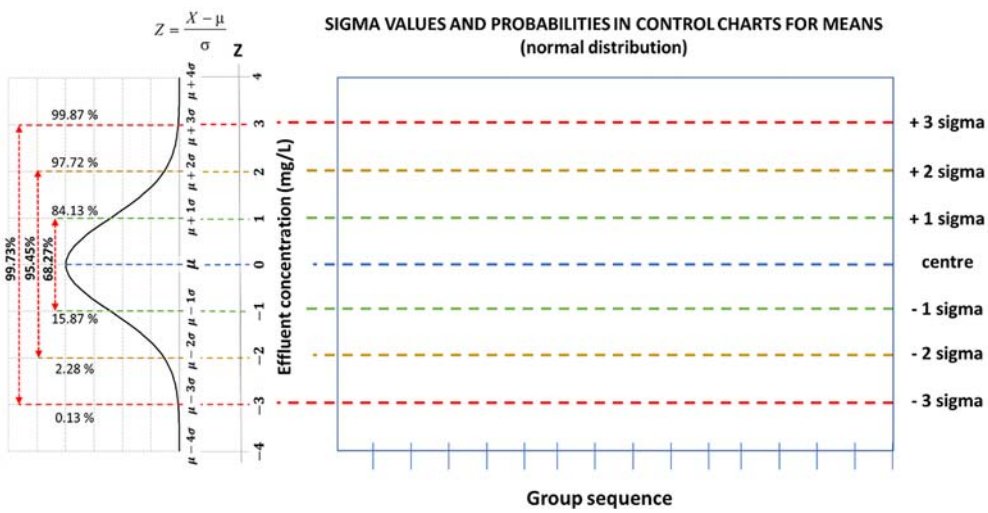


Figure 9.10 Control charts for means with the prediction intervals (sigma values) used to monitor the percentage of data points from future samples that will fall within a particular range, assuming a normal distribution.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

periodic sample means ($\bar{x}$) to fall between the limits defined by UCL and LCL (Montgomery, 2009):

$$\text{UCL} = \mu + Z_{\alpha/2} \cdot \sigma_{\bar{x}} \tag{9.22}$$

$$\text{LCL} = \mu - Z_{\alpha/2} \cdot \sigma_{\bar{x}} \tag{9.23}$$

Note that in this case, UCL and LCL are **upper and lower confidence limits**, since we are monitoring control based on periodic sample means ($\bar{x}$).

Now, after having discussed these basic concepts, we can go back to the data from our example. Using a value of $Z_{\alpha/2} = 3$, we obtain the values of the control limits by applying Equations 9.22 and 9.23 (using a sample mean of 10.0 mg/L and a standard error of 1.0):

$$\text{UCL} = \bar{x} + Z_{\alpha/2} \cdot \sigma_{\bar{x}} = 10.0 + 3 \times 1.0 = 13.0 \text{ mg/L}$$

$$\text{LCL} = \bar{x} - Z_{\alpha/2} \cdot \sigma_{\bar{x}} = 10.0 - 3 \times 1.0 = 7.0 \text{ mg/L}$$

C. 5

Recall from Figure 4.2a and the discussion in Chapter 5 that, for a normal distribution, ~68% of future data points fall within one standard deviation of the mean, ~95% of future data points fall within two standard deviations of the mean, and ~99% of future data points fall within three standard deviations of the mean. The same is true for standard errors and future sample means. That is, ~68% of future sample means will fall within one standard error of the true population mean, ~95% of future sample means fall within two standard errors of the true population mean, and ~99% of future sample means fall within three standard errors of the true population mean.

(c) Control limits and associated probabilities for individual data points

S. 4.5.4

C. 8

S. 8.2.5

Based on the concept described above, you can see that control charts can also be used to monitor the performance of individual data points collected in a time series. The probabilities of the occurrence of a single future data point (assuming a normal distribution) is defined by the prediction interval (see Section 4.5.4). Figure 9.10 shows the prediction interval used with a typical control chart for means. Part of this figure is based on Figure 8.8 and Table 8.1 (Chapter 8), when we presented the properties of normal distribution and standard normal variable Z (Section 8.2.5). In Figure 9.10, we present the control chart with different values of $\boldsymbol{Z}$ (+3, +2, +1, 0, −1, −2, −3), which is directly associated with σ (in the quality control literature, this is often spelled out as '**sigma**').

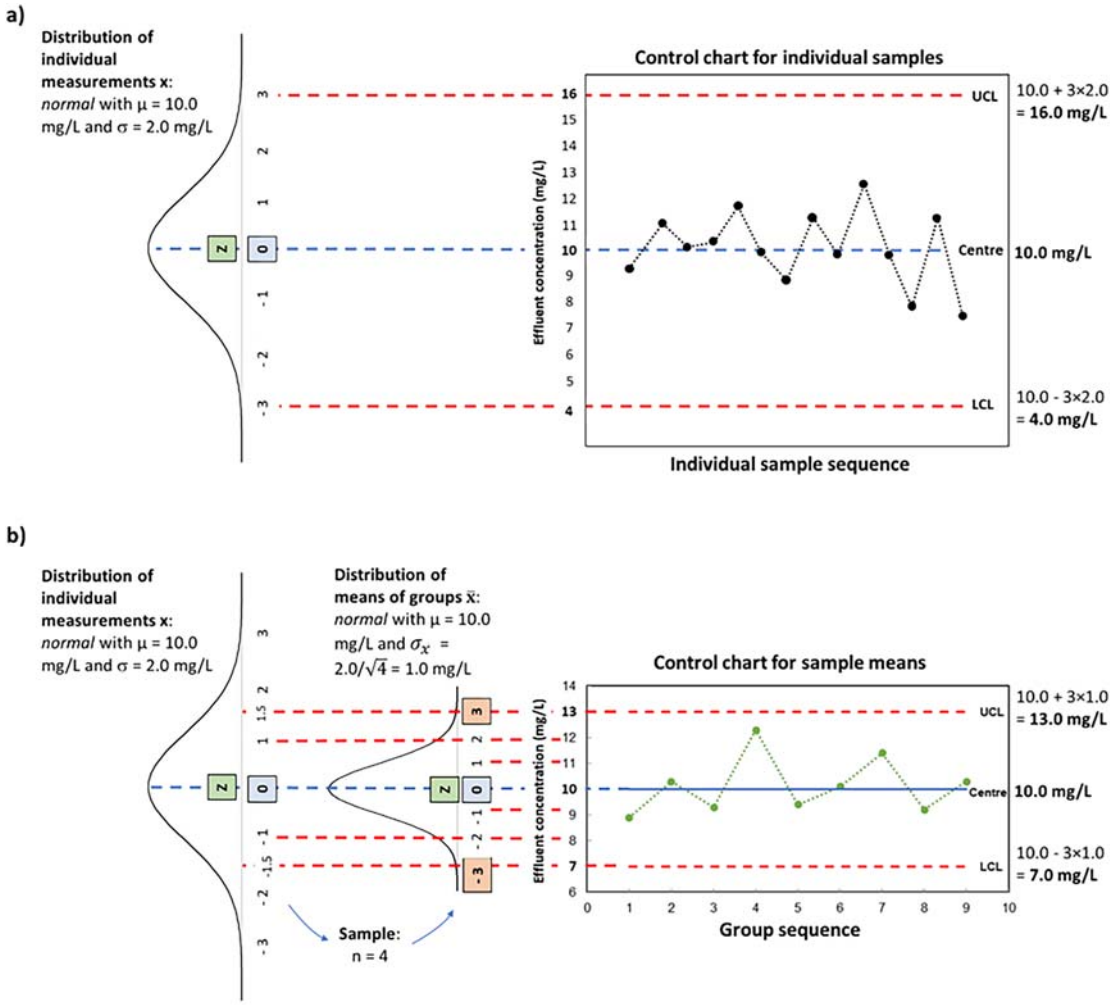
From Figure 9.10, we see that for a **normal distribution prediction interval**, we expect that **>99%** of the values of future samples will be **inside** the limits defined by the *mean* $\pm$ 3 × σ (or Z between –3 and + 3). In other words, we expect that, if data are normally distributed, almost all of the future data points will be within the boundaries defined in what is called a **three-sigma control chart** in the quality control literature. Here, the UCL and LCL are defined using σ instead of $\sigma_{\bar{x}}$.

$$\text{UCL} = \mu + Z\sigma \tag{9.24}$$

$$\text{LCL} = \mu - Z\sigma \tag{9.25}$$

We can analyse the graph for any Z value. For instance, we can expect that ~95% of the data will be between the boundaries of $Z = -2$ and $Z = +2$, and that ~68% should fall between $Z = -1$ and $Z = +1$. Since we want to have a high probability of occurrence of data inside our control limits, traditionally the **three-sigma control chart** is adopted, with the upper control limit set at $\mathbf{UCL} = +\mathbf{3}\boldsymbol{\sigma}$ and the lower control limit set at $\mathbf{LCL} = -\mathbf{3}\boldsymbol{\sigma}$.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest



S. 9.3.2

Figure 9.11 Explanation of the underlying concepts behind the control chart for means derived from the data presented in Section 9.3.2, assumption of a normal distribution. Process mean $\mu = 10.0$ mg/L and process standard deviation $\sigma = 2.0$ mg/L shown for (a) individual samples and (b) sample means from groups, with number of samples per group $n = 4$. Source: Inspired by a figure presented in Montgomery (2009).

Figure 9.11 summarizes the sequence of our calculations to establish the control limits. On the left-hand side, we have the expected distribution of all our *individual measurements*. Our long-term mean concentration is 10.0 mg/L with a standard deviation of 2.0 mg/L. Since our control graph is based on groups comprising internal samples, we move to the second distribution shown in the figure. This is the distribution of the *means of the various groups*. We know that each group is composed of four samples ($n = 4$), and therefore, the standard error of the sample average is $2.0/\sqrt{4} = 1.0$ mg/L (as described in Equation 9.21). The upper and lower control limits, for a three-sigma control chart for means, have just been calculated in the preceding paragraph, using

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Equations 9.22 and 9.23. It is important that you really see that the **control limits have been calculated using the distribution of the group means** (second distribution on the figure) and *not based on the distribution of the original individual measurements* (first distribution on the left-hand side of the figure). The distribution of the group means is less widespread than the distribution of the individual measurements, because of the factor $\sqrt{n}$. The figure also shows, for the sake of information, the lines associated with $Z \pm 1$ and $Z \pm 2$.

At this point, we need to clarify the concepts associated with **control limits** and **specifications limits**, which are frequently used interchangeably by some authors but are inherently different (Montgomery, 2009):

- **Control limits** are *internally* driven by the natural variability of the process, as measured by the process standard deviation.
- **Specification limits** are determined *externally* and may be set by managers or regulators. One such specification could be the standard defined by the regulatory agency or the target established by the management, as widely discussed in the introductory sections of this chapter.
- There is *no mathematical or statistical relationship between the control limits and specification limits*.
- We should *not incorporate specification limits in control charts for means* (these can be included in control charts with individual observations, not averages, as presented in Section 9.8.5).

S. 9.8.5

Even though we can consider that a **treatment plant** could be managed like an **industrial process**, whose end-product is the final effluent, there are some underlying differences when interpreting a control chart:

- The **raw material** that is used in a treatment plant cannot be controlled. An industry can control the quality of the raw materials it buys, but a treatment plant, with few exceptions, must treat whatever comes in. Nevertheless, management programmes that require the pre-treatment of industrial wastewater collected in a municipal system can help reduce some of the variability associated with the raw wastewater inputs. Domestic sewage quality is often quite predictable if it is only originating from households. The diurnal fluctuations in influent quality can be dampened by adding an equalization basin (see Section 2.2.4 and Example 2.3). In the case of water treatment plants, raw water sources may be influenced by environmental events, such as rainfall. Still, the temporal variability of water and wastewater sources is often not controllable by the plant operator.
- Treatment plants usually operate without any control of **environmental conditions**, such as temperature and rainfall, which may affect treatment performance. An industry is expected to have more control on external factors.
- Some wastewater treatment systems, especially those based on natural technologies (e.g., wetlands or ponds), have very few or no **manipulated variables** that can be used to control the process. Only large-scale and long-term maintenance activities, such as desludging a pond or unclogging a wetland, may help restore a deteriorating effluent quality.
- As a result of the points mentioned above, **variability of effluent quality** in treatment plants is likely to be greater than that of end products from industrial processes. Nevertheless, if you have a good understanding about the variability of your system over a long period of time, control charts can still provide some useful insight about the treatment plant's performance.

S. 2.2.4

(d) Other control limits and control zones

We have emphasized the widespread utilization of control charts based on 3σ (three-sigma) control limits, a range which should account for $>99\%$ of future samples or sample means.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

If values are outside these limits more than 1% of the time, then we can assume that the system is out of control and corrective actions are required. Likewise, if we see a non-random (e.g., increasing) trend, we can assume that there may be a problem with the operation of the system that may require corrective actions.

We can also specify **warning limits** that differ from the control limits based on the fact that they use a different ***Z* value**; for instance, we can specify an upper and lower limit based on $\mu +/- 2\sigma$ and call them the **upper warning limit (UWL)** and the **lower warning limit (LWL).** Table 9.5 presents a summary of these possibilities.

As mentioned before, we should remember that the control of an end-product by an industry involves different objectives in comparison with the end-product from a treatment plant (final effluent):

- **Industry**: End-product quality should be **as close as possible to the centre line**, indicating little variability of the product. Provided the centre line represents a value that reflects well the specifications established by the industry, we can infer that *the closer the values in the control chart are to the centre line*, *the better the process operation and control*.
- **Treatment plant**: If we think in terms of a **pollutant**, in principle, we can judge that *the lower its concentration in the end-product (final effluent)*, *the better the process operation and control*.

Therefore, for **treatment plants**, it would be unfair to say that *values below the centre line, and especially below the lower warning limit (LWL) and lower control limit (LCL)*, indicate a system in or approaching an out-of-control state. Our conclusion would be exactly the opposite: we have indications that the *system is working well, in fact even better than expected based on past performance under normal operating conditions*, which is what we used to establish our control limits (i.e., the long-term running averages and standard deviations).

Based on these considerations, the phrases used in the figures presented in Table 9.5 could be reworded. Figure 9.12 shows some proposed different wording that can be used as a more appropriate nomenclature for control charts applied to a treatment plant. Remember that **regular operation**, with its average value and standard deviation, *is the one we used to establish our control limits* and should be set based on long-term running averages and standard deviations from samples collected during normal operating conditions.

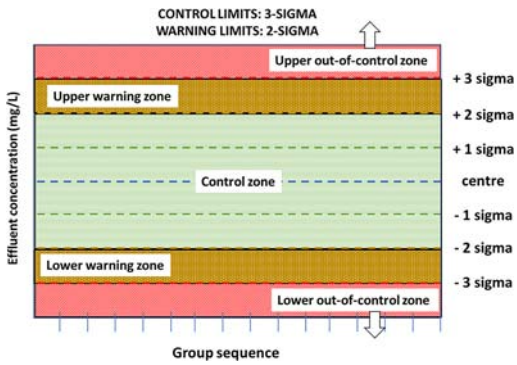
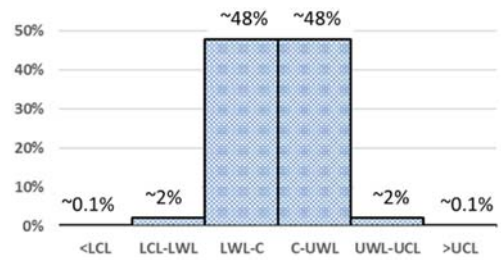
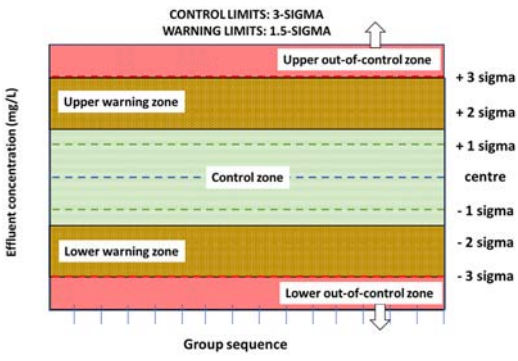
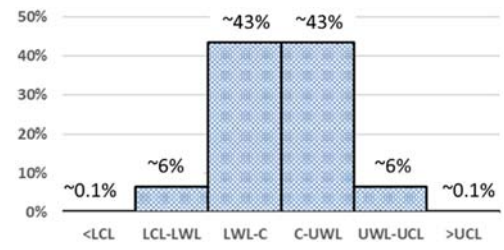
You may have noted that we are mentioning very little about **water bodies**, and most of our examples are based on treatment plants. This is because the applications for control charts in a water body are more limited (there is more natural variation and less control over concentrations in this natural system). However, the underlying concepts can also potentially be used for the water quality in some water bodies (such as reservoirs used for water supply).

Also, note that we are concentrating on the description of the monitoring of a constituent that is a pollutant. If we were studying a **constituent to be preserved** (e.g., dissolved oxygen in a water body), our zones would be inverted in both charts in Figure 9.4: the upper zones would indicate better quality. A similar comment can be made for **removal efficiencies in treatment plants**: higher values indicate better performance.

In all control charts shown in this section, if we are referring to **concentrations**, we **cannot plot negative values**, since they have no physical meaning. Even if in the calculation of our lower limits (LWL or LCL), we obtain negative values, our graph must start from the value of zero. A similar comment applies to **removal efficiencies** in treatment plants: our charts **must not have control lines above 100%**.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Table 9.5 Description of possible specifications for the warning limits.

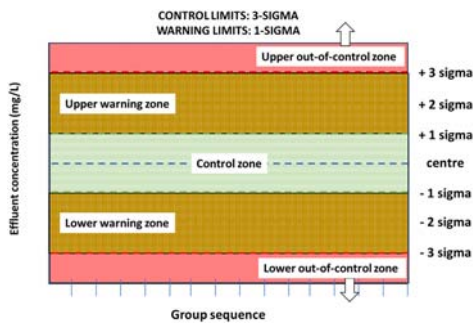
Control and Warning Limits	Expected Percentage of Data Inside Each Zone
Warning Limit Set at 2σ • Widely used limit for UWL and LWL • Relatively narrow range for warning zone, especially for environmental data (only ~2% + ~2% = ~4 – 5% of the data are expected to be inside the warning zone) • Control zone covers ~48% + ~48% = ~95% of the expected data	
	
Warning Limit Set at 1.5σ • Uses the same sum of sigma values of the control zone (control zone uses 1.5 + 1.5σ and warning zone also uses 1.5 + 1.5σ) • Allows a small increase in the warning zone, that can now contain more values (~6% + ~6% = ~12 – 13% of the data are expected to be inside the warning zone) • Control zone covers ~43% + ~43% = ~86 to 87% of the expected data	
	

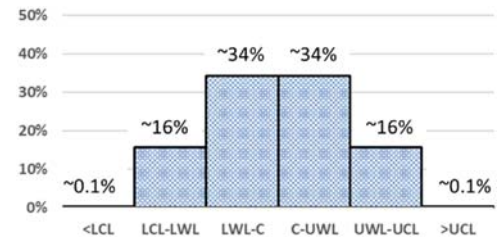
(*Continued*)

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Table 9.5 Description of possible specifications for the warning limits (*Continued*).

Control and Warning Limits	Expected Percentage of Data Inside Each Zone
Warning Limit Set at 1σ • Control zone covers an interval of $1+1=2\sigma$, equal to the interval of the upper warning zone (2σ) and lower warning zone (2σ) • Allows a more balanced coverage between control and warning zones • Warning zone can now contain ~16% + ~16% = ~32% of the data that are expected to be inside the warning zone • Control zone covers a lower probability range (~34% + ~34% = ~68% of the expected data) compared with the other options	





Notes: UCL, upper control limit; UWL, upper warning limit; Centre, central line; LWL, lower warning limit; LCL, lower control limit.
UCL is always set at $+3\sigma$ and LCL is always set at -3σ.

Advanced

9.8.3 Setting up a control chart for means (assumption of a normal distribution)

(a) Elements in the control chart for means

S. 9.8.2

Initially, we will mention that this section describes the procedure for setting up a control chart for means under the assumption that the **data are normally distributed**. The concepts presented in the preceding section (Section 9.8.2) are all based on the normality of the data. However, we have mentioned several times that environmental data are frequently asymmetrical and in many cases can be better represented by a log-normal distribution. *Control charts for log-normally distributed data are presented in Section 9.8.4*.

S. 9.8.4

S. 9.8.2

The basic elements for the control charts for means have been described in Section 9.8.2. However, in Section 9.8.2, we used the concepts of long-term **process mean** and long-term **process standard deviation**, which are descriptors of the **population of the data**, and as mentioned previously, should be calculated from many data points obtained during long-term operation of a system under control, reflecting its regular operating conditions.

But now, we have the practical challenge to construct a control chart for means based on **samples derived from our monitoring programme**. The following elements are integral parts of this control chart:

- Individual data obtained in the monitoring programme with the system in a regular operation under control

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

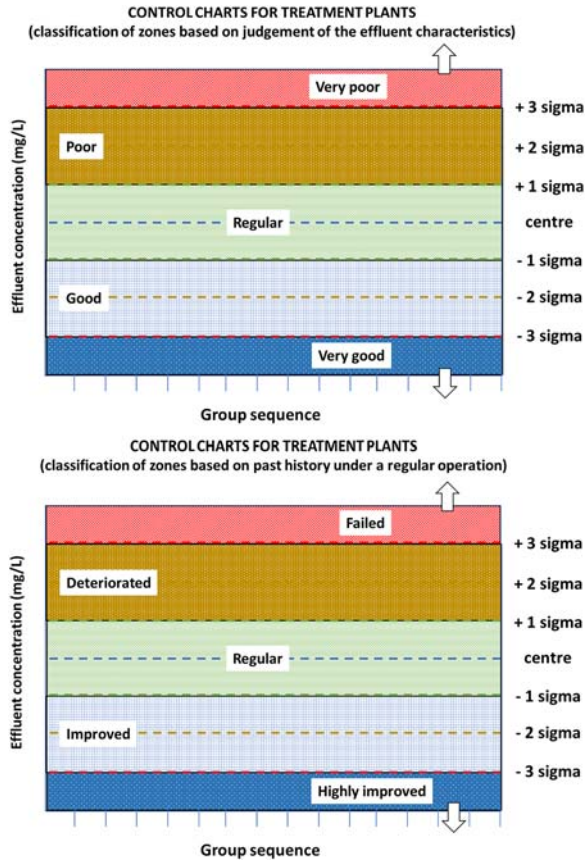


Figure 9.12 Possible classification of operating zones for the effluent characteristics from a treatment plant. You may choose different nomenclatures (based on your own concepts) and different limits for each zone (based on the selection of sigma values).

- Number of subgroups (k) (e.g., days, weeks, months, etc.)
- Number of samples per subgroup (n) (i.e., number of samples collected per day, per week, per month, etc.)
- Mean of the means ($\overline{\overline{x}}$)
- Standard error of the mean ($\sigma_{\bar{x}}$)

With these values we are able to establish the centre line and the control limits, using the theory presented in Section 9.8.2 and the practical considerations included here. The control limits are obtained from the equations summarized in Table 9.6.

S. 9.8.2

The calculations of $\overline{\overline{x}}$ and $\sigma_{\bar{x}}$ are presented in subsections to follow.

(b) Individual data obtained in the monitoring programme

In order to derive the limits of our control chart, the monitoring data we will use need to **reflect the system operating under control**. From our historical data, we may want to exclude periods in which we consider that the system was not operating according to its regular mode. By doing this, we assume that the variability in our data is explained only by non-assignable causes. This means

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Table 9.6 Summary of equations for setting up a control chart for means under the assumption of normal distribution.

Control limit	Formula	Value of Z_p (Sigma)	Equation number
UCL	$\mathrm{UCL} = \overline{\overline{x}} + Z_p \cdot \sigma_{\bar{x}}$	3.0	(9.26)
UWL	$\mathrm{UWL} = \overline{\overline{x}} + Z_p \cdot \sigma_{\bar{x}}$	1.0, 1.5, or 2.0	(9.27)
Centre line	$\overline{\overline{x}}$	0	(9.28)
LWL	$\mathrm{LWL} = \overline{\overline{x}} - Z_p \cdot \sigma_{\bar{x}}$	1.0, 1.5, or 2.0	(9.29)
LCL	$\mathrm{LCL} = \overline{\overline{x}} - Z_p \cdot \sigma_{\bar{x}}$	3.0	(9.30)

Note: $\overline{\overline{x}}$, mean of the means; Z_p, sigma values; $\sigma_{\bar{x}}$, standard error of the mean.

we may have to eliminate values that we feel were outliers because of their excessive variability resulting from assignable causes.

After we set our control chart, reflecting a controlled operation, we will use the graph with newly obtained data, and we will interpret whether the system is remaining under control. The variability in the data from an out-of-control process is assumed to be associated with assignable causes and should not be used for deriving the long-term running average or standard deviation used to set the control and warning limits. In some cases, we may need to re-establish new control limits, if conditions change considerably. An example would be if a treatment plant starts to receive an increased influent load, and this becomes the new regular operation, even if the treatment performance decreases. If this is the case, we should calculate new control limits based on long-term averages and standard deviations calculated from data collected under these new operating conditions.

(c) Determining the number of samples to be collected per time interval (*n*)

If you are working with a control chart for sample means, and not individual measurements, you might be wondering what an appropriate sample size for each time interval is (e.g., for each day, or each week, or each month, etc.). In industrial operations, the *n* values that comprise each sample are generally samples collected from a batch of 'widgets' being produced, for example. In water and wastewater treatment plants, there is not a direct parallel, since the product of this process is the continuous flow of water. Nevertheless, we can define sample periods and sample sizes based on our general knowledge about temporal trends in the performance of treatment plants.

In principle, for us to use statistical quality control methods, we need to accept the assumptions that the data analysed are **statistically independent** (*not autocorrelated*) and originate from a population that is normally distributed (this assumption has already been discussed in this section). For environmental pollutants, Gilbert (1987) states the difficulty in rigorously complying with these assumptions but states that quality control charts can still provide useful information for purposes of process control. In the study of treatment plants, the first assumption may be met in some cases, depending on the frequency used to collect the samples. Berthouex and Hunt (1975) comment that effluent data from wastewater treatment plants collected at time intervals between two and three days may be considered independent enough so that treatment plants that apply such sampling intervals, or even larger ones, may use control charts with relative safety. For treatment plants that adopt a more intensive monitoring programme, such as large treatment plants, industries and plants located in environmentally sensitive areas, and the dependence and autocorrelation of data, should be considered. There are advanced techniques for addressing the case of autocorrelated data, but these will not be covered here (see Chapter 11 for the concept of autocorrelation).

C. 11

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

As a simple approach for organizing our subgroups, we could consider the following possibilities, bearing in mind that we want to adopt small subgroups ($n < 10$):

- If the monitoring data are obtained on a **daily basis**, we may consider forming *subgroups representing every week*, what will lead us to have $\boldsymbol{n = 7}$ (if all days in the week are monitored) or $\boldsymbol{n = 5}$ (if no weekend monitoring is practiced).
- If the monitoring data are obtained on a **weekly basis**, we may form *subgroups representing every month*, what will result in $\boldsymbol{n = 4}$ (four weeks per month).
- Apply any other grouping criterion we may find justifiable and explain this in the report.

An additional factor to consider when establishing a control chart programme is the **frequency** *with which we obtain our monitoring data*. This will vary widely from plant to plant, from a research project to research project, and also from constituent to constituent. Data collection frequency (period between consecutive water samples collected) may be obtained on a *monthly*, *weekly*, *daily*, or *hourly* basis if it involves sampling and laboratory analysis, or on a *near-continuous* basis if it involves sensors.

As a summary, we propose subgroups with $\boldsymbol{n}$ **between 4 and 7** for the purposes of the quality control charts discussed here. For the sake of simplicity, we will consider here only **fixed sizes of subgroups** (n adopted as a fixed value, not varying during the study time).

(d) Mean value versus the mean of the mean values

The centre line on a control chart represents the long-term mean value (μ) of the process. This value is usually unknown but can be estimated by averaging a large number of sample means obtained when the process was in control. For instance, suppose we are interested in making a control chart for individual samples (e.g., Figure 9.11a). If we have a total of 200 individual and independent random samples collected during normal operating conditions, we would calculate the average, which will define the **centre line** of our control chart for means.

However, if instead of making a control chart for sample means (groups of samples; e.g., Figure 9.11b), then the centre line of our control chart should be based on the mean of mean values (from samples collected during normal operating conditions). For instance, suppose we collect $n = 4$ samples per day and our control chart is based on the mean value from the four daily readings. To set the centre line of our control chart, we should calculate the mean of means as follows:

$$\text{Centre line} = \bar{\bar{x}} = \frac{\sum_{i=1}^{k} \bar{x}_i}{k} \tag{9.31}$$

where

$\bar{\bar{x}}$ = mean of the means (mean of the k values of subgroup means ($\bar{x}_i$))
$\bar{x}_i$ = mean of each subgroup (mean of the n values that comprise a subgroup)
k = number of subgroups (a subgroup can be a day, a week, a month, etc.)

(e) Process standard deviation and standard deviation of the means

The process standard deviation is also usually unknown to us, but we can estimate it from our long-term large sample of data obtained when the process was in control. Even though this is a sample standard deviation (usually denoted with s), we will call it σ for the purposes of a control chart, if it is calculated from a very large and long-term running sample from conditions where the system was operating in control.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

If we are interested in making a control chart for individual samples (e.g., Figure 9.11a) and we have a total of 200 individual and independent random samples collected during normal operating conditions, then we would calculate σ as the standard deviation of those 200 values.

However, if we are making a control chart for sample means (groups of samples; e.g., Figure 9.11b), then the standard deviation of the whole sample s is not an unbiased estimator of σ (Montgomery, 2009). To circumvent this, traditionally, in most statistical textbooks and software, the standard deviation of each subgroup is computed using an estimate based on the amplitude or range R (difference between the largest and smallest value) of the values inside each subgroup. The random variable $W = R/\sigma$ is called the *relative range*. The distribution of W has been well studied. It indicates that the mean of W is a constant d_2 that depends on the size of the sample. Therefore, the factor d_2 represents the relation between the amplitude and the standard deviation.

Based on all k values of R, we estimate the **process standard deviation σ** based on the mean of R ($\bar{R}$) and the constant d_2.

$$\hat{\sigma} = \frac{\bar{R}}{d_2} = \frac{\left(\sum_{i=1}^{k} R_i\right)/k}{d_2} \tag{9.32}$$

where

$\hat{\sigma}$ = estimated process standard deviation
R_i = amplitude (or range) of the values inside each subgroup i (difference between the largest and smallest values inside each subgroup i)
$\bar{R}$ = mean of the k values of amplitude (or range) R_i
d_2 = tabulated constant, representing the relation between the amplitude and the standard deviation.

The statistical textbooks that cover quality control charts traditionally include a table with the values of d_2 in an appendix. The d_2 values are reproduced in Table 9.7, for different values of n.

Montgomery (2009) states that this traditional approach of estimating the process standard deviation based on the range $\bar{R}$ and d_2 loses its efficiency when the sample size n gets very large, because the range method ignores all the information in the sample between the two extremes of maximum and minimum in the subgroup. Although Table 9.7 presents the values of d_2 for n up to

Table 9.7 Values of the factor d_2 for control charts for means, as a function of the number of data in each subgroup (n).

n	d_2	n	d_2	n	d_2
2	1.128	10	3.078	18	3.640
3	1.693	11	3.173	19	3.689
4	2.059	12	3.258	20	3.735
5	2.326	13	3.336	21	3.778
6	2.534	14	3.407	22	3.819
7	2.704	15	3.472	23	3.858
8	2.847	16	3.532	24	3.895
9	2.970	17	3.588	25	3.931

Note: Obs, not valid for $n > 25$. Wider applicability of the procedure is for $\boldsymbol{n \leq 10}$.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

25, its best utilization would be for **small sample sizes**. Mendenhall and Sincich (1988) suggest the following approaches for *estimating the process standard deviation* $\hat{\sigma}$, based on the number of data points inside each subgroup (n):

- ***For n ≤ 15***: use the average of the ranges of each subgroup ($\bar{R}$) and the tabulated value of d_2. Some references (Levine *et al.*, 1998; Montgomery, 2009) suggest the limit of $n \leq 10$. This concern may not be a problem in our case, since we are proposing to use small subgroups (n *between 4 and 7*).
- ***For n > 15***: use the standard deviation s of the whole data set.

In the spreadsheet associated with Example 9.6, we have worksheets using both approaches, for you to compare the results.

Since the chart we are studying is a control chart for *means*, we need to estimate the **standard error of the mean** $\sigma_{\bar{x}}$ based on the process standard deviation σ:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \tag{9.33}$$

where

$\sigma_{\bar{x}}$ = standard error of the mean
σ = process standard deviation
n = number of data points inside each subgroup.

Combining Equations 9.32 and 9.33 for small sample sizes, we obtain the estimation of the **standard error of the mean ($\sigma_{\bar{x}}$)** based on $\bar{R}$, d_2, and n as

$$\boxed{\sigma_{\bar{x}} = \frac{(\bar{R}/d_2)}{\sqrt{n}}} \tag{9.34}$$

(f) Amplitude of values in each subgroup

Strictly speaking, we should analyse the behaviour of the sequence of amplitudes (ranges) of each subgroup over time. The control chart for means should be only applied if the amplitudes

Table 9.8 Equations for estimating appropriate values of the control lines of a control chart for means under the assumption of normal distribution (*small sample sizes in each subgroup;* ***n ≤ 15***).

Control limit	Formula	Value of Z_p (Sigma)	Equation number
UCL	$\mathrm{UCL} = \bar{\bar{x}} + Z_p \cdot \frac{(\bar{R}/d_2)}{\sqrt{n}}$	3.0	(9.35)
UWL	$\mathrm{UWL} = \bar{\bar{x}} + Z_p \cdot \frac{(\bar{R}/d_2)}{\sqrt{n}}$	1.0, 1.5, or 2.0	(9.36)
Centre line	$\bar{\bar{x}}$	0	(9.37)
LWL	$\mathrm{LWL} = \bar{\bar{x}} - Z_p \cdot \frac{(\bar{R}/d_2)}{\sqrt{n}}$	1.0, 1.5, or 2.0	(9.38)
LCL	$\mathrm{LCL} = \bar{\bar{x}} - Z_p \cdot \frac{(\bar{R}/d_2)}{\sqrt{n}}$	3.0	(9.39)

Note: $\bar{\bar{x}}$, mean of the means (mean of the k values of mean $\bar{x}_i$ for each subgroup); Z_p, sigma values; $\bar{R}$, mean of the k values of amplitude (or range) R_i; n, number of data in each subgroup.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Table 9.9 Equations for calculating the control lines of a control chart for means under the assumption of normal distribution (*large sample sizes in each subgroup;* ***n*** $\geq$ ***15***).

Control limit	Formula	Value of Z_p (Sigma)	Equation number
UCL	$\mathrm{UCL} = \overline{\overline{x}} + Z_p \cdot \frac{s}{\sqrt{n}}$	3.0	(9.40)
UWL	$\mathrm{UWL} = \overline{\overline{x}} + Z_p \cdot \frac{s}{\sqrt{n}}$	1.0, 1.5, or 2.0	(9.41)
Centre line	$\overline{\overline{x}}$	0	(9.42)
LWL	$\mathrm{LWL} = \overline{\overline{x}} - Z_p \cdot \frac{s}{\sqrt{n}}$	1.0, 1.5, or 2.0	(9.43)
LCL	$\mathrm{LCL} = \overline{\overline{x}} - Z_p \cdot \frac{s}{\sqrt{n}}$	3.0	(9.44)

Note: $\overline{\overline{x}}$, mean of the means (mean of the k values of mean $\bar{x}_i$ for each subgroup); Z_p, sigma values; s, standard deviation of the whole data set, that is, all $n \cdot k$ measurements; n, number of data in each subgroup.

are relatively stable. As a matter of fact, there are control charts for ranges (**R-charts**), but they will not be covered here, aiming at a simplified coverage of the subject.

(g) Control limits for the chart for means

Based on the considerations made on the preceding subsections, we can now summarize the equations used to choose appropriate values for the control lines (Tables 9.8 and 9.9).

Example

EXAMPLE 9.6 BUILD A CONTROL CHART FOR MEANS UNDER THE ASSUMPTION OF A NORMAL DISTRIBUTION

Using the data below (same data from other examples used in this chapter), build a control chart for means. Assume the data follow a normal distribution.

Data (values are in mg/L and are the same as in Example 6.3):

2.8	4.2	3.9	3.3	2.8	1.7	1.9	2.5	3.1	3.8	2.7	4.1
4.3	4.8	5.1	5.2	3.9	3.5	2.7	3.1	2.8	3.4	4.9	2.8
2.8	1.8	2.1	2.6	2.3	2.4	2.5	1.3	2.9	2.4	2.1	3.6

Excel

Note: This example is also available as an Excel spreadsheet.

Solution:

S. 9.8.3

(a) Decide on the number of data per subgroup (*n*)

Based on the considerations summarized in Section 9.8.3, we will use a small sample size per subgroup ($n = 4$). Since we have a total of 36 data points, the number of subgroups is $k = 36/4 = 9$ subgroups.

Ideally, we should have a larger number of subgroups (k) in order to derive our control limits. However, we will keep this example with only nine subgroups, in order to show you the entire procedure of setting up a control chart for means and allow comparisons with all other examples we carried out using the same dataset.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

(b) Calculate the mean and amplitude of each subgroup and the average values for the *k* subgroups

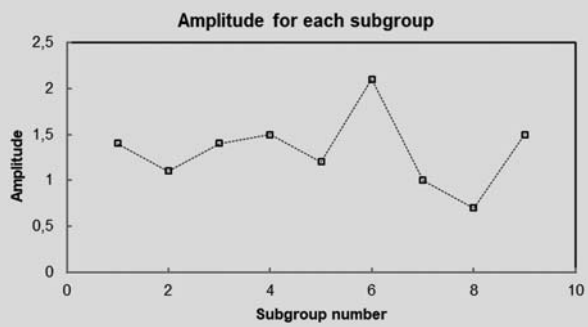
We organize our computational spreadsheet with one subgroup per row. In each row, we include four data values, since we decided on having $n = 4$. Since our number of data points per subgroup (n) is small, we will use the computations based on ranges R and factor d_2.

Subgroup	Measured data ($n = 4$)				Statistics for each subgroup			
					Mean ($\bar{x}_i$)	Maximum	Minimum	Amplitude (R_i) (max – min)
1	2.8	4.2	3.9	3.3	3.55	4.2	2.8	1.4
2	2.8	1.7	1.9	2.5	2.23	2.8	1.7	1.1
3	3.1	3.8	2.7	4.1	3.43	4.1	2.7	1.4
4	4.3	4.8	5.6	5.8	5.13	5.8	4.3	1.5
5	3.9	3.5	2.7	3.1	3.30	3.9	2.7	1.2
6	2.8	3.4	4.9	2.8	3.48	4.9	2.8	2.1
7	2.8	1.8	2.1	2.6	2.33	2.8	1.8	1.0
8	2.3	2.4	2.5	1.8	2.25	2.5	1.8	0.7
9	2.9	2.4	2.1	3.6	2.75	3.6	2.1	1.5
Mean values				$\bar{\bar{x}}=$	3.16		$\bar{R}=$	1.32

In the Excel spreadsheet that is associated with this example, there is no need for you to prepare manually the table as above – everything is done automatically.

(c) Graph for amplitudes (ranges)

At this point, it is instructive for us to plot a graph with the amplitudes to see whether they do not present any marked tendency or unusual behaviour. The most complete approach would be to construct a control chart for ranges but, in order to simplify this approach, these are not described in our book. However, they are easy to construct, and you can consult statistical textbooks to see how to build and interpret them. Based on a simple visual interpretation of the graph below, we do not detect any abnormal behaviour, and thus carry on with the construction of the control chart for means.



(d) Calculation of the mean of the means and the standard error of the mean

From the table shown above, we obtain two important values that will be directly used to calculate the values of the control lines:

- *Mean of the means*: $\bar{\bar{x}} = \mathbf{3.16\ mg/L}$
- *Mean of the amplitudes (ranges)*: $\bar{R} = \mathbf{1.32\ mg/L}$

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

For $n = 4$, we obtain the value of the factor $\boldsymbol{d_2}$ from Table 9.7 as **2.059**. With the values of the mean of the ranges ($\bar{R}$) and the factor d_2, we can estimate the process standard deviation ($\hat{\sigma}$) using Equation 9.32:

$$\hat{\sigma} = \frac{\bar{R}}{d_2} = \frac{1.32}{2.059} = 0.642\ \text{mg/L}$$

However, for calculating the control lines, we need the value of the *standard error of the mean* ($\sigma_{\bar{x}}$). This can be calculated by dividing the value of the process standard deviation $\hat{\sigma}$ by $\sqrt{n}$ or by using Equation 9.34:

$$\sigma_{\bar{x}} = \frac{\hat{\sigma}}{\sqrt{n}} = \frac{0.642}{\sqrt{4}} = 0.321\ \text{mg/L}$$

or

$$\sigma_{\bar{x}} = \frac{(\bar{R}/d_2)}{\sqrt{n}} = \frac{(1.32/2.059)}{\sqrt{4}} = 0.321\ \text{mg/L}$$

Note: We have adjusted some numerical values here to match those obtained using the more accurate calculations using the Excel spreadsheet.

(e) Calculation of the control limits

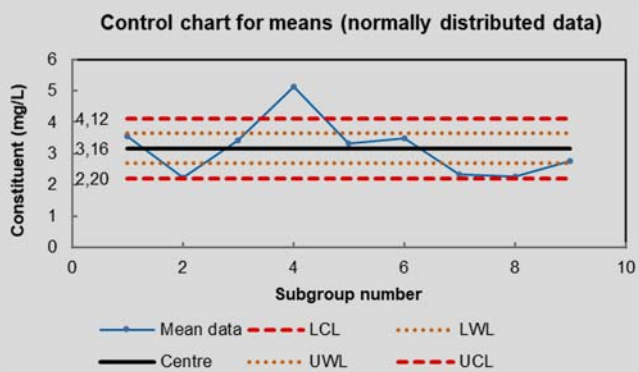
For calculating the control limits, we need to decide on the sigma (σ) values to be adopted. For the upper (UCL) and lower (LCL) **control** limits, we will use the traditional value of 3σ ($\boldsymbol{Z_p = 3.0}$). For the upper (UWL) and lower (LWL) **warning** limits, in this example, we will use 1.5σ ($\boldsymbol{Z_p = 1.5}$). We can change the sigma values very easily in the Excel spreadsheet, and everything will be calculated automatically.

We now have all the elements for calculating the control limits. We can either use the direct equations in Table 9.6 or the detailed equations in Table 9.8 (they are essentially the same).

- Upper control limit: $\text{UCL} = \bar{\bar{x}} + Z_p.\ \sigma_{\bar{x}} = 3.16 + 3.0 \times 0.321 = 4.12\ \text{mg/L}$
- Upper warning limit: $\text{UWL} = \bar{\bar{x}} + Z_p.\ \sigma_{\bar{x}} = 3.16 + 1.5 \times 0.321 = 3.64\ \text{mg/L}$
- Centre : $\bar{\bar{x}} = 3.16\ \text{mg/L}$
- Lower warning limit: $\text{LWL} = \bar{\bar{x}} - Z_p.\ \sigma_{\bar{x}} = 3.16 - 1.5 \times 0.321 = 2.68\ \text{mg/L}$
- Lower control limit: $\text{LCL} = \bar{\bar{x}} - Z_p.\ \sigma_{\bar{x}} = 3.16 - 3.0 \times 0.321 = 2.20\ \text{mg/L}$

(f) Control chart for means

The resulting control chart for means is plotted below, including the control lines and the average values of each subgroup ($\boldsymbol{\bar{x}_i}$). Since the graph is based on the assumption of normal distribution, you can see that the upper and lower control lines are symmetrical around the centre line.



Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

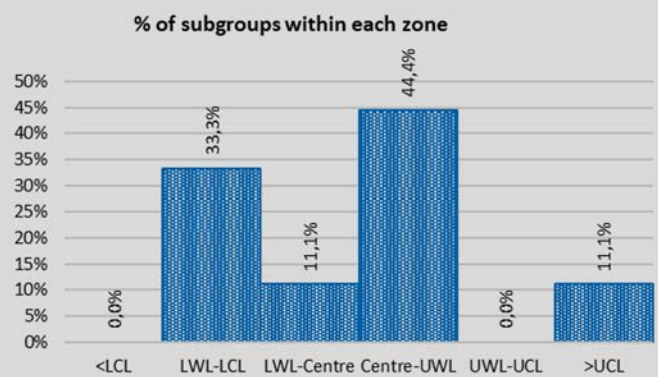
We can also see that one of the values (subgroup 4) is above the upper control limit (UCL), what could characterize that the system was not under control when we derived the control limits. At this stage, we should examine the possible assignable causes related to subgroup 4. We could rebuild the control chart obtaining only measurements indicating that the system was fully under control, or we could remove subgroup 4 from the analysis and calculate the control limits again.

Another additional aspect in the interpretation of the chart is that there is no clear upward or downward trend, and we could assume that the data are approximately randomly distributed around the mean (we could carry out statistical tests for supporting this assumption, but these types of tests will not be applied here in this example to keep things simple).

With the exception of subgroup 4, the other values situated above the centre line are still all below the upper warning limit (UWL). Therefore, they can be considered to be associated with only random variation and are not a matter of concern. Now, of course, if we were to have a large sample size, for example, of 100 subgroups, just by random variation, it would be reasonable to expect that approximately 12–13% of the subgroup means will be within the warning zone just by random non-assignable variability (see Table 9.5). If we had 1000 subgroups, we would expect to see approximately 1 of them show a mean value outside of the control limits, just by random chance (e.g., $0.1\% \times 1000 = 1$). It is all a matter of which sigma values we use and how large of a sample size we use.

Back to our example, from the nine values plotted, four of them are below the centre line, in the zone which indicates good performance. Note that the concepts of *poor* and *good* performance are only related to the internal history of the system and the control limits set for it, and not to any external evaluation or specification, such as a standard or target level. What is considered good or poor for one system may not be good or poor for another system, since the limits reflect only the internal behaviour of each treatment plant.

These comments are also supported by the column chart below, which presents the percentage of subgroup mean values that are included in each of the control zones.



Advanced

9.8.4 Setting up a control chart for means (assumption of a log-normal distribution)

(a) Concepts for a control chart for means under the assumption of a log-normal distribution

Throughout this chapter, we kept a balance between the normal and the log-normal distributions. Whenever possible, we presented the theory and applications for quality assessment based on both distributions.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

In the field of quality control charts, the traditional literature concentrates on charts based on the assumption of normality, as we described in the previous section. The control lines were symmetrical around the centre line.

Now, we will introduce the less-commonly applied concept of control charts based on asymmetrical distributions and the **assumption of log-normality** of the data. Some researchers, such as Ferrell (1958), Morrison (1958), Joffe and Sichel (1968), and Cheng and Xie (2000), have studied this previously and have proposed different approaches. The method proposed by Morrison (1958) continues to be widely cited and adopted by authors interested in the application of statistical process control to data from populations that follow a log-normal distribution (Burr, 1976; Cheng & Xie, 2000; Gilbert, 1987; Shaban, 1988; Shore, 1998; Shore, 2000). However, the application of this method requires the use of tabulated constants, some of which are not easy to understand how they have been developed.

Oliveira and von Sperling (2009) proposed a basic approach for control charts based on **simple properties of the log-normal distribution**. This is the approach presented here, with the incorporation of some additional concepts, so that it maintains the same fundamental structure as the one adopted for the normal distribution.

Our coverage here will be very direct, since the conceptual concepts of control charts for means have already been presented in the preceding sections.

The definition of the *number of samples per subgroup (n)* and the *number of subgroups (k)* will be the same.

S. 5.6.4

(b) Geometric mean and geometric standard deviation

The calculation of the geometric mean (M_g) and geometric standard deviation (s_g) will use the concepts previously described in Chapter 5 (Sections 5.6.4 and 5.7.e) and Chapter 8 (Section 8.3.1):

S. 5.7

$$\text{Geometric mean } \bar{X}_g = 10^{\text{(arithmetic mean of the log10 of the original values)}} \tag{9.45}$$

$$\text{Geometric standard deviation } s_g = 10^{\text{(standard deviation of the log10 of the original values)}} \tag{9.46}$$

S. 8.3.1

Therefore, initially, we need to calculate the *log$_{10}$ of all our original observations*. We then *split our log-transformed data into different subgroups*.

For each subgroup, we calculate the **geometric mean $\bar{x}_g$** using Equation 9.45. The mean of the geometric means ($\bar{\bar{x}}_g$) will be the centre line of the chart, given by

$$\text{Centre line} = \bar{\bar{x}}_g = \frac{\sum_{i=1}^{k} \bar{x}_{g_i}}{k} \tag{9.47}$$

where

$\bar{\bar{x}}_g$ = mean of the geometric means of the different subgroups i (mean of the k values of the subgroup geometric means ($\bar{x}_{g_i}$))

$\bar{x}_{g_i}$ = geometric mean of each subgroup i (geometric mean of the n values that comprise a subgroup i)

k = number of subgroups.

To calculate the **geometric standard deviation s_g**, which will allow us to calculate the sigma control lines, we need to take into account the two approaches described in Section 9.8.3.e for control charts assuming normal distribution, but now with the relevant adaptations for the log-transformed data:

S. 9.8.3

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

- *For small sizes of the subgroups* ($\boldsymbol{n \leq 15}$), we estimate the geometric standard deviation based on the amplitude of the log-transformed data and the factor d_2

$$s_g = 10^{\frac{\overline{R}\ \log 10\ \text{data}}{d_2}} = 10^{\frac{\left(\sum_{i=1}^{k} R_i\ \log 10\ \text{data}\right)/k}{d_2}} \tag{9.48}$$

where

s_g = estimated geometric standard deviation
R_i $\log_{10}$ data = amplitude (or range) of the $\log_{10}$-transformed values inside each subgroup (difference between the largest and smallest values inside each group)
$\bar{R}$ $\log_{10}$ data = mean of the k values of amplitude (or range) R_i
d_2 = tabulated constant, representing the relation between the amplitude and the standard deviation (see Table 9.7).

- *For large sizes of the subgroups* ($\boldsymbol{n > 15}$), we use the geometric standard deviation based on the whole data set (all $n \cdot k$ observations), according to Equation 9.46, adapted:

$$s_g = 10^{(\text{arithmetic standard deviation of the log10 transformed data})} \tag{9.49}$$

You remember that, for the control charts for normal distribution, after calculating the process standard deviation, we needed to divide it by $\sqrt{n}$ in order to obtain the standard error of the mean. For the log-normal distribution, we will do something similar but with some additional details, as explained in subsection 'c'.

These concepts will be well clarified in Example 9.7.

(c) Calculation of the control limits

S. 8.3.5

To calculate the sigma control limits, we will adapt the approach used for the normal distribution control chart, based on the comments we formulated in Chapter 8 (Section 8.3.5, describing measures of central tendency and variation in the log-normal distribution).

Table 9.10 Unified approach for calculating the control lines of a control chart for means under the assumptions of normal and log-normal distributions.

Control Limit	Normal Distribution	Log-normal Distribution	Value of Z_p (Sigma)	Equation Number
UCL	$\text{UCL} = \overline{\overline{x}} + s\,\frac{Z_p}{\sqrt{n}}$	$\text{UCL} = \overline{\overline{x}}_g \times s_g^{(Z_p/\sqrt{n})}$	3.0	(9.50)
UWL	$\text{UWL} = \overline{\overline{x}} + s\,\frac{Z_p}{\sqrt{n}}$	$\text{UWL} = \overline{\overline{x}}_g \times s_g^{(Z_p/\sqrt{n})}$	1.0, 1.5, or 2.0	(9.51)
Centre line	$\overline{\overline{x}}$	$\overline{\overline{x}}_g$	0	(9.52)
LWL	$\text{LWL} = \overline{\overline{x}} - s\,\frac{Z_p}{\sqrt{n}}$	$\text{LWL} = \overline{\overline{x}}_g \div s_g^{(Z_p/\sqrt{n})}$	1.0, 1.5, or 2.0	(9.53)
LCL	$\text{LCL} = \overline{\overline{x}} - s\,\frac{Z_p}{\sqrt{n}}$	$\text{LCL} = \overline{\overline{x}}_g \div s_g^{(Z_p/\sqrt{n})}$	3.0	(9.54)

Notes: n, number of data in each subgroup; Z_p, sigma values; $\overline{\overline{x}}$, mean of the means of the different subgroups i (mean of the k values of mean $\bar{x}_i$ for each subgroup); $\overline{\overline{x}}_g$, mean of the geometric means of the different subgroups i (mean of the k values of the subgroup geometric means ($\bar{x}_{g_i}$)); s, arithmetic standard deviation (calculated from the amplitude method or based on the standard deviation of the whole data set, depending on the size of the subgroups n); s_g, geometric standard deviation (calculated from the amplitude method or based on the geometric standard deviation of the whole data set, depending on the size of the subgroups n). The values of the arithmetic standard deviation (s) and geometric standard deviation (s_g) can be calculated as judged more appropriate, based on the sample size ($n \leq 15$ or $n > 15$), using the amplitude method or the standard deviation of the whole data set (see relevant text in Sections 9.8.3.e and 9.8.4.b).

S. 9.8.3
S. 9.8.4

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

For the *normal distribution*, we saw that the dispersion of the data around the mean μ for different quantities of standard deviation σ (standard normal variable Z) depended on an additive relationship: $\mu \pm \sigma$. For the **log-normal distribution**, the relations are multiplicative: $\boldsymbol{\mu_g \times / \div \sigma_g}$. Simply stated, we have

- What is 'addition' in a normal distribution is 'multiplication' in a log-normal distribution.
- What is 'subtraction' in a normal distribution is 'division' in a log-normal distribution.
- What is 'multiplication' in a normal distribution is 'raising to a power' in a log-normal distribution.

Based on this, Oliveira and von Sperling (2009) organized the equations for the control chart for means under the assumption of **log-normal distribution**, which are summarized in Table 9.10. For the sake of completeness, we also present the equations for **normal distribution** so that you can compare both approaches and see the unified concept. We have also standardized the notations for the parameters in both distributions and reorganized the equations for normal distribution so that it is easier for you to make the comparisons. The values of the *arithmetic standard deviation* (s) and *geometric standard deviation* (s_g) can be calculated as judged more appropriate, based on the sample size ($n \leq 15$ or $n > 15$), using the amplitude method or the standard deviation of the whole data set (see relevant text in Sections 9.8.3.e and 9.8.4.b).

S. 9.8.3

S. 9.8.4

Example

EXAMPLE 9.7 BUILD A CONTROL CHART FOR MEANS UNDER THE ASSUMPTION OF A LOG-NORMAL DISTRIBUTION

Using the same data from the examples used in this chapter, and especially Example 9.6, build a control chart for means. Assume a log-normal distribution for the data.

Data (values are in mg/L):

2.8	4.2	3.9	3.3	2.8	1.7	1.9	2.5	3.1	3.8	2.7	4.1
4.3	4.8	5.1	5.2	3.9	3.5	2.7	3.1	2.8	3.4	4.9	2.8
2.8	1.8	2.1	2.6	2.3	2.4	2.5	1.3	2.9	2.4	2.1	3.6

Excel

Note: This example is also available as an Excel spreadsheet.

Solution:

(a) Prepare a table with the $\log_{10}$ transformation of the data

We calculate the $\log_{10}$ of the original data and obtain the following table:

0.447	0.623	0.591	0.519	0.447	0.230	0.279	0.398	0.491	0.580	0.431	0.613
0.633	0.681	0.748	0.763	0.591	0.544	0.431	0.491	0.447	0.531	0.690	0.447
0.447	0.255	0.322	0.415	0.362	0.380	0.398	0.255	0.462	0.380	0.322	0.556

(b) Decide on the number of data points per subgroup (n)

We will use the same number of data points per subgroup adopted in Example 9.6: $n = 4$. Also, from Example 9.6, we see that we have a total of 36 data and obtain the same value of $k = 36/4 = 9$ subgroups.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

(c) Calculate the geometric mean and amplitude of each subgroup and the average geometric mean and amplitude for the *k* subgroups

We organize our computational table with one subgroup per row, in the same way as we did in Example 9.6. In each row, we include four data values, since we decided on having $n = 4$. Since our number of data per subgroup (n) is small, we will use the computations based on ranges and factor d_2. In the table, we insert the values of the $\log_{10}$-transformed data.

Subgroup	$\log_{10}$ of Measured Data ($n = 4$)				Statistics for Each Subgroup ($\log_{10}$-transformed data)				
					Mean	Geomean ($\bar{x}_{g_i}$)	Maximum	Minimum	Amplitude (R_i)
1	0.447	0.623	0.591	0.519	0.545	3.507	0.623	0.447	0.176
2	0.447	0.230	0.279	0.398	0.339	2.181	0.447	0.230	0.217
3	0.491	0.580	0.431	0.613	0.529	3.379	0.613	0.431	0.181
4	0.633	0.681	0.748	0.763	0.707	5.088	0.763	0.633	0.130
5	0.591	0.544	0.431	0.491	0.514	3.269	0.591	0.431	0.160
6	0.447	0.531	0.690	0.447	0.529	3.381	0.690	0.447	0.243
7	0.447	0.255	0.322	0.415	0.360	2.290	0.447	0.255	0.192
8	0.362	0.380	0.398	0.255	0.349	2.232	0.398	0.255	0.143
9	0.462	0.380	0.322	0.556	0.430	2.693	0.556	0.322	0.234
Mean values					$\bar{\bar{x}}_g =$	3.114		$\bar{R} =$	0.186

Geomean of each subgroup $= 10^{(\text{arithmetic mean of log10-transformed data})}$.

In the Excel spreadsheet that is associated with this example, there is no need to prepare manually the table as above – everything is done automatically.

(d) Calculation of the mean of the geometric means and the geometric standard deviation

From the table shown above, we obtain two important values, which will be directly used in the calculation of the control lines:

- *Mean of the geometric means*: $\bar{\bar{x}} = \mathbf{3.114\ mg/L}$ (centre line)
- *Mean of the amplitudes (ranges)*: $\bar{R} = \mathbf{0.186}$

For $n = 4$, we obtain the value of the factor $\boldsymbol{d_2}$ from Table 9.7 as **2.059**. With the values of the mean of the ranges ($\bar{R}$) and the factor d_2, we can estimate the geometric standard deviation (s_g) using Equation 9.48:

- *Geometric standard deviation*: $s_g = 10^{(\overline{R}\log 10\ \text{data}/d_2)} = 10^{(0.186/2.059)} = 1.231$

(e) Calculation of the control limits

For calculating the control limits, we need to decide on the sigma values to be adopted. We will use the same ones used in Example 9.6. For the upper (UCL) and lower (LCL) **control** limits, we will use the traditional value of 3 sigma ($\boldsymbol{Z_p = 3.0}$). For the upper (UWL) and lower (LWL) **warning** limits, we will use 1.5 sigma ($\boldsymbol{Z_p = 1.5}$). We can change the sigma values very easily in the Excel spreadsheet, and everything will be calculated automatically.

We now have all the elements for calculating the control limits. We will use the equations summarized in Table 9.10 for the log-normal distribution.

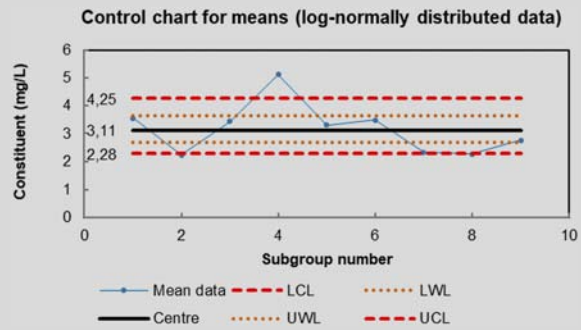
- Upper control limit: $\text{UCL} = \bar{\bar{x}}_g \times s_g^{\left(\frac{Z_p}{\sqrt{n}}\right)} = 3.114 \times 1.231^{\left(3.0/\sqrt{4}\right)} = 4.25\ \text{mg/L}$
- Upper warning limit: $\text{UWL} = \bar{\bar{x}}_g \times s_g^{\left(\frac{Z_p}{\sqrt{n}}\right)} = 3.114 \times 1.231^{\left(1.5/\sqrt{4}\right)} = 3.64\ \text{mg/L}$
- Centre: $\bar{\bar{x}} = 3.11\ \text{mg/L}$

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

- Lower warning limit: $LWL = \overline{\overline{x}}_g \div s_g^{\left(\frac{Zp}{\sqrt{n}}\right)} = 3.114 \div 1.231^{\left(1.5/\sqrt{4}\right)} = 2.66\ mg/L$
- Lower control limit: $LCL = \overline{\overline{x}}_g \div s_g^{\left(\frac{Zp}{\sqrt{n}}\right)} = 3.114 \div 1.231^{\left(3.0/\sqrt{4}\right)} = 2.28\ mg/L$

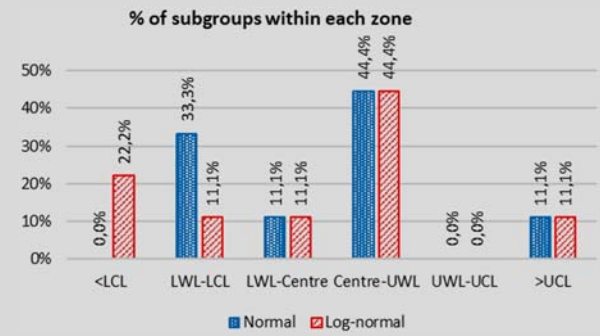
(f) Control chart for means

The resulting control chart for means is plotted below, including the control lines and the average values of each subgroup ($\bar{x}_i$). Since the graph is based on the assumption of log-normal distribution, we can see that the upper and lower control lines are not symmetrical around the centre line.



We can compare this chart for means, based on the assumption of log-normality of the data, with the one derived in Example 9.6, which was based on the assumption of normality of the data. There is not much difference in them in this particular example, because the original data were not strongly asymmetrical. However, as a general rule, in the log-normal graph, because of the asymmetry in the distribution, the **upper control limits are farther apart from the centre line**, indicating that there is more space for the points to be inside the control lines, that is, there is less chance for us to conclude that the process may be out of control. On the other hand, **the lower control lines are closer to the centre line**.

Similar to Example 9.6, we also present here a graph showing the percentage of subgroup mean values that are included in each of the control zones. In order to be more illustrative, we plot together the results from Example 9.6 (normal distribution) and those from the current example (log-normal distribution). Since the original data were not markedly asymmetrical, both results are not substantially different. The largest differences occurred in the lower control zones, which are more squeezed in the log-normal chart.



When doing this type of analysis on your data set, you might want to first conduct a test of goodness-of-fit to the normal versus the log-normal distribution (see Section 8.2.8), then use those results as a basis for your assumption about normality versus log-normality when establishing your control chart.

S. 8.2.8

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

9.8.5 Control chart for individual measurements (normal and log-normal distributions)

Advanced

When dealing with environmental monitoring data, there are several situations in which we need to work with individual measurements, that is, we establish our sample size as $\boldsymbol{n = 1}$. A common reason is when monitoring data become available slowly, or when they are costly to obtain, resulting in low monitoring frequencies, and the long interval between samples could cause problems in rational subgrouping (Hines *et al.*, 2003; Montgomery, 2009). If this is the case, **control charts for individual measurements** are a useful alternative.

The simplest version of this chart is when we use the **moving range (amplitude) of two successive measurements**:

$$\text{MR} = \text{ABS}\,(x_i - x_{i-1}) \tag{9.55}$$

where

MR = moving range of two consecutive values
ABS = absolute value (positive value of the difference between the two consecutive values)
x_i = measurement i, x_{i-1} = measurement $i - 1$.

The **centre line** is equal to the **arithmetic mean** $\bar{x}$ (for *normal distribution*) or **geometric mean** $\bar{x}_{\mathbf{g}}$ (for *log-normal distribution*) of the individual measurements (whole data set):

$$\text{Normal distribution: centre line} = \bar{\boldsymbol{x}} \tag{9.56}$$

$$\text{Log-normal distribution: centre line} = \bar{\boldsymbol{x}}_{\mathbf{g}} \tag{9.57}$$

The **standard deviation** is based on the mean amplitude of the moving ranges ($\bar{R}$) and the tabulated value of d_2. Since the moving ranges are based on two successive values, we obtain the value of d_2 from Table 9.7 as $\boldsymbol{d_2 = 1.128}$. Adapting Equations 9.34 and 9.48 for $n = 1$, we obtain

$$\text{Normal distribution: standard deviation: } s = \frac{\bar{R}}{d_2} \tag{9.58}$$

$$\text{Log-normal distribution: geometric standard deviation: } s_g = 10^{\frac{\bar{R}\log 10\ \text{data}}{d_2}} \tag{9.59}$$

where

s = estimated arithmetic standard deviation
s_g = estimated geometric standard deviation
$\bar{R}$ = mean of the moving ranges calculated from the original individual measurements
$\bar{R}$ = $\log_{10}$ data is the mean of the moving ranges calculated from the $\log_{10}$-transformed values of the individual measurements
d_2 = tabulated constant, representing the relation between the standard deviation and the amplitude. In this case, for two consecutive values, $d_2 = 1.128$ (see Table 9.7).

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Example

EXAMPLE 9.8 BUILD A CONTROL CHART FOR INDIVIDUAL MEASUREMENTS UNDER THE ASSUMPTIONS OF NORMAL AND LOG-NORMAL DISTRIBUTIONS

*Using the same data from the examples used in this chapter, and especially Examples 9.6 and 9.7, build a control chart for **individual measurements**. Consider the assumptions of normal and log-normal distributions for the data. Data (values are in mg/L):*

2.8	4.2	3.9	3.3	2.8	1.7	1.9	2.5	3.1	3.8	2.7	4.1
4.3	4.8	5.1	5.2	3.9	3.5	2.7	3.1	2.8	3.4	4.9	2.8
2.8	1.8	2.1	2.6	2.3	2.4	2.5	1.3	2.9	2.4	2.1	3.6

Excel

Note: This example is also available as an Excel spreadsheet.

Solution:

(a) Prepare a table with all the individual measurements, their $\log_{10}$ transformations, the moving range of the original values, and the moving range of the $\log_{10}$ values

Data sequence	Original data			$\log_{10}$ transformed data	
	Value	Moving range		$\log_{10}$ value	Moving range of $\log_{10}$ values
1	2.8	-	ABS(4.2-2.8)	0.447	-
2	4.2	1.4		0.623	0.176 ABS(0.623-0.447)
3	3.9	0.3	ABS(3.9-4.2)	0.591	0.032
4	3.3	0.6		0.519	0.073
5	2.8	0.5	$\log_{10}$(2.8)	0.447	0.071
6	1.7	1.1		0.230	0.217
7	1.9	0.2		0.279	0.048
8	2.5	0.6		0.398	0.119
9	3.1	0.6		0.491	0.093
10	3.8	0.7		0.580	0.088
11	2.7	1.1		0.431	0.148
12	4.1	1.4		0.613	0.181
13	4.3	0.2		0.633	0.021
14	4.8	0.5		0.681	0.048
15	5.6	0.8		0.748	0.067
16	5.8	0.2		0.763	0.015
17	3.9	1.9		0.591	0.172
18	3.5	0.4		0.544	0.047
19	2.7	0.8		0.431	0.113
20	3.1	0.4		0.491	0.060
21	2.8	0.3		0.447	0.044
22	3.4	0.6		0.531	0.084
23	4.9	1.5		0.690	0.159
24	2.8	2.1		0.447	0.243
25	2.8	0.0		0.447	0.000
26	1.8	1.0		0.255	0.192
27	2.1	0.3		0.322	0.067
28	2.6	0.5		0.415	0.093
29	2.3	0.3		0.362	0.053
30	2.4	0.1		0.380	0.018
31	2.5	0.1		0.398	0.018
32	1.8	0.7		0.255	0.143
33	2.9	1.1		0.462	0.207
34	2.4	0.5		0.380	0.082
35	2.1	0.3		0.322	0.058
36	3.6	1.5		0.556	0.234
Mean	**3.16**	**0.70**		**0.478**	**0.100**

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

(b) Calculate the means and the standard deviations

The arithmetic mean has been already calculated in the table:

$$\bar{x} = 3.16\,\text{mg/L}$$

The geometric mean is calculated from the mean of the $\log_{10}$-transformed data:

$$\bar{x}_g = 10^{0.478} = 3.01\,\text{mg/L}$$

The arithmetic standard deviation is calculated from Equation 9.58, based on the mean of the ranges ($\bar{R}$) and knowing that $d_2 = 1.128$:

$$s = \frac{\bar{R}}{d_2} = \frac{0.70}{1.128} = 0.62\,\text{mg/L}$$

The geometric standard deviation is calculated from Equation 9.59, based on the mean of the ranges of the $\log_{10}$-transformed data ($\bar{R}$ of $\log_{10}$ data) and knowing that $d_2 = 1.128$:

$$s_g = 10^{\frac{\overline{R}\,\log 10\,\text{data}}{d_2}} = 10^{0.100/1.128} = 1.225$$

(c) Control limits

As in Examples 9.8 and 9.9, for the upper (UCL) and lower (LCL) **control** limits, we will use the traditional value of 3 sigma ($\boldsymbol{Z_p = 3.0}$). For the upper (UWL) and lower (LWL) **warning** limits, we will use 1.5 sigma ($\boldsymbol{Z_p = 1.5}$).

Using Equations 9.60–9.64 presented in Table 9.11, we calculate our control limits:

Normal distribution

- Upper control limit: $\text{UCL} = \bar{x} + s \cdot Z_p = 3.16 + 0.62 \times 3.0 = 5.03\,\text{mg/L}$
- Upper warning limit: $\text{UWL} = \bar{x} + s \cdot Z_p = 3.16 + 0.62 \times 1.5 = 4.09\,\text{mg/L}$
- Centre line: $\text{Centre} = \bar{x} = 3.16\,\text{mg/L}$
- Lower warning limit: $\text{LWL} = \bar{x} - s \cdot Z_p = 3.16 - 0.62 \times 1.5 = 2.22\,\text{mg/L}$
- Lower control limit: $\text{LCL} = \bar{x} - s \cdot Z_p = 3.16 - 0.62 \times 3.0 = 1.29\,\text{mg/L}$

Log-normal distribution

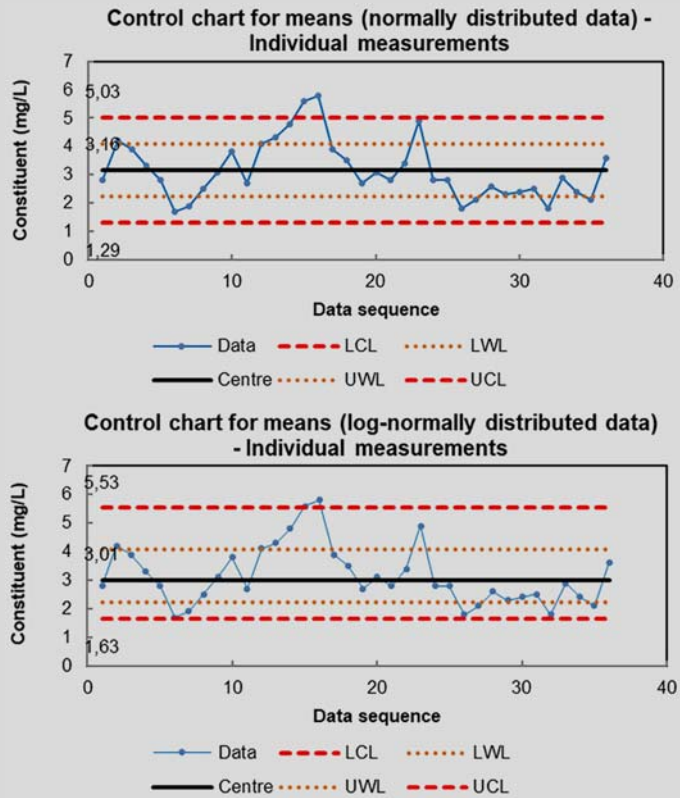
- Upper control limit: $\text{UCL} = \bar{x}_g \times s_g^{Z_p} = 3.01 \times 1.225^{3.0} = 5.53\,\text{mg/L}$
- Upper warning limit: $\text{UWL} = \bar{x}_g \times s_g^{Z_p} = 3.01 \times 1.225^{1.5} = 4.08\,\text{mg/L}$
- Centre line: $\text{Centre} = \bar{x}_g = 3.01\,\text{mg/L}$
- Lower warning limit: $\text{LWL} = \bar{x}_g \div s_g^{Z_p} = 3.01 \div 1.225^{1.5} = 2.22\,\text{mg/L}$
- Lower control limit: $\text{LCL} = \bar{x}_g \div s_g^{Z_p} = 3.01 \div 1.225^{3.0} = 1.63\,\text{mg/L}$

Note: There may be small differences between the values calculated directly in this example and those calculated in the Excel spreadsheet, due to rounding errors.

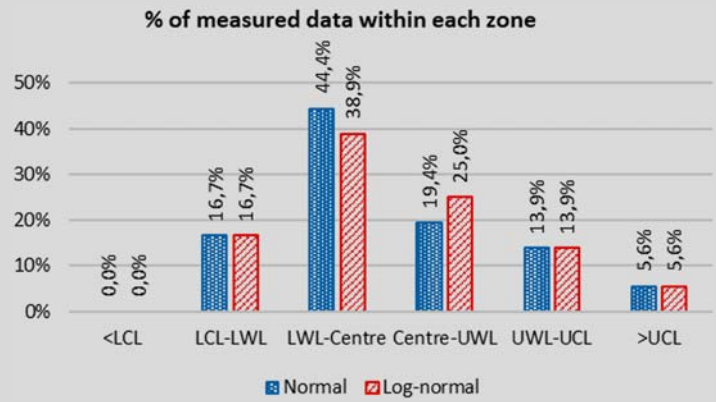
(d) Control charts for individual measurements

The resulting control charts, with the individual measurements and the control lines, are plotted below, for the assumptions of normal and log-normal distributions. The interpretation of the main elements has already been presented in Examples 9.6 and 9.7, with the difference here that all individual data points are plotted, and the control lines are different. But note that now the control lines are more distant from the centre line, in comparison when the data have been grouped into the rational subgroups, as was done with the control charts for means.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest



Similar to Examples 9.6 and 9.7, we also present a graph showing the percentage of values inside each of the control zones for the normal and log-normal distributions. Since the original data were not markedly asymmetrical, both results are not substantially different.



Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Table 9.11 Unified approach for calculating the control lines of a control chart for **individual measurements** under the assumptions of normal and log-normal distributions.

Control Limit	Normal Distribution	Log-normal Distribution	Value of Z_p (sigma)	Equation Number
UCL	$UCL = \bar{x} + s \cdot Z_p$	$UCL = \bar{x}_g \times s_g^{Z_p}$	3.0	(9.60)
UWL	$UWL = \bar{x} + s \cdot Z_p$	$UWL = \bar{x}_g \times s_g^{Z_p}$	1.0, 1.5, or 2.0	(9.61)
Centre line	$\bar{x}$	$\bar{x}_g$	0	(9.62)
LWL	$LWL = \bar{x} - s \cdot Z_p$	$LWL = \bar{x}_g \div s_g^{Z_p}$	1.0, 1.5, or 2.0	(9.63)
LCL	$LCL = \bar{x} - s \cdot Z_p$	$LCL = \bar{x}_g \div s_g^{Z_p}$	3.0	(9.64)

Notes: For individual measurements, $n = 1$. These equations are similar to those presented in Table 9.10 (control chart for means), substituting n by 1. Z_p, sigma values; $\bar{x}$, arithmetic mean of all the individual measurements; $\bar{x}_g$, geometric mean of all the individual data measurements; s, arithmetic standard deviation (calculated from the moving ranges of two consecutive values) (see Equation 9.58); s_g, geometric standard deviation (calculated from the moving ranges of the log10-transformed data of two consecutive values) (see Equation 9.59).

S. 9.8.3
S. 9.8.4

The control limits can be defined as presented in Section 9.8.3 (for *normal distribution*) and Section 9.8.4 (for *log-normal distribution*). We present in Table 9.11 a summary of the relevant equations. Note that these equations are similar to those presented in Table 9.10, taking into account that $n = 1$.

9.8.6 Control chart for the proportion of failures (p-chart)

Advanced

In this chapter, so far, we have interpreted compliance with targets or standards using several different statistical approaches. Now, we will use a control chart to analyse the **proportion of data** that is in **non-conformity** with the target or standard. We will consider that the non-conformity values are those that represent **failures** to comply with the specifications.

S. 9.4

Remember that in Section 9.4 we covered this subject, when we used the *Z-test for proportions* in the evaluation of compliance based on the proportion of non-conformity (failures) or conformity with the standard. We suggest that you go back to that section to reinforce your concepts on this topic.

Note that the control charts for means, analysed in Sections 9.8.1–9.8.5, were not associated with pre-defined specifications. The control limits were established purely based on an **internal evaluation** of the past history of the process data and the variability around the mean. Now, we will introduce an **external specification**, that is, the target established by the management staff or a standard specified by a regulatory agency.

Since the chart we are analysing now is based on the **proportion (p) of failures**, we can also call it a **p-chart**.

S. 9.8.3

In this chart, we also separate our dataset into ***k* subgroups**, each of them containing ***n* data** (see Section 9.8.3.c). For a p-chart developed for an *industrial process*, ideally, we should have large values of n so that a representative proportion of failures is obtained in each subgroup. However, for the case of *treatment plants*, we have the limitations discussed previously, and frequently we work with subgroups with relatively small n values.

The distribution of the proportion p can be represented by a **binomial distribution**. When we deal with **categorical data**, such as in our case (*failure* or *non-failure* with the standard), we could assign a category for each result, for instance: failure = 1, non-failure = 0. If we have a sample with five data, of which three are failures and two are non-failures, the *arithmetic mean* of this categorized sample will be $(1 + 1 + 1 + 0 + 0)/5 = 3/5 = 0.60$. This mean is equal to the proportion of the sample that has the

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

characteristic we are analysing. Since we are analysing failure, *the mean of 0.60 is equal to the proportion of failure in our sample* ($p = 3/5 = 0.60$). Note that we are defining failure as non-conformity with the standard. In statistical terms, success or failure is related to the occurrence of data with the characteristics we specify. To avoid confusion, we will stick to our practical concept of failure as non-conformity with the standard.

Based on these considerations, in **each subgroup**, let us consider the proportion of failure with the standard (p) as

$$p = \frac{X}{n} \tag{9.65}$$

where

p = proportion of data in each subgroup that is not conforming with the standard
X = number of data points in the subgroup that are not conforming with the standard, and n is the total number of data points in the subgroup.

The **mean of all the *k* values of *p*** (denoted as $\bar{p}$) will be an estimate of the population mean of the proportions ($\mu_p = \bar{p}$) and, therefore, will define the **centre line** for our p-chart.

The **standard error** of the mean values of the proportion p is given by

$$s_p = \sqrt{\frac{\bar{p}(1-\bar{p})}{n}} \tag{9.66}$$

where

s_p = standard error of the mean proportion p
n = total number of data points in the subgroups (assumes all subgroups have the same number of data points).

The control limits of the p-chart will be obtained from (Hines *et al.*, 2003; Levine *et al.*, 1998; Mendenhall & Sincich, 1988; Montgomery, 2009):

- **Centre line:**

$$\boxed{\text{Centre} = \bar{p}} \tag{9.67}$$

- **UCL and UWL:**

$$\boxed{\text{UCL and UWL} = \bar{p} + Z_p \cdot s_p = \bar{p} + Z_p \sqrt{\frac{\bar{p}(1-\bar{p})}{n}}} \tag{9.68}$$

- **LCL and LWL:**

$$\boxed{\text{LCL and LWL} = \bar{p} - Z_p \cdot s_p = \bar{p} - Z_p \sqrt{\frac{\bar{p}(1-\bar{p})}{n}}} \tag{9.69}$$

where

Z_p = sigma values to be adopted in our graph. For the **UCL** and **LCL**, the traditional value is 3 sigma ($\boldsymbol{Z_p = 3.0}$). For the **UWL** and **LWL**, the values of Z_p can be adopted as 1.0, 1.5, or 2.0.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Example

EXAMPLE 9.9 BUILD A CONTROL CHART FOR THE PROPORTION OF FAILURES (P-CHART)

Using the same data from the examples shown in this chapter, and especially Examples 9.6–9.8, build a control chart for the proportion of failure (p-chart). Consider that the regulatory standard specified by the agency is a maximum value of 4.0 mg/L. Samples with concentrations greater than the standard are considered to be failures (in non-conformity) and concentrations less than or equal to the standard are considered to be non-failures (in conformity).

Data (values are in mg/L):

2.8	4.2	3.9	3.3	2.8	1.7	1.9	2.5	3.1	3.8	2.7	4.1
4.3	4.8	5.1	5.2	3.9	3.5	2.7	3.1	2.8	3.4	4.9	2.8
2.8	1.8	2.1	2.6	2.3	2.4	2.5	1.3	2.9	2.4	2.1	3.6

Excel

Note: This example is also available as an Excel spreadsheet.

Solution:

(a) Decide on the number of data points per subgroup (*n*)

We will use the same number of data points per subgroup that was adopted in Examples 9.6 and 9.7: $n = 4$. Also, from these examples, we see that we have a total of 36 data points, and we obtain the same value of $k = 36/4 = 9$ subgroups.

(b) Calculate the number of failures and the percentage of failure in each of the *k* subgroups

We organize our computational table with one subgroup per row. In each row, we include four data values, since we decided on having $n = 4$. We will compare each of the four values in the subgroups with the standard and specify the number of data which are not complying with the standard (failure is defined if the concentration is $>$4.0 mg/L).

Subgroup	Measured data ($n = 4$)				Statistics for each subgroup	
					Number of failures	Proportion of failures (p)
1	2.8	**4.2**	3.9	3.3	1	0.25
2	2.8	1.7	1.9	2.5	0	0.00
3	3.1	3.8	2.7	**4.1**	1	0.25
4	**4.3**	**4.8**	**5.6**	**5.8**	4	1.00
5	3.9	3.5	2.7	3.1	0	0.00
6	2.8	3.4	**4.9**	2.8	1	0.25
7	2.8	1.8	2.1	2.6	0	0.00
8	2.3	2.4	2.5	1.8	0	0.00
9	2.9	2.4	2.1	3.6	0	0.00
Summary values					$\sum$ ***failures*** $= 7$	$\overline{p} = 0.194 = 19.4\%$

- Measured values in non-conformity are shown in **bold**.
- Proportion of failure may be expressed as a relative number or percentage. For instance, in subgroup 1, we have one failure among the four samples, that is, $p = ¼ = 0.25 = 25\%$.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

- In the Excel spreadsheet that is associated with this example, there is no need for you to prepare manually the table as above – everything is done automatically.

From the table, we see that the mean of the proportions of failure is $\overline{\boldsymbol{p}} = \mathbf{0.194} = \mathbf{19.4\%}$. This can be obtained by dividing the total number of failures (7) by the total number of data (36), which is $7/36 = 0.194 = 19.4\%$. Since all subgroups have the same size ($n = 4$), this value of $0.194 = 19.4\%$ is also the mean of the nine values of p, as shown in the table.

Because the number of data points in our subgroups is small ($n = 4$), we have few possibilities for the values of p. Depending on the number of failures, for $n = 4$, we will have proportions of $0/4$, $1/4$, $2/4$, $3/4$, and $4/4$, that is, 0.00, 0.25, 0.50, 0.75, and 1.00. This is one limitation to using such a small sample size per subgroup.

(c) Control limits

As in Examples 9.6–9.8, for the upper (UCL) and lower (LCL) **control** limits, we will use the traditional value of 3 sigma ($\boldsymbol{Z_p = 3.0}$). For the upper (UWL) and lower (LWL) **warning** limits, we will use 1.5 sigma ($\boldsymbol{Z_p = 1.5}$).

Using Equations 9.67–9.69, we calculate our control limits:

- Upper control limit: UCL

$$= \bar{p} + Z_p \sqrt{\frac{\bar{p}(1-\bar{p})}{n}} = 0.194 + 3.0\sqrt{\frac{0.194(1-0.194)}{4}} = 0.788 = 78.8\%$$

- Upper warning limit: UWL

$$= \bar{p} + Z_p \sqrt{\frac{\bar{p}(1-\bar{p})}{n}} = 0.194 + 1.5\sqrt{\frac{0.194(1-0.194)}{4}} = 0.491 = 49.1\%$$

- Centre line: Centre $= \bar{p} = 0.194 = 19.4\%$

- Lower warning limit: LWL

$$= \bar{p} - Z_p \sqrt{\frac{\bar{p}(1-\bar{p})}{n}} = 0.194 - 1.5\sqrt{\frac{0.194(1-0.194)}{4}} = -0.102 = -10.2\%$$

- Lower control limit: LCL

$$= \bar{p} - Z_p \sqrt{\frac{\bar{p}(1-\bar{p})}{n}} = 0.194 - 3.0\sqrt{\frac{0.194(1-0.194)}{4}} = -0.399 = -39.9\%$$

Please observe that the calculations of LWL and LCL led to negative values. They have no physical meaning and will not be plotted in the graph, since we do not have negative proportions of failure. This is because by adopting this approach using a Z value, we are using the normal approximation to the binomial distribution (see discussion above in Section 9.4).

S. 9.4

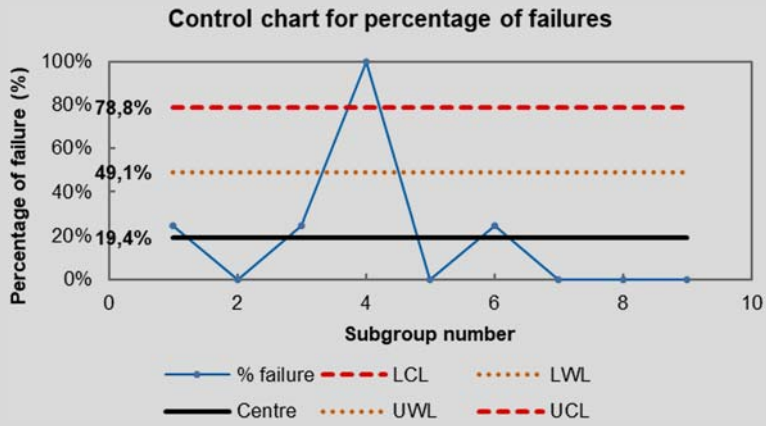
Also note that there may be small differences between the values calculated directly in this example and those calculated in the Excel spreadsheet, due to rounding errors.

(d) Control chart for the proportion of failure

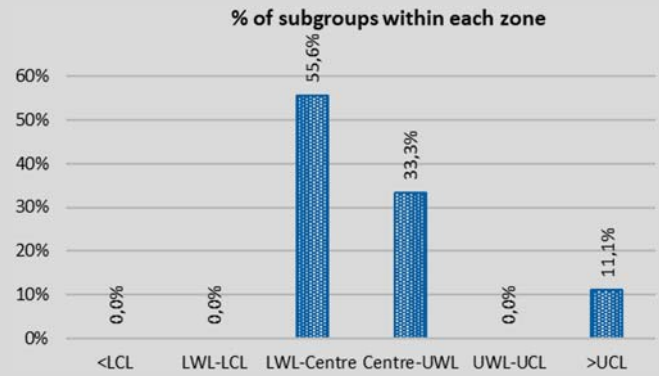
The resulting control chart is presented below. Notice that we have not plotted LCL and LWL, because their calculation led to negative values. Subgroup 4 had 100% failure rate, and you should consider whether this is enough evidence for you to determine that the process is out of

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

control, and whether you should intervene in some way, or exclude this specific subgroup for some reason.



We also present below a graph showing the percentage of values inside each of the control zones. Almost all subgroups were inside the boundaries defined by the *warning* limits (LWL-Centre and Centre-UWL). However, 11.1% of the subgroups (one out of nine) were above the UCL. We already know that this was subgroup 4.



Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

9.9 CHECK-LIST FOR YOUR REPORT

- ✓ Have you defined clearly whether you are dealing with an internally specified target or with a quality standard specified by a regulatory agency? If the latter is the case, have you made clearly which regulation are you referring to, including region or country?
- ✓ Have you made it clear whether the target or standard to be complied with is based on average values, maximum permissible values, minimum allowable values, or percentage of conformity?
- ✓ Are you presenting suitable graphs that compare your data with the target or standard value?
- ✓ If the specifications of the target or standard are based on average values of the monitoring data, are you taking into account the variability of your data and incorporating hypotheses tests to support your claim of conformity or non-conformity?
- ✓ Have you specified your null (H_0) and alternative (H_a) hypotheses in a clear way, so that the result of your test allows you to take a strong conclusion?
- ✓ Have you taken into consideration the distribution (mainly symmetry) of the data to decide on whether you should apply a parametric or a non-parametric hypothesis test?
- ✓ From the various one-sample hypothesis tests described in this chapter, have you selected the one that will best cover your needs?
- ✓ Are you taking into consideration the sample size and the requirements for each hypothesis test?
- ✓ Have you specified your resulting *p*-value and its interpretation in comparison with the significance level you specified for the test ($\alpha = 0.01$, 0.05, or 0.10)?
- ✓ Are you making the right conclusion from your hypothesis test, that is, you should only say that the null hypothesis is rejected and the alternative hypothesis is accepted (and not that your null hypothesis is accepted, unless you want to deepen into a more complex analysis or errors)?
- ✓ If you are reporting percentage of compliance, have you made it clear whether it is based on your original monitoring data or on a distribution that you fit to the data (such as normal or log-normal distributions)? If the latter is the case, have you mentioned which frequency distribution have you used?
- ✓ Have you considered doing a more advanced analysis with the monitoring data, such as frequency analysis, reliability analysis, or statistical quality control (control charts)?
- ✓ If you are doing these more advanced analyses, have you made it clear whether you are using the assumptions of normal or log-normal distributions?
- ✓ If you undertook a reliability analysis, did you present clearly the values of the CV you have calculated from your monitoring data and the reliability level (percentage of compliance) you have adopted?
- ✓ If you have developed control charts, have you considered carefully whether the monitoring data you used was adequate for representing a process under control? Have you made a suitable decision on the number of data to include in each subgroup and the number of subgroups?

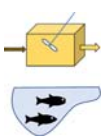
Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Chapter 10

Making comparisons with your monitoring data. Tests of hypotheses

This chapter will help you compare monitoring results from different treatment units, plants, water bodies, phases, or conditions. We continue the discussion we started in Section 9.3 on the topic of one-sample hypothesis tests, and we expand it, presenting a more thorough conceptual background for hypothesis testing and associated errors. Then, we discuss the comparison of two samples to infer whether differences exist between the means of two populations sampled. We present methods for some statistical tests that assume that the underlying distribution is close to normal (parametric tests, including the *t* test), and we present other tests which do not require assumptions regarding the distribution (non-parametric tests, including the Wilcoxon–Mann–Whitney *U*-test and the Wilcoxon Signed-Rank test). After that, we expand our analysis to include methods for the comparison of more than two samples, using the parametric analysis of variance (ANOVA) approach and the non-parametric Kruskal–Wallis test, both followed by multiple comparison procedures (Tukey and Dunn tests, respectively).



The contents in this chapter are applicable to both treatment plant monitoring and water quality monitoring.

CHAPTER CONTENTS

10.1 Introduction **318**
10.2 Inferences about Population Central Values **323**
10.3 One-sample Parametric Tests for a Population Mean (*Z* Test and *t* Test) **338**
10.4 Inferences Comparing Two Population Central Values **348**
10.5 Comparing the Central Values of More Than Two Samples **371**
10.6 Check-List for Your Report **395**

© 2020 The Authors. This is an Open Access book chapter distributed under the terms of the Creative Commons Attribution Licence (CC BY-NC-ND 4.0), which permits copying and redistribution for non-commercial purposes with no derivatives, provided the original work is properly cited (https://creativecommons.org/licenses/by-nc-nd/4.0/). This does not affect the rights licensed or assigned from any third party in this book. The chapter is from the book Assessment of Treatment Plant Performance and Water Quality Data: A Guide for Students, Researchers and Practitioners, Marcos von Sperling, Matthew E. Verbyla and Sílvia M. A. C. Oliveira (Authors).
doi: 10.2166/9781780409320_0317

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

10.1 INTRODUCTION

10.1.1 Types of hypothesis tests

Basic

S. 5.2

In Section 5.2, we presented different types of *descriptive* studies in treatment plants and water bodies that would require the use of different types of descriptive statistics. Here, we present different types of *comparative* studies in treatment plants and water bodies that require different types of statistical hypothesis tests. Specifically, we will cover methods used to make *comparisons between treatment unit processes, treatment plants, water bodies, phases, or conditions*. Figure 10.1 shows examples of such studies for which we may want to **compare the central values** (**means** or **medians**) of one or more samples, taking into account the variability in the data (e.g., comparing the central value of one sample to a fixed value or compare the central values of two samples to each other). The top part of the figure is for treatment plants, and the bottom part is for water bodies, but their structure is similar. See Section 5.2 for a description of each of these typical studies.

10.1.2 Decisions that need to be made before testing hypotheses

Before doing a hypothesis test, we first need to decide upon the following elements (see Figure 10.1):

- number of data sets to be compared
- dependence or independence among data sets
- parametric or non-parametric test

(a) Number of data sets to be compared

When deciding how many data sets you want to compare, you have the following options

S. 9.3

- **One data set**: one-sample hypothesis test is used here for the comparison of monitoring results of a *single* data set with a reference value, such as a regulatory standard (not shown in the figure; covered in Section 9.3).
- **Two data sets**: two-sample tests are used, if there are *only two* treatment units, plants, water bodies, phases, or conditions.
- **More than two data sets**: tests for multiple comparisons are used, if there are *more than two samples* (e.g., more than two treatment units, plants, water bodies, phases, or conditions).

S. 9.3

S. 10.4

S. 10.5

Defining the number of groups is simple. You need only to decide on the variables you want to compare. For example, in **Section 9.3**, when we studied compliance with standards, we used only **one sample** (one-sample hypothesis test), since we wanted to compare one single data set with a fixed value, established in a regulatory norm. In this chapter, we revisit one-sample tests, providing additional theoretical background. After that, we will devote our attention to analyse tests with **two samples** (**Section 10.4**) and **more than two samples** (**Section 10.5**).

(b) Dependence or independence among the data sets

When performing tests with two or more than two samples, you need to determine if the data sets are independent from each other, or if there is some sort of dependency that links the data sets together in some way. To determine if your data sets are independent or dependent, use the logic described in the subsequent box.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

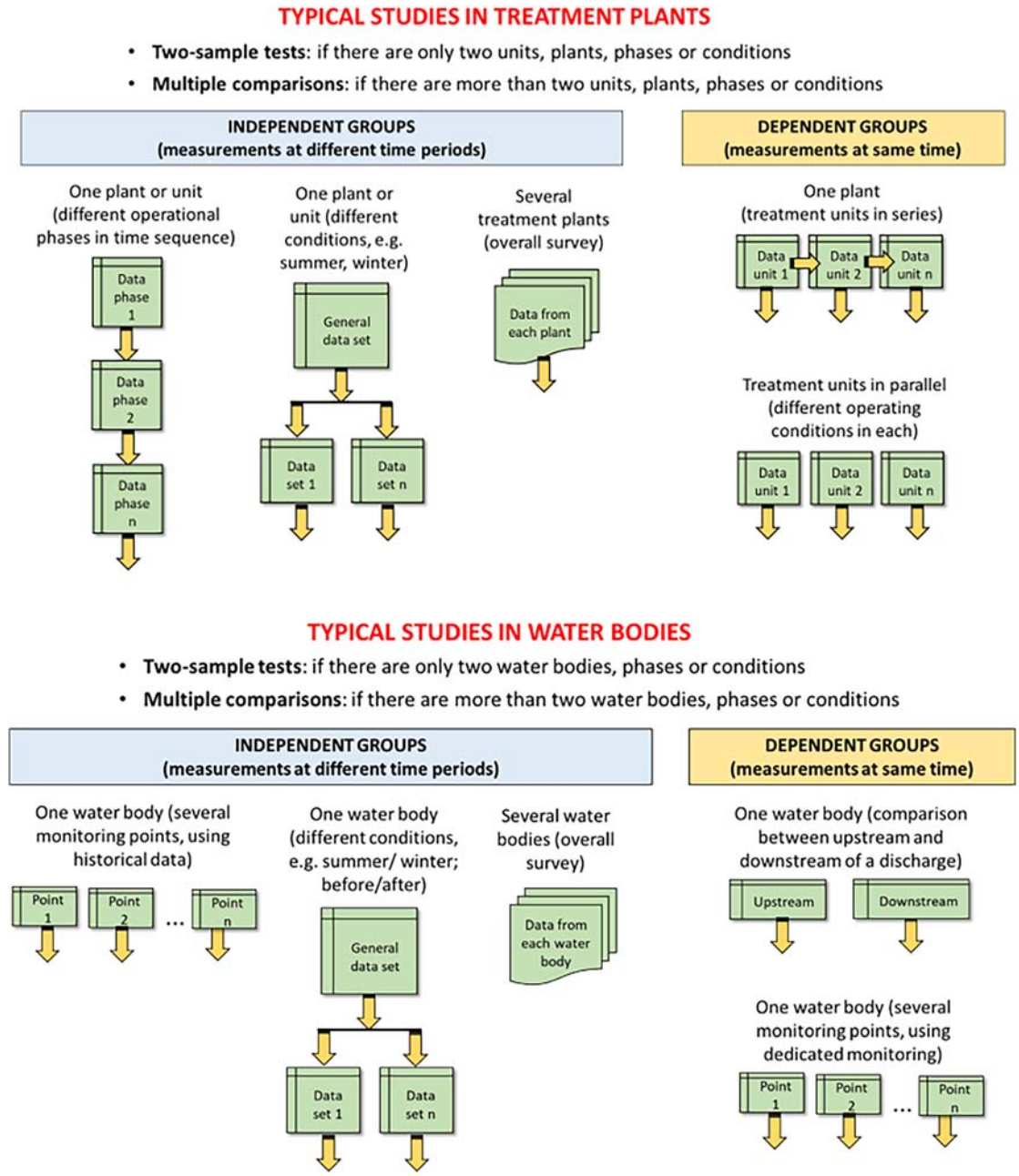


Figure 10.1 Typical studies in treatment plants (top) and water bodies (bottom) that require comparisons among data sets.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

- **Independent groups**: use if the data sets to be compared are independent from each other. In our case, this typically occurs when measurements are made at different time periods or in surveys done in different treatment plants or water bodies.
- **Dependent groups**: use if the data sets to be compared have some degree of dependence, for instance, if they have been obtained at the same time (e.g., comparisons of treatment units in parallel or sampling points upstream and downstream of a discharge in a river). These are called matched data, or when we have only two data sets, they are called paired samples or **matched pairs**, giving a clear indication of dependence. In our book, we only cover two dependent samples (matched pairs) and not multiple dependent samples.

Figure 10.2 illustrates the concept behind hypothesis tests used for independent versus dependent samples. With the **independent samples**, there is no correspondence or connection between one data point from sample 1 and another data point from sample 2. We only compute the means from each sample and compare them to know whether they are equal or not. However, in the case of **dependent samples**, each data point from sample 1 is associated in some way with a data point from sample 2, that is, they are a pair. To conduct hypothesis test with paired samples, we calculate the differences between all *matched pairs*, and test whether the mean of the differences is equal to zero (which would be the case if the means from samples 1 and 2 were equal). It is not difficult to understand that a matched-pairs experiment is likely to provide stronger conclusions about the equality of means compared with the independent-sample experimental design. The comments we made for *means* (in parametric tests) apply for *medians* (in non-parametric tests).

Establishing whether the groups are dependent or independent is not trivial and, in many cases, may be misleading. Some researchers argue that, in our field of *environmental statistics, it is very difficult to assume that we have truly dependent data sets*, even if measurements are made at the

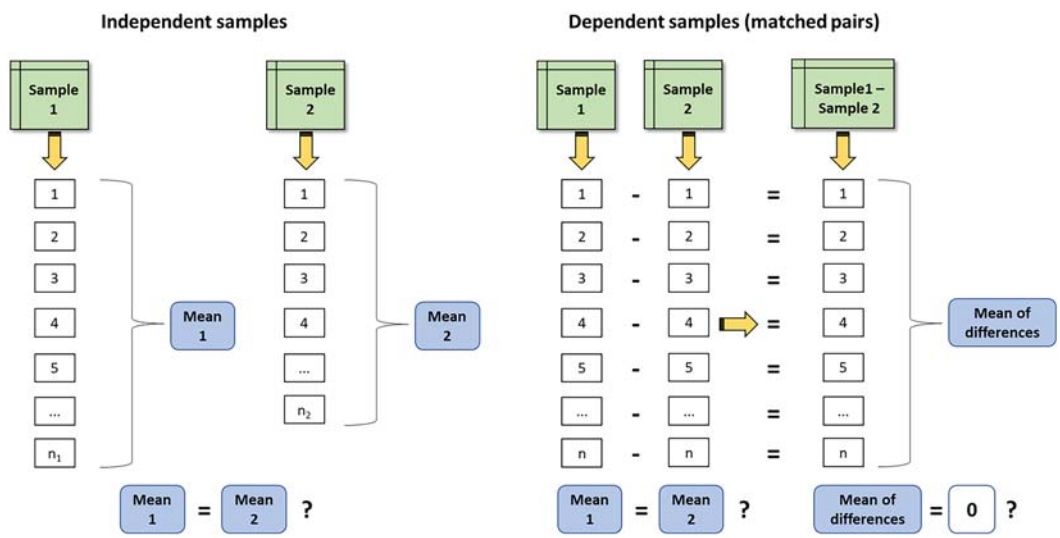


Figure 10.2 A visual representation of the structure of independent and dependent two-sample hypothesis tests.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

same time. Other unaccounted for environmental factors, different from the explicit factor we are studying, may cause the data sets to lose their degree of dependence.

However, there are some clear examples of samples that *should* be treated as matched pairs. For instance, consider a study of a river where you are measuring the concentration of some contaminants in samples collected immediately upstream and downstream from a suspected source of contamination. Your sample size is $n = 12$ for each sample, with upstream and downstream samples collected monthly. In this case, the ambient background concentration of the pollutant may change drastically throughout the year, but what you are interested in is if there is a significant increase in the concentration between the upstream and downstream locations. In this case, the samples are clearly matched pairs.

In this book, we will follow the classical structure of presenting the statistical methods for independent and dependent groups – it is up to you to decide which approach to use (independent or dependent samples) based on your knowledge of the system you are studying. **If you are really in doubt, we suggest that you use tests for independent data sets, even if they lose some statistical power compared with the tests for dependent sets.**

(c) Parametric or non-parametric test

Another decision we need to take is whether we should adopt **parametric or non-parametric tests**:

- **Parametric tests**: These tests require the assumption that the *underlying distribution of the population is a **normal** distribution*. They use the **original data** or a transformation of the data (such as a $\log_{10}$ transformation) and make inferences about their **mean** values.
- **Non-parametric tests**: For these tests, *you do not have to make assumptions regarding the distribution*. They work with the **ranked data** and make inferences about their **median** values.

C. 8

S. 10.2.4

S. 8.2.8

Let us discuss this a bit more, using concepts dealt with in Chapter 8. Some hypotheses tests shown in this Chapter 10 are based on the *assumption that the data are **normally distributed***, or that they are, at least, *symmetrical around the mean*. If this assumption is fulfilled, you can apply the **parametric tests**. You will understand more about why the normality assumption is required after reading the discussion about *rejection limits* and *probability levels* (p-values) associated with the normal distribution (Section 10.2.4). To review how to test whether your distribution is normal, see Section 8.2.8.

However, if your data are ***not normally distributed*** or are *not symmetrical around the mean*, you have two choices: (a) transform the data to make them normal and use parametric tests or (b) do not transform the data and use non-parametric tests.

- **Transform the data and use parametric tests**. If you suspect that your data follow a **log-normal distribution**, you could attempt to **transform the data set** to make it *normally distributed* (if your data follow a log-normal distribution and you take the log_{10} of each original value, the resulting log-transformed values will be normally distributed). If the distribution of your transformed data is similar to the normal distribution, you could then apply a **parametric test** using the transformed data set. We are presenting this choice, given the importance of log-normal distributions for environmental data. However, many researchers will prefer to go directly to the following alternative and use non-parametric tests.
- **Do not transform the data and use non-parametric tests**. If the distribution of your data shows **departures from normality** (especially if it is highly skewed or has a different shape), even after

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

data transformations are applied, then you must use **non-parametric tests**. These tests work with *ranked data*, and so they *do not depend on the original data having a normal distribution*. Another application of a non-parametric test is if your original data are not susceptible to measurement, but rather can be reported only as *ranked values* in increasing or decreasing order. The non-parametric statistical tests use information of ranked data, such as nominal or ordinal observations, rather than metric data required by the conventional tests. *No assumptions about the form of the parent population are required*, hence the name 'non-parametric.'

As a summary, we can make the following recommendations (see Figure 10.3):

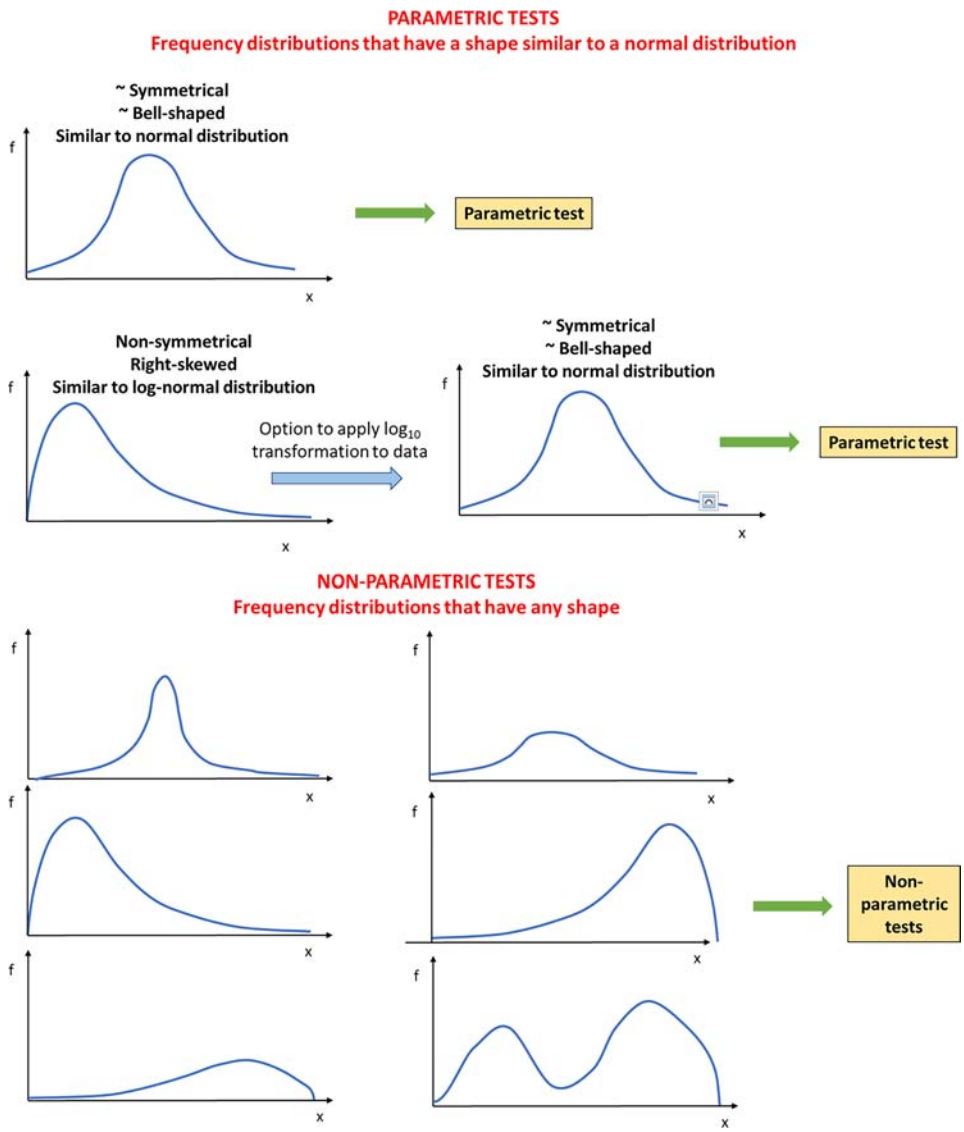


Figure 10.3 Decision on the adoption of parametric or non-parametric hypothesis tests.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

- If your data follow a **normal distribution** or are symmetrically distributed, you can use **parametric tests.**
- If your data are skewed and follow a **log-normal distribution**, you could try to convert the original values to their **$\log_{10}$ values** and use **parametric tests** with the log-transformed data set (there are other transformations you can also try, but to maintain simplicity in this book, we only discuss the $\log_{10}$ transformation).
- If you are **not sure about the distribution of your data**, do not want to make transformations in your original data, or simply cannot or do not want to decide on this regard, you can apply **non-parametric tests.**
- In most cases, the loss of power of non-parametric tests compared with parametric tests will be small, and non-parametric tests can perform better in the case of skewed distributions, particularly from samples with few data points.

In this chapter, we present the parametric tests followed by their non-parametric alternative. It is up to you to decide on which approach to follow, based on the considerations made above.

10.1.3 Summary of the different hypothesis tests

C. 9

There are several different hypothesis tests that can be employed, as shown in Table 10.1. In the table, we present in **bold** and with asterisk (*) those that are covered in this chapter (with a mention that one-sample tests are also dealt-with in Chapter 9). Figure 10.4 summarizes the tests that are covered in our book. The methods are classified according to (a) number of samples, (b) dependence or independence between the groups, and (c) parametric or non-parametric procedure.

10.2 INFERENCES ABOUT POPULATION CENTRAL VALUES

10.2.1 Establishing the test hypotheses

Basic

Some of the concepts presented in this section have already been presented in Section 9.3, when we studied one-sample hypothesis tests. Here, we present a more formal coverage of these concepts.

S. 9.3

Hypothesis testing is based on information about the sample(s) and is a part of statistical inference. If the original hypothesis is framed appropriately, we can make inferences about a **population parameter** through the *analysis of the differences between the observed results* (*sample statistics*) *and the expected results*. Also, we can compare the population parameters of two or more data sets, and this is what we cover here in this chapter.

For the discussion below, let us assume that we want to infer whether the means of two samples are different or not (see Figure 10.1 for possible applications of this concept). To determine if the difference between the means of our two data set is statistically significant, we need to use **hypothesis testing**. To conduct a hypothesis test, we must start by defining a **null hypothesis H_0** and an **alternative hypothesis H_a**. Think of them like this:

- The ***null hypothesis*** is typically *the one that you do not believe to be true, the situation that you believe you can invalidate with your study.*
- The ***alternative hypothesis*** is *the one that you believe to be true or that you want to try to validate*.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Table 10.1 Hypothesis tests for one, two or more samples, highlighting those that are covered in this chapter.

Applicability	Number of Samples	Parametric Test	Non-Parametric Test
Comparison with a reference value	One	• **Student *t* test (*)** • **Z test for proportions (*)**	• **Wilcoxon signed-rank test (*)** • **Sign test (*)**
Differences among independent groups	Two	• **Student *t* test (*)**	• Wald–Wolfowitz runs test • **Mann–Whitney *U*-test (*)** • Kolmogorov–Smirnov two-sample test • Chi-square test (homogeneity)
	Multiple	• **Analysis of variance (ANOVA) (*)** • **Multiple comparisons: extension to Tukey test (*)**	• **Kruskal–Wallis analysis of ranks (*)** • **Multiple comparisons: extension to Dunn test (*)** • Median test • Chi-square
Differences among dependent groups (paired samples)	Two	• **Student *t* test (*)**	• Friedman test • Sign test • **Wilcoxon's matched-pairs test (*)** • Dichotomous variables: McNemar's Chi-square test
	Multiple	• ANOVA for repeated measurements	• Friedman ANOVA • Friedman's two-way analysis of variance • Cochran Q test

Note: (*) and **bold:** test described in this chapter.
One-sample tests are mainly covered in Chapter 9, but the *t* test is also analysed in Chapter 10.

C. 9
C. 10

Suppose we want to validate that there is a significant difference between the average values of the two samples we are analysing. Therefore, we should construct our null hypothesis to be that the true mean concentrations are equal, and our alternative hypothesis that the true mean concentrations are not equal (i.e., one is greater than or less than the other). This hypothesis test will result in one of the following two conclusions; either:

- we **reject the null hypothesis** (*in favour of our alternative hypothesis*); or
- we **fail to reject the null hypothesis** (*which does not mean that the null hypothesis is true*, by the way!).

For our example, if we **reject the null hypothesis**, this means that there is enough evidence to say that there is a *significant difference* between the mean values of the two samples and:

- If the **average value of Sample 1 is less than the average value of Sample 2**, then we can say that the mean value of Sample 1 is *significantly* less than the mean value of Sample 2.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

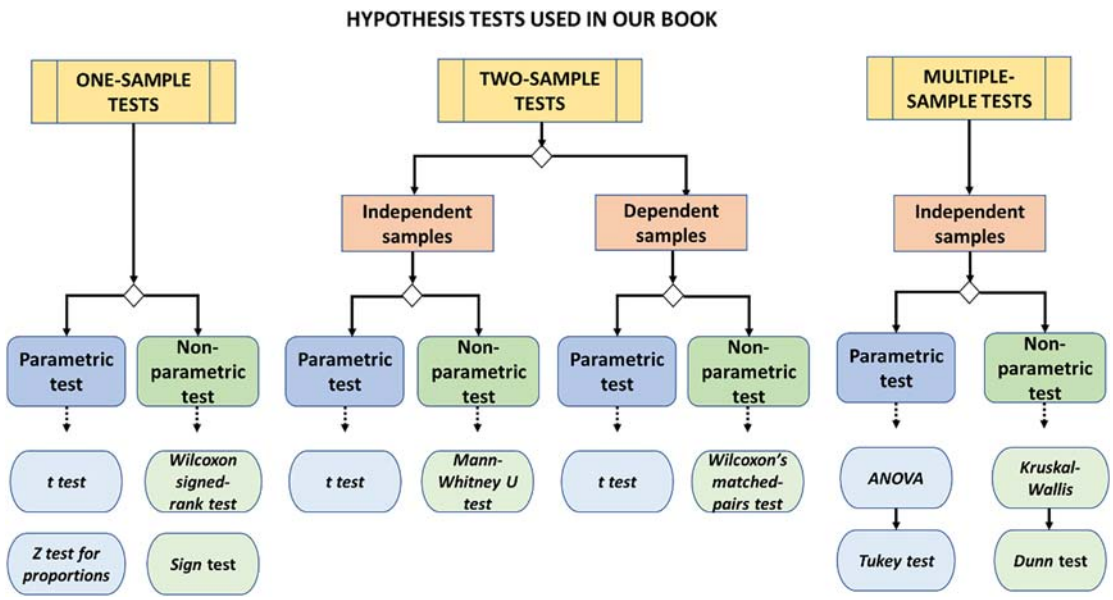


Figure 10.4 Hypothesis tests covered in our book.

- If the **average value of Sample 1 is greater than the average value of Sample 2**, then we can say that the mean value of Sample 1 is *significantly* greater than the mean value of Sample 2.

If we **fail to reject the null hypothesis**, this means that there is ***not*** enough evidence to say that there is a *significant difference* between the mean values of the two samples. In other words:

- we do not have enough confidence to say whether the true average values are above or below one another (it could be either)
- we cannot say that the null hypothesis is true, but we cannot say it is false either
- we also cannot draw any conclusions about the alternative hypothesis in this case!
- it also often means that we may need to collect more data (the more data we collect, the more likely we are to be able to reject the null hypothesis)

S. 9.3

As we mentioned in Section 9.3, do not worry if it takes you a while to understand these concepts. The logic of hypothesis testing is not so straightforward and can be difficult to comprehend at first. One analogy that may help is the 'presumption of innocence' principle used in law, which is where a person is considered innocent until there is enough evidence to prove that they are guilty. With statistical hypothesis testing, we assume 'innocence' (that the null hypothesis is true) until we have enough evidence to find it to be 'guilty' (that the null hypothesis is false and the alternative hypothesis is true). As a researcher, you can think of yourself as the prosecution team – your goal is to try to find the 'plaintiff' (your experiment) to be found 'guilty' (i.e., as a researcher, you want to find *statistically significant* results that disprove the null hypothesis and favour your alternative hypothesis).

To formalize these concepts and express them using mathematical notation, we can reinforce that the hypothesis test starts with a theory, request, or statement about a particular parameter of a population, called the **null hypothesis (H_0)**. This hypothesis establishes the absence of a difference between the

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

parameters. It is always the first hypothesis to be formulated, and it is the one that we will test. This hypothesis assumes that the observed results are entirely due to chance. For example, a null hypothesis can assume that the mean of a population (μ) is not different from zero (or $\mu = 0$), and the mathematical notation would be H_0: $\mu = 0$. We can also assume that the population mean is no different than, say, 3.0 mg/L or is no different than 10.0 kg/d, and our hypothesis would be specified as H_0: $\mu = 3.0$ mg/L or H_0: $\mu = 10.0$ kg/d, respectively. Using general mathematical notation, the null hypothesis can be stated as follows:

$$H_0: \mu = \mu_0 \tag{10.1}$$

where μ_0 is the value with which we want our population mean μ to be compared.

If the null hypothesis is indeed true, any observed difference between the means of your sample data is merely a consequence of natural sampling variability.

In the next step, an **alternative hypothesis (H_a)** is formulated, against the null hypothesis H_0. This hypothesis will be accepted if H_0 is rejected. In other words, *if we reject the null hypothesis (H_0 is false), then the alternative hypothesis (H_a) is assumed to be true.* For a given null hypothesis H_0, one of several alternative hypotheses H_a could be chosen. For example, if: H_0: $\mu = 0$, then the alternative hypotheses could be H_a: $\mu \neq 0$ or H_a: $\mu > 0$ or H_a: $\mu < 0$.

In mathematical notation, if we have a **one-sample test**, with the null hypothesis:

$$H_0 : \mu = \mu_0$$

then the alternative hypothesis could be

$$H_a: \mu \neq \mu_0 \quad \text{or} \quad H_a: \mu > \mu_0 \quad \text{or} \quad H_a: \mu < \mu_0 \tag{10.2}$$

Note that it is important when testing hypotheses to establish clearly both the null and alternative hypotheses because they will determine which type of statistical test is used.

If we are comparing the means of **two samples** (μ_1 and μ_2), then we could have:

$$H_0: \mu_1 = \mu_2 \tag{10.3}$$

$$H_a: \mu_1 \neq \mu_2 \quad \text{or} \quad H_a: \mu_1 > \mu_2 \quad \text{or} \quad H_a: \mu_1 < \mu_2 \tag{10.4}$$

Key points on the establishment of test hypotheses

- The **null hypothesis** (H_0) is the one to be tested.
- The **alternative hypothesis** (H_a) is developed as the opposite of H_0. It represents the **conclusion that will be supported if H_0 is rejected**.
- H_0 refers to the **population parameter** (ex, refers to the population mean μ and not the sample mean $\bar{x}$).
- The formulation of H_0 **always** contains a **sign of equality (=)** in relation to the specified parameter.
- The formulation of H_a **never** contains a **sign of equality** with respect to the parameter (in traditional statistical hypothesis testing, you can never prove something to be equal, you can only prove that they are not equal).

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Advanced

10.2.2 The four potential outcomes to a statistical test

Now let us talk about **errors**. A statistical hypothesis test has four potential outcomes – two outcomes each for the following two possible situations, or realities, that may occur; either:

- **Reality #1:** the null hypothesis is true, or
- **Reality #2:** the null hypothesis is false (and the alternative hypothesis is true).

This actual reality is always unbeknownst to us! We will never actually know with 100% certainty what the true reality is between the two options presented above. We can only make an inference (i.e., an 'educated guess') about the reality based on our sample data. Statistics allows us to place a confidence level on that 'educated guess.' With this in consideration, you need to understand the **four potential outcomes for any statistical analysis**. The first two outcomes pertain to *reality #1* – that the null hypothesis is indeed true (unbeknownst to us). The second two outcomes pertain to *reality #2* – that the null hypothesis is indeed false.

Reality #1: The null hypothesis is indeed true (unbeknownst to us)

S. 9.3

- ***Outcome 1a: Erroneously rejecting the null hypothesis* (*also known as a 'Type I' or 'alpha' error*).** In the example of assessing compliance with a regulatory standard (dealt with in Section 9.3), this would be when you erroneously conclude (based on your data and a statistical test) that the sample is in compliance with the standard, when in reality, the true mean concentration exceeds the standard level. This outcome will happen on average with a probability chosen by us as the *significance level* (α). For example, if we choose the usual significance level of 0.05, then this outcome will occur for us during one out of every 20 experiments on average ($1/20 = 0.05$) for the situation where the reality (unbeknownst to us) is that the null hypothesis is indeed true and the alternative hypothesis is not. If we choose a significance level of 0.01, then this outcome will only occur one out of every 100 experiments on average ($1/100 = 0.01$).
- ***Outcome 1b: Correctly failing to reject the null hypothesis* (*also called 'non-significant' findings*).** In the example of assessing compliance, this would be like when you correctly conclude (based on your data and a statistical test) that you do not have enough information to determine if the sample is in compliance with the standard or not. These are the type of findings that researchers often refer to as 'non-significant.' If the significance level is set at 0.05, then this outcome will occur 19 out of every 20 experiments on average ($19/20 = 0.95$).

Reality #2: The null hypothesis is indeed false (unbeknownst to us)

- ***Outcome 2a: Correctly rejecting the null hypothesis* (*power*).** This outcome is generally the stuff of your dreams. Correctly rejecting a null hypothesis in favour of an alternative one is referred to as the *power of an experiment*. In the example of assessing compliance, this is the outcome where you correctly determine that your sample is in compliance with the standard limit.
- ***Outcome 2b: Erroneously failing to reject the null hypothesis* (*'Type II' or 'beta' error; i.e., you need a larger sample size*).** This outcome represents a situation where the alternative hypothesis is actually true in reality, but the sample size was too small to detect it with significance. For example, if we are assessing compliance, it would be the situation where the true concentration does indeed comply with the standard limit, but either the difference between the true concentration and the limit is very small or the sample size (number of replicate samples) you collected was very small.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Table 10.2 The two types of errors in hypothesis testing.

Decision Made (Based on Sample Data and Statistical Test)	**True State of Nature (Reality) for the Population(s) Studied**	
	H_0 True (H_a False)	**H_0 False (H_a True)**
Reject H_0	Type I (alpha) Error	Correct decision (significant findings, statistical power)
Do not reject H_0	Correct decision (non-significant findings)	Type II (beta) Error

Table 10.2 summarizes these four outcomes.

The probability of incurring a type I (alpha) error is denoted by the symbol α, while the probability of incurring a type II (beta) error is denoted by the symbol β. In order to **minimize the type I error**, the values of the **significance level $\boldsymbol{\alpha}$** must be established prior to the application of the hypothesis tests. Usually, selected values of α are **0.01, 0.05, and 0.10**. This implies a probability of avoiding this error equal to **$(1-\alpha)$**, which is called **confidence level**, and, as a result, is equal to **0.99, 0.95, 0.90**, respectively. The values selected for α determine the regions of rejection and non-rejection of statistical hypotheses (see Section 10.2.4). If α is small, the decision to reject H_0 is quite convincing, but if α is large, the decision to reject H_0 will entail a greater risk of error.

S. 10.2.4

The probability associated with the occurrence of **type II error** is more complicated to calculate, since α and β are inversely related. That is, for a fixed value of n, when we change the rejection region to increase α, the value of β decreases and vice versa. This will be discussed further below.

As mentioned above, the complementary probability of **type II error $(1-\beta)$** is called the **power of the statistical test** to detect a real difference and represents the probability of correctly stating that there is a difference when it actually exists.

Summary of errors in hypothesis tests

- **Rejecting the null hypothesis H_0 if it is true is a Type I error**. *The probability of making a Type I error is given by α.*
- **Failing to reject the null hypothesis H_0 if it is false is a Type II error**. *The probability of making a Type II error is given by β.*
- The **confidence level $(1-\alpha)$** is the *probability that H_0 will not be rejected when, in fact, H_0 is true*.
- The **power of a statistical test $(1-\beta)$** is the *probability of rejecting the null hypothesis H_0 when, in fact, H_0 is false*.

Other comments on hypothesis tests

- The **significance level α** is specified before the hypothesis test is performed. So, the risk of incurring in a type I error, α, is under your control.
- Traditionally, levels of **$\alpha \leq 0.10$** (10%) are adopted. Frequently, **$\alpha = 0.05$** (5%) is used.
- Traditionally, **levels of $\beta \leq 0.20$** (20%) are adopted. Consequently, a **power > 0.80** (80%) is obtained.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

- Once α is defined, the size of the rejection region is known.
- The statistical power (and the β region) is only known if you know the α value, the effect size (the observed difference between samples), and the sample size.

10.2.3 One-tailed and two-tailed hypotheses tests

Basic

S. 9.3
S. 10.4
S. 10.5

In the context of Section 9.3, hypotheses tests were used to compare a single sample to a quality standard value. This was called a **single-sample test**. Remember, here we are talking about one statistical sample, which is a single data set comprising values obtained from many water samples collected at the same point. Hypothesis testing can also be used to compare **two or more samples** to each other, and these tests are presented in Sections 10.4 and 10.5.

For these tests, there are three approaches you can use: **two-tailed**, **left-tailed (one-tailed)**, and **right-tailed (one-tailed)**. We will give more emphasis to the two-tailed test in this chapter, but will also describe the logic behind *a left-tailed and a right-tailed test*. Below, we give a description of the *three approaches for a* ***two-sample test*** (comparing the mean values of two data sets, associated with the populational means μ_1 and μ_2):

- **Two-tailed test.** The first approach is to assume a *null hypothesis* H_0: $\mu_1 = \mu_2$ and an *alternative hypothesis* H_a: $\mu_1 \neq \mu_2$. This would lead to a 'two-tailed' test (both right- and left-tailed). If you use a two-tailed *t* test, then you may have one of the following three outcomes:
  - The mean concentration of sample 1 (μ_1) is *less than* the mean concentration of sample 2 (μ_2) (and the *p*-value is less than α, so the difference is significant)→$\boldsymbol{\mu_1 < \mu_2}$.
  - The mean concentration of sample 1 (μ_1) is *greater than* the mean concentration of sample 2 (μ_2) (and the *p*-value is less than α, so the difference is significant)→$\boldsymbol{\mu_1 > \mu_2}$.
  - The mean concentration of sample 1 (μ_1) may be less than or greater than the mean concentration of sample 2 (μ_2) (but the *p*-value is greater than α, so the difference is **not** significant). This means that, based on the data, you do not have enough confidence to know if the means are different or not. This result may indicate that you need to collect more samples.
- **Left-tailed test.** If you assume a *null hypothesis* H_0: $\mu_1 = \mu_2$ and an *alternative hypothesis* H_a: $\mu_1 < \mu_2$, it would lead to a one-sided, 'left-tailed' or 'inferior' test. Use this test if you have a strong fundamental reason to believe that sample 1 *should have a lower mean* than sample 2.
- **Right-tailed test.** Likewise, if you assume a *null hypothesis* H_0: $\mu_1 = \mu_2$ and an *alternative hypothesis* H_a: $\mu_1 > \mu_2$, it would lead to a one-sided, 'right-tailed' or 'superior' test. Use this test if you have a strong fundamental reason to believe that sample 1 *should have a greater mean* than sample 2.

If you have *a priori* indications that the mean of sample 1 is expected to be lower or greater than the mean of sample 2 (as when comparing the effluent concentration from a treatment plant with the influent concentration, or the upstream sampling point in a river that receives a point-source pollution with the downstream sampling point), you may consider adopting a **one-tailed test**. However, if you have no *a priori* indication whether the mean of sample 1 should be greater or lower than the mean of sample 2, you will probably prefer a **two-tailed test**. In most cases, the latter approach will be more useful for you, and we give priority to it in this chapter.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Advanced

10.2.4 Rejection and non-rejection regions

(a) The concept of rejection regions

In the previous sections in this chapter we have used the expression '**rejection region**.' Each *hypothesis test has its own critical test statistics* ($Z_{critical}$, $t_{critical}$, $F_{critical}$, or other), which define critical values in the distribution that are associated with the chosen significance level (α). These critical values establish the boundaries of the rejection and non-rejection regions, as we will see below.

S. 9.3

Let us start with the example analysed in Section 9.3. There, we wanted to *compare whether our sample mean (3.16 mg/L) was significantly different than the stipulated regulatory standard of 4.00 mg/L*. Therefore, we will use a *two-tailed, one-sample test* (as was done in Section 9.3). Our null hypothesis is H_0: mean = standard and our alternative hypothesis is H_a: mean ≠ standard.

Imagine a normal distribution centred on the value of the standard, i.e., $\mu = 4.00$ mg/L. Let us assume that we have calculated the rejection limits (we will show soon how to calculate them), and we found the lower value of the rejection region to be 3.65 mg/L and the upper value of the rejection region to be 4.35 mg/L. If our sample mean falls within this interval, we cannot reject the hypothesis that the mean is equal to the standard (4.00 mg/L). However, if our sample mean falls outside this region, it will now be in a rejection region. This is the case in our example, because we see that our sample mean of 3.16 mg/L is lower than 3.65 mg/L, and therefore it fell in the lower rejection region. The conclusion is that we have to *reject the null hypothesis* that the sample mean (3.16 mg/L) is equal to the standard (4.00 mg/L), and thus we *support the alternative hypothesis* that the sample mean is significantly different from the standard, in fact, we can say that it is *significantly less than* the standard. This is illustrated in Figure 10.5.

Now, let us suppose that our sample mean was 3.90 mg/L. This value is inside the non-rejection region, and so we cannot reject the null hypothesis that our sample's mean is equal to the standard. A similar comment could be made if our sample mean was equal to 4.30 mg/L. Both situations are illustrated in Figure 10.6.

Now, let us suppose that we had the same mean values (3.16, 3.90, and 4.30 mg/L), the same standard (4.00 mg/L) and the same null and alternative hypotheses. However, we calculated our

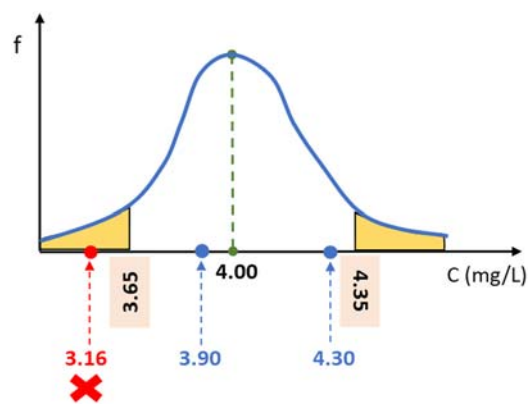


Figure 10.5 Interpretation of a one-sample two-tailed hypothesis tests in terms of the rejection regions (rejection regions are shaded).

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

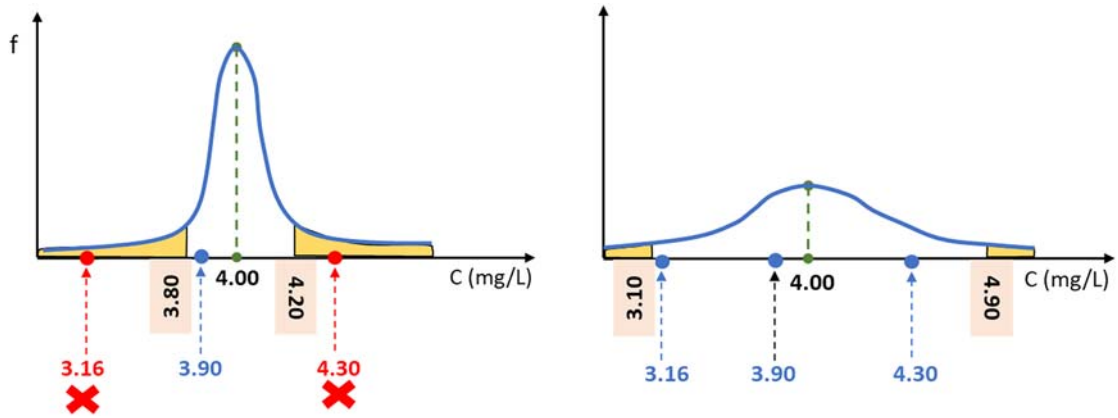


Figure 10.6 Interpretation of one-sample two-tailed hypothesis tests with different rejection regions.

non-rejection region to be narrower (only from 3.80 to 4.20 mg/L). Now, two out of the same three mean values (3.16 and 4.30 mg/L) would be in the rejection regions. For these values, we would reject the null hypothesis that the mean is equal to the standard. This is illustrated in Figure 10.5 (left side).

In the sequence, we keep the same test conditions, but now we have a wider non-rejection region (3.10–4.90 mg/L). In this situation, none of the sample means fall in the rejection regions. As a consequence, we cannot reject the null hypothesis that the sample mean is equal to the standard for any of the three mean values. See Figure 10.5 (right side).

The **width of the non-rejection region is directly influenced by the sample size** (n = number of data points in your sample). For the same values of mean and standard deviation, the higher the value of n, the narrower the non-rejection region (and the more likely we are to reject the null hypothesis). In Section 10.3.3 we discuss this for the specific case of the t test.

S. 10.3.3

Now that we have seen the concept of rejection regions using an applied case study, let us consider a generic case where we state our null hypothesis to be H_0: $\mu = \mu_0$ (e.g., Equation 10.1). In Figure 10.7 (top), a **two-tailed test** is illustrated. Suppose again that we want to test whether the mean μ is significantly different from some value μ_0. Therefore, we have rejection regions on both sides of the distribution. The null hypothesis is H_0: $\mu = \mu_0$ and the alternative hypothesis is H_a: $\mu \neq \mu_0$. We calculate the test statistics (e.g., Z value or other) and, if the test statistic falls within one of the **rejection regions**, we ***reject*** *the null hypotheses* that $\mu = \mu_0$, and accept the alternative hypothesis H_a: $\mu \neq \mu_0$. However, if the test statistic *does not fall in the rejection region*, we ***cannot reject*** *the hypothesis that* $\mu = \mu_0$ and we also cannot accept the alternative hypothesis $\mu \neq \mu_0$. In this case, we probably need to collect more data.

In Figure 10.7 (middle), a **left-tailed** test is illustrated. In this test, again, we calculate the test statistic, and if it falls within the **rejection region**, we ***reject*** *the null hypotheses* that $\mu = \mu_0$ and accept the alternative hypothesis H_a: $\mu < \mu_0$. However, if the test statistic *does not fall within the rejection region*, we ***cannot reject*** *the hypothesis that* $\mu = \mu_0$ and we also cannot accept the alternative hypothesis $\mu < \mu_0$. We probably need more data.

Finally, a similar situation is shown in Figure 10.7 (bottom), with a **right-tailed test**, using the same null hypothesis H_0: $\mu = \mu_0$, but now with the alternative hypothesis, H_a: $\mu > \mu_0$. The reasoning is the same as above for the left-tailed test, but now instead of having the rejection region on the

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

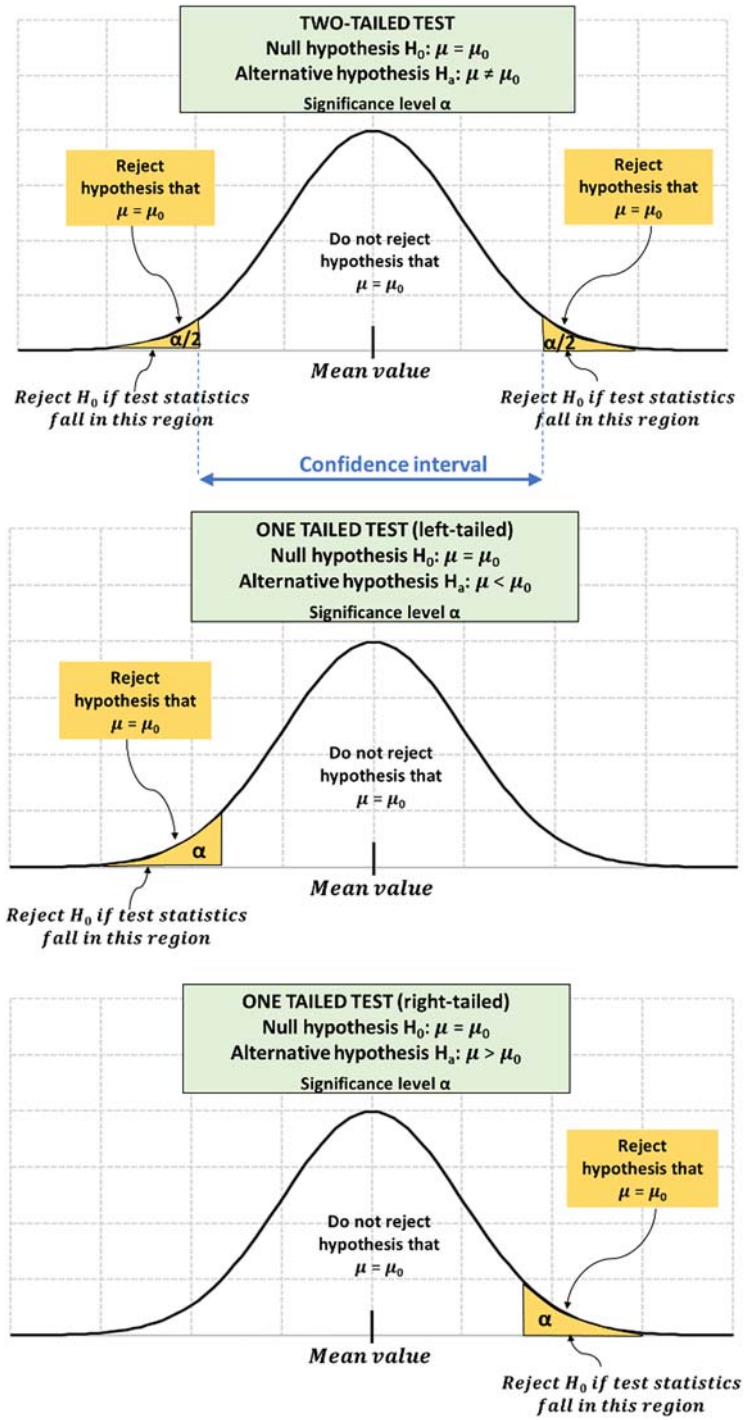


Figure 10.7 Rejection regions in two-tailed (top) and one-tailed (middle: left tail; bottom: right tail) tests.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

left-hand side, it is on the right-hand side of the distribution. We calculate the test statistic and, if it falls within the **rejection region**, we ***reject*** *the null hypotheses* that $\mu = \mu_0$ and accept the alternative hypothesis H_a: $\mu > \mu_0$. If the test statistic *does not fall in the rejection region*, we ***cannot reject*** *the hypothesis that* $\mu = \mu_0$ but we also cannot accept the alternative hypothesis $\mu > \mu_0$.

Basic

S. 8.2.5

(b) How to establish critical values that define the rejection regions

The next step is for you to learn how to establish the critical values that define the boundaries of the rejection regions. Let us remember some basic fundamental concepts about the normal distribution. In Section 8.2.5, we introduced you to the concept of the *standard normal variable* (Z) and showed you how it is calculated. In Figure 8.8 and Table 8.1, we showed you the percentage of data points that fall inside the limits defined by Z, between –1 and +1 (~68%), between − 2 and +2 (~95%), and between −3 and +3 (>99%). Now, we will use the same concepts, but will calculate what are the Z values that lead non-rejection regions of 90%, 95%, and 99%. For this, we use the Excel function **NORM.S.INV(probability)**. By doing so, we end up with the plot shown in Figure 10.8, for three common confidence levels ($=1-\alpha$). The non-rejection regions shown here are for two-tailed tests, but you can develop a similar chart for one-tailed tests.

The values in the X-axis are the **critical values of Z ($Z_{critical}$)** that define the boundaries between the rejection and non-rejection regions. For instance, if you are using a significance level of $\alpha = 0.05 = 5\%$, then your confidence level is $1 - \alpha = 0.95 = 95\%$. For a two-tailed test, we see that we have one rejection region with 2.5% to the left ($Z_{crit} = -1.960$) and one rejection region with 2.5% to the right ($Z_{crit} = +1.960$), comprising a total rejection region of 5% and a total non-rejection region of 95% (Figure 10.8). If you calculate the test statistic and it is less than –1.960 or greater than +1.960, it will fall in the rejection region and will lead to rejection of the null hypothesis.

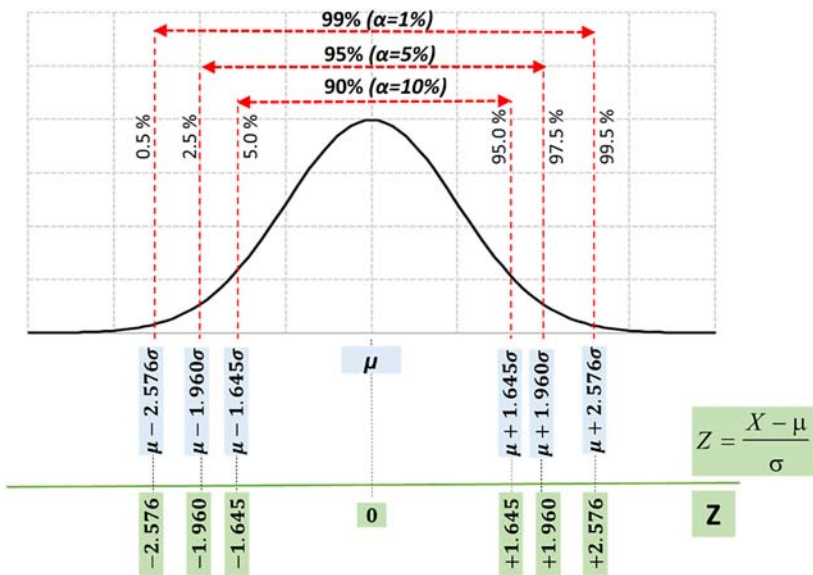


Figure 10.8 Indication of the critical values of Z ($Z_{critical}$) for different values of significance level α and confidence level ($1 - \alpha$).

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Table 10.3 Rejection regions based on the Z values for a normal distribution, the significance levels and the test conditions.

Significance Level (α)	Confidence Level ($1-\alpha$)	Rejection Regions		
		Two-Tailed Test	Left-Tailed Test	Right-Tailed Test
10%	90%	$Z<-1.645$ or $Z>+1.645$	$Z<-1.282$	$Z>+1.282$
5%	95%	$Z<-1.960$ or $Z>+1.960$	$Z<-1.645$	$Z>+1.645$
1%	99%	$Z<-2.576$ or $Z>+2.576$	$Z<-2.326$	$Z>+2.326$

Based on this concept, Table 10.3 presents a summary of the rejection regions associated with the significance levels of 10%, 5%, and 1% ($\alpha = 0.10$, 0.05 and 0.01) and the associated values of Z_{critical}. We can also use the expression $\boldsymbol{Z_\alpha}$ instead of Z_{critical} to make clear the association with the α levels.

(c) Example: rejection regions for significance level $\boldsymbol{\alpha = 0.05}$

In order to clarify even further the concepts presented above, Figure 10.9 presents the **rejection and non-rejection regions** for the most common significance level of 5% **($\boldsymbol{\alpha = 0.05}$)**, for two-tailed and one-tailed tests.

As you can see in Figure 10.9, the rejection regions are defined by the critical Z values, which are calculated based on the significance level and the number of tails in the test. When the significance level is 0.05, for a left-tailed test (Figure 10.9 middle), the critical Z value is calculated as **NORM.S.INV(0.05) $= -1.645$**. For a right-tailed test (Figure 10.9 bottom), the critical Z value is calculated as **NORM.S.INV(1 − 0.05 = 0.95) = 1.645**. Finally, for the two-tailed test, we need to calculate two critical Z values, a lower one and an upper one. Since we are splitting our 5% rejection region into both sides, each side gets a probability of $0.05/2 = 0.025 = 2.5\%$. Therefore, the limit of the lower rejection region (the lower critical Z value) is calculated as **NORM.S.INV(0.025) $= -1.960$**. Likewise, the limit of the upper rejection region (the upper critical Z value) is calculated as **NORM.S.INV(1 − 0.025) = NORM.S.INV(0.975) = 1.960**.

The purpose of this section was to give you a basic understanding of the association between rejection regions and significance levels, based on a normal distribution. However, please note that:

> Each hypothesis test has its **own statistic (*Z*, *t*, *F*, and others), associated with its underlying distribution**. You should use them accordingly, but the concept of the association between a rejection region and a significance level is similar for the other hypothesis tests.

Traditionally, statistics textbooks present the **critical values of the test statistic** in **look-up tables**. But **Excel** also has built-in **functions** for many distributions. In our book, we try as much as possible to use these functions or to develop the calculations without resorting to the use of look-up tables. The examples to follow in this chapter present the relevant **test statistic** and also the associated ***p*-value**, which is covered in the following section.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

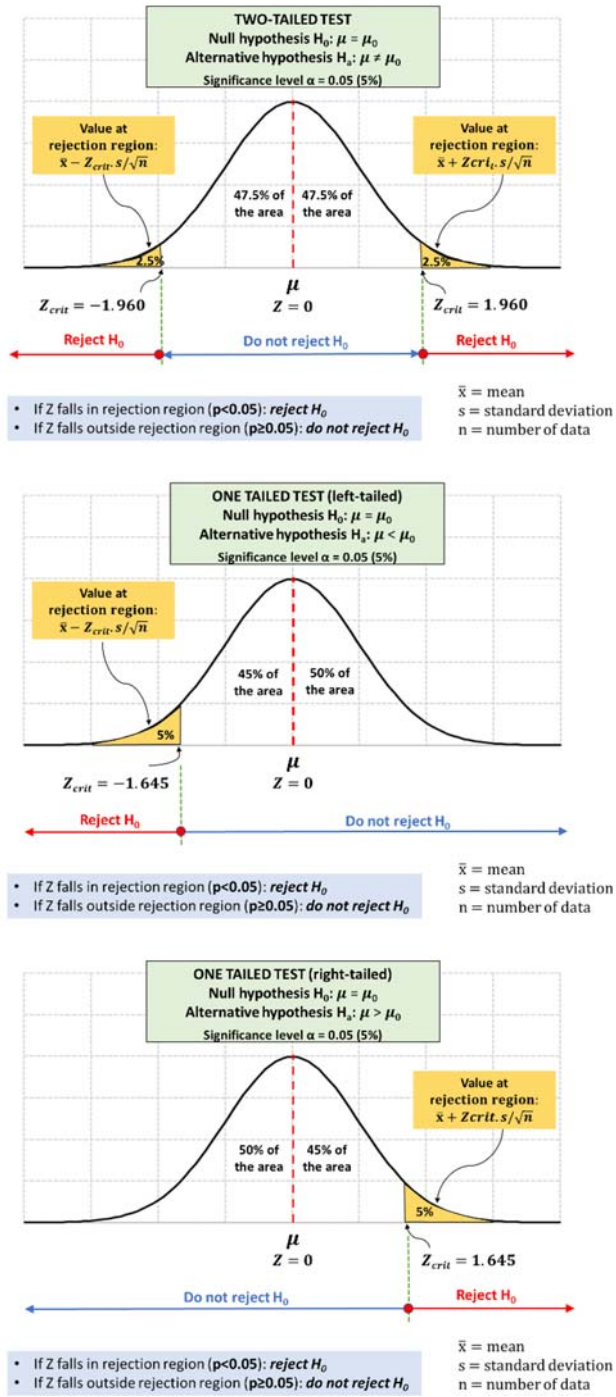


Figure 10.9 Rejection and non-rejection areas, for the *Z* statistics of a normal distribution and a significance level of 0.05 (two-tailed and one-tailed tests).

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

10.2.5 Probability levels (*p*-values) and effect size estimates

Basic

(a) Calculating and interpreting the *p*-value

A practical way (and perhaps the most common way) of expressing results from hypothesis tests is to present a ***p*-value** (also called the *probability level* or the *observed significance level*). The *p*-value is the *probability of incorrectly rejecting the null hypothesis when it is actually true* (i.e., finding misleading results by chance, as described in Outcome 1a from Section 10.2.2). The *p*-value is interpreted in comparison with a prespecified **significance level** (*α level* or *type I error level*, as discussed above):

S. 10.2.2

- **If the *p*-value is less than the significance level α, then we reject the null hypothesis.**
- **If the *p*-value is greater than or equal to α, then we do not reject the null hypothesis.**

The calculation of the *p*-value is an integral part of the statistical test, and our Excel files and practical examples show you how to do this calculation. For the normal distribution, the Excel function **NORM.S.DIST**(Z; TRUE for cumulative) returns the ***p*-value** for a given value of *Z*. Note that the *Z* value we are using here in this Excel function is not the same as the critical *Z* value that was described above in Sections 10.2.4(b) and 10.2.4(c) and calculated using the function NORM.S.INV. The *Z* value we need to use now is called the ***test statistic***, or for the normal distribution, the test *Z* value. It is calculated using Equation 10.5, described below in Section 10.3.1. We will get into more detail about that calculation later in this chapter. Keep in mind that the test statistic *Z* is used with the normal distribution, but there are other test statistics for other distributions, such as the *t* test statistic (for the *t* distribution), the *F* test statistic (for the *F* distribution), etc. Likewise, these other distributions also have critical values (e.g., the critical *t* statistic, the critical *F* statistic, etc.). When doing a hypothesis test, we are always comparing our test statistic to the critical statistic, or alternatively comparing our *p*-value to the significance level α. Both approaches lead to the exact same outcome.

S. 10.2.4

S. 10.3.1

(b) Going beyond the *p*-value: effect size estimates and their precision

While placing emphasis on the *p*-value has traditionally been the most widely used method for interpreting the results of a statistical test, recently, attention has been brought to the limitations associated with only reporting the *p*-value (without any context about the statistical power associated with the test).

For a thought-provoking discussion on this topic, consult Sullivan and Feinn (2012) and Halsey *et al*. (2015). In short, scientists have now argued that you should also report your estimate of the **effect size**, and the precision associated with this estimate. So, what is the effect size? Let us use our familiar example from Section 9.3, where we are interested in comparing our sample mean of 3.16 mg/L to the stipulated regulatory standard value of 4.00 mg/L. In this case, the effect size is equal to $4.00 - 3.16 = 0.84$ mg/L. In other words, based on our sample mean of 3.16 mg/L, our best guess is that the population's mean concentration is 0.84 mg/L lower than the regulatory limit of 4.00 mg/L. However, there is a precision associated with that estimate of 0.84 mg/L. This precision can be characterized using a confidence interval, such as the 95% confidence interval. To calculate this, we follow this sequence:

S. 9.3

- We calculate the difference between 4.00 mg/L and each of our sample values (the average of these differences will be 0.84 mg/L).

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

- From those differences, we calculate the standard deviation and then divide it by the square root of the sample size to get the standard error (SE).
- After that we calculate the margin of error by multiplying the SE by the critical Z value based on our significance level (for $\alpha = 0.05$, the critical Z is computed in Excel using NORM.S.INV $(1 - 0.05/2) =$ NORM.S.INV$(0.975) = 1.96$).
- Finally, we first add, then subtract this margin of error from the estimated effect size of 0.84 mg/L to get the upper and lower 95% confidence limits on our effect size.

S. 9.3

In Example 9.1 from Section 9.3, the data set showed an effect size of 0.84 mg/L with a standard deviation of 1.0 mg/L on a sample size of $n = 36$. Therefore, we can calculate the **lower and upper 95% confidence limits** as follows:

$$\text{Lower 95\% Confidence Limit} = 0.84 - 1.96\frac{1.0}{\sqrt{36}} = 0.51\ \text{mg/L}$$

$$\text{Upper 95\% Confidence Limit} = 0.84 + 1.96\frac{1.0}{\sqrt{36}} = 1.17\ \text{mg/L}$$

Thus, we can say that our sample's mean concentration is 0.84 mg/L lower than the limit of 4.00 mg/L. We can also say that this difference is significant (recall from Example 9.1 in Section 9.3 that the p-value for the hypothesis test was less than the α value of 0.05). Finally, we can go beyond simply presenting the p-value to say that *though our best estimate of the difference is 0.84 mg/L,* ***we have 95% confidence that this difference is between 0.51 and 1.17 mg/L***. You can see how this is providing your audience with much more information than simply presenting a p-value by itself.

(c) Summary of important information for your hypothesis tests

In summary, here is the most important information you need to know about hypothesis testing in order to complete the applied examples in this chapter:

- When doing hypothesis tests, we need to **obtain strong conclusions**, and our ability to do so will depend on how we formulate our hypotheses.
- The **significance level (α) directly is inversely related with the confidence level** of the hypothesis test; typically, we use a value of $\alpha = 0.05$, but if you use a lower value, it will make for a more rigorous hypothesis test.
- The hypothesis test produces a p-value, which allows us to draw a conclusion about the test: **if p-value $< \alpha$, we reject H_0; if p-value $\geq \alpha$, we fail to reject H_0**.
- The ***p*-value is the probability of incorrectly rejecting the null hypothesis when it is actually true** (i.e., finding misleading results by chance).
- The p-value is **only associated with Type I error, and gives no indication regarding a Type II error**.
- **Rejecting the null hypothesis H_0 is a strong conclusion.**
- **Not rejecting the null hypothesis H_0 is generally a weak conclusion** (it usually suggests that we need to collect more data to draw a stronger conclusion).
- **Not rejecting the null hypothesis *does not mean that we can accept the null hypothesis***; we can only say that *the null hypothesis cannot be rejected.*
- The **alternative hypothesis H_a is usually the theory we want to support**; we typically do not believe the null hypothesis to be true, and we are attempting to provide evidence against it (in favour of our alternative hypothesis).

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Advanced

10.3 ONE-SAMPLE PARAMETRIC TESTS FOR A POPULATION MEAN (*Z* TEST AND *T* TEST)

10.3.1 One-sample *Z* test (when σ is *known*)

S. 9.3

One-sample parametric tests have already been presented in Section 9.3, applied to the problem of assessing compliance with quality standards. One-sample tests will also be presented here, albeit in a more complete way. We begin this section with the *Z* test and then move onto the *t* test.

The *Z* test statistic is

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \tag{10.5}$$

where

Z = test statistic
$\bar{x}$ = sample mean
μ = population mean
σ = population standard deviation
n = number of data points in the sample

As stated previously, the rejection region for a significance level α is

- *Two-tailed test*: reject H_0 if $z < z_{\alpha/2}$ or $z > z_{\alpha/2}$
- *One-tailed test (left-tailed)*: reject H_0 if $z < z_{\alpha}$
- *One-tailed test (right-tailed)*: reject H_0 if $z > z_{\alpha}$

In most situations, if $n \geq 30$, the Central Limit Theorem allows us to use these procedures when the population distribution is non-normal. Also, if $n \geq 30$, then we can replace σ with the sample standard deviations, *s*. For further information, see Ott and Longnecker (2010). See Example 10.1 for a simplified application of the one-sample *Z* test when σ is known (Example 10.1). Keep in mind that in real-world settings, we usually never know the true value of the population standard deviation (σ), we typically only have an estimate of it from our sample (i.e., the sample standard deviation, *s*). If we do not know the true population standard deviation, we would use the approach presented in Section 10.3.2.

S. 10.3.2

Example

EXAMPLE 10.1 COMPARING THE MEAN VALUE OF YOUR SAMPLE WITH A FIXED VALUE, WHEN σ IS KNOWN

Suppose that the data refer to 36 BOD (biochemical oxygen demand) values obtained from a water body flowing through a new housing development. In a previous study, completed several years before the new housing development was created, it was found that the mean and standard deviation of BOD were 4.0 and 1.2 mg/L, respectively. You now want to check whether the conditions of the water body have changed or remain the same.

Note: these monitoring data are the same as those used in Example 6.3 and the examples in Section 9.3.

S. 9.3

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Data (*values are in mg/L*):

2.8	4.2	3.9	3.3	2.8	1.7	1.9	2.5	3.1	3.8	2.7	4.1
4.3	4.8	5.6	5.8	3.9	3.5	2.7	3.1	2.8	3.4	4.9	2.8
2.8	1.8	2.1	2.6	2.3	2.4	2.5	1.8	2.9	2.4	2.1	3.6

Sample data:

- *Number of data*: $\boldsymbol{n = 36}$
- $\bar{x} = \boldsymbol{3.16\ mg/L}$

Previous information (*presumed to represent the population*):

- $\mu = \boldsymbol{4.0\ mg/L}$
- $\sigma = \boldsymbol{1.20\ mg/L}$

Is the average concentration equal to the expected value of 4.0 mg/L?

Excel

Note: this example is also available as an Excel spreadsheet.

Solution:

First, we establish the test hypotheses:

- H_0: $\mu = 4.0$ mg/L.
- H_a: $\mu \neq 4.0$ mg/L.

We assume that σ is known (1.2 mg/L). Based on the Central Limit Theorem, which tell us that the distribution of the sample mean is a normal distribution, the test statistic Z would be calculated as follows (Equation 10.5):

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Adopting a value of $\alpha = 0.05$, for a two-tailed test, $\alpha/2 = 0.05/2$, we will reject the null hypothesis if $\bar{X}$ lies more than 1.96 standard errors above $\mu = 4.0$ (see also Table 10.3).

Thus, the decision rule will be

Reject H_0 if $z > +1.96$ or $Z < -1.96$. Otherwise, do not reject H_0. Recall, that we find the values 1.96 and -1.96 based on our significance level α and the fact that we are using a two-tailed test, i.e., $\alpha/2 = 0.05/2 = 0.025$. To calculate these values, we use the Excel function NORM.S.INV(0.025) $= -1.96$ and NORM.S.INV(0.975) $= 1.96$.

Now, for $\bar{X} = 3.16$ mg/L and $\sigma = 1.2$ mg/L:

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{3.16 - 4.0}{\frac{1.2}{\sqrt{36}}} = \boldsymbol{-4.2}$$

IMPORTANT: Take note that in this example, because we are assuming that the true population standard deviation is known to be $\sigma = 1.2$, we ignore the standard deviation of our data (which is $s = 1.0$) in the calculation of the *Z* value, instead we opt to use what we assume is the true standard deviation of the population, $\sigma = 1.2$, in our calculation of the test statistic *Z*. This is seldom the case in our field (we very rarely know the value of the population's standard deviation).

In this example, since –4.2 < –1.96, the decision is to **reject H_0**.

In other words, the difference between the current BOD mean and the BOD mean prior to the housing development is significant and cannot be attributed to chance. The new housing development has modified the average concentrations of BOD of the river.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Another way to report the results of hypothesis testing is to present the ***p*-value**, a statistic that is compared directly with the level of significance, α, for rejection or not of H_0.

The *p*-value is the probability of obtaining a value of the test statistic that is as likely or more likely to lead to rejection of H_0, assuming that the null hypothesis is true.

- If *p-value* $< \alpha$, then reject H_0
- If *p-value* $\geq \alpha$, then do not reject H_0.

The smaller the *p*-value, the stronger the evidence for rejecting H_0. Thus, a *p*-value reports the test results on a continuous scale, rather than just the dichotomous decision 'reject H_0' or 'do not reject H_0.'

The *p*-value can be computed with the test statistic *Z*, which was calculated above as $Z = -4.2$. To calculate the *p*-value, we use the Excel function NORM.S.DIST (−4.2;TRUE), that calculates the area under a normal curve to the left of a value *x* that is *Z* standard deviations away from the mean. The value obtained is 0.000013.

For two-tailed tests, we need to multiply this value by 2:

$$\textbf{\textit{p}-value} = 2 \times P(Z \geq |\text{computed } Z|) = 2P(Z \geq |-4.2|) = 2 \times (0.000013) = \mathbf{0.000026}$$

Because the *p*-value (0.000026) is lower than α (0.05), we **reject H_0** and conclude that there is sufficient evidence to support the alternative hypothesis (H_a).

Now, we can take it a step further, and report the effect size and the 95% confidence interval on this effect size. Our best estimate of the effect size is 4.00–3.16 = 0.84 mg/L. The confidence interval around this effect size is computed as follows (note again that we use the true standard deviation of the population, $\sigma = 1.2$, instead of the standard deviation of our sample's effect size, which is $s = 1.0$).

$$\text{Lower 95\% Confidence Limit} = 0.84 - 1.96\frac{1.2}{\sqrt{36}} = 0.45\,\text{mg/L}$$

$$\text{Upper 95\% Confidence Limit} = 0.84 + 1.96\frac{1.2}{\sqrt{36}} = 1.23\,\text{mg/L}$$

Basic

10.3.2 One-sample *t* test (when σ is unknown)

In the previous section, the hypothesis tests used were based on the assumption that the population standard deviation was known (or that we had enough historical data that we felt that the sample standard deviation (s) was a reasonably precise estimate of the population standard deviation (σ)).

However, in practice, **the population standard deviation (σ)** is seldom known, which means that we need to use a different approach for the one-sample hypothesis test. William Sealy Gosset, a researcher of the Guinness Brewery in Ireland, faced a similar problem in the early twentieth century. He was interested in making inferences about the mean quality of various brews, but he was not supplied with a large sample to reach his conclusions. Gosset concluded that when he used the test statistic with s in the place of σ for *small sample sizes*, he was falsely rejecting H_0 at a higher rate than that specified by α, increasing the probability of Type I error. To solve this problem, he proposed the following approach, which uses the *t* statistic instead of the *Z* statistic.

Suppose we have a simple random sample of size n (<30) drawn from a normal population with a mean μ and a standard deviation σ. However, we do not know the values of μ or σ. So we randomly extract a sample of size n and obtain the values of $\bar{x}$ and s (estimates of μ or σ) and then perform the following calculation:

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \tag{10.6}$$

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

If we were to repeat this sampling process several times, obtaining different estimates $\bar{x}$ and s each time, we would calculate different values of t (t_1, t_2, t_3, …), using the standard deviations from each sample (s_1, s_2, s_3, …). If we plotted a histogram of these values of t, it would show a curve that approaches a normal curve as n approaches infinity (with zero mean). When n is smaller, the curve looks slightly flatter, and has fatter tails. When Gosset discovered this, he published the results in the journal *Biometrika*, in 1908, under a pseudonym of Student. Therefore, the *t statistic* and its distribution are called the *Student's t distribution* or, simply, **Student's *t***.

Note that there is a different t distribution for each sample size. When we speak of a specific t distribution, we have to specify the **degrees of freedom (df)**. The degrees of freedom for this t statistic come from the sample standard deviation s in the denominator of Equation 10.6, and for a one-sample t test, it is equal to the sample size (n) minus one, i.e., $df = n - 1$.

The *t distribution curves* are symmetric and bell-shaped like the normal distribution and have their peak at a t value of 0, just like the normal distribution has its peak at the Z value of 0. However, the spread of the t distribution is a bit broader than the spread of the standard normal distribution, especially for smaller sample sizes. The larger the sample size, the closer the t distribution is to the normal distribution. This reflects the fact that the standard deviation s approaches σ for large sample size n.

In Equation 10.6, the term $s/\sqrt{n}$ is the **standard error** (**SE**).

A summary of the Student's t distribution is presented below (extracted from comments by Mendenhall & Sincich, 1988; Ott & Longnecker, 2010), and also illustrated in Figure 10.10 for a typical confidence level of 95%.

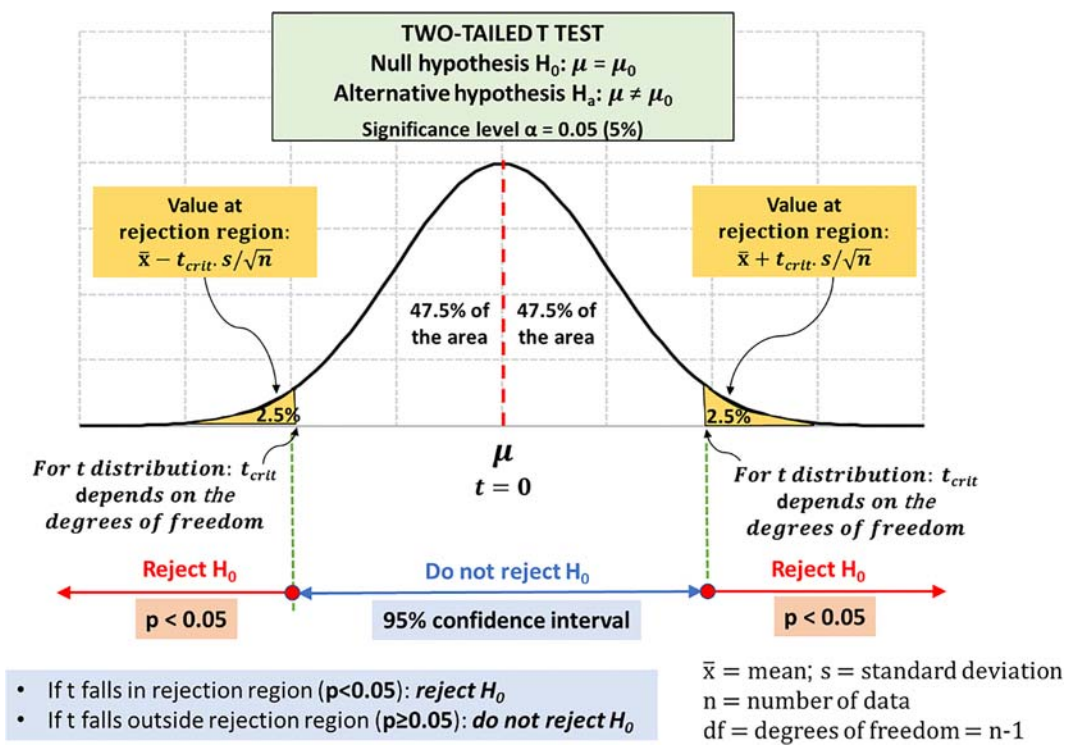


Figure 10.10 Schematics of a one-sample two-tailed t test for a significance level of 0.05.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Properties of Student's *t* distribution

- There are *many different t distributions*, each slightly different based on the sample size, which determines the value of a parameter called **degrees of freedom (df)**.
- The *t* distribution is **symmetrical around 0** and hence has a **mean equal to** 0, similar to the *Z* value with the normal distribution.
- With $t = \dfrac{(\bar{x} - \mu)}{\left(\dfrac{s}{\sqrt{n}}\right)}$ we conclude that *t* has a distribution with $\mathbf{df} = \boldsymbol{n} - \mathbf{1}$, and *as n approaches infinity, the t distribution approaches the normal distribution.*

***t* Test (one sample)**

- **Description**: compares the **mean** value of the sample with a specified reference value.
- **Type**: *parametric* test.
- **Input data required**: number of data points, mean and standard deviation, and value of the reference value we want to use for the comparison, plus the specification of the desired significance level for the test (α).
- **Output data produced**: *t* statistic, *p*-value.
- **Test hypotheses**:
  - *Null hypothesis* H_0: $\mu = \mu_0$.
  - *Alternative hypothesis* H_a: $\mu \neq \mu_0$ (*two-tailed*); $\mu < \mu_0$ (*left-tailed*) *or* $\mu > \mu_0$ (*right-tailed*).
- **Test statistic**: $t = \dfrac{(\bar{x} - \mu)}{\left(\dfrac{s}{\sqrt{n}}\right)}$ and the *t* distribution are based on $n - 1$ degrees of freedom.
- **Rejection region**: $t < t_{\alpha/2}$ or $t > t_{\alpha/2}$ (two-tailed); $t < -t_\alpha$ or $t > t_\alpha$ (one-tailed).
- **Rejection with *p*-value**: *p*-value $< \alpha$.
- **Assumptions**: the data have been obtained independently and represent a random sample from a population that is *normally distributed.*
- **Comment**: if the sample size is small ($n < 30$) and if we cannot ascertain that the population from which the sample was obtained is normally distributed, non-parametric tests may be needed. If the sample size is large, the *t* test is relatively robust to small departures from normality.

Example

S. 9.3

EXAMPLE 10.2 COMPARING THE MEAN VALUE OF YOUR SAMPLE WITH THE VALUE OF A REGULATORY STANDARD USING THE T TEST

We will use again the same monitoring data from Example 10.1, which came from Example 6.3 and the examples in Section 9.3. We want to check what is the probability that our sample mean is significantly equal to the regulatory standard of 4.0 mg/L.

Sample data:

- Number of data: $n = \mathbf{36}$
- Sample mean: $\bar{x} = \mathbf{3.16\ mg/L}$
- Sample standard deviation: $s = \mathbf{1.04\ mg/L}$

Regulatory standard:

- $\boldsymbol{\mu} = \mathbf{4.0\ mg/L\ (standard)}$
- $\boldsymbol{\sigma} = \mathbf{unknown}$

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Excel

Note: this example is also available as an Excel spreadsheet.

Solution:

We establish the following hypotheses for the two-tailed test:

- Null hypothesis H_0: $\mu = 4.0\ mg/L$.
- Alternative hypothesis H_a: $\mu \neq 4.0\ mg/L$.

For $\alpha = 0.05$, the **critical value of the distribution *t***, with $36 - 1 = 35$ degrees of freedom is **2.030**, that is, $t_{0.05;35} = 2.030$, obtained by Excel function T.INV.2T(probability; deg_freedom) or T.INV.2T (0.05,35).

The *t*-test statistic will be

$$t = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} = \frac{3.16 - 4.0}{\frac{1.04}{\sqrt{36}}} = -4.846$$

Thus, the decision rule will be
Reject H_0 if $t > t_{35} = 2.030$ or if $t < t_{35} = -2.030$
Otherwise, do not reject H_0.

Since $-4.846 < -2.030$, the decision is to **reject H_0**. In other words, we **reject the hypothesis that the mean is equal to the standard in favour of the alternative hypothesis that the mean is not equal to the standard** (in this case, it is significantly lower than the standard).

We may *convert the t statistics that define the rejection limits to the original scale of concentration values* (*mg/L*), so that we can more easily compare them with the mean (3.16 mg/L) and regulatory standard (4.00 mg/L).

The basic equation is

$$X = \mu \pm t \cdot \frac{s}{\sqrt{n}}$$

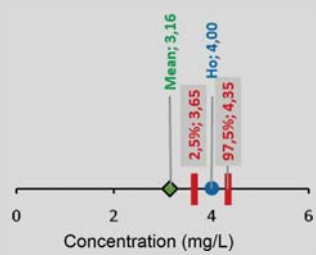
The concentration value of the lower rejection limit is

$$\text{LRL} = \mu - t_{0.05,35} \times \frac{s}{\sqrt{n}} = 4.00 - (2.030) \times \frac{1.04}{\sqrt{36}} = 3.65\ \text{mg/L}$$

The concentration value of the upper rejection limit is

$$\text{URL} = \mu + t_{0.05,35} \times \frac{s}{\sqrt{n}} = 4.00 + (2.030) \times \frac{1.04}{\sqrt{36}} = 4.35\ \text{mg/L}$$

The following scheme shows the value of the mean compared with the rejection limits and the regulatory standard. We can see that the mean (3.16 mg/L) is located outside of the non-rejection region, which is defined by the interval [3.65; 4.35]. Therefore, the *mean concentration* (*3.16 mg/L*) *is in the rejection region*, and the **null hypothesis needs to be rejected**.



Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

An alternative way of doing the analysis is by obtaining the ***p*-value**. The *p*-value may be calculated using the Excel function TDIST(x; deg_freedom; tails). Therefore, TDIST(ABS(−4.846);36−1;2) = **0.000026**.

Since *p*-value (0.000026) < significance level (0.05), we **reject the null hypothesis that the mean is equal to the standard**. Therefore, we can **accept the alternative hypothesis** that the mean is different from the standard. Since the mean (3.16 mg/L) is lower than the standard (4.00 mg/L), we conclude that the **mean is significantly lower than the regulatory standard**.

Again, if we were to go further than simply providing the *p*-value, we would calculate the mean effect size (4.00 − 3.16 = 0.84 mg/L) and then find the 95% confidence interval of this effect size. However, this time, we would use the *t* statistic for our particular significance level (α value) and degrees of freedom (df = 35) instead of the *Z* statistic, as we have done previously. The results obtained are only slightly different, the effect size is 0.84 mg/L with a 95% confidence interval of 0.50 − 1.18 mg/L.

$$\text{Lower 95\% Confidence Limit} = \bar{x} - t_{0.05,35} \cdot \frac{s}{\sqrt{n}} = 0.84 - 2.030 \cdot \frac{1.04}{\sqrt{36}} = 0.50\,\text{mg/L}$$

$$\text{Upper 95\% Confidence Limit} = \bar{x} + t_{0.05,35} \cdot \frac{s}{\sqrt{n}} = 0.84 + 2.030 \cdot \frac{1.04}{\sqrt{36}} = 1.18\,\text{mg/L}$$

10.3.3 Sample size and the *t* test

Advanced

(a) Effect of sample size on the results from the *t* test

S. 4.5.3

The determination of the **rejection regions and the confidence interval are influenced by the sample size (number of data points *n*)**, as you can see from the equations for the test. Go back to Section 4.5.3 in Chapter 4 where we give a clear introduction to the concept of confidence intervals and the influence of the sample size.

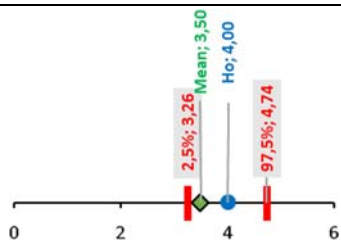
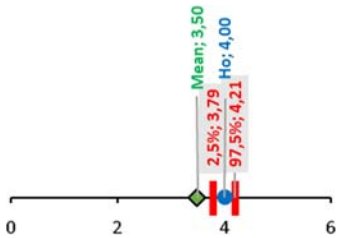
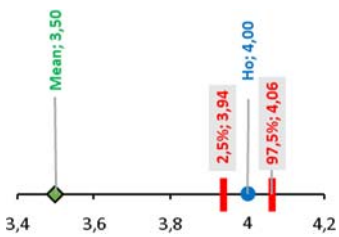
Let us now assume that for the same regulatory standard (4.00 mg/L), we want to check whether a sample with a *mean value* of 3.50 mg/L is significantly lower than the standard. We state the null and alternative hypotheses in a similar way as we did for Example 10.2, that is null hypothesis H_0: $\mu = 4.0\ mg/L$; alternative hypothesis H_a: $\mu \neq 4.0\ mg/L$. We will use the same standard deviation from Example 10.2, $s = 1.04$. Now, let us analyse the test results for three different sample sizes: $\boldsymbol{n = 10}$**, **$\boldsymbol{n = 100}$**, and **$\boldsymbol{n = 1000}$**. If we use the Excel spreadsheet for Example 10.2 (or redo the calculations presented here for Example 10.2), we will get the rejection regions and test results for a one-sample two-tailed *t* test shown in Table 10.4.

The interpretation of Table 10.4 clearly shows the **influence of the sample size (*n*) on the results and conclusions from the *t* test**. With a small sample ($n = 10$), the non-rejection interval was wide, and the mean of 3.50 mg/L stayed inside it, which led to the non-rejection of H_0. When we increased the sample size (to 100 and 1000), the non-rejection interval decreased, which led to the rejection of H_0, and the confirmation of the alternative hypothesis that mean ≠ standard.

The result with a **very large sample size** leads us to reflect on the result of a statistical test. The non-rejection interval is so narrow [3.94–4.06 mg/L] that, in practice, practically all water quality samples will probably fall outside of it. Considering that we are working with environmental systems that often have large inherent variability (unlike an industrial process, where operating conditions are tightly controlled), we need to treat these results with an understanding that our system is likely quite dynamic and may not always be in a steady state. *We usually always consider larger sample sizes to be better*. However, in some cases, if you have extremely large

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Table 10.4 Example of results from the *t* test under the same conditions, but varying the sample size ($n = 10, 100, 1000$).

Sample Size (*n*)	Rejection Region	Conclusion
10		Mean outside the rejection region (inside the non-rejection interval of 3.26–4.74 mg/L). *p*-value = **0.163** **Do not reject H_0**
100		Mean in the rejection region (outside the non-rejection interval of 3.79–4.21 mg/L) *p*-value = $\mathbf{5.454 \times 10^{-6}}$ **Reject H_0**
1000		Mean in the rejection region (outside the non-rejection interval of 3.94–4.06 mg/L) *p*-value = $\mathbf{4.11 \times 10^{-47}}$ **Reject H_0**

Mean$n = 3.50$ mg/L; standard deviation$n = 1.04$ mg/L;
Null hypothesis H_0: $\mu = 4.0\ mg/L$; alternative hypothesis H_a: $\mu \neq 4.0\ mg/L$.

sample sizes, you may experience **issues associated with a precision that appears very high**. However, keep in mind that if the system is not in steady state, what used to be the mean concentration can change. It is up to you, based on your knowledge of the system, to interpret this with consideration of all of the elements that impact the behaviour of your system, and not based solely on numbers resulting from statistical tests. *Statistics is a tool that can help you, but you are ultimately in control of the interpretation of your results and at times, you may also need to use your best judgement and some common sense when drawing conclusions.*

(b) Determination of the required sample size

Advanced

The analysis provided above prompts us to think about the following question:

How many data points should be collected in order to detect a desired difference from the population mean (a desired effect size), considering our desired statistical power?

S. 3.5

In Section 3.5, we dealt with this problem and employed a power calculation to lead us to the definition of the required sample size. Go to that section for a review about this method.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

S. 3.2.2

S. 3.5

Here, we will use a slightly different and perhaps more straightforward procedure for calculating the required sample size. We can estimate the required sample size (n) if we consider that our sample standard deviation (s) is a good predictor of the population standard deviation (σ). We may perform a t test after specifying the probability α of making a Type I error, and a probability β of incurring a Type II error (see Section 3.2.2 for a discussion of test errors). In Section 3.5, we mentioned that a conventional approach is to use **0.05 for the α error and 0.20 for the β error (i.e., 80% power).**

We can then state that we want to be able to detect a specified difference between μ (actual population mean) and μ_0 (mean specified in H_0). This is called the effect size. In order to be able to detect a significant difference at the significance level α with a power of $1-\beta$, the minimum sample size required can be calculated using the following equation (Zar, 1999):

$$n = \frac{s^2}{(\mu_A - \mu_0)^2} \times (t_\alpha + t_\beta)^2 \tag{10.7}$$

where

n = required sample size (required number of data points)
s^2 = sample variance (standard deviation s, squared)
$\mu_A - \mu_0$ = the desired effect size; that is, the difference that we want to be able to detect with significance, between our sample and the fixed value or the presumed population mean
t_α = critical value of t for α and df $= n - 1$
t_β = critical value of t for β and df $= n - 1$

α can be either $\alpha(1)$ or $\alpha(2)$, respectively, depending on whether a one-tailed or a two-tailed test is to be used.

Note that Equation 10.7 may be rearranged and expressed in the following way, for you to be able to see more clearly the influence of t_α and t_β:

$$\frac{(\mu_A - \mu_0)}{\frac{s}{\sqrt{n}}} = t_\alpha + t_\beta \tag{10.8}$$

S. 3.5

On the left-hand side, the term $(\mu_A - \mu_0)/s$ is what we called **Cohen's d** in Section 3.5, that dealt with power analysis. It is a standardized effect size. Also, you can see that the denominator on the left-hand side ($s/\sqrt{n}$) is the **standard error** (**SE**).

We saw that the calculation of the t statistic depends on the degrees of freedom (df). Since df depends directly on n (df $= n - 1$), n cannot be calculated directly from Equation 10.7, but must be obtained by iteration.

The values of t_α and t_β can be obtained using the following Excel functions:

- t_α (two-tailed): **T.INV.2T(probability; deg_freedom)**; where probability is α.
- t_β (one-tailed)**: T.INV(probability; deg_freedom)**; where probability is $1 - \beta$.

This procedure is illustrated in Example 10.3.

S. 3.5

In summary, you can use the methods presented here and in Section 3.5 to determine the required sample size for your experiment, based on your desired effect size and your tolerance for Type I and Type II errors. However, choosing an *appropriate* sample size for your experiment or study is going to depend on other considerations, such as *available funding, resources, and time for your monitoring, as well as all the logistics involved in the experimental set-up*. Just keep in mind that if you plan an experiment with a low statistical power, your chances of success (i.e., an

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

S. 10.2.2

outcome where you successfully detect a significant difference, i.e., Outcome 2a from Section 10.2.2) will be very low, and you may be more likely to have inconclusive results (i.e., Outcome 2b from Section 10.2.2). Therefore, you might assess the situation, and determine if it is better to spend a small amount of funds, resources, and time to most likely obtain inconclusive results, or if it is worth it to invest sufficient funds, resources, and time to have a higher chance of finding significant results.

Another consideration is you might want to consider adjusting your desired effect size. Ask yourself – would the study be just as impactful if you used a smaller effect size? Is it worth spending all of the time, effort, and resources to collect and analyse so many samples just to be able to detect a very small effect size? **There is a difference between finding results that are *significant* from a statistical perspective versus results that are *meaningful* from a practical perspective.**

Let us consider an example to illustrate this concept. Suppose you are testing a modification to a wastewater treatment process to understand its effect on phosphorus removal. You set up two reactors, one with and one without the modification, and you measure the effluent phosphorus concentration in both of them. You are interested in seeing if the modified process produces effluent with a significantly lower concentration of phosphorus than the unmodified process. Your null hypothesis is that the mean difference is equal to zero (H_0: $\mu = 0$) and your alternative hypothesis is that the mean difference is greater than zero (H_a: $\mu > 0$). To determine your required sample size for the experiment, you need to assume a desired effect size. Suppose your sample size calculations indicate that $n = 30$ samples are required to detect a significant difference of 5 mg/L in the phosphorus concentration, but a sample size of $n = 300$ would be required to detect a significant difference of 0.05 mg/L. You might decide that it is not worth to collect and analyse 300 samples because you do not care if the modification reduces the effluent concentration by 0.05 mg/L. That small reduction is not worth for you to collect and analyse 300 samples. However, if the modification were to cause a decrease of 5 mg/L, it would be meaningful. It would be worth the effort, especially if it only requires analysing 30 samples!

Example

EXAMPLE 10.3 ESTIMATING THE REQUIRED SAMPLE SIZE TO BE ABLE TO DETECT A DESIRED EFFECT SIZE (ONE-SAMPLE T TEST)

We will use again the same data from Example 10.2. Suppose we wish to test at the 0.05 significance level ($a = 0.05$) with an 80% power ($\beta = 1 - 0.80 = 0.20$) to detect an effect size of 0.5 mg/L (that is, to detect a significant difference even if the true mean concentration is 3.50 mg/L compared to the standard of 4.00 mg/L; $4.0 - 3.5 = 0.5$). In Example 10.2 we saw that the sample standard deviation was 1.04 mg/L. Of course, when you are performing a power calculation, you will not have the data, so you will not be able to calculate the standard deviation from your sample (because, presumably, you have not collected it yet!). In this case, you should assume a standard deviation based on samples analysed previously for the same constituent in a similar system.

Excel

Note: this example is also available as an Excel spreadsheet.

Solution:

We start with an initial estimate of the required sample size. Let us start with $n = 10$. Thus, $df = 10 - 1 = \mathbf{9}$.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

With $\boldsymbol{\alpha = 0.05}$ (probability of 0.05) and df = 9, we calculate t_α as $t_{0.05;9} = 2.262$ (two-tailed test) using the Excel function **T.INV.2T(probability; deg_freedom)** or T.INV.2 T(0.05; 9) = 2.262. Therefore, $\boldsymbol{t_\alpha = 2.262}$.

With $\boldsymbol{\beta = 0.20}$, we have a probability of $1 - 0.20 = 0.80$ (**80% power**). With a probability of 0.80 and df = 9, we calculate t_β as $t_{0.80;9} = 0.883$ (one-tailed test, Excel function **T.INV(probability; deg_freedom)** or T.INV(0.80; 9) = 0.883). Therefore, $\boldsymbol{t_\beta = 0.883}$.

Using Equation 10.7, we obtain:

$$n = \frac{s^2}{(\mu_A - \mu_0)^2} \times (t_\alpha + t_\beta)^2 = \frac{1.04^2}{(0.5)^2} \times (2.262 + 0.883)^2 = 42.8$$

We now use the next integer above the calculated value of n. Therefore, we adopt $n = 43$ as an estimate, and obtain the following values, calculated as above: **df = 42**; $t_\alpha = t_{0.05;42} = \mathbf{2.018}$; $t_\beta = t_{0.80;42} = \mathbf{0.850}$.

Using these new values, we obtain:

$$n = \frac{s^2}{(\mu_A - \mu_0)^2} \times (t_\alpha + t_\beta)^2 = \frac{1.04^2}{(0.5)^2} \times (2.018 + 0.850)^2 = 35.6$$

You can now try with $n = 36$, and will see that the calculation converges in $\boldsymbol{n = 36}$ **as the required sample size**. If not, you could go through another iteration. You can also use the **Solver tool** to obtain direct convergence, as illustrated in the Excel file associated with this example.

You can try with different values of α and β and see what the impact in the resulting value of n is. Remember, this is just an **estimate to give you an idea of your required sample size**, and not a final value to be implemented in your experiments. To take this decision you will need to *analyse the available funds, resources and time for your monitoring, as well as all the logistics involved in the experimental set-up*. Just keep in mind that if you plan an experiment with lower statistical power, your chances of success will be lower, and you will be more likely to have inconclusive results.

10.4 INFERENCES COMPARING TWO POPULATION CENTRAL VALUES

10.4.1 Two-sample tests covered in this chapter

Basic

So far, we have covered the use of hypothesis tests to make inferences on a parameter from a single population. However, a *more commonly employed statistical procedure* is the **comparison of two samples**, to infer *whether there are significant differences between the central values of the two populations sampled*. For instance, we might wish to compare the mean (or median) nitrogen content of two lakes or the mean (or median) removal efficiencies of two wastewater treatment plants. In Figure 10.11, we present the different hypothesis tests covered in this chapter. You can see that we have tests for independent and dependent samples, and options for parametric (testing for the means) and non-parametric (testing for the medians) tests:

S. 10.4.2
S. 10.4.3
S. 10.4.4
S. 10.4.5

- **Parametric *t* test** for *independent samples* (Section 10.4.2)
- **Non-parametric Mann–Whitney *U*-test** for *independent samples* (Section 10.4.3)
- **Parametric *t* test** for *dependent samples* (matched pairs) (Section 10.4.4)
- **Non-parametric Wilcoxon's matched-pairs test** for *dependent samples* (Section 10.4.5)

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

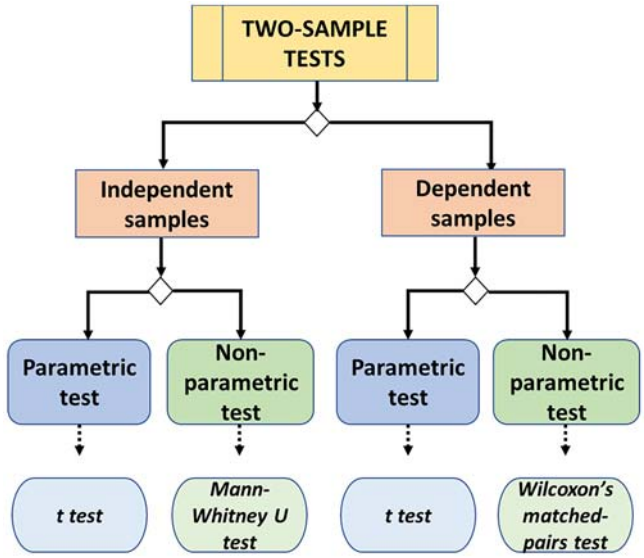


Figure 10.11 Two-sample tests covered in this chapter.

10.4.2 Inferences about the population means: parametric *t* test for two independent samples

Basic

In this section, we will consider a situation where we are comparing **independent random samples** from **two populations that have normal distributions** with *different means* μ_1 and μ_2, but *identical standard deviations* σ_1 and σ_2. Because the standard deviation of the populations (σ) is unknown in most cases, we *must estimate its value*. This estimate is denoted by s_p and is formed by combining (*pooling*) the two independent estimates of σ (s_1 and s_2). This is called assuming a common variance:

$$s_p = \sqrt{\frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{n_1 + n_2 - 2}} \tag{10.9}$$

The ***t* statistic** (test statistic) assuming *common variance* will then be calculated as follows:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{s_p^2 \left(\frac{1}{n_1} - \frac{1}{n_2} \right)}} \tag{10.10}$$

If we use a test in which our null hypothesis is that the means are equal (as we are doing here), then $(\mu_1 - \mu_2) = 0$ and the calculated t value can be obtained from a simplification of Equation 10.10:

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{s_p^2 \left(\frac{1}{n_1} - \frac{1}{n_2} \right)}} \tag{10.11}$$

As already mentioned, in most cases the true standard deviation of the two tested populations is not known. *The only available information is the means* ($\bar{x}_1, \bar{x}_2$) *and standard deviations* (s_1, s_2) *of the*

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

samples. The test may be **one-tailed** or **two-tailed**, depending on whether or not you have a strong reason to believe that the mean of one population should be larger than the mean of the other.

In fact, the s_p^2 *is a weighted average of the sample's variances,* s_1^2 *and* s_2^2. The process of pooling the two sample variances costs an additional degree of freedom, since two parameters (s_1^2 and s_2^2) are estimated. The *degrees of freedom* for this type of two-sample t test is calculated as $\mathbf{df} = \boldsymbol{n_1} + \boldsymbol{n_2} - \mathbf{2}$.

Three assumptions are necessary to perform this test:

- Both samples were selected at *random.*
- The populations from which the samples were drawn are *normally distributed.*
- The *variances* of the two populations are *equal.*

The *randomness* assumption of the samples is mandatory. If the samples are not randomly collected, then this statistical test cannot be used. The verification of the *normality* assumption can be made through basic descriptive statistics and graphical analysis, such as box-whisker diagrams and normal probability graph (see Section 8.2.8). The third assumption, that the *variances* of the two populations are equal, can be checked using an ***F*-test for the equality of variances**.

S. 8.2.8

Below we present the sequence for performing the F-test.

(a) Testing the equality of variances (*F*-test)

Tests to determine equality of variances are based on a probability distribution called the ***F*-distribution**. This is the *theoretical distribution of values that would be expected by randomly sampling from a normal population of values,* and it has a non-symmetrical shape, extending from 0 to $+\infty$.

When $\sigma_1^2 = \sigma_2^2$, $\sigma_1^2/\sigma_2^2 = 1$, s_1^2/s_2^2 follows an F distribution with $\mathrm{df}_1 = n_1 - 1$ and $\mathrm{df}_2 = n_2 - 1$. The test procedure is summarized below.

Test hypotheses:

$\mathrm{H}_0: \sigma_1^2 = \sigma_2^2$

$\mathrm{H}_a: \sigma_1^2 > \sigma_2^2$ or $\mathrm{H}_a: \sigma_1^2 < \sigma_2^2$

Test statistic:

$$F = \frac{s_1^2}{s_2^2} \tag{10.12}$$

For a specified value of α and with $\mathrm{df}_1 = n_1 - 1$ and $\mathrm{df}_2 = n_2 - 1$, the outcomes of the test are:

- Reject H_0 if $F > F_{\alpha,\ \mathrm{df1},\ \mathrm{df2}}$
- Reject H_0 if $F < F_{\alpha,\ \mathrm{df1},\ \mathrm{df2}}$

As mentioned above, the F-distribution is *constrained on the left by zero* and has a *long tail to the right*. If we always place the larger variance in the numerator, the ratio will always be greater than 1.0 and the calculated test statistics will always fall on the right. We then can test for significance using a **one-tailed critical region on the right-side of the distribution**.

If the **calculated value of *F* exceeds the critical value of *F*** with df degrees of freedom and a determined significance level, the **null hypothesis is rejected,** and we have enough evidence to conclude that the variances are significantly different from each other. If the calculated F value is lower than the critical value, then we cannot reject the null hypothesis, and we would operate under the assumption that the variances are equal to each other.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

The **critical value of *F*** is calculated by using the Excel function that returns the inverse of the (right-tailed) F probability distribution:

F.INV.RT(probability, **deg_freedom1**, **deg_freedom2**) → F.INV.RT ($\alpha; n_1 - 1; n_2 - 1$).

(b) Testing the equality of means (*t* test with equal variances)

If the variances are not significantly different, the next step in the procedure is to test the equality of means. For obvious reasons, the significance level adopted for this test cannot be higher than the significance adopted for the F-test for the equality of variances.

The test hypotheses are:

Two-tailed test:

- H_0: $\mu_1 = \mu_2$ or $(\mu_1 - \mu_2) = 0$
- H_a: $\mu_1 \neq \mu_2$ or $(\mu_1 - \mu_2) \neq 0$

One-tailed (left-tailed):

- H_0: $\mu_1 = \mu_2$ or $(\mu_1 - \mu_2) = 0$
- H_a: $\mu_1 < \mu_2$ or $(\mu_1 - \mu_2) < 0$

One-tailed (right-tailed):

- H_0: $\mu_1 = \mu_2$ or $(\mu_1 - \mu_2) = 0$
- H_a: $\mu_1 > \mu_2$ or $(\mu_1 - \mu_2) > 0$

Sometimes, for the sake of clarity, some people prefer to state the null hypothesis **H_0** in **one-tailed tests** using the signs '>' or '<'. Even if this allows for an easier understanding of the role of H_0 in one-tailed tests, formally speaking, H_0 should only have ' = ' signs, and H_a is the hypothesis that should accommodate '>' or '<' signs. This consideration does not influence the test results, it is only related to how we report our hypotheses.

For a level α, Type I error, we have the following possible outcomes:

- *Two-tailed test*: Reject H_0 if $t < t_{\alpha/2}$ or $t > t_{\alpha/2}$
- *One-tailed test* (*left-tailed*): Reject H_0 if $t < t_\alpha$
- *One-tailed test* (*right-tailed*): Reject H_0 if $t > t_\alpha$

The **critical value of *t*** (t_{crit}) can be calculated using Excel function:

T.INV.2T(probability; **deg_freedom**)

(c) Testing the equality of means (*t* test with unequal variances)

The comparison of two means from normal populations without assuming equal variances is known as the 'Behrens–Fisher problem,' referring to the solution provided originally by Behrens (1929) and Fisher (1939), but also by numerous other studies. One of the easiest of such procedures is attributed to Smith (1936), who proposed an adaptation of the t test statistic (compare it with Equation 10.11):

$$t'_{calc} = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}} \tag{10.13}$$

(d) Excel function for direct calculation of the *p*-value of the *t* test

Excel has a built-in **function** for returning the ***p*-value** in the case of one or two-tailed tests, and also equal variances (homoscedastic) or unequal variances (heteroscedastic), without the need for doing intermediate calculations:

T.TEST(array1, **array2**, **tails**, **type**)

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

where

- **Array1**. The first data set.
- **Array2**. The second data set.
- **Tails**. Specifies the number of distribution tails. If tails = 1, T.TEST uses the one-tailed distribution. If tails = 2, T.TEST uses the two-tailed distribution.
- **Type**. The kind of *t* test to perform:
  - 1 for paired *t* test (covered in Section 10.4.4)
  - 2 for two-sample test with equal variances (homoscedastic)
  - 3 for two-sample test with unequal variances (heteroscedastic)

S. 10.4.4

(e) Summary of the two-sample *t* test

***t* test (two independent samples; comparison of means)**

- **Description**: compares the **mean** value of two samples
- **Type**: *parametric* test
- **Input data required**: number of data points in each sample (n_1 and n_2), mean ($\bar{X}_1$ and $\bar{X}_2$), and standard deviation (s_1 and s_2) of each of the two samples, plus the specification of the desired significance level for the test (α)
- **Output data produced**: *t* statistic, *p*-value
- **Test hypotheses:**
- *Null hypothesis* H_0: $\mu_1 = \mu_2$
- *Alternative hypothesis* H_a: $\mu_1 \neq \mu_2$ (*two-tailed*); $\mu_1 < \mu_2$ (*left-tailed*) *or* $\mu_1 > \mu_2$ (*right-tailed*)
- **Test statistic** (for equal variances):

$$t_{\text{calc}} = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{s_p^2\left(\frac{1}{n_1} - \frac{1}{n_2}\right)}}$$

where

$$s_p = \sqrt{\frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{n_1 + n_2 - 2}}$$

and distribution of *t* is based on $(n_1 + n_2 - 2)$ degrees of freedom

- **Test statistic** (for unequal variances):

$$t'_{\text{calc}} = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}}$$

and distribution of *t* is based on $(n_1 + n_2 - 2)$ degrees of freedom

- **Rejection region**: $t < t_{\alpha/2}$ or $t > t_{\alpha/2}$ (two-tailed); $t < -t_\alpha$ or $t > t_\alpha$ (one-tailed)
- Rejection with *p*-value: *p*-value $< \alpha$
- **Assumptions**: the relative frequency distribution of the populations from which the samples are selected are both approximately *normally distributed*; the variances of the populations are equal (if not, apply test for unequal variances); the random samples are selected independently from the two samples.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Example

EXAMPLE 10.4 COMPARISON BETWEEN THE MEANS OF TWO SAMPLES USING THE T TEST FOR INDEPENDENT SAMPLES

You monitored a certain constituent in the effluent from two treatment plants (or in two water bodies). Use the two-tailed t test for independent samples to analyse if the means of the two samples are significantly different from each other. Note that the samples have different numbers of data points.

Sample 1	2.8	4.2	3.9	3.3	2.8	1.7	1.9	2.5	3.1
	3.8	2.7	4.1	4.3	4.8	5.6	5.8	3.9	
	3.5	2.7	3.1						
Sample 2	2.8	3.4	4.9	2.8	2.8	1.8	2.1	2.6	2.3
	2.4	2.5	1.8	2.9	2.4	2.1	3.6		

Excel

Note: this example is also available as an Excel spreadsheet.

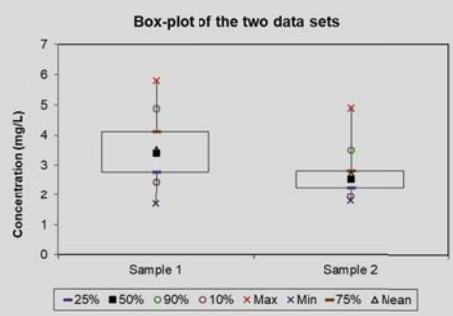
Solution:

- **Descriptive statistics**

  From these data, we have the following basic descriptive statistics:

Statistic	Sample 1	Sample 2
Number of data (n)	20	16
$\bar{x}$	3.53	2.70
s	1.09	0.77

Before performing the calculations, you may be interested in analysing the box-plots of the two samples first, in order to have an initial impression about their measures of central tendency and relative variability. Visually speaking, you notice that the concentrations of Sample 1 seem to be greater than those from Sample 2. You also observe that the mean concentration from Sample 1 is 3.53 mg/L, and it is greater than the mean concentration from Sample 2, which is 2.70 mg/L. Both concentrations appear to follow a normal distribution, though Sample 2 shows some slight departures from normality. Now, you want to know whether these differences are significant, and for this you decide to use the *t* test.



Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

- **Testing the equality of the variances**

  For **testing the equality of variances** in the two samples sets, the null and alternative hypothesis are:

  $H_0: \sigma_1^2 = \sigma_2^2$

  $H_a: \sigma_1^2 > \sigma_2^2$

  One-tailed test: reject H_0 if $\sigma_1^2 > \sigma_2^2$

  $$F = \frac{S_1^2}{S_2^2} = \frac{1.09^2}{0.77^2} = \frac{1.188}{0.593} = \mathbf{2.00}$$

  The critical F value depends on the size of each sample (and its resulting degree of freedom $n-1$) and the significance level (adopted as 0.05 here):

  $df_1 = n_1 - 1 = 20 - 1 = 19; \quad df_2 = n_2 - 1 = 16 - 1 = 15$

  Using the Excel function F.INV.RT(probability,deg_freedom1,deg_freedom2), we can calculate the critical F:

  $$\mathbf{F.INV.RT}\ (\boldsymbol{\alpha}; \mathbf{n_1 - 1}; \mathbf{n_2 - 1}) = \text{F.INV.RT}\ (0.05; 19; 15) = 2.34$$

  Since $F_{calc} = 2.00 < F_{0.05;19;15} = 2.34 \Rightarrow$ **do not reject H_0**

  Therefore, we have no evidence for concluding that the variances are different. As a result, the ***t* test with equal variances can be applied**.

- **Two-sample *t* test assuming equal variances**

  To perform the t test, we specify:
  - Null hypotheses **H_0:** $\mu_1 = \mu_2 \Rightarrow (\mu_1 - \mu_2 = 0)$
  - Alternative hypothesis **H_a:** $\mu_1 \neq \mu_2 \Rightarrow (\mu_1 - \mu_2 \neq 0)$
  - Significance level for the test $(\alpha) = 0.05$ (confidence level of 0.95 or 95%)

  Total degrees of freedom: $df = n_1 + n_2 - 2 = 20 + 16 - 2 = 34$

  Because σ is unknown, we must estimate its value, by means of the s_p, which is formed by pooling the two independent estimates of σ (s_1 and s_2) (Equation 10.9):

  $$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{(20 - 1)(1.09)^2 + (16 - 1)(0.77)^2}{20 + 16 - 2} = 0.926$$

  This means that the pooled standard deviation is $\sqrt{0.926} = 0.96$ (this value will be used in Example 10.5).

  The calculated t value can be obtained by Equation 10.10:

  $$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

  Since our null hypothesis states equality of means, we have $(\mu_1 - \mu_2) = 0$ and the **calculated *t*** value can be obtained from:

  $$t_{calc} = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{3.53 - 2.70}{\sqrt{0.926\left(\frac{1}{20} + \frac{1}{16}\right)}} = \mathbf{2.557}$$

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

The **critical value of *t*** ($t_{0.05;34}$) is obtained from the Excel function:

T.INV.2T(probability; deg_freedom) = T.INV.2T(0.05; 34) = **2.032**

Comparison between the calculated and the critical *t* values and test result (note that we use the absolute value of t_{calc}):

Since $|t_{calc}| = 2.557 > t_{0.05;34} = 2.032 \Rightarrow$ reject H_0

Conclusion: accept the alternative hypothesis. ***The means are significantly different from each other.***

- ***p*-value for the *t* test**

An alternative way of doing the analysis is by obtaining the ***p*-value**. The *p*-value is calculated for a two-tailed test using Excel function:

$$\textbf{T.DIST.2T}(\boldsymbol{x}; \textbf{deg_freedom}) = \text{T.DIST.2T}(t_{calc}; n_1 + n_2 - 2) = \text{T.DIST.2T}(2.557; 20 + 16 - 2) = \mathbf{0.0152}$$

Since *p*-value $<$ significance level α ($0.0152 < 0.05$), we **reject the null hypothesis**. The conclusions are as stated above.

Note that we could also have used directly the Excel function **T.TEST** to perform the direct calculation of the *p*-value, without the need of intermediate calculations (see description in the text just before this example):

T.TEST(array1,array2,tails,type) = **T.TEST(**array of the 20 data points from Sample 1; array of the 16 data points from Sample 2; 2 for a two-tailed test; 2 for equal variances) = 0.015 (same value calculated above)

Therefore, once again, ***p*-value = 0.015**, and we come to the same conclusion of rejecting the null hypothesis.

If we had the problem of **unequal variances**, we should select number '**3**' for the test type in the function T.TEST.

- **Effect size and precision of the estimate**

Now, as discussed previously, we should take our analysis a bit further and provide some information about the effect size and the precision of this estimate. Our estimate of the effect size is $3.53 - 2.70 = 0.83$ mg/L. The 95% confidence interval is calculated using the same approach we used in previous examples from this chapter (Examples 10.1 and 10.2), only now we use the Standard error based on the pooled standard deviation $\sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$:

$$\text{Lower 95\% Confidence Limit} = (\bar{X}_1 - \bar{X}_2) - t_{0.05,34} \cdot \sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

$$= (3.53 - 2.70) - 2.032 \cdot \sqrt{0.926\left(\frac{1}{20}+\frac{1}{16}\right)} = \mathbf{0.17\ mg/L}$$

$$\text{Upper 95\% Confidence Limit} = (\bar{X}_1 - \bar{X}_2) + t_{0.05,34} \cdot \sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

$$= (3.53 - 2.70) + 2.032 \cdot \sqrt{0.926\left(\frac{1}{20}+\frac{1}{16}\right)} = \mathbf{1.49\ mg/L}$$

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

So, the difference between the two samples is significant, and we can say that sample 1 is greater than sample 2 by 0.83 mg/L with a 95% confidence interval of 0.17 to 1.49 mg/L. Note that *when the 95% confidence interval of the difference between the samples does not include the value of 0, the p-value will be below the 0.05 threshold, and the results will be significant.* If our 95% confidence interval included 0, then we would find that the *p*-value is greater than 0.05.

(f) Required sample size for the *t* test to detect a desired difference between means (samples with unequal sizes)

S. 10.3.3

In Section 10.3.3.b and in Example 10.3 we showed how to estimate the **sample size** for a **one-sample *t* test**. We will cover a similar topic here, applied for the case of a **two-sample *t* test**, in order to be able to *detect a specified difference between the two means, under a certain power*. Remember the concept of **test power** and *type II error* in Section 10.2.2.

S. 10.2.2

S. 10.3.3

The estimation of the sample size is an *iterative procedure*, employing a series of successively improving estimates of the required number of data points n in each sample. The required total sample size (n_1 and n_2) can be calculated by Equation 10.14, which has the same structure of Equation 10.7, which was described in Section 10.3.3.b for a one-sample test. The modification here is that the numerator is multiplied by 2 (since we have two samples) and the standard deviation s_p reflects the within-sample variability.

$$n = n_1 = n_2 = \frac{2s_p^2}{(\mu_A - \mu_0)^2} \times (t_\alpha + t_\beta)^2 \tag{10.14}$$

This calculation will lead to the same sample size in both groups, that is $n = n_1 = n_2$. We do the iteration in a similar way we did in Example 10.3 and obtain again an equal number of sample sizes ($n = n_1 = n_2$). For a given total number of data ($n_1 + n_2$), the two-sample t test has *maximum power and robustness when* $n_1 = n_2$.

However, if $\boldsymbol{n_1 \neq n_2}$ (which is *frequently the case*), we may *initially propose the value for* n_1 that we consider to be adequate for the monitoring programme. After that, we then find the *required size of the second sample* (n_2) using Equation 10.15:

$$n_2 = \frac{n \times n_1}{2n_1 - n} \tag{10.15}$$

Example

EXAMPLE 10.5 REQUIRED SAMPLE SIZES FOR A TWO-SAMPLE T TEST TO DETECT A SPECIFIED DIFFERENCE BETWEEN MEANS

You want to know the required sample size (number of data points) for the two samples presented in Example 10.4 so that you are able to detect a significant difference between the two means of at least 0.5 mg/L (i.e., your desired effect size is 0.5 mg/L). Adopt a 0.05 significance level and a power of 80%. To make the calculations, you need the value of the within-population standard deviation. You decided to use the same value of the pooled standard deviation calculated in Example 10.4 ($s = 0.96$ mg/L).

Excel

Note: this example is also available as an Excel spreadsheet.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Solution:

Let us guess initially that a sample size of 50 is required for each sample ($n_1 = 50$ and $n_2 = 50$). Then, df $= 2\ (n - 1) = 2\ (50 - 1) = 98$.

With $\boldsymbol{\alpha = 0.05}$ (probability of 0.05) and df $= 98$, we calculate t_α as $t_{0.05;98} = 1.984$ (two-tailed test, Excel function **T.INV.2 T(probability; deg_freedom)** or T.INV.2 T(0.05; 98) $= 1.984$. Therefore, $\boldsymbol{t_\alpha = 1.984}$.

With $\boldsymbol{\beta = 0.20}$, we have a probability of $1 - 0.20 = 0.80$ (**80% power**). With a probability of 0.80 and df $= 98$, we calculate t_β as $t_{0.80;98} = 0.845$ (one-tailed test, Excel function **T.INV(probability; deg_freedom)** or T.INV(0.80; 98) $= 0.845$. Therefore, $\boldsymbol{t_\beta = 0.845}$.

Using Equation 10.14, we obtain the following result:

$$n = n_1 = n_2 = \frac{2s_p^2}{(\mu_A - \mu_0)^2} \times (t_\alpha + t_\beta)^2 = \frac{2 \times (0.96)^2}{(0.5)^2} \times (1.984 + 0.845)^2 = 59.03$$

We now use the next integer above the calculated value of n. Therefore, we adopt $n = 60$ as an estimate, and obtain the following values, calculated as above: **df** $= 2 \times (60 - 1) = \mathbf{118}$; $t_\alpha = t_{0.05;118} = \mathbf{1.980}$; and $t_\beta = t_{0.80;118} = \mathbf{0.844}$.

Using these new values, we obtain:

$$n = n_1 = n_2 = \frac{2s_p^2}{(\mu_A - \mu_0)^2} \times (t_\alpha + t_\beta)^2 = \frac{2 \times (0.96)^2}{(0.5)^2} \times (1.980 + 0.844)^2 = 58.8$$

We can now try with $n = 59$, and will see that the calculation converges in $\boldsymbol{n = n_1 = n_2 = 59}$ **as the required number of data in each sample.** If we sum up the two sample sizes, we end up with $n_1 + n_2 = 59 + 59 = \mathbf{118}$.

We can also use the **Solver tool** to obtain direct convergence, as illustrated in the Excel file associated with this example.

Let us imagine that, for some reason, you are not able to collect the required number of data points for sample 1 ($n = 59$), and are constrained by a practical situation that you can get only, say, 45 data points for sample 1 ($n_1 = 45$). Then we will have to recalculate the required sample size for n_2. Note that n_2 is not simply $118 - 45 = 73$. We lost some power by the fact that both sample sizes are not equal, and will have to recalculate n_2 by using Equation 10.15:

$$n_2 = \frac{n \times n_1}{2n_1 - n} = \frac{59 \times 45}{2 \times 45 - 59} = 85.6 = \mathbf{86}$$

In summary, we will have $\boldsymbol{n_1 = 45}$ and $\boldsymbol{n_2 = 86}$, with a total number of data equal to $45 + 86 = \mathbf{131}$. As expected, this value is larger than the one that was obtained with equal-sized samples (118), because of the need to compensate for the loss of power associated with the fact that the sample sizes are not equal. By increasing the total number of data, you are able to keep the power of 80%.

It is important to comment again (as we did in Example 10.3) that this is just an **estimate to give you an idea of your required sample sizes**, and not a final value to be implemented in your experiments. You will still need to *analyse the available funds, resources and time for your monitoring, as well as all the logistics involved in the experimental set-up*. Also, consider if 0.5 mg/L is a *meaningful* effect size for your study, or if you would also be happy with a slightly larger effect size (which would allow you to use a smaller sample size while keeping the same statistical power).

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

10.4.3 Inferences about the population medians: non-parametric Mann Whitney *U*-test (Wilcoxon–Mann–Whitney *U*-test) for two independent samples

Advanced

The Wilcoxon–Mann–Whitney *U*-test, most commonly called the Mann–Whitney *U*-test, was first developed by Wilcoxon in 1945 to compare medians from two independent samples of equal size. In 1947, Mann and Whitney generalized the test for samples of different sizes. For this test, the *actual values of the data are not used*, but rather the **order of the values** (or positions assumed when *ranking the data in order of magnitude*). The ***U*-test is a non-parametric alternative** *to the parametric t test for independent samples* (covered in Section 10.4.2).

S. 10.4.2

When the population distribution is highly skewed or heavily tailed, the **median is more appropriate** *than the mean as a representation of the centre of the population*. Furthermore, the *t* test procedures are not appropriate when applied to random samples from populations with small sample sizes.

The non-parametric statistical tests use information of a lower rank, such as nominal or ordinal observations, rather than the metric data required by conventional tests. No assumptions about the form of the parent population are required, hence the name 'non-parametric,' as discussed in Section 10.1.c.

S. 10.1

Please find below a summary of the Mann–Whitney *U*-test. The sequence of calculations will be better explained directly through an application, in Example 10.6.

Mann–Whitney *U*-test (for independent samples)

- **Description**: compares the **median** values of two samples
- **Type**: *non-parametric test*
- **Input data required**: number of data, ranking of data, specification of the desired significance level for the test
- **Output data produced**: critical values for assessment of rejection regions, *p*-value
- **Test hypotheses:**
- *Null hypothesis* H_0: $median_1 = median_2$
- *Alternative hypothesis* H_a: $median_1 \neq median_2$ (*two-tailed*); $median_1 < median_2$ (*left-tailed*) *or* $median_1 > median_2$ (*right-tailed*)
- **Test statistics:** Z_{calc} for large samples and U_{calc} for small samples (see Example 10.6)
- **Large-sample approach used**: when the larger of the two samples has at least 20 observations
- **Small-sample approach used**: when the larger of the two samples has fewer than 20 observations
- **Rejection region**: Large samples: $Z < -Z_{\alpha/2}$ or $Z > Z_{\alpha/2}$ (two-tailed); $Z < -Z_{\alpha}$ or $Z > Z_{\alpha}$ (one-tailed). Small samples: $U < U_{crit}$, with both statistics being expressed only as positive numbers
- **Rejection with *p*-value**: *p*-value $< \alpha$
- **Assumptions**: the data have been obtained independently and represent random samples from their populations; no assumptions have to be made about the shapes of the population probability distributions
- **Comments**: (i) if the sample size is small ($n < 30$) and if we cannot ascertain that the population, from which the sample was obtained, is normally distributed, non-parametric tests may be preferred. If the sample size is large, the parametric *t* test is relatively robust to small departures from normality. (ii) Tied observations are assigned ranks equal to the average of their ranks (for instance, if two observations are tied in the eight position, they will be placed between 8 and 9, and both are assigned the rank of 8.5).

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Example

EXAMPLE 10.6 COMPARISON BETWEEN THE MEDIANS OF TWO SAMPLES USING THE NON-PARAMETRIC MANN–WHITNEY U-TEST FOR INDEPENDENT SAMPLES

You monitored a certain constituent in the effluent from two treatment plants (or in two water bodies). Use non-parametric Mann–Whitney U-test for independent samples for analysing whether the medians of the two samples are significantly different from each other. Note that the samples have different numbers of data points. These data are the same as those from Example 10.4, in which the t test was applied.

Sample 1	2.8	4.2	3.9	3.3	2.8	1.7	1.9	2.5	3.1
	3.8	2.7	4.1	4.3	4.8	5.6	5.8	3.9	
	3.5	2.7	3.1						
Sample 2	2.8	3.4	4.9	2.8	2.8	1.8	2.1	2.6	2.3
	2.4	2.5	1.8	2.9	2.4	2.1	3.6		

Excel

Note: this example is also available as an Excel spreadsheet.

Solution:

Initially, we should get a visualization of our data. A suitable graph is the box-plot. Give a look at the box-plot presented in Example 10.4 – it will be the same here, since we are using the same data.

The procedure involves ranking the data and then perform the test calculations, which are divided into two possibilities: (i) large samples (in which the larger sample has at least around 20 observations) or (ii) small samples (in which the larger sample has fewer than 20 observations). The advantage of the large-sample test variant is that the test statistic approaches a normal distribution, and the Z statistic can be used.

(a) Ranking of the data

Sample 1 has 20 data points ($\boldsymbol{n_1 = 20}$) and sample 2 has 16 points ($\boldsymbol{n_2 = 16}$). Our total sample size ($n_1 + n_2$) is $\boldsymbol{N = 20 + 16 = 36}$.

The data are ranked in *ascending order,* and the ranking applies to both samples together. The lowest value in either of the two groups is given rank 1, the second lowest value is assigned rank 2, and so on, with the highest value being assigned rank N. When there is a tie (values exactly equal to each other), the rank for each tied value is the average of the ranks that would be occupied by them.

We then get R_1 and R_2, which are the sum of the ranks in samples 1 and 2, respectively.

Data		Order		Rank	
Sample 1	**Sample 2**	**Sample 1**	**Sample 2**	**Sample 1**	**Sample 2**
2.8	2.8	15th	17th	17	17
4.2	3.4	31st	24th	31	24
3.9	4.9	28th	34th	28.5	34
3.3	2.8	23rd	18th	23	17
2.8	2.8	16th	19th	17	17

(*Continued*)

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Data		Order		Rank	
Sample 1	**Sample 2**	**Sample 1**	**Sample 2**	**Sample 1**	**Sample 2**
1.7	1.8	1st	2nd	1	2.5
1.9	2.1	4th	5th	4	5.5
2.5	2.6	10th	12th	10.5	12
3.1	2.3	21st	7th	21.5	7
3.8	2.4	27th	8th	27	8.5
2.7	2.5	13th	11th	13.5	10.5
4.1	1.8	30th	3rd	30	2.5
4.3	2.9	32nd	20th	32	20
4.8	2.4	33rd	9th	33	8.5
5.6	2.1	35th	6th	35	5.5
5.8	3.6	36th	26th	36	26
3.9		29th		28.5	
3.5		25th		25	
2.7		14th		13.5	
3.1		22nd		21.5	
Sum of ranks				$R_1 = 448.5$	$R_2 = 217.5$

If the ranking is correct, $(R_1 + R_2)$ may be equal to $N\,(N+1)/2$, where $N = n_1 + n_2$.
Since $(217.5 + 448.5) = 666 = (17 \times 21/2)$, the ranking is correct.
As a matter of fact, you do not need to present your data in ascending form in the table. You can use the Excel function RANK.AVG to rank your data and apply the average criterion for the tied data:
RANK.AVG(number; ref; [order])

- **Number**. The number whose rank you want to find.
- **Ref**. Array of all values in the samples.
- **Order**. A number specifying how to rank number (0 or omitted: descending order; any non-zero value: ascending order).

Our test hypotheses and test conditions are:

For performing the test, we specify:

- Null hypotheses $\mathbf{H_0}$: $\text{median}_1 = \text{median}_2$
- Alternative hypothesis $\mathbf{H_a}$: $\text{median}_1 \neq \text{median}_2$
  - Significance level for the test $(\alpha) = 0.05$ (confidence level of 0.95 or 95%)

(b) Mann–Whitney test for large samples

When there are at least 20 observations in the larger sample ($n > 20$), the U statistic is assumed to follow a normal distribution. In this example, our sample size is just borderline for this criterion, and we will do the calculations for large samples and for small samples, in order to show you the steps involved. Now, we will use the procedure for large samples, knowing that it can be done on an automated basis, without the need for lookup tables for the critical values of the test.

We check which one is our smaller sample. In this case, it is sample 2, which has only 16 observations. For this sample, we get the sum of ranks, which was found to be 217.5 (see table above).

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

The mean of the ranks in the smaller sample is given by

$$\text{Mean of ranks in smaller sample} = \frac{n_{\text{smaller}} \times (N+1)}{2} = \frac{16 \times (36+1)}{2} = 296$$

The standard deviation of the ranks in the smaller sample is

$$\text{Standard deviation of ranks in smaller sample} = \sqrt{\frac{n_1 \times n_2 \times (N+1)}{12}}$$

$$= \sqrt{\frac{20 \times 16 \times (36+1)}{12}} = 31.4$$

For large samples, the U distribution follows a normal distribution, and we can use the Z statistics. The Z statistics can be calculated as follows:

$$Z = \frac{R_{n\,\text{smaller}} - \text{Mean ranks}_{n\,\text{smaller}}}{\text{Standard deviation ranks}_{n\,\text{smaller}}} = \frac{217.5 - 296}{31.4} = \mathbf{-2.499}$$

The critical Z is the one traditionally obtained using Excel function NORM.S.INV(probability). In our case, since we have a two-tailed test, the probability is $\alpha/2 = 0.05/2 = 0.025$. The result of this is $-\mathbf{1.960}$. See Section 10.2.4 for a discussion on the determination of the rejection regions for the normal distribution.

S. 10.2.4

Since $\mathbf{Z} < \mathbf{Z_{crit}}$ ($-2.499 < -1.960$), we fall in the rejection region, and therefore **reject the null hypothesis** that the medians of the two samples are equal.

Another way of analysing the test hypothesis is by finding the ***p*-value**, as done in the other examples. The *p*-value can be obtained using the Excel function **NORM.S.DIST**(ABS(Z); TRUE for cumulative).

For a **two-tailed test** we use the syntax shown below. Notice that we use the absolute value (ABS) of the Z statistics obtained above.

p-value = 2 × (1 − NORM.S.DIST(ABS(Z); TRUE) = 2 × (1 − NORM.S.DIST(ABS(−2.499); TRUE) = **0.0125**

Since this *p*-value is lower than the significance level (*p*-value $< \alpha$, or $0.0125 < 0.05$), we **reject our null hypotheses** that the medians of both samples are equal. Note that in Example 10.4, in which we used the same data and applied the parametric *t* test, we obtained *p*-value = 0.0152, which also indicated rejection of the null hypotheses (in that example, the comparisons were between the means of the two samples, and here we compare the medians of both samples).

If we had one-tailed tests (which is not the case here), we would have:

Left-tailed: *p*-value = NORM.S.DIST(ABS(Z))

Right-tailed: *p*-value = (1-NORM.S.DIST(ABS(Z))

(c) Mann–Whitney test for small samples

As we saw before, if you have small sample sizes (in which the larger sample has less than 20 observations), we need to use the procedure for small sample sizes in the Mann–Whitney *U*-test. But please note that the test will have little power if the samples are really small, and, if so, it may come to the point of not being worthwhile to do a hypothesis test.

The procedure is as follows.

Based on the number of observations in each sample (n_1 and n_2) and the sum of the ranks calculated before (R_1 and R_2), we calculate U and U':

$$U = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1 = 20 \times 16 + \frac{20 \times (20+1)}{2} - 448.5 = 81.5$$

$$U' = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2 = 20 \times 16 + \frac{16 \times (16+1)}{2} - 217.5 = 238.5$$

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

The lower of these two values is called U_{calc}. In our case, $\boldsymbol{U_{calc} = 81.5}$.

The value of the **critical *U*** statistic can be obtained from the following **look-up table**, for this case of small samples.

Table of critical values of the Mann–Whitney *U* distribution, with $\alpha = 0.05$ for two-sided tests and $\alpha = 0.025$ for one-sided tests.

n2 → n1↓	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
1												
2	0	0	0	1	1	1	1	1	2	2	2	2
3	2	3	3	4	4	5	5	6	6	7	7	8
4	4	5	6	7	8	9	10	11	11	12	13	13
5	7	8	9	11	12	13	14	15	17	18	19	20
6	10	11	13	14	16	17	19	21	22	24	25	27
7	12	14	16	18	20	22	24	26	28	30	32	34
8	15	17	19	22	24	26	29	31	34	36	38	41
9	17	20	23	26	28	31	34	37	39	42	45	48
10	20	23	26	29	33	36	39	42	45	48	52	55
11	23	26	30	33	37	40	44	47	51	55	58	62
12	26	29	33	37	41	45	49	53	57	61	65	69
13	28	33	37	41	45	50	54	59	63	67	72	76
14	31	36	40	45	50	55	59	64	67	74	78	83
15	34	39	44	49	54	59	64	70	75	80	85	90
16	37	42	47	53	59	64	70	75	81	86	92	98
17	39	45	51	57	63	67	75	81	87	93	99	105
18	42	48	55	61	67	74	80	86	93	99	106	112
19	45	52	58	65	72	78	85	92	99	106	113	119
20	48	55	62	69	76	83	90	98	105	112	119	127

For a two-tailed test with $\alpha = 0.05$, $n_1 = 20$, and $n_2 = 16$, we get, from the look-up table: $\boldsymbol{U_{crit} = U_{\alpha;n1;n2} = 98}$.

Unlike other tests, with this test, we reject H_0 if U_{calc} is less than or equal to the critical value (U_{crit}), with both statistics being expressed only as positive numbers:

$U_{calc} = 81.5 < U_{crit} = 98$

Since $U_{calc} < U_{crit}$ **we reject H_0**. In other words, we support the alternative hypothesis H_a, concluding that the medians of both populations are significantly different.

Note that for this particular example, all approaches (parametric, non-parametric, small-sample, large-sample) led to the same conclusion. However, you should note that this may not always be the case.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

10.4.4 Inferences about the population means: parametric *t* test for two dependent samples (paired data)

Advanced

Paired (dependent) samples can be found in several research studies. In Figure 10.1, we showed some typical situations in which you could consider that there is a degree of dependency between the samples. For example, if you would like to see whether the addition of a certain chemical product improves the performance of a treatment plant, you may run two pilot units in parallel, one with the addition of the chemical product and the other without addition (you call this one a *control* unit). Both units operate at nearly the same conditions, and the only difference is the addition of the chemical product. Since the influent to both units is the same, you have only one sampling point. For the effluent from both units, your monitoring programme specifies the collection of samples at approximately the same time, and because of this you consider that both data sets are a matched pair. Another example could be if you want to study the impact of a certain point-source pollution in a river. You collect samples upstream and downstream of the discharge point in the river at approximately the same time and, again, you could consider that the upstream and downstream samples are matched, that is, they are dependent.

An experimental design using paired samples is almost always better than one based on independent samples. The *pairing technique increases the efficiency of the statistical test*, making it more sensitive to small differences between treatments. Observations are collected in pairs, so that the two elements of each pair are homogeneous in all directions, except for the factor to be compared. The groups are organized so that the intervening variables have the same frequency in the two groups.

Similar to what we discussed in the previous sections, here we can also apply parametric and non-parametric tests:

S. 10.4.4

- **Parametric**: *t test for matched pairs* (testing for equality of **means**) – this Section 10.4.4

S. 10.4.5

- **Non-parametric**: *Wilcoxon signed-rank test for matched pairs* (testing for equality of **medians**) – Section 10.4.5

S. 10.1

Let us recall here what we mentioned in Section 10.1, that *establishing whether the groups are dependent or independent is not trivial* and, in many cases, may be misleading. Some researchers argue that, in our field of *environmental statistics, it is very difficult to assume that we have truly dependent data sets*, even if measurements are made at the same time. Other environmental factors for which we cannot account may cause our datasets to lose their degree of dependence. *If you are in doubt, use tests for independent data sets* (parametric *t* test – Section 10.4.2 – or non-parametric Mann–Whitney *U*-test – Section 10.4.3), even if they lose some power compared with the tests for dependent sets.

S. 10.4.2

S. 10.4.3

In paired hypothesis tests, both samples need to have the **same number of data**, organized in *pairs*. If you are conducting a paired experimental design, and on one day, you lose a sample from one of the pairs, you have to throw the other data from the pair out. The method only works for paired samples if you have both pairs for each time you sampled.

We will now cover the *t* test for dependent samples. In our case here, there are samples of pairs (X_1, Y_1; X_2, Y_2, … , X_n, Y_n), and for each pair we can calculate the difference between their values $D = X - Y$. These differences comprise a new single variable (D_1, D_2, … , D_n). We can use these differences with the same approach we used previously for the one sample *t* test described in Section 10.3.2. Go to that section to remember the fundamentals of this test. We could expect that the mean of the differences would be zero if the samples are equal.

S. 10.3.2

Typically, we would state the test hypothesis as follows:

- **Null hypothesis H_0**: $\mu_1 = \mu_2$ *or* H_0: $\mu_D = 0$
- **Alternative hypothesis H_a**: $\mu_1 \neq \mu_2$ *or* H_a: $\mu_D \neq 0$

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

The test statistic is:

$$t_{calc} = \frac{\bar{x}_D - \mu_D}{\frac{s_D}{\sqrt{n}}} \tag{10.16}$$

where

$\bar{x}_D$ = mean value of the differences between X_i and Y_i
μ_D = mean value specified in the null hypothesis for D (or $\mu_1 = \mu_2$)
s_D = standard deviation of the differences between X_i and Y_i
n = number of data points in the statistical sample of the differences

In the case we want to analyse the **equality of means**, we make μ_D equal to zero in Equation 10.16, and obtain:

$$\boxed{t_{calc} = \frac{\bar{x}_D}{\frac{s_D}{\sqrt{n}}}} \tag{10.17}$$

The **critical value of *t*** (t_{crit}) can be calculated using the following **Excel** functions, applied to the sample with the differences:

T.INV.2T(probability; deg_freedom) for a two-sample test;
T.INV(probability; deg_freedom) for a left-tailed test.

You can perform the *t* test for matched pairs directly, without needing to calculate the sample with the differences. For this, you can use the following **Excel function** for returning the ***p*-value**:

T.**TEST**(array1; array2; tail; type)

where

- **Array1**. The first data set.
- **Array2**. The second data set.
- **Tails**. Specifies the number of distribution tails. If tails = 1, T.TEST uses the one-tailed distribution. If tails = 2, T.TEST uses the two-tailed distribution.
- **Type**. The kind of *t* test to perform: 1 (for paired *t* test)

You can find below a summary of the matched-pairs *t* test. The application can be found in Example 10.7.

***t* test for matched pairs (sample with the differences between the two original matched samples)**

- **Description**: compares the **mean** value of the sample of differences with a specified reference value μ_D (usually zero)
- **Type**: *parametric* test
- **Input data required**: number of data points in the sample with differences (n), mean ($\bar{d}$) and standard deviation (s_D) of the sample with differences, value of the reference value we want to use for the comparison (usually zero), plus the specification of the desired significance level for the test (α)
- **Output data produced**: t statistic, *p*-value
- **Test hypotheses:**
- *Null hypothesis* H_0: $\mu = \mu_D$ (*usually* $\mu_D = 0$)

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

- *Alternative hypothesis* H_a: $\mu \neq \mu_D$ (*two-tailed*); $\mu < \mu_D$ (*left-tailed*) or $\mu > \mu_D$ (*right-tailed*) – *usually* $\mu \neq 0$ (*two-tailed*); $\mu < 0$ (*left-tailed*) or $\mu > 0$ (*right-tailed*)
- **Test statistic**: $t = \dfrac{\bar{x}_D - \mu_D}{s_D/\sqrt{n}}$
- **Rejection region**: $t < t_{\alpha/2}$ or $t > t_{\alpha/2}$ (two-tailed); $t < -t_\alpha$ or $t > t_\alpha$ (one-tailed)
- **Rejection with *p*-value**: *p*-value $< \alpha$
- **Assumptions**: the relative frequency distribution of the population of differences is approximately *normally distributed*; the paired differences are randomly selected from the population of differences

Example

EXAMPLE 10.7 COMPARISON BETWEEN THE MEANS OF TWO SAMPLES USING THE T TEST FOR DEPENDENT SAMPLES (MATCHED PAIRS)

You monitored a certain constituent upstream and downstream of a suspected point-source pollution in a river. The upstream and the downstream samples were collected approximately at the same time, and since all other conditions were the same (apart from the discharge), you concluded that the samples could be considered dependent. Use the t test for matched pairs for analysing the following question: with a significance level $\alpha = 0.05$, can it be said that samples 1 and 2 have equal means? Or are the means significantly different?

Matched pair no.	***1***	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***	***11***	***12***	***13***	***14***	***15***	***16***	***17***	***18***
Sample 1	2.8	4.2	3.9	3.3	2.8	1.7	1.9	2.5	3.1	3.8	2.7	4.1	4.3	4.8	5.6	5.8	3.9	3.5
Sample 2	2.7	3.1	2.8	3.4	4.9	2.8	2.8	1.8	2.1	2.6	2.3	2.4	2.5	1.8	2.9	2.4	2.1	3.6

Excel

Note: this example is also available as an Excel spreadsheet.

Solution:

For the computations, we create a new variable composed of the paired differences between the values in samples 1 and 2:

Matched pair no.	***1***	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***	***11***	***12***	***13***	***14***	***15***	***16***	***17***	***18***
Sample 1	2.8	4.2	3.9	3.3	2.8	1.7	1.9	2.5	3.1	3.8	2.7	4.1	4.3	4.8	5.6	5.8	3.9	3.5
Sample 2	2.7	3.1	2.8	3.4	4.9	2.8	2.8	1.8	2.1	2.6	2.3	2.4	2.5	1.8	2.9	2.4	2.1	3.6
Difference	0.1	1.1	1.1	−0.1	−2.1	−1.1	−0.9	0.7	1.0	1.2	0.4	1.7	1.8	3.0	2.7	3.4	1.8	−0.1

We then formulate our test hypotheses:

H_0: $\mu_1 = \mu_2$ or H_0: $\mu_D = 0$
H_a: $\mu_1 \neq \mu_2$ or H_a: $\mu_D \neq 0$

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

The descriptive statistics of the two samples and the sample with the differences are:

Statistics	Sample 1	Sample 2	Sample with Differences (sample 1–Sample 2)
Number of data	18	18	18
Mean	3.59	2.72	$\bar{X}_D = 0.87$
Standard deviation	1.13	0.73	$s_D = 1.44$

Degrees of freedom: the sample with the differences has 18 values. Therefore, $df = n - 1 = 18 - 1 = 17$.

The test statistics (t_{calc}) using the sample with the differences is given by Equation 10.17:

$$t_{calc} = \frac{\bar{x}_D}{\frac{s_D}{\sqrt{n}}} = \frac{0.87}{\frac{1.44}{\sqrt{18}}} = \mathbf{2.561}$$

For $\alpha = 0.05$, the **critical value of the distribution *t* (t_{crit})**, with $df = 17$ is **2.110**, that is, $t_{0.05;17} =$ **2.110**, obtained by Excel function T.INV.2T(probability; deg_freedom) or T.INV.2T(0.05;17).

Decision: $t_{calc} = 2.561 > t_{0,05;17} = 2.110 \Rightarrow$ we reject H_0 and we conclude that the means between the samples are significantly different, for $\alpha = 0.05$.

An alternative way of doing the analysis is by obtaining the ***p*-value**. The *p*-value may be calculated using the Excel function **TDIST**(*x*; deg_freedom; tails). Therefore, TDIST(ABS(2.561);18 − 1;2) = **0.0202**.

Since *p*-value (0.0202) < significance level (0.05), we **reject the null hypothesis that the mean of the differences is equal to zero (**or, in other words, we **reject the hypothesis that the means of both samples are equal)**.

Still another way of doing this whole analysis, without the need for computing the differences, is to use the Excel function **T.TEST**(array1;array2;tails;type), where

- **Array1**: first data set.
- **Array2**: second data set.
- **Tails**: 2 for a two-tailed test.
- **Type**: 1 for a paired *t* test

We then obtain ***p*-value = 0.0202**, which is the same value as the one calculated above, and we come, again, to the same conclusion.

10.4.5 Inferences about the population medians: non-parametric Wilcoxon signed-rank test for two dependent samples (matched pairs)

Advanced

The **Wilcoxon *T* test** is a non-parametric alternative to the Student's *t* test for paired samples when data are not normally distributed. It was developed by Wilcoxon in 1945 and is based on the positions of the differences between the pairs of data. It considers that when two dependent variables have the same median, the positive and negative differences should be zero.

We start with a similar concept that we used for the *t* test, that is, we produce a new sample which is composed of the differences between each of the pairs of the two original samples. We then use the

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

S. 9.3.6

one-sample **Wilcoxon signed-rank test** to analyse the new sample made up of the differences. This test was previously described in Section 9.3.6, and the main elements are presented here.

Typically, we state the test hypotheses as follows:

- **Null hypothesis H_0**: $median_1 = median_2$ or H_0: $median_D = 0$
- **Alternative hypothesis H_a**: $median_1 \neq median_2$ or H_a: $median_D \neq 0$

Basically, the test involves finding the differences between paired values (let us call this the 'sample of differences'), and then calculating the **difference between each value of this 'sample of differences' and M_0** (in our case, since we want to test whether the two sample medians are equal, **M_0 is adopted as equal to zero**). Some differences will be positive (when the value is greater than zero) and others will be negative (when the value is lower than zero). All differences are ranked, and the *sum of the ranks of the positive differences* (**R^+**) and the *sum of the ranks of the negative differences* (**R^-**) are calculated. The *smallest of the two values* (R^+ *and* R^-) is called **R** (or R_{calc}) and is used for the calculation of the test statistic.

If the **sample is small**, we need to use a look-up table (Table 10.5) to consult the critical values of the R statistic of the Wilcoxon test. We compare the value of R_{calc} with R_{crit}. **If $R_{calc} < R_{crit}$ we reject the null hypothesis H_0.**

However, if the sample is relatively large (**$n \geq 20$**), the distribution of R is approximately normal and we can use the **Z statistic**, according to the following equation (Hines *et al.*, 2003):

$$Z_{calc} = \frac{R - n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}} \tag{10.18}$$

where

R = smallest value between R^+ (sum of the ranks of the positive differences) and R^- (sum of the ranks of the negative differences)
n = number of data of your sample

You then compare this value of **Z_{calc}** with the critical value of Z (**Z_{crit}**). The rejection regions are those already shown in Table 10.3 (with a discussion on their interpretation). From the table, we see that, for two-tailed tests at the 5% significance level, the rejection region is

- For $\alpha = 0.05$, reject null hypothesis H_0 if $Z_{calc} < -1.960$ or $Z_{calc} > 1.960$.

Additionally, from the value for Z_{crit}, you can then calculate the ***p*-value** using the Excel function NORM.S.DIST:

- Null hypothesis H_0: **Sample median $= M_0$** (*two-tailed*):

  p-value $= 2 \times (1 -$ NORM.S.DIST(ABS(Z_0); TRUE))

- Null hypothesis H_0: **Sample median $\geq M_0$** (*left-tailed*):

  p-value $=$ NORM.S.DIST (ABS(Z_0); TRUE)

- Null hypothesis H_0: **Sample median $\leq M_0$** (*right-tailed*):

  p-value $= 1 -$ NORM.S.DIST (ABS(Z_0); TRUE)

You **reject the null hypothesis H_0 if *p*-value $< \alpha$**.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Table 10.5 Critical values of R_0 (R_{crit}) in the Wilcoxon matched-pairs signed-ranks test.

n	α (two-tailed) = 0.10 α (one-tailed) = 0.05	α (two-tailed) = 0.05 α (one-tailed) = 0.025	α (two-tailed) = 0.01 α (one-tailed) = 0.005
4			
5	0		
6	2	0	
7	3	2	
8	5	3	0
9	8	5	1
10	10	8	3
11	13	10	5
12	17	13	7
13	21	17	9
14	25	21	12
15	30	25	15
16	35	29	19
17	41	34	23
18	47	40	27
19	53	46	32
20	60	52	37
21	67	58	42
22	75	65	48
23	83	73	54
24	91	81	61
25	100	89	68
26	110	98	75
27	119	107	83
28	130	116	91
29	140	126	100
30	151	137	109

A summary of the Wilcoxon signed-rank test is presented below (extracted from comments by Hines *et al.*, 2003):

Wilcoxon Signed-Rank Test for matched pairs using a normal approximation for the test statistic (for a large sample, $n \geq 20$)

- **Description**: compares the sum of the ranks of the positive differences (R^+) and the sum of the ranks of the negative differences (R^-), where the differences are between the values of the sample and a specified value (in our case, this value is zero)
- **Type**: *non-parametric test*
- **Input data required**: data from your sample, which will be further processed to calculate the ranks, plus the specification of the desired significance level for the test

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

- **Output data produced**: *p*-value
- **Test hypotheses:**
  - *Null hypothesis* H_0: $median_1 = median_2$ *or* H_0: $median_D = 0$
  - *Alternative hypothesis* H_a: $median_1 \neq median_2$ *or* H_a: $median_D \neq 0$
- **Test statistics**: $Z_{calc} = \left(\dfrac{(R - n(n+1)/4)}{\left(\sqrt{n(n+1)(2n+1)/24}\right)} \right)$
- **Rejection region**: $Z < Z_{\alpha/2}$ or $Z > Z_{\alpha/2}$ (two-tailed); $Z < -Z_{\alpha}$ or $Z > Z_{\alpha}$ (one-tailed)
- **Rejection with *p*-value**: *p*-value $< \alpha$
- **Assumptions**: No assumptions have to be made about the shape of the probability distribution
- **Comments:** The version of this test that uses the normal approximation for the test statistic is to be used for large samples ($n \geq 20$). If there are less than 20 pairs of data, you should use the version with the look-up table, described in this section.

Example

EXAMPLE 10.8 COMPARISON BETWEEN THE MEDIANS OF TWO SAMPLES USING THE WILCOXON TEST FOR DEPENDENT SAMPLES (MATCHED PAIRS)

The problem here is the same as the one in Example 10.7*, with the difference that in Example 10.7 we tested for the equality of* ***means****, and in this Example 10.8 we test for the equality of* ***medians****.*

You monitored a certain constituent upstream and downstream of a point-source pollution in a river. The upstream and the downstream samples were collected approximately at the same time, and since all other conditions were the same (apart from the discharge), you concluded that the samples could be considered dependent. Use the non-parametric Wilcoxon test for matched pairs for analysing the following question: with a significance level $\alpha = 5\%$*, can it be said that samples 1 and 2 have equal medians?*

Matched pair no.	***1***	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***	***11***	***12***	***13***	***14***	***15***	***16***	***17***	***18***
Sample 1	2.8	4.2	3.9	3.3	2.8	1.7	1.9	2.5	3.1	3.8	2.7	4.1	4.3	4.8	5.6	5.8	3.9	3.5
Sample 2	2.7	3.1	2.8	3.4	4.9	2.8	2.8	1.8	2.1	2.6	2.3	2.4	2.5	1.8	2.9	2.4	2.1	3.6

Excel

Note: this example is also available as an Excel spreadsheet.

Solution:

Procedure for small samples **(*n* < *20*)**
Test hypotheses:

- H_0: there are no differences between the medians
- H_a: the medians are different

The sequence of calculation is listed below, and further exemplified in the table to follow.

(a) The difference (d) in each pair of data is calculated, keeping the difference sign ($+$ or $-$).
(b) Discard the differences equal to zero and use the number of differences remaining. In the example, there are no differences equal to zero, so $n = 18$.
(c) Differences are ranked by value, from the smallest to the largest, ignoring the sign. In the case of ties, the average of the tie positions is used. If the ordering is correct, the sum of the absolute

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

values of the two samples must be equal to $n\ (n+1)/2$. In the example, $n = 18$ and Σ ranks $= 171 = 18 \times 19/2$.

(d) Transfer the difference sign to the corresponding sample rank.

(e) $R\,(+)$ will be the sum of the ranks with a positive sign and $R\,(-)$ will be the sum of the ranks with a negative sign. The lowest absolute value of these sums will be the $R_{\text{calc.}}$ As in the example, $R(+) = 137$ and $R(-) = -34$, $R_{\text{calc.}} = |-34| = 34$. Therefore, $\boldsymbol{R_{\text{calc}} = 34}$.

(f) $R_{\alpha,n}$ is the critical value for the test, obtained from Table 10.5. In this example, $\boldsymbol{R_{\text{crit}} = R_{0.05,18} = 40}$ (two-sided).

(g) **Test decision:** $\boldsymbol{R_{\text{calc}} < R_{\text{crit}}}$ $(34 < 40)$. Therefore, we reject the null hypothesis that the medians from both samples are equal.

Computational table

n	**Sample 1**	**Sample 2**	**Differ. Samp 1 – Samp 2 (*d*)**	**Absolute *d* |*d*|**	**Rank absolute *d***	**Rank (with sign)**	**Sum *R* (+)**	**Sum *R* (−)**
1	2.8	2.7	0.1	0.1	1	1	1	
2	4.2	3.1	1.1	1.1	9.5	9.5	9.5	
3	3.9	2.8	1.1	1.1	9.5	9.5	9.5	
4	3.3	3.4	−0.1	0.1	2.5	−2.5		2.5
5	2.8	4.9	−2.1	2.1	15	−15		15
6	1.7	2.8	−1.1	1.1	8	−8		8
7	1.9	2.8	−0.9	0.9	6	−6		6
8	2.5	1.8	0.7	0.7	5	5	5	
9	3.1	2.1	1.0	1	7	7	7	
10	3.8	2.6	1.2	1.2	11	11	11	
11	2.7	2.3	0.4	0.4	4	4	4	
12	4.1	2.4	1.7	1.7	12	12	12	
13	4.3	2.5	1.8	1.8	13.5	13.5	13.5	
14	4.8	1.8	3.0	3	17	17	17	
15	5.6	2.9	2.7	2.7	16	16	16	
16	5.8	2.4	3.4	3.4	18	18	18	
17	3.9	2.1	1.8	1.8	13.5	13.5	13.5	
18	3.5	3.6	−0.1	0.1	2.5	−2.5		2.5
$\sum$ R				**171**			**137**	**−34**

***Procedure for large samples* ($n \geq 20$)**

In our case here, $n = 18$, which is lower than the suggested minimum value of 20 for characterizing a large sample. However, we will carry out the calculations for you to see how the procedure is used.

If the sample is relatively large ($\boldsymbol{n \geq 20}$), the distribution of R is approximately normal and we can use the Z statistic, according to the following equation. The inputs for the equation are the R_{calc} value

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

(calculated above as 34) and the number of data points ($n = 18$).

$$Z_{calc} = \frac{R_{calc} - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} = \frac{34 - \frac{18(18+1)}{4}}{\sqrt{\frac{18(18+1)(36+1)}{24}}} = \frac{34 - 85.5}{\sqrt{541.5}} = \mathbf{-2.243}$$

From Table 10.3, we see that, for $\alpha = 0.05$ and a two-tailed test, $\boldsymbol{Z_{crit} = -1.960}$.

Test decision: $\boldsymbol{Z_{calc} < Z_{crit}}$ ($-2.243 < -1.960$). Therefore, we **reject the null hypothesis** that the medians from both samples are equal.

We can also obtain the test decision using the ***p*-value** for large sample tests. The *p*-value can be obtained using the Excel function NORM.S.DIST(Z; TRUE for cumulative). In our case, we have:

p-value $= 2 \times [1 -$ NORM.S.DIST(ABS(-2.243))$] = 0.0249$

Since ***p*-value** $<$ **significance level** α ($0.0249 < 0.05$) we reject the null hypothesis H_0 that the sample medians are equal.

10.5 COMPARING THE CENTRAL VALUES OF MORE THAN TWO SAMPLES

10.5.1 Types of multiple-sample tests covered in this chapter

Advanced

S. 10.3

S. 10.5.2

S. 10.5.3

S. 10.5.4

S. 10.5.5

In Section 10.3, we dealt with one-sample tests. In Section 10.4, we covered two-sample tests. In this section, we will cover hypothesis tests used to make comparisons between more than two samples simultaneously. The tests we will describe are all for independent samples and, as we did in the other sections, we will present parametric and non-parametric alternatives (see Figure 10.12):

- **Parametric test to compare central values from more than two samples:** *Analysis of Variance – ANOVA* (Section 10.5.2) followed by the *post hoc Tukey test* for multiple comparisons (Section 10.5.3)
- **Non-parametric test to compare central values from more than two samples:** *Kruskal–Wallis test* (Section 10.5.4) followed by the *post hoc Dunn test* for multiple comparisons (Section 10.5.5)

10.5.2 Parametric test for more than two population central values. ANOVA

Advanced

ANOVA is a *parametric* statistical technique developed by R. A. Fisher. The procedure consists of the decomposition of the total variation between the values obtained in an experiment for several identifiable components.

ANOVA is an extension of Student's t test, but for comparisons between more than two data sets. It determines whether all data groups have the same mean values or if at least one of them is different from the others. This is done by comparing estimates of the overall variance of the data set, specifically analysing the variation of the data **between the groups** compared to the variation of the data **within the groups**.

For instance, if four groups (four samples) are being compared, would it be correct to perform multiple t tests between the groups, to compare them in a stepwise manner, two samples at a time? Let us analyse this.

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$$

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

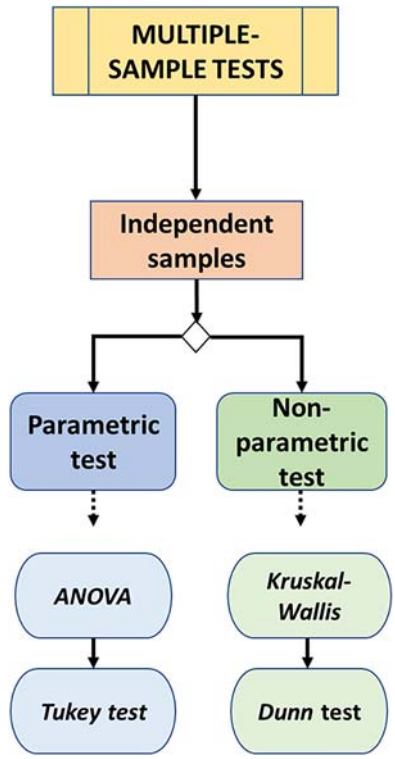


Figure 10.12 Multiple-sample tests covered in this chapter.

Since we have a combination of 4 elements, 2 by 2, we have 6 possibilities:

$$C = \frac{n!}{x!(n-x)!} = \frac{4!}{2!(4-2)!} = 6$$

Therefore, it would be necessary to check:

• $\bar{x}_1$ versus $\bar{x}_2$	• $\bar{x}_2$ versus $\bar{x}_3$
• $\bar{x}_1$ versus $\bar{x}_3$	• $\bar{x}_2$ versus $\bar{x}_4$
• $\bar{x}_1$ versus $\bar{x}_4$	• $\bar{x}_3$ versus $\bar{x}_4$

Pearson (1942) proved that the probability of incurring a Type I error (erroneously concluding that there is a significant difference that does not exist) increases with the number of means being compared. Table 10.6 shows how the Type I error increases depending on the level of significance used in the test and the number of means being compared.

As it can be seen from the table, for $\alpha = 0.05$, the probability is 5% if the comparison is between only two samples, but it becomes 14% if it is between three samples and 26% if it is between four samples.

To avoid this compounded potential for error, the correct procedure is to use a **single test to compare all means** in one step, and to *identify the existence of at least one difference between groups, if any exist*. Then,

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Table 10.6 Probability of making at least one type I error by using two-sample *t* tests for all *k* pairwise comparisons of *k* means.

Number of Means (*k*)	Level of Significance (α) Used in the *t* Tests		
	0.05	**0.01**	**0.001**
2	0.05	0.01	0.001
3	0.14	0.03	0.003
4	0.26	0.06	0.006
5	0.40	0.10	0.010
6	0.54	0.14	0.015
10	0.90	0.36	0.044

Adapted from Zar (1999).

one of several existing techniques of **multiple comparisons** between any two samples may be applied later, using something we call a *post hoc* test.

When we use ANOVA, we subdivide the total variation into (i) differences between the groups, which are attributed to the treatment effects and (ii) differences within the groups, which are attributed to chance (or inherent group variations) due to simple random experimental error. Thus, the total variation in the data is subdivided into two fractions:

- **Between-groups variation**. Variation between the means of the various groups (samples) when compared to the general average of all data points (effect of different treatments).
- **Within-groups variation**. Variation inside each group (each sample) relative to the mean of that group (individual or random differences in responses).

The total variation is equal to the sum of the between-groups variation and the within-groups variation.

When using ANOVA, we decompose the total variation between the values into several identifiable elements, and each component assigns the variation to a different cause or source of variation. The number of causes of variation or 'factors' depends on the design of your experiment. For instance, we have one-way (single factor) ANOVA and two-way (two factor) ANOVA. In our book, we will cover only the **Single-Factor Analysis of Variance (one-way ANOVA)**, which is described below. You should consult additional references to learn about other types of analyses that involve more than one experimental factor.

Excel includes add-ins that you can obtain from: Tools > Excel Add-ins > Analysis ToolPak > ANOVA. The ANOVA analysis tool has the following options:

- ANOVA: *Single Factor*
- ANOVA: *Two-Factor with Replication*
- ANOVA: *Two-Factor without Replication*

The Excel analysis tools are useful because they perform the calculations as a statistical software would. However, they are not dynamic like the Excel functions we have been using so far. Therefore, if you change anything in your input data, you will need to run the add-in function again. This may not be a problem for you, considering that you have already spent so much time obtaining and organizing your monitoring data. However, another limitation of the Excel add-ins is that the individual calculations are not shown. Therefore, it is somewhat of a 'black-box' tool that provides you with the answer, but does not show you how the calculations are performed.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

With **Single-Factor Analysis of Variance** (**one-way ANOVA**), several groups are compared relative to a single factor of interest. The comparison test between group effects assumes that k groups A, B, … k can generate different means, but the variance (s^2) between individual samples within a group is the same for all populations being compared. We follow the **assumption** that the groups or levels of the factor under study represent populations whose outcomes are randomly and independently drawn, follow a normal distribution, and have equivalent variances ($\sigma_A^2 = \sigma_B^2 = \cdots = \sigma_k^2 = \sigma^2$).

For instance, if five groups are compared and the assumptions of normality and homoscedasticity (equal variances) are valid, we will have the graph shown in Figure 10.13 (top), in case the means of the five populations are equal. Our null hypothesis H_0 is then: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$. However, if H_0 is false and we have $\mu_4 > \mu_1 > \mu_2$, but $\mu_2 = \mu_3 = \mu_5$, then the populations will be equal in shape, but displaced among themselves (bottom graph).

In order to perform ANOVA, the total variation is divided into two fractions, one attributed to the differences between groups and another due to inherent variations within the groups. The **total variance** is usually represented by the **Total Sum of Squares or SS_{total}.**

The test hypotheses are:

- **Null hypothesis H_0**: *all population means are equal* ($\mu_1 = \mu_2 = \ldots \mu_k$).
- **Alternative hypothesis H_a**: *at least one of the population means is different from the others.*

Note that, in the alternative hypothesis, we state that at least one of the population means is different from the others. However, when we get the initial result of the ANOVA test, even if the result is significant, we do not know which of the groups or how many of the groups are different. To arrive at this conclusion, we need to undertake *post hoc* multiple comparison tests, such as the Tukey test (described in Section 10.5.3).

S. 10.5.3

Under the null hypothesis, which assumes that the arithmetic means of the groups are equal, a measure of total variation can be obtained between all observations by adding the squared differences between each observation and the **grand mean** ($\bar{x}$), which is based on *all observations in all groups*, combined.

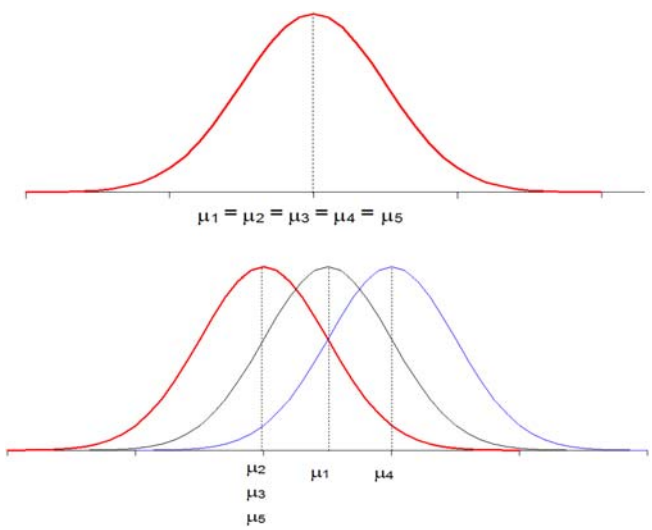


Figure 10.13 Five normal distributions with equal variance. Top: equal means. Bottom: μ_1 and μ_4 have unequal means, but μ_2, μ_3, and μ_5 have equal means.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

The **total variation** ($\mathbf{SS_{total}}$) can be obtained by

$$\mathbf{SS_{total}} = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (X_{ij} - \bar{X})^2 \tag{10.19}$$

where

$$\bar{X} = \sum_{j=1}^{k} \sum_{i=1}^{n_j} \frac{X_{ij}}{n} = \text{grand mean} \tag{10.20}$$

X_{ij} = ith observation in the group or level j
n_j = number of observations in group j
n = total number of observations in all groups ($n = n_1 + n_2 + \cdots + n_k$)
k = number of groups or levels being compared
grand mean = mean of all observations combined

The **between-groups variation**, generally called the **sum of squares between groups** ($SS_{between}$), is measured by summing the squared differences between the arithmetic mean of the sample of each group $\bar{x}_j$ and the grand mean $\bar{x}$, weighted by the size of the sample n_j in each group. The variation between groups is calculated using the following equation:

$$\mathbf{SS_{between}} = \sum_{j=1}^{k} n_j (\bar{X}_j - \bar{X})^2 \tag{10.21}$$

where

k = number of groups or levels being compared
n_j = number of observations in group j
$\bar{x}_j$ = arithmetic mean of the sample or group j
$\bar{x}$ = grand mean of the observations

The **within-group variation**, commonly called the **sum of squares within groups** (SS_{within}), measures the difference between each observation and the arithmetic mean of its own group, and accumulates the squares of these differences over all groups. The variation within the group can be calculated as follows:

$$\mathbf{SS_{within}} = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (X_{ij} - \bar{X}_j)^2 \tag{10.22}$$

where

X_{ij} = ith observation in the group or level j
$\bar{x}_j$ = arithmetic mean of the sample or group j

Because k levels are being compared, there are $k - 1$ degrees of freedom, associated with the sum of squares between groups. Since each of the levels k contributes $n_j - 1$ degrees of freedom, the total number of degrees of freedom associated with the sum of the squares within the groups is

$$\sum_{j=1}^{k} (n_j - 1) = n - k \tag{10.23}$$

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Thus, there are $n - k$ degrees of freedom associated with the sum of the squares within the groups. There are also $n - 1$ degrees of freedom associated with the sum of squares, because each observation X_{ij} is being compared to the general mean $\bar{x}$, based on all n observations.

If each of these sums of squares is divided by their corresponding degrees of freedom, three variances or **quadratic terms of means** will be obtained, namely the mean square value between samples ($MS_{between}$), the mean square value within samples (MS_{within}), and the total mean square value (MS_{total}):

$$MS_{between} = \frac{SS_{between}}{k - 1} \tag{10.24}$$

$$MS_{within} = \frac{SS_{within}}{n - k} \tag{10.25}$$

$$MS_{total} = \frac{SS_{total}}{n - 1} \tag{10.26}$$

Since the variance is calculated by dividing the sum of the squared differences by their appropriate degrees of freedom, **all quadratic terms of averages are variances**.

The main interest is to compare the arithmetic means of groups k or the levels of a factor, to determine if there is a treatment effect between the k groups and, for this, we analyse the **variances**. If the null hypothesis is true and there are no real differences in the means of the k groups, all three quadratic terms of means, $MS_{between}$, MS_{within}, and MS_{total}, provide estimates of the variance (σ^2) that are only inherent to the data.

Thus, to test the null hypothesis, we calculate the **test statistic F (F_{calc})** which is the ratio between $MS_{between}$ and MS_{within}:

$$\boxed{F = \frac{MS_{between}}{MS_{within}}} \tag{10.27}$$

Essentially, what we are saying here is that if the mean square value between samples is much larger than the mean square within samples, then the ratio of the two of them (the F test statistic or the calculated F value) will be very large, and if it becomes large enough to surpass the critical F value, we will say that the results are significant (this is analogous to the F-test for variances that we presented in Section 10.4.2(a)). This implies that the variation between different samples is much larger than the natural variation within the samples (which would indicate that at least one of the sample means is different from the others).

S. 10.4.2

The decision rule again depends on the *significance level* (α) adopted and on the degrees of freedom. The F statistic has two values for the degrees of freedom. These are denoted df_1 and df_2, and called the numerator and denominator degrees of freedom, respectively, $\boldsymbol{k - 1}$ and $\boldsymbol{n - k}$ (see Equations 10.24 and 10.25).

For a given significance level (α), one can **reject the null hypothesis** if the test statistic $\boldsymbol{F}$ **exceeds the critical value of** $\boldsymbol{F_{S(k-1;\ n-k)}}$ of the F distribution. Note that the *rejection region for the F test is **always** in the right tail of the distribution*.

Excel has the following function for returning the inverse of the right-tailed F probability distribution, and thus obtaining the critical F statistic (F_{crit}):

F.INV.RT(probability;degfreedom1;degfreedom2)

In our case, we have

F.INV.RT($\alpha; k - 1; n - k$)

If the null hypothesis is true, the calculated F statistic (or F_{calc}) is expected to be approximately equal to 1, since the variation between samples would be no different than the variation within samples (in other words,

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Table 10.7 Typical ANOVA table, showing the calculations for the sum of squares, the degrees of freedom, the mean square values, the calculated F test statistic, and the p-values.

Source of Variation	Sum of Squares (SS)	Degrees of Freedom	Mean Square Value (MS)	F_{calc}	p-Value
Between groups	$\sum_{j=1}^{k} n_j(\bar{X}_j - \bar{X})^2$	$k-1$	$\frac{SS_{between}}{k-1}$	$F = \frac{MS_{between}}{MS_{within}}$	
Within groups	$\sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ij} - \bar{X}_j)^2$	$n-k$	$\frac{SS_{within}}{(n-k)}$	–	
Total	$\sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ij} - \bar{X})^2$	$n-1$	–	–	

Note: see Equations 10.19, 10.21, 10.22, 10.24, 10.25, 10.26.

the data all appear to come from the same distribution with the same mean). On the other hand, if H_0 is rejected (and there are real differences in the means), the F statistic is expected to be substantially greater than 1, since the numerator $MS_{between}$ would be calculating the treatment effect or the differences between groups, which is greater than the variability that is naturally inherent to the data. The denominator, MS_{within}, would be measuring only the inherent variability. Thus, the ANOVA procedure generates an F test, in which the null hypothesis can be rejected at a selected significance level only if the calculated F statistic is large enough to exceed $F_{S(k-1;\ n-k)}$, the critical value of the upper tail of the F distribution.

The ***p*-value** can be obtained using the following Excel function:

F.DIST.RT(x;degfreedom1;degfreedom2)

In our case, we have

F.DIST.RT($F_{calc}; k-1; n-k$)

The calculations are displayed in an ANOVA table, typically presented as shown in Table 10.7.

Recalling the assumptions of ANOVA

The three main assumptions in the analysis of variance are

- Randomness and independence of errors (differences between each observed value and the arithmetic mean of its own group must be independent)
- Normality
- Homogeneity of variance (homoscedasticity)

However, we can make the following comments about these assumptions:

- ANOVA is a **robust** statistical procedure and provides **reliable** results even with considerable heteroscedasticity, provided that the sample sizes are equal or approximately equal.
- ANOVA is also **reasonably robust** even though the variable under study shows *departures from the normal distribution*, especially for *large samples*.
- However, if these **assumptions are seriously violated**, the solution is to use a transformation of the data or a non-parametric test. In this case, the test indicated would be the *non-parametric Kruskal–Wallis test* (Section 10.5.3).

S. 10.5.3

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Example

EXAMPLE 10.9 COMPARING THE MEANS OF THREE SAMPLES USING ANOVA

You have monitored the concentration of a certain constituent in three different water bodies, A, B, and C. Analyse whether any of the means of the three data sets are significantly different from the others, at a significance level $\alpha = 0.05$. The data (values are in mg/L) are shown in the table below:

Water Body	A	B	C
	3	5	2
Data	2	4	3
	3	7	4
		8	1
n	3	4	4
$\bar{X}$	**2.7**	**6.0**	**2.5**

Note: this example is purely didactic, for you to be able to see how the calculations are done, since in practice there is not enough data to apply a conclusive ANOVA. Normally, your data sets will be much larger. However, the methods remain the same.

Excel

Note: this example is also available as an Excel spreadsheet.

Solution:

Initially, we should prepare **box-plots** to visualize the measures of central tendency and variation in the samples. We will not produce them here because the available data are clearly insufficient for constructing a meaningful chart. However, for large samples that you will encounter in practice, you should prepare box-plots such as the ones shown in Example 10.7. You will include the number of samples you have under each individual box. In our example here, we would have three boxes since there are three samples.

From the table above, we can see that there are differences between the means of the sample concentrations for the three water bodies, but we do not know if these mean values are significantly different. For that, we need to run an ANOVA.

Therefore, we postulate the following test hypotheses:

H_0: $\mu_A = \mu_B = \mu_C$
H_a: the means are not all equal

The final results are presented in the form of an ANOVA summary table, as follows.

Summary of the Analysis of Variance – ANOVA

Source of Variation	df	Sums of Squares, SS	Mean Squares, MS	F_{calc}
Between groups	2	29.97	14.98	7.65
Within groups	8	15.67	1.96	
Total	10	45.64	–	–

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Now let us see how we can construct the ANOVA summary table:
Number of groups: $\mathbf{k = 3}$
Number of data: $\mathbf{n = 11}$

- **Calculation of the grand mean**

$$\bar{X} = \sum_{j=1}^{k}\sum_{i=1}^{n_j}\frac{X_{ij}}{n} = \frac{3+2+3+5+4+7+8+2+3+4+1}{11} = \mathbf{3.8}$$

- **Sum of squares**

  Variation between groups SS ($SS_{between}$):

$$SS_{between} = \sum_{j=1}^{k} n_j(\bar{X}_j - \bar{X})^2 = (3)(2.7-3.8)^2 + (4)(6.0-3.8)^2 + (4)(2.5-3.8)^2 = \mathbf{29.97}$$

  Variation within group SS (SS_{within}):

$$SS_{within} = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ij} - \bar{X}_j)^2 = [(3-2.7)^2 + (2-2.7)^2 + (3-2.7)^2]$$
$$+ [(5-6.0)^2 + (4-6.0)^2 + (7-6.0)^2 + (8-6.0)^2]$$
$$+ [(2-2.5)^2 + (3-2.5)^2 + (4-2.5)^2 + (1-2.5)^2] = \mathbf{15.67}$$

  Total sum of squares (SS_{total}):

$$SQT = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ij} - \bar{X})^2 = [(3-3.8)^2 + (2-3.8)^2 + (3-3.8)^2]$$
$$+ [(5-3.8)^2 + (4-3.8)^2 + (7-3.8)^2 + (8-3.8)^2]$$
$$+ [(2-3.8)^2 + (3-3.8)^2 + (4-3.8)^2 + (1-3.8)^2] = \mathbf{45.64}$$

- **Mean squares**

$$MS_{between} = \frac{SS_{between}}{k-1} = \frac{29.67}{3-1} = \mathbf{14.98}$$
$$MS_{within} = \frac{SS_{within}}{n-k} = \frac{15.67}{11-3} = \mathbf{1.96}$$

- **Test statistic F (F_{calc})**

$$F = \frac{MS_{between}}{MS_{within}} = \frac{14.98}{1.96} = \mathbf{7.65} \rightarrow \boldsymbol{F_{calc}} = \mathbf{7.65}$$

- **Critical value of F (F_{crit})**

  Degrees of freedom:

df_N = df of the numerator $= k - 1 = 3 - 1 = 2$
df_D = df of the denominator $= n - k = 11 - 3 = 8$

  The critical value of $F_{S(k-1;\ n-k)}$ can be obtained with the Excel function:

  F.INV.RT(probability;degfreedom1;degfreedom2)

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

In our case, we have:

F.INV.RT $(a; k-1; n-k)$ = F.INV.RT $(0.05; 2; 8) = 4.46 \rightarrow$ $\boldsymbol{F_{crit} = 4.46}$

- **Test result**

  $\boldsymbol{F_{calc} = 7.65 > F_{crit} = 4.46}$

  Since the test statistic (F_{calc}) is greater than the critical value (F_{crit}), we **reject the null hypothesis** *that the population means are all equal, and conclude that there is at least one mean value that has a* (*statistically*) *significant difference from the other population means*.

- ***p*-value**

  We can obtain a similar conclusion by calculating the *p*-value, which is obtained from the inverse of the *F* distribution for the value of F_{calc}. We can use the following Excel function:

  F.DIST.RT(x;degfreedom1;degfreedom2)

  In our case, we have

  F.DIST.RT($F_{calc}; k-1; n-k$) = F.DIST.RT$(7.65; 2; 8) = 0.013889 \rightarrow$ ***p*-value = 0.013889**

  Since ***p*-value** $< \alpha$ $(0.013889 < 0.05)$, we **reject the null hypothesis**.

- **Summary table using Excel add-in for ANOVA in the Analysis ToolPak**

  If we use the Excel add-in for ANOVA (see instructions in Excel), we get the following summary table. As expected, the values are the same as those calculated above.

  Anova: Single Factor

  SUMMARY

Groups	**Count**	**Sum**	**Average**	**Variance**		
A	3	8	2.666667	0.333333		
B	4	24	6	3.333333		
C	4	10	2.5	1.666667		
ANOVA						
Source of variation	**SS**	**df**	**MS**	**F**	***p*-value**	***F* crit**
Between groups	29.9697	2	14.98485	7.651838	0.013889	4.45897
Within groups	15.66667	8	1.958333			
Total	45.63636	10				

10.5.3 *Post hoc* multiple comparison analysis following ANOVA: the parametric Tukey test

Advanced

ANOVA is used to determine if there is a difference between several groups, but a significant result from this test does not indicate *which sample*(*s*) significantly differ from the others, when compared two by two. It only shows that there is at least one difference between the groups studied.

The identification of significant differences between means, based on pairwise comparisons, made between two samples at a time, should be done after completing the ANOVA. There are several different tests that can be used for these multiple comparisons between means; some examples are the Tukey test, the Bonferroni test, Dunnet's test, the Scheffe test, the Student–Newman–Keuls method, etc. Only the

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Tukey test will be demonstrated in this book. Further information about the use of the other tests can be found in Zar (1999) and Levine *et al.* (2012). Such *post hoc* procedures can be used **if and only if** the result of the ANOVA (the F test) was statistically significant.

We will now describe the **Tukey test**. In all multiple comparisons' tests, equal sample sizes are desirable for maximum power and robustness, but we will show the procedures that can be used for analysis with unequal samples sizes between groups.

The Tukey test considers the null hypothesis H_0: $\mu_A = \mu_B$ versus the alternative hypothesis H_0: $\mu_A \neq \mu_B$, where the subscripts A and B denote any possible pair of groups. For k groups, $k\,(k-1)/2$ different pairwise comparisons can be made.

The differences between the means of all pairs are calculated and the standard error (SE) is estimated as follows:

$$SE = \sqrt{\frac{MS_{within}}{2}\left(\frac{1}{n_A} + \frac{1}{n_B}\right)} \tag{10.28}$$

where MS_{within} is the mean square within groups, calculated previously in the ANOVA test.

Table 10.8 Critical values of the q distribution (Studentized range) for a significance level of 0.05, as a function of the number of groups (k from 2 to 8) and the degrees of freedom (df).

df	Number of Groups (k)						
	2	**3**	**4**	**5**	**6**	**7**	**8**
5	3.64	4.60	5.22	5.67	6.03	6.33	6.58
6	3.46	4.34	4.90	5.30	5.63	5.90	6.12
7	3.34	4.16	4.68	5.06	5.36	5.61	5.82
8	3.26	4.04	4.53	4.89	5.17	5.40	5.60
9	3.20	3.95	4.41	4.76	5.02	5.24	5.43
10	3.15	3.88	4.33	4.65	4.91	5.12	5.30
11	3.11	3.82	4.26	4.57	4.82	5.03	5.20
12	3.08	3.77	4.20	4.51	4.75	4.95	5.12
13	3.06	3.73	4.15	4.45	4.69	4.88	5.05
14	3.03	3.70	4.11	4.41	4.64	4.83	4.99
15	3.01	3.67	4.08	4.37	4.59	4.78	4.94
16	3.00	3.65	4.05	4.33	4.56	4.74	4.90
17	2.98	3.63	4.02	4.30	4.52	4.70	4.86
18	2.97	3.61	4.00	4.28	4.49	4.67	4.82
19	2.96	3.59	3.98	4.25	4.47	4.65	4.79
20	2.95	3.58	3.96	4.23	4.45	4.62	4.77
24	2.92	3.53	3.90	4.17	4.37	4.54	4.68
30	2.89	3.49	3.85	4.10	4.30	4.46	4.60
40	2.86	3.44	3.79	4.04	4.23	4.39	4.52
60	2.83	3.40	3.74	3.98	4.16	4.31	4.44
120	2.80	3.36	3.68	3.92	4.10	4.24	4.36
200	2.77	3.31	3.63	3.86	4.03	4.17	4.29

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

After that, for each difference between means, the test statistic q is calculated (q_{calc}):

$$q_{calc} = \frac{\bar{X}_B - \bar{X}_A}{SE}$$

If the calculated q value is greater than the critical value, $q_{(\alpha;k;n-k)}$, then H_0: $\mu_A = \mu_B$ is rejected. The critical value is dependent upon α (the significance level), df (degrees of freedom within groups for the analysis of variance), and k (the total number of means being tested). The critical value in this test is known as a 'Studentized range' (abbreviated q) and can be found in specific tables in statistical books. In our book we present a short version, only for $\alpha = 0.05$, and for k varying from 2 to 8, as seen in Table 10.8.

Summary of Tukey's procedure

- Arrange and number all sample means in order of decreasing magnitude
- Tabulate the pairwise differences
- Calculate the standard error (SE) of each difference between means:

$$SE_{B,A} = \sqrt{\frac{MS_{within}}{2}\left(\frac{1}{n_B} + \frac{1}{n_A}\right)}$$

- Calculate q value: $q_{calc} = \left(\frac{\bar{X}_B - \bar{X}_A}{SE}\right)$
- Rejection region: $|q_{calc}| > q_{(\alpha;k;n-k)}$
- Rejection with p-value: p-value $< \alpha$

The procedure is illustrated in Example 10.10.

Example

EXAMPLE 10.10 USING THE TUKEY TEST TO DETECT DIFFERENCES BETWEEN SAMPLES, FOLLOWING ANOVA

In Example 10.9, we completed an ANOVA with three samples, each one originating from a different water body (A, B, and C). The test resulted in a rejection of the null hypothesis, indicating that the means of the concentrations were not all equal.

In this example, we will perform a post hoc multiple comparison test (specifically, a parametric Tukey test) to test if there are significant differences between each of the three samples, analysing them two by two. We will perform the tests at a significance level $\alpha = 0.05$.

Look for the original input data in Example 10.9.

Note: this example is purely didactic, for you to be able to see how the calculations are done, since in practice there is not enough data to apply a conclusive ANOVA and the post hoc Tukey test. Normally, your samples will be much larger.

Excel

Note: this example is also available as an Excel spreadsheet.

Solution:

The first step is to arrange and number all sample means in order of decreasing magnitude.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Water Bodies	Means	*n*
B	6.0	4
A	2.7	3
C	2.5	4

The differences between the means of all pairs are then calculated:

$\bar{X}_B - \bar{X}_A$	$\bar{X}_B - \bar{X}_C$	$\bar{X}_A - \bar{X}_C$

Estimate the standard error (SE) of each difference between averages, using Equation 10.28:

$$SE = \sqrt{\frac{MS_{within}}{2}\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}$$

where MS_{within} is the mean square within groups, calculated previously in the ANOVA test in Example 10.9 (A and B indicate any two samples, in this case, water bodies).

$$SE_{B,A} = \sqrt{\frac{MS_{within}}{2}\left(\frac{1}{n_B} + \frac{1}{n_A}\right)} = \sqrt{\frac{1.96}{2}\left(\frac{1}{4} + \frac{1}{3}\right)} = 0.756$$

$$SE_{B,C} = \sqrt{\frac{MS_{within}}{2}\left(\frac{1}{n_B} + \frac{1}{n_C}\right)} = \sqrt{\frac{1.96}{2}\left(\frac{1}{4} + \frac{1}{4}\right)} = 0.700$$

$$SE_{A,C} = \sqrt{\frac{MS_{within}}{2}\left(\frac{1}{n_A} + \frac{1}{n_C}\right)} = \sqrt{\frac{1.96}{2}\left(\frac{1}{3} + \frac{1}{4}\right)} = 0.756$$

For each difference between means, the test statistic q is calculated (q_{calc}):

$$q_{calc} = \frac{\bar{X}_B - \bar{X}_A}{SE}$$

$$\boldsymbol{q_{calcB,A}} = \frac{6.0 - 2.7}{0.756} = \frac{3.3}{0.756} = \mathbf{4.411}$$

$$\boldsymbol{q_{calcB,C}} = \frac{6.0 - 2.5}{0.700} = \frac{3.5}{0.700} = \mathbf{5.002}$$

$$\boldsymbol{q_{calcA,C}} = \frac{2.7 - 2.5}{0.756} = \frac{0.2}{0.756} = \mathbf{0.221}$$

From Example 10.9, we saw that the total number of data points is $n = 11$, the number of groups is $k = 3$ and the degrees of freedom are equal to $n - k = 11 - 3 = 8$ (df = 8). With $\alpha = 0.05$, $k = 3$, df = 8, we go to Table 10.8 and obtain the value of $q_{crit} = 4.04$. In the Excel spreadsheet associated with this example, the consultation to the table is automatic, using the VLOOKUP function.

In the Tukey test, **the critical value (q_{crit}) is the same for all comparisons** between means (for all q_{calc} values).

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

If the calculated value $|q_{calc}|$ is greater than $q_{(0.05;3;8)}$ (q_{crit}), the null hypothesis H_0: $\mu_A = \mu_B$ is rejected (comment valid for all two-by-two comparisons).

The final results are presented in the summary table, as follows.

Comparison	$\bar{X}_2 - \bar{X}_1$	n_2; n_1	SE	q_{calc}	q_{crit}	Conclusion	Interpretation
B versus A	3.3	4;3	0.756	**4.411**	**4.04**	***Reject* H_0**	$\mu_B \neq \mu_A$
B versus C	3.5	4;4	0.700	**5.002**	**4.04**	***Reject* H_0**	$\mu_B \neq \mu_C$
A versus C	0.2	3;4	0.756	**0.221**	**4.04**	*Do not reject* H_0	μ_A not $\neq \mu_C$

The interpretation of the table can be exemplified by

B vs A: $|q_{calc}| = 4.411 > q_{crit} = 4.04 \Rightarrow$ reject H_0

Conclusion for B versus A: the mean concentration of the constituent in water body B is significantly different from the mean concentration in water body A (and since $\bar{X}_2 - \bar{X}$ is positive, we can say that $\mu_B > \mu_A$). Since the mean concentrations in water bodies A and C do not differ significantly from each other, we can also say that water body B has a different mean concentration than water body C (in this case, $\mu_B > \mu_C$).

Below we show different ways in which you can present the **summary of your test results**.

Option *A*, using letters to indicate significant difference or not

Constituent	Mean Values of the Constituent (mg/L) and Letter for Assessing Significant Difference		
	Water Body A	**Water Body B**	**Water Body C**
Constituent 1	2.7 **a**	6.0 **b**	2.5 **a**

- Samples with the same letters (a-a, b-b, c-c…) do not differ significantly from each other.
- Samples with different letters (a-b, b-c, a-c, …) differ significantly from each other.
- From the comparisons between the values of the mean concentrations, we can infer whether a significantly different sample is greater than or lower than the other sample.

Option B, using arrows to indicate whether a sample mean is significantly greater than (or lower than) another sample mean

	Water Body A	**Water Body B**	**Water Body C**
Water body A	–	↑	<>
Water body B	↓	–	↓
Water body C	<>	↑	–

<> **Bold group** (top) not significantly different from the *italic group* (left).
↑ **Bold group** (top) significantly greater than the *italic group* (left).
↓ **Bold group** (top) significantly lower than the *italic group* (left).

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

10.5.4 Non-parametric Kruskal–Wallis test for more than two population central values

Advanced

This test is considered an extension of the Mann–Whitney *U*-test for two independent samples. The non-parametric Kruskal–Wallis test may be employed in instances where the parametric ANOVA is not applicable. In this case, the non-parametric test may in fact be more powerful. The non-parametric analysis is especially desirable when the k samples do not come from normal populations, and it may also be applied when the k population variances are somewhat heterogeneous.

The **Kruskal–Wallis test** is often used to *test whether groups of independent samples were drawn from populations having equal medians*.

Then, the test hypotheses, based on the medians, are

- **Null hypothesis H_0**: $M_1 = M_2 = \cdots = M_k$
- **Alternative hypothesis H_a**: *not all medians are equal*

As in other non-parametric tests, we do not use the population parameters in the formulation of the hypotheses, and neither parameters nor sample statistics are used in the test calculations. The Kruskal–Wallis test statistic, H, is calculated as follows:

$$H = \left[\frac{12}{N(N+1)}\sum_{j=1}^{k}\frac{R_i^2}{n_i}\right] - 3(N+1) \tag{10.29}$$

where

N = total number of observations in all k groups
n_j = number of observations in group i
R_i = sum of the ranks of the n_i observations in group i

S. 10.4.3

The procedure for ranking data is identical to that presented in Section 10.4.3 for the Mann–Whitney test. You can use the Excel function RANK.AVG to rank your data and apply the average criterion for the tied data:

RANK.AVG(number; ref; [order])

- **Number**. The number whose rank you want to find.
- **Ref**. Array of all values in the samples.
- **Order**. A number specifying how to rank numbers (0 or omitted: descending order; any non-zero value: ascending order)

To check whether the ranks have been assigned correctly, the **sum of all ranks must be equal to $N(N+1)/2$**, where $N = n_1 + n_2 + \cdots + n_k$.

If there are tied ranks, H is a little smaller than it should be, and a correction factor (CF) may be computed as follows:

$$\mathrm{CF} = 1 - \frac{\sum t}{n^3 - n} \tag{10.30}$$

$$\sum t = \sum_{i=1}^{n}(t_i^3 - t_i) \tag{10.31}$$

where t_i is the number of ties in the group and n is the number of groups of tied ranks.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

The corrected value of H is

$$\boxed{H_c = \frac{H}{CF}} \tag{10.32}$$

Interestingly enough, H or H_c could also be computed by applying the procedures of ANOVA to the ranks of the data in order to obtain the groups SS and total MS, as we can see below.

$$H = \frac{SS_{between}}{MS_{total}} \tag{10.33}$$

where

$$SS_{between} = \sum_{j=1}^{k} n_j(\bar{X}_j - \bar{X})^2$$

and

$$MS_{total} = \frac{SS_{total}}{N-1} = \frac{\sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ij} - \bar{X})^2}{N-1}$$

Critical values of H for small sample sizes in each group ($n \leq 5$) may be used, but for larger samples in each group ($n > 5$), *H may be considered to be approximated by the* **Chi-square (χ^2)** distribution, with $\boldsymbol{k-1}$ **degrees of freedom**. A chi-square distribution is an asymmetric distribution, whose format depends only on the number of degrees of freedom (the greater the number of degrees of freedom, the more symmetrical the distribution becomes).

Critical values for H and χ^2 are found in look-up tables in most statistics textbooks. Table 10.9 reproduces a partial table for a significance level of 0.05, for different combinations of numbers of data (n) in each of the groups (varying from three to five subgroups, or $k = 3$ to $k = 5$). This table should be used if you have **small sample sizes**.

If you have **larger sample sizes and/or greater number of groups ($k > 5$)**, you may use the **Chi-square (χ^2)** approximation to the H distribution, and use the following Excel function:

CHISQ.INV.RT(probability; degfreedom) = CHISQ.INV.RT ($\alpha; k-1$)

where

- the function returns the inverse of the right-tailed probability of the chi-squared distribution
- probability: significance level (α)
- degrees of freedom: $k-1$

For a given level of significance (α), we can reject the null hypothesis if the test statistic H (H_{calc}) exceeds the critical value of $H_\alpha; n_1; n_2 \ldots$ or χ^2_{k-1}.

The ***p*-value** can be calculated from H_{calc} using the Excel function for the Chi-Square distribution CHISQ.DIST.RT:

***p*-value** = **CHISQ.DIST.RT**(x; degfreedom) = CHISQ.DIST.RT($H_{calc}; k-1$)

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Table 10.9 Critical values of H for small samples, for a significance level of 0.05, as a function of the number of data points in each group (for three, four and five groups).

n_1	n_2	n_3	H_{crit}	n_1	n_2	n_3	n_4	n_5	H_{crit}
2	2	2	–	2	2	1	1		–
3	2	1	–	2	2	2	1		5.679
3	2	2	4.714	2	2	2	2		6.167
3	3	1	5.143	3	1	1	1		–
3	3	2	5.361	3	2	1	1		–
3	3	3	5.600	3	2	2	1		5.833
4	2	1	–	3	2	2	2		6.333
4	2	2	5.333	3	3	1	1		6.333
4	3	1	5.208	3	3	2	1		6.244
4	3	2	5.444	3	3	2	2		6.527
4	3	3	5.791	3	3	3	1		6.600
4	4	1	4.967	3	3	3	2		6.727
4	4	2	5.455	3	3	3	3		7.000
4	4	3	5.598	4	1	1	1		–
4	4	4	5.692	4	2	1	1		5.833
5	2	1	5.000	4	2	2	1		6.133
5	2	2	5.160	4	2	2	2		6.545
5	3	1	4.960	4	3	1	1		6.178
5	3	2	5.251	4	3	2	1		6.309
5	3	3	5.648	4	3	2	2		6.621
5	4	1	4.985	4	3	3	1		6.545
5	4	2	5.273	4	3	3	2		6.795
5	4	3	5.656	4	3	3	3		6.984
5	4	4	5.657	4	4	1	1		5.945
5	5	1	5.127	4	4	2	1		6.386
5	5	2	5.338	4	4	2	2		6.731
5	5	3	5.705	4	4	3	1		6.635
5	5	4	5.666	4	4	3	2		6.874
5	5	5	5.780	4	4	3	3		7.038
6	1	1	–	4	4	4	1		6.725
6	2	1	4.822	4	4	4	2		6.957
6	2	2	5.345	4	4	4	3		7.142
6	3	1	4.855	4	4	4	4		7.235
6	3	2	5.348	2	1	1	1	1	–
6	3	3	5.615	2	2	1	1	1	–
6	4	1	4.947	2	2	2	1	1	6.750
6	4	2	5.340	2	2	2	2	1	7.133
6	4	3	5.610	2	2	2	2	2	7.418

(*Continued*)

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Table 10.9 Critical values of H for small samples, for a significance level of 0.05, as a function of the number of data points in each group (for three, four and five groups) (*Continued*).

n_1	n_2	n_3	H_{crit}	n_1	n_2	n_3	n_4	n_5	H_{crit}
6	4	4	5.681	3	1	1	1	1	–
6	5	1	4.990	3	2	1	1	1	6.583
6	5	2	5.338	3	2	2	1	1	6.800
6	5	3	5.602	3	2	2	2	1	7.309
6	5	4	5.661	3	2	2	2	2	7.682
6	5	5	5.729	3	3	1	1	1	7.111
6	6	1	4.945	3	3	2	1	1	7.200
6	6	2	5.410	3	3	2	2	1	7.591
6	6	3	5.625	3	3	2	2	2	7.910
6	6	4	5.724	3	3	3	1	1	6.576
6	6	5	5.765	3	3	3	2	1	7.769
6	6	6	5.801	3	3	3	2	2	8.044
7	7	7	5.819	3	3	3	3	1	8.000
8	8	8	5.805	3	3	3	3	2	8.200
				3	3	3	3	3	8.333

Source: Zar (1999), modified.
Note that there are missing values (denoted with –), indicating insufficient number of data points in a specific group for conducting the test.

Example

EXAMPLE 10.11 COMPARING THE MEDIANS OF THREE SAMPLES USING THE NON-PARAMETRIC KRUSKAL–WALLIS TEST

You have monitored the concentration of a certain constituent in three different water bodies, A, B, and C. Analyse whether the medians of the three data sets are significantly different from each other, at a significance level of $\alpha = 0.05$. The data (values are in mg/L) are shown in the table below:

Water Body	A	B	C
	3	5	2
Data	2	4	3
	3	7	4
		8	1
Number of data (n)	3	4	4
Medians	3.0	6.0	2.5

Note: this example is purely didactic, for you to be able to see how the calculations are done, since in practice there is not enough data to apply a conclusive Kruskal–Wallis test. Normally, your samples will be much larger.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Excel

Note: this example is also available as an Excel spreadsheet.

Solution:

Our test hypotheses are:
- H_0: $M_1 = M_2 = M_3$
- H_a: not all medians are equal

The data are arranged in ascending order within each water body and the ranks are determined.

Constituent (mg/L)			Ranks		
A	**B**	**C**	**A**	**B**	**C**
2	4	1	2.5	7.5	1
3	5	2	5	9	2.5
3	7	3	5	10	5
	8	4		11	7.5
Number of data in each sample (n_i):			3	4	4
Sum of ranks in each sample (R_i):			12.5	37.5	16.0

The **total number of data** (N) is: $3 + 4 + 4 = 11 \rightarrow \boldsymbol{N = 11}$
The **total sum of the ranks** is: $12.5 + 37.5 + 16.0 = \mathbf{66}$
A practical way to check if there was no error in the sum of all ranks is to compute:

$$\frac{N(N+1)}{2} = \frac{11(11+1)}{2} = \mathbf{66}$$

Ok, this value of 66 matches the value previously calculated of the total sum of the ranks (66).
The test statistic H can be calculated using Equation 10.29:

$$H = \left[\frac{12}{N(N+1)}\sum_{j=1}^{k}\frac{Ri^2}{n_i}\right] - 3(N+1)$$

$$H = \left\{\frac{12}{11(11+1)}\left[\frac{(12.5)^2}{3} + \frac{(37.5)^2}{4} + \frac{(16.0)^2}{4}\right]\right\} - 3(11+1) = \mathbf{6.513}$$

Since there were tied ranks, we must make the correction of the value of H. From the table above, we compute the number of tied ranks.

Number of tied ranks: 2 in rank 2.5; 3 in rank 5; and 2 in rank 7.5. We apply Equation 10.30 to allow for a further correction of H.

$$\sum t = \sum_{i=1}^{n}(t_i^3 - t_i) = (2^3 - 2) + (3^3 - 3) + (2^3 - 2) = 36$$

The correction factor may be computed using Equation 10.30:

$$\text{CF} = 1 - \frac{\sum(t^3 - t)}{N^3 - N} = 1 - \frac{36}{11^3 - 11} = \mathbf{0.973}$$

The corrected value of H comes from Equation 10.32:

$$H_c = \frac{H}{\text{CF}} = \frac{6.513}{0.973} = \mathbf{6.696}$$

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Therefore, $\boldsymbol{H_{calc} = 6.696}$.

We now need to obtain the critical value of H (H_{crit}). Since the samples are small, we will obtain the value from Table 10.9. We look at the combinations of n equal to 4, 4, 3. For this combination, we get the value of $H_{0,05;4;4;3} = 5.598$.

Therefore, $\boldsymbol{H_{crit} = 5.598}$.

If we had used the **Chi-square approximation to the H distribution**, and applied the Excel function CHISQ.INV.RT (probability; deg_freedom) = CHISQ.INV.RT (α; $k-1$) = CHISQ.INV.RT (0.05; 11 − 1), we would have obtained $H_{crit} = 5.991$.

Since $\mathbf{H_{calc} > H_{crit}}$ (6.696 > 5.598), we **reject the null hypothesis of equal population medians and conclude that there is a statistically significant difference among the population medians**. A similar decision can be obtained using the p-value, which can be calculated from H_{calc} using the Excel function for the Chi-Square distribution CHISQ.DIST.RT:

$$\boldsymbol{p\text{-value}} = \mathbf{CHISQ.DIST.RT}(x; \text{degfreedom}) = \text{CHISQ.DIST.RT}(H_{calc}; k-1)$$
$$= \text{CHISQ.DIST.RT}(6.696; 3-1) = \mathbf{0.0352}$$

Since p-value < significance level α (0.0352 < 0.05), we **reject the null hypothesis of equal population medians.**

However, we do not know which median(s) is(are) different from the others. In order to know this, we need to complete a *post hoc* multiple comparison test. The non-parametric method for making these *post hoc* comparisons is presented next.

10.5.5 *Post hoc* multiple comparison analysis following Kruskal–Wallis: the non-parametric Dunn test

Advanced

S. 10.5.3

As we saw in the previous sections, a significant result from the ANOVA test and the Kruskal–Wallis test does not indicate specifically which group(s) have significant differences from other groups. In Section 10.5.3, we saw the application of Tukey's test to identify significant differences between groups on a pairwise basis, for the parametric case. For the non-parametric case, there are also several *post hoc* multiple comparison tests that can be performed after a Kruskal–Wallis test. Several of these tests are described in Zar (1999). In this book, we will demonstrate the use of one of them, namely the non-parametric **Dunn test** for *multiple comparisons between groups*.

Unlike Tukey's test, in the Dunn test, the *sum of the ranks is used instead of the means*. Thus, the following steps should be followed to implement the Dunn test for each of the possible $[k\,(k-1)/2]$ pairs:

- Arrange the sample in ascending order of the sum of the ranks (R_1, R_2, … , R_k), the lowest being identified as 3 and the largest as 1.
- Determine the sample sizes: n_1, n_2, … , n_k.
- Tabulate the differences between pairs, starting with the difference between the highest and the lowest value of the sums of the ranks (Ri).
- For each pair, calculate the standard error (SE), using the following equation:

$$\text{SE} = \sqrt{\frac{N(N+1)}{12}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \tag{10.34}$$

where $N = (n_1 + n_2 + \cdots + n_k)$

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Table 10.10 Critical values for the *Q* distribution (Dunn test) at a significance level of 0.05 for different number of groups (*k*).

k	2	3	4	5	6	7	8	9	10
Q_{crit}	1.960	2.394	2.639	2.807	2.936	3.038	3.124	3.197	3.261

- If tied ranks are present, then the following equation may be used for SE:

$$\mathrm{SE} = \sqrt{\left(\frac{N(N+1)}{12} - \frac{\sum t}{12(N-1)}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \tag{10.35}$$

  where $\sum t$ is the same used in the Kruskal–Wallis test when ties are present (see Equation 10.31).

- For the test statistic, for each pair, we use:

$$Q = \frac{\bar{R}_2 - \bar{R}_1}{\mathrm{SE}} \tag{10.36}$$

  where $\bar{R}$ indicates a mean rank: $\bar{R}_1 = R_1/n_1$ and $\bar{R}_1 = R_2/n_2$.

  The critical values for this test ($Q_{\alpha,k}$) are given in Table 10.10. You can consult the table directly, or use the VLOOKUP function.

- Test hypotheses (to be done for each pair of groups being compared):
  - **Null hypothesis $\mathbf{H_0}$**: $M_1 = M_2$
  - **Alternative hypothesis $\mathbf{H_a}$**: $M_1 \neq M_2$

- Test decision (for each pair):

  If $\boldsymbol{Q_{calc} > Q_{crit}}$ **→ reject the null hypothesis** that the medians of the pair you are analysing are equal.

Example

EXAMPLE 10.12 USING THE DUNN TEST TO DETECT DIFFERENCES BETWEEN SAMPLES, FOLLOWING THE KRUSKAL–WALLIS TEST

In Example 10.11, we showed the non-parametric Kruskal–Wallis test that was carried out with three samples, each one originating from a different water body (A, B, and C). The test resulted in a rejection of the null hypothesis, indicating that the medians of the concentrations are not all equal.

Now, perform a post hoc multiple comparison test (non-parametric Dunn test) to detect differences between each of the three samples, analysing them two by two. Use a significance level of $\alpha = 0.05$. The original data can be found in Examples 10.9 and 10.11.

Note: this example is purely didactic, for you to be able to see how the calculations are done, since in practice there is not enough data to apply conclusive Kruskal–Wallis and the post hoc Dunn tests. Normally, your samples will be much larger.

Excel

Note: this example is also available as an Excel spreadsheet.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Solution:

(a) Initial calculations

Concentrations (mg/L) in the Three Water Bodies, A, B, and C)			Ranks of the Values of the Constituent in the Three Water Bodies, A, B, and C		
A	B	C	A	B	C
3	5	2	**5**	9	**2.5**
2	4	3	**2.5**	**7.5**	**5**
3	7	4	**5**	10	**7.5**
	8	1		11	1
n			$n = 3$	$n = 4$	$n = 4$
Medians			3.0	6.0	2.5
$\sum(R_i)$			12.5	37.5	16.0

(b) Standard error (SE) for each pair

We now calculate the standard errors (SE) according to Equation 10.34. In this equation, we need the value of $\sum t$, which was calculated in Example 10.11 using Equation 10.31. If you review this example, you will recall that we found $\sum t$ to be equal to 36.

$$\text{SE} = \sqrt{\left(\frac{N(N+1)}{12} - \frac{\sum t}{12(N-1)}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

We now perform these calculations of the SE on a pairwise basis: Sample B versus Sample C; Sample B versus Sample A; and Sample C versus Sample A.

- For $n_B = 4$ and $n_C = 4$ (B versus C):

$$\text{SE} = \sqrt{\left(\frac{11(11+1)}{12} - \frac{36}{12(11-1)}\right)\left(\frac{1}{4} + \frac{1}{4}\right)} = 2.313$$

- For $n_B = 4$ and $n_A = 3$ (B versus A):

$$\text{SE} = \sqrt{\left(\frac{11(11+1)}{12} - \frac{36}{12(11-1)}\right)\left(\frac{1}{4} + \frac{1}{3}\right)} = 2.498$$

- For $n_A = 4$ and $n_C = 3$ (C versus A):

$$\text{SE} = \sqrt{\left(\frac{11(11+1)}{12} - \frac{36}{12(11-1)}\right)\left(\frac{1}{4} + \frac{1}{3}\right)} = 2.498$$

(c) Test statistics for each pair (Q_{calc} for each pair)

We now need to calculate the test statistic (Q_{calc}) for each pair. In the test procedure, note that the means of the ranks ($\bar{R}_i$), rather than the sums of the ranks (R_i), are arranged in order of magnitude. We then prepare this computational table in order to obtain each mean rank.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

	Water Bodies		
	A	**B**	**C**
Rank sums (R_i):	12.5	37.5	16.0
Sample sizes (n_i):	3	4	4
Mean ranks (R_i/n_i):	**4.2**	**9.4**	**4.0**
Samples ranked by mean ranks (i):	3	1	2

$\mathbf{Q_{calc}}$ is computed using Equation 10.36, based on the differences between the ranks divided by the standard error (SE):

$$Q = \frac{\bar{R}_2 - \bar{R}_1}{\text{SE}}$$

We apply this equation for each pair of groups, in order to obtain $\mathbf{Q_{calc}}$ for the corresponding pair:

B versus A:

$$Q = \frac{(9.4) - (4.2)}{2.498} = \mathbf{2.085}$$

B versus C:

$$Q = \frac{(9.4) - (4.0)}{2.313} = \mathbf{2.324}$$

A versus C:

$$Q = \frac{(4.2) - (4.0)}{2.498} = \mathbf{0.067}$$

(d) Critical value of the distribution ($\mathbf{Q_{crit}}$ for all groups)

The next step is to determine the critical value of the test ($\mathbf{Q_{crit}}$). This value can be obtained from Table 10.10, for $k = 3$ and significance level $= 0.05$. The resulting value is $\mathbf{Q_{crit}} = \mathbf{2.394}$. Note that this critical value is the same for all the pairwise comparisons.

(e) Test results and decisions

Our **test hypotheses** are made on a pairwise basis, and are structured as follows:

- Null hypothesis H_0: $M_1 = M_2$
- Alternative hypothesis H_a: $M_1 \neq M_2$

The **test decision** is

- $\mathbf{Q_{calc}} > \mathbf{Q_{crit}}$: **reject** the null hypothesis (medians are different)
- $\mathbf{Q_{calc}} \leq \mathbf{Q_{crit}}$: **do not reject** the null hypothesis (medians are not different)

The final results are shown in the summary table presented below.

Comparison	Difference ($\bar{R}_2 - \bar{R}_1$)	SE	Q_{calc}	Q_{crit}	Conclusion
B versus A	5.2	2.498	2.085	2.394	Do not reject H_0: *Medians are not different*
B versus C	5.4	2.313	2.324	2.394	Do not reject H_0: *Medians are not different*
A versus C	0.2	2.498	0.067	2.394	Do not reject H_0: *Medians are not different*

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

(f) Summary tables

The summary can also be presented in the two styles shown in Example 10.10.

Option A, using letters to indicate significant difference or not

Constituent	Median Values of the Constituent (mg/L) and Letter for Assessing Significant Difference		
	Water Body A	**Water Body B**	**Water Body C**
Constituent 1	3.0 **a**	6.0 **a**	2.5 **a**

- Samples with the same letters (a-a, b-b, c-c …) do not differ significantly from each other.
- Samples with different letters (a-b, b-c, a-c, …) differ significantly from each other.
- From the comparisons between the values of the mean concentrations, we can infer whether a significantly different sample is greater than or lower than the other sample.

Option B, using arrows to indicate whether a sample median is significantly greater than (or lower than) another sample mean

	Water Body A	**Water Body B**	**Water Body C**
Water body A	–	<>	<>
Water body B	<>	–	<>
Water body C	<>	<>	–

<> **Bold group** (top) not significantly different from the *italic group* (left)
↑ **Bold group** (top) significantly greater than the *italic group* (left)
↓ **Bold group** (top) significantly lower than the *italic group* (left)

(g) Additional comments

You may have found it strange that the Dunn test did not indicate any significant differences in the medians, in contrast with the Kruskal–Wallis test, whose result showed at least one significantly different median from the group.

Remember that our example here is very simplistic, with a very limited number of data points in each sample. We did this in order to be able to demonstrate more clearly how the calculations are done (these comments also apply for Examples 10.9–10.12). However, we are aware that conducting hypothesis tests with such a limited number of data points may lead to inconclusive results. This is one example of a result that is inconclusive.

But you may also find it interesting to know that if we had used a different non-parametric *post hoc* test (namely, the Bonferroni test, not shown here), we would find that the median from sample B was found to be significantly different from samples A and C, and that samples A and C were not significantly different. This outcome is similar to the one obtained by the Tukey test (see Example 10.10).

Recall that there are many different versions of *post hoc* multiple comparison tests that can be used after ANOVA and Kruskal–Wallis tests. Here, we are only showing one example each for the parametric and non-parametric cases. Naturally, because each of the different *post hoc* comparison tests use slightly different approaches, they may sometimes produce different results, even for the same data sets, especially when the sample size is small. This is a reminder that you can use statistical methods as a tool, but in the end, the interpretation of the

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

results will require some of your expertise, common sense, and judgement. In some cases, it might be worthwhile to collect some additional data points in order to confirm a result that you think might end up being significant.

10.6 CHECK-LIST FOR YOUR REPORT

- ✓ Make sure that you have described clearly the different samples you are comparing (the different groups).
- ✓ Confirm that you have analysed the distributions of your samples, to see whether you need to use parametric or non-parametric statistical tests.
- ✓ Check whether you have taken into account your sample sizes and the resulting implications in terms of statistical power for the desired effect size.
- ✓ Ascertain that you have considered the applicability of each test, together with the fulfilment of their required assumptions (e.g., normality of data, independence of samples, etc.)
- ✓ Check that you have presented suitable graphs to aid in the interpretation of the results, such as box-plots.
- ✓ Verify that, in the hypothesis test you are using, you have specified your null (H_0) and alternative (H_a) hypotheses in a clear way, so that the result of your test allows you to make a strong conclusion.
- ✓ Confirm that you have also specified the significance level and whether you are comparing means or medians.
- ✓ Check that you have mentioned clearly which hypothesis test you applied.
- ✓ Verify that you have specified your resulting *p*-value and its interpretation in comparison with the significance level you specified for the test ($\alpha = 0.01$, 0.05, or 0.10).
- ✓ If applicable, verify whether you have also presented the calculated and the critical values of the distribution, associated with the rejection regions.
- ✓ Reflect whether you have gone a step beyond simply presenting the *p*-value and the determination of significant versus not significant results, but you reported in addition the estimated effect size and its confidence interval.
- ✓ Check that you have made the correct and appropriate conclusion from your hypothesis test, that is, you can only say that either the null hypothesis is rejected and the alternative hypothesis is accepted or that you cannot reject the null hypothesis. Remember, you *cannot say* that your null hypothesis has been accepted!
- ✓ In case you are comparing more than two samples, verify that you have done a *post hoc* multiple comparison test to analyse which of the samples is significantly different from the others (if this is the case, as previously indicated by the previous ANOVA or Kruskal–Wallis test). Remember, you should not use multiple two-sample *t* tests when you have more than two samples to compare simultaneously!

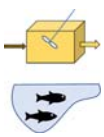
Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Chapter 11

Relationship between monitoring variables. Correlation and regression analysis

This chapter shows you how to analyse the relationship between two or more variables from your monitoring programme (influent and effluent concentrations, environmental conditions, removal efficiencies, applied loading rates, or others). The topics include correlation and regression analysis between variables. Correlation studies encompass correlation coefficients, correlation matrices, cross-correlation, and autocorrelation, including parametric Pearson and non-parametric Spearman correlation methods. For regression analysis, we place emphasis on the linear regression model, which is covered in detail. Other regression models (multiple linear regression and non-linear regression) are also addressed in this chapter.



The contents in this chapter are applicable to both treatment plant monitoring and water quality monitoring.

CHAPTER CONTENTS

11.1 Introduction **398**
11.2 Correlation Coefficient. **402**
11.3 Correlation Matrix **424**
11.4 Cross-correlation and Autocorrelation. **429**
11.5 Simple Linear Regression. **440**
11.6 Multiple Linear Regression **470**
11.7 Non-linear Regression **473**
11.8 Check-List for Your Report **476**

© 2020 The Authors. This is an Open Access book chapter distributed under the terms of the Creative Commons Attribution Licence (CC BY-NC-ND 4.0), which permits copying and redistribution for non-commercial purposes with no derivatives, provided the original work is properly cited (https://creativecommons.org/licenses/by-nc-nd/4.0/). This does not affect the rights licensed or assigned from any third party in this book. The chapter is from the book Assessment of Treatment Plant Performance and Water Quality Data: A Guide for Students, Researchers and Practitioners, Marcos von Sperling, Matthew E. Verbyla and Sílvia M. A. C. Oliveira (Authors).
doi: 10.2166/9781780409320_0397

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

11.1 INTRODUCTION

Basic

C. 10

C. 12-15

C. 7

C. 13

In our book, we have encouraged you to do more with your data – instead of *simply reporting monitoring data*, we are advising you to try to gain a deeper **understanding of the behaviour of the system** you are studying. As an example, in Chapter 10, we described how you could compare two variables to know whether their central values (means or medians) are equal. In Chapters 12–15, we will show you how to integrate statistics with process analysis, covering water and mass balances, loading rates, reaction kinetics, reactor hydraulics, and process modelling.

In this chapter, we describe how you can study the **relationship between two or more variables** that are part of your monitoring programme. These variables can be influent and effluent concentrations, environmental conditions, removal efficiencies (see Chapter 7), applied loading rates (see Chapter 13), or any other variable that may be considered to play an important role in your water body or treatment plant.

We will cover '**correlation**' and '**regression analysis**' in this chapter, including the following items:

Correlation	Regression Analysis
• Correlation (simple correlation) (Section 11.2)	• Simple linear regression (Section 11.5)
• Correlation matrix (Section 11.3)	• Multiple linear regression (Section 11.6)
• Autocorrelation and cross-correlation (Section 11.4)	• Non-linear regression (Section 11.7)

Note that we use the expressions **correlation** and **regression**. A simplified difference between them can be stated as follows:

- **Correlation**: Used to represent the strength of the *linear* relationship between two variables. In a correlation, there is *no concept of dependent and independent variables*, that is, the correlation between x and y is the same as the correlation between y and x.
- **Regression analysis**: Describes how a *dependent variable* (y) is numerically related to the *independent variable* (x) or independent variables (x_1, x_2, …, x_n) via a regression equation with coefficients in its structure. The regression model may be *linear* or *non-linear*.

Figure 11.1 shows the concept of **correlation between two variables**. A *scatter plot* is always a useful way of visually analysing the type of relationship between the variables. If the data points seem to be positioned over an 'imaginary' straight line (even if not perfectly), then we can suppose that there may be a linear relationship between the two variables. We measure the strength of the linear relationship by a **linear correlation coefficient** (***r***). In Section 11.2, we will show you how to calculate and interpret the correlation coefficient.

S. 11.2

Now we will introduce the concept of **regression analysis**, which is illustrated in Figure 11.2 for the same data points from Figure 11.1. The figure also shows several elements of importance in a regression analysis. You can see clearly that the major difference from *correlation* is that now we have a fitting of a line to the data points and an associated equation, which allows us to predict the value of Y (dependent variable or response variable) based on a value of X (independent variable or predictor variable). A linear regression (fitting of a straight line) is illustrated. Since the data points are the same as in Figure 11.1, the coefficient of correlation (r) is the same. Because we have a model and the resulting predictions, we may also have prediction errors if the fitting is not perfect. We analyse the goodness of fit using the concept of the Coefficient of Determination (correlation coefficient raised to the power two, r^2 or R^2).

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

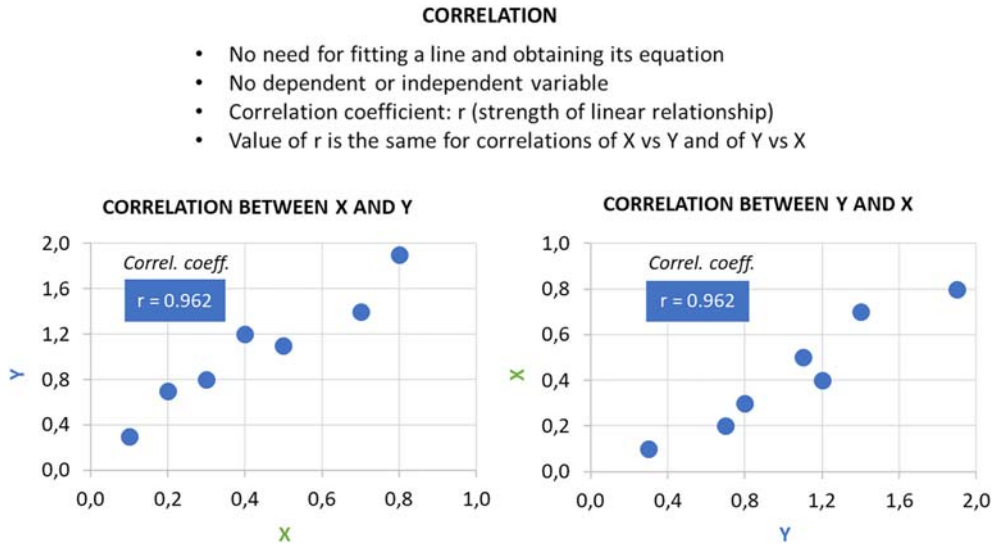


Figure 11.1 Example of the concept of correlation between two variables *X* and *Y*.

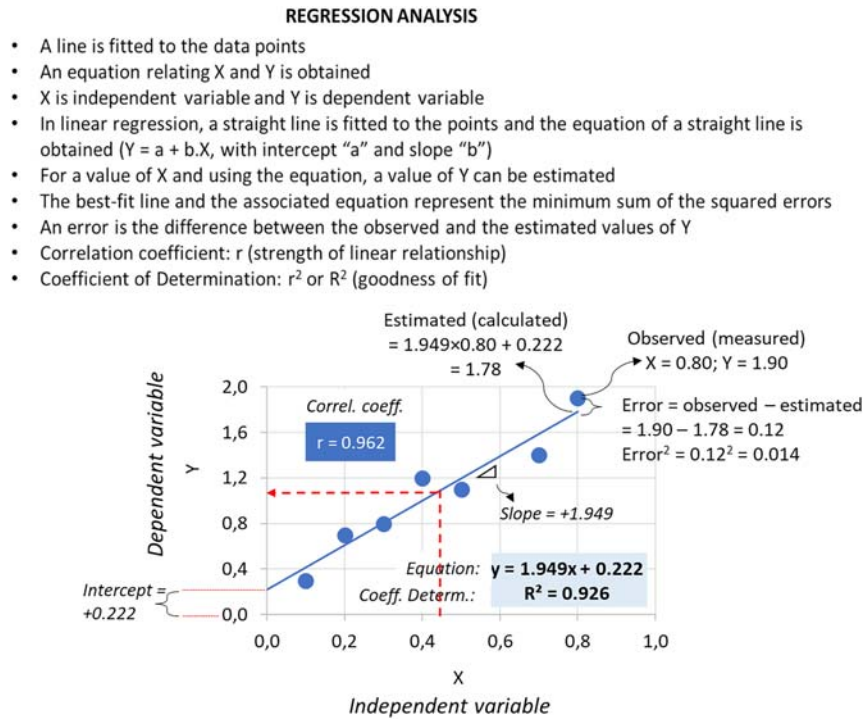


Figure 11.2 Example of the concept of regression analysis between *X* and *Y* and important elements associated with it. A linear regression is illustrated.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

S. 11.5

The figure may seem, at this stage, a bit complex, with many elements, but all of them will be duly explained in Section 11.5.

Let us discuss some **basic concepts of regression analysis**. Regression analysis is a statistical technique to model and investigate the relationship between two or more variables. It is mainly used for forecasting purposes (i.e., predicting future responses). A statistical model is developed to predict the values of a dependent variable or a response variable, based on the values of at least one independent or explanatory variable (Montgomery, 2005).

The way the pairs of data points are related defines the type of relationship between the variables and the type of regression model. The purpose of the analysis is to fit a curve to the data points, and by fitting a curve, we mean defining a curve that *passes as close to the points as possible*. After fitting a curve, we can determine the values of the **coefficients of the model**. Thus, it will be possible to evaluate a possible dependence of y in relation to x and to express mathematically this behaviour by means of an equation.

There are several models that can be tried to fit the data, involving one or more independent variables (Table 11.1). Figure 11.3 illustrates examples of a linear and a non-linear model.

S. 11.5

S. 11.6

Most of the concepts in this chapter are related to **linear regression models** (see Section 11.5 for simple linear regression and Section 11.6 for multiple linear regression). Linear models, especially simple linear regression, are extremely important for the assessment of monitoring data. They are usually the first model we attempt to fit the data, in order to explore the quantitative relationship between our variables. But remember that this approach assumes a **linear** relationship between the variables, which frequently may not be the case, especially considering that we are dealing with environmental systems. For some environmental systems, **non-linear** relationships may be more applicable. Non-linear regression is also covered in this chapter (Section 11.7).

S. 11.7

In other chapters of this book, we discuss other modelling approaches not directly associated with regression analysis (e.g., non-regression-based models). These other mechanistic models require an

Table 11.1 Types of regression analysis.

Factor	Type of Regression	Characteristics
Regarding the relationship between each independent variable x and the dependent variable y	Linear regression	• The variables are linearly related. • In an x–y scatter plot, the points should lie approximately in a straight line. • The least squares solution leads to a unique solution (minimization of the sum of the squared errors).
	Non-linear regression	• The variables are not linearly related. • In the scatter plot, the points are not distributed over a straight line. • The equation of a straight line is not used. • If there is no explicit solution, obtained by transformation (e.g. log-transformation) of variables, the solution must be obtained by iterative numerical methods of minimizing the error function (sum of the squared errors). There is no guarantee that the numerical methods may converge to the same solution.
Regarding the number of independent variables	Simple regression	• There is only one independent variable (x)
	Multiple regression	• There is more than one independent variable (x_1, x_2, …, x_n).

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

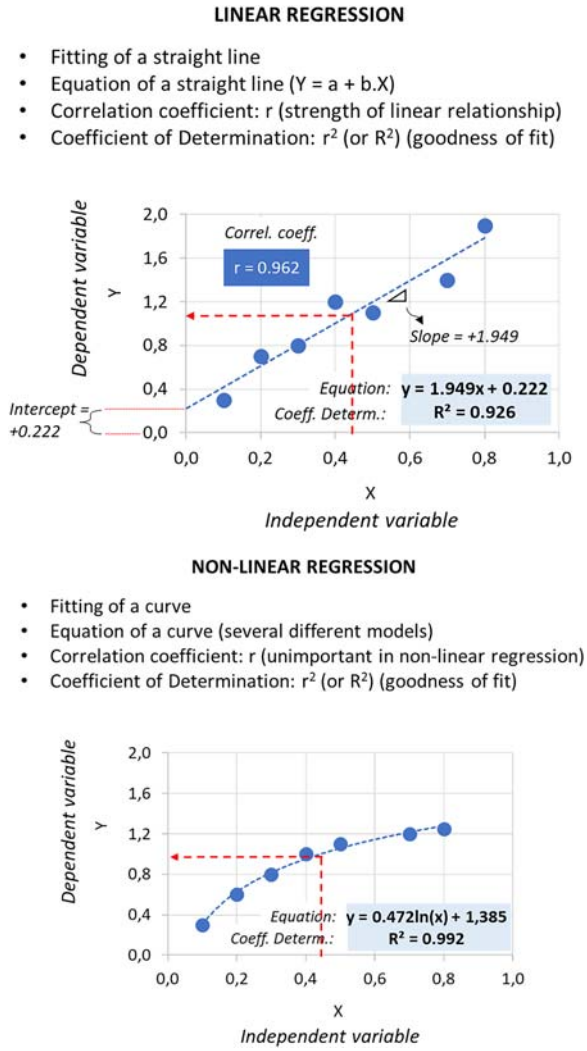


Figure 11.3 Example of the concepts of a linear regression (top) and a non-linear regression (bottom).

C. 12
C. 14
C. 15

understanding of the principles of mass balance (Chapter 12), the kinetics of the reactions (Chapter 14), the reactor hydraulics (Chapter 14), and other process-based considerations (Chapter 15).

At this point, we should recognize the importance of models that are based on **regression analysis** (linear and non-linear), as well as models that are not based on **regression** (see the previous paragraph). In this book, we show you how to construct both types of models, and we stress that you should use both approaches in a judicious manner – in other words, one approach can complement the other. We hope you agree that we do not want to simply fit any equation to our data, without considering a more fundamental understanding of the relationships between the variables involved. Using models in a thoughtful manner, with sound engineering judgment, is necessary if **we want our model to be useful to others**. You have the tools and so make the best use of them, contribute to the advancement in the knowledge of your system, and make this knowledge useful to others!

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

11.2 CORRELATION COEFFICIENT

11.2.1 Pearson's linear correlation coefficient

(a) The concept of the correlation coefficient (*r*)

Basic

The **correlation coefficient, *r***, is a *dimensionless measure of the strength of the linear relationship between two quantitative variables, x and y.*

Some examples in water quality and water treatment studies, where a linear relationship might exist, are electrical conductivity versus salinity; total suspended solids concentration versus turbidity (e.g., in a river); particulate chemical oxygen demand (COD) versus volatile suspended solids (e.g., in the effluent of a wastewater treatment plant); and the calibration of a laboratory equipment with known added amounts of a reagent.

Pearson's linear correlation coefficient, or simply the **correlation coefficient**, is a *dimensionless number*, that is, *independent of the unit of measurement* of the variables. The measurement of the intensity of the association between two quantitative variables is given by ***r***, which **ranges from −1 to +1**, and which has a **sign that indicates the direction of the relationship between *x* and *y***:

- ***r* > 0: *y* increases with increasing *x***
- ***r* < 0: *y* decreases with increasing *x***

The visualization of the relationship between the two variables is best seen in a **scatter plot**, a chart widely covered in other sections of this book. This is simply a plot of each pair of *x* and *y* data points, with the following variables:

- **Variable (*x*)**: values typically shown on the *horizontal axis*
- **Variable (*y*)**: values typically shown on the *vertical axis*

Figure 11.4 gives some examples of different relationships between the two variables, as depicted by their scatter plots, lines of best fit (with the **slope *b***), and ranges of values of the correlation coefficient (*r*). The *slope of the best-fit straight-line (**b**) found using regression has the same sign of the correlation coefficient r*. The slope will be discussed in more detail in Section 11.5.

S. 11.5

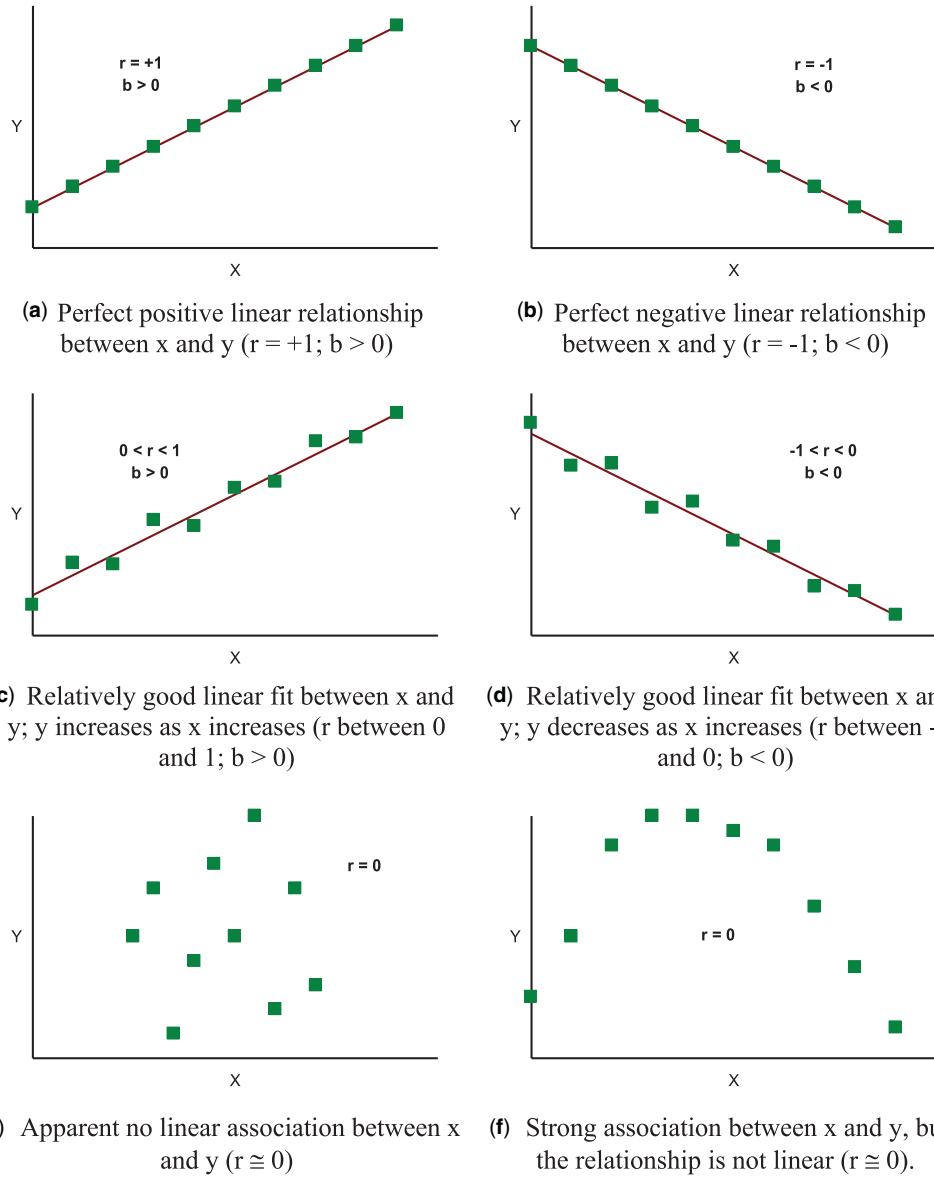
(b) Considerations about the linear correlation coefficient (*r*)

Basic

We can make the following considerations related to the interpretation of the correlation coefficient *r*:

- The coefficient *r* is a correlation coefficient that expresses a **linear relationship**. It is possible that even when $r = 0$, a non-linear model may provide an excellent fit to the experimental data (e.g., Figure 11.4f).
- The correlation coefficient *r* is unaffected by *linear transformations* in the two variables. If you **multiply or divide a variable by a constant**, *this does not change the correlation coefficient between that variable and the other variable*. For instance, the correlation between travelling time and distance in a river does not depend on whether the time is expressed as minutes, days, or hours, and if distance is measured in kilometres or miles. Also, if you **add or subtract a constant value** to your variable, *this does not change the correlation coefficient between the two variables*.
- You must be **careful about the direct interpretation of *r***. A high value of *r* can be obtained even if the points are not distributed along a line. A simple example to understand this is when you calculate the *linear* correlation coefficient for the *quadratic* function $y = x^2$, for x ranging from 1 to 10, and obtain $r = 0.975$. The function is non-linear, but you still obtain a high value of a linear correlation coefficient *r*. If you do not consider the $x-y$ scatter plot in addition to the correlation coefficient, the results can be a bit deceiving.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest



(a) Perfect positive linear relationship between x and y (r = +1; b > 0)

(b) Perfect negative linear relationship between x and y (r = -1; b < 0)

(c) Relatively good linear fit between x and y; y increases as x increases (r between 0 and 1; b > 0)

(d) Relatively good linear fit between x and y; y decreases as x increases (r between -1 and 0; b < 0)

(e) Apparent no linear association between x and y (r ≅ 0)

(f) Strong association between x and y, but the relationship is not linear (r ≅ 0).

Figure 11.4 Examples of scatter plots and the association between variables.

- Atypical values (**outliers**) may introduce *distortions* in the interpretation of correlation and regression analyses, forcing an increase in the value of *r* and changing the parameter values in the regression analysis. Atypical values can be important to include in your analysis, as long as they represent reliable measurements. In this case, they may provide important information about the behaviour of your system. However, if outliers are obtained due to laboratory error or atypical conditions, then they might mislead your interpretation of correlation or regression analysis results. The consideration about discarding outliers or not is

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

complex, but it is nevertheless an important component of the analysis of the experimental data. Discarding is only justified when there is suspicion about the reliability of the specific experimental data (see Section 5.5 for a detailed discussion on outliers).

S. 5.5

- **A high correlation does not imply causality**. A high positive or negative value of r does not allow us to conclude that a change in x will cause a change in y. The only valid conclusion we may have is that there may be a linear relationship or trend between x and y.

(c) Equation used to calculate Pearson's linear correlation coefficient

Basic

As already mentioned, the correlation coefficient r is a measure of the strength of the linear relationship between two variables, x and y. It is computed for a sample of n measurements on x and y using the following equation:

$$r = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\left[\sum x^2 - \frac{(\sum x)^2}{n}\right]\left[\sum y^2 - \frac{(\sum y)^2}{n}\right]}} \tag{11.1}$$

In Excel, we can calculate the correlation coefficient directly using the function **CORREL**(array1, array2). In Section 11.5.2, when we discuss regression analysis, we present Equation 11.46, which rewrites this equation based on the analysis of variance. It is also applied in Section 11.5.7 and Example 11.7.

(d) Interpretation of the correlation coefficient r and testing for significance

Basic

You already know that an r value of $+1$ or -1 indicates a perfect linear correlation and an r value equal to zero indicates the absence of a linear correlation between the two variables. However, in your study, you are likely to obtain values that lie between these extreme situations. How can you interpret these intermediate values? Of course, you also know that the closer r is to $+1$ or -1, the stronger is the linear correlation, and the closer it is to zero, the weaker is the linear correlation. This helps, but you may feel that it does not provide a clear answer to your most important questions: 'is my correlation coefficient high or low', or 'is my correlation strong or weak'?

Unfortunately, there is *no single straightforward answer to these questions*, and the interpretation and subsequent conclusion will depend on the knowledge you have about your system, the relative accuracy you expect from such an estimate, the number of data points you have, and the shape of the spread of the data in your scatter plot. Remember that you are dealing with the environmental data, and thus, in many cases, you should not expect high correlations. For instance, a value of $r = 0.5$ might suffice in some cases, while a value of $r = 0.7$ might be considered a better indicator of high correlation in other cases. Still, if you are calibrating a piece of lab equipment with known added values of a calibration solution, you should expect a much higher value of r, e.g., $r > 0.9$.

In the informal literature (e.g. websites), you will probably find several proposed '**rules of thumb**' that indicate the strength of a correlation based on a classification system for r values, such as the one shown below (there are other variants):

- $r = 0$: *no correlation*
- $0 < r < 0.4$ (or $-0.4 < r < 0$): *weak correlation*
- $0.4 \leq r < 0.7$ (or $-0.7 < r \leq -0.4$): *moderate correlation*
- $0.7 \leq r < 1.0$ (or $-1.0 < r \leq -0.7$): *strong correlation*
- $r = 1$ (or $r = -1$): *perfect correlation*

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

This approach of presenting 'rules of thumb' is not adopted by most statistical textbooks, because of several intervening factors mentioned above. We are not saying that you should not use them (it is difficult to avoid the use of such a simple classification); we are only providing a *word of caution* and suggesting that you do not rely exclusively on such fixed ranges, without a deeper consideration of the system you are studying. There are other options that you can use to interpret your value of r.

Our proposal here – which is the same one adopted in most statistical textbooks – is that you perform a **hypothesis test** *on your correlation coefficient* to test whether it is significantly different from zero. The theory of hypothesis tests has already been extensively discussed in Chapter 10, and you should look for the theoretical support there. A basis for a hypothesis test on the correlation coefficient is detailed below.

C. 10

Our correlation coefficient $\boldsymbol{r}$ measures the correlation between x and y values in the **sample**, and we expect that a similar linear correlation coefficient exists for the population from which our samples were extracted. The **population** correlation coefficient is represented by $\boldsymbol{\rho}$ (Greek letter *rho*). We estimate that the population correlation coefficient using a sample statistic, the correlation coefficient r (e.g., Equation 11.1).

In this case, the traditional test hypothesis involves working with a distribution of r for a situation where $\rho = 0$, thus enabling the formulation of the following test hypotheses:

- **Null hypothesis H_0**: $\rho = 0$ (there is no correlation between the variables)
- **Alternative hypothesis H_a**: $\rho \neq 0$ (there is a significant correlation between the variables)

To measure the intensity of the association observed between the two variables, we need to test whether this correlation is greater than could be expected by chance. Therefore, we test the null hypothesis, H_0: $\rho = 0$, against the alternative hypothesis, H_a: $\rho \neq 0$.

The test statistic to determine the existence of a significant linear correlation is given by:

$$t = \frac{r - \rho}{\sqrt{\frac{1 - r^2}{n - 2}}} \tag{11.2}$$

where
r = correlation coefficient of the sample
ρ = correlation coefficient of the population (adopted as zero in the null hypothesis)
n = number of pairs of x–y data.

The t statistic follows a t distribution with $n - 2$ degrees of freedom (df $= n - 2$). We use a *one-sample two-tailed t test*.

The test decision is as follows: reject H_0 if the absolute value of $\boldsymbol{t_{calc}}$ (given by Equation 11.2) is greater than $\boldsymbol{t_{crit}}$ (function of the significance level α and the degrees of freedom). In other words, **reject H_0 if $|t_{calc}| > t_{\alpha;n-2}$**.

The value of $\boldsymbol{t_{crit}}$ can be obtained from the Excel function **T.INV.2T**(probability; deg_freedom).

We can also calculate the associated ***p*-value**. For this, we use the Excel function **T.DIST.2T**(x; deg_freedom) = T.DIST.2T(t_{calc}; $n - 2$).

A detailed calculation of r is shown in Example 11.1. The Excel spreadsheet associated with the example performs the calculations automatically.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

S. 11.5

In Section 11.5, when we present the simple regression analysis, the discussion will be similar to the one above, but we will focus instead on the **slope of the regression line**. We will see that *both approaches lead to equivalent results in terms of the significance test*.

Here, we come back again to the limitations of the commonly used 'rules of thumb'. If you carry out a hypothesis test as shown above, you may find one of the following situations: (i) it is possible to have a ***strong*** *correlation that is* ***not significant*** and (ii) it is possible to have a ***weak*** correlation that ***is significant***.

Figure 11.5 plots the distribution of r for $\rho = 0$. You can see how increasing sample size (n) makes the distribution taller and skinnier, and more likely to lead to a significant result when using the t test, even if the correlation is 'weak'. This example is also provided as an Excel spreadsheet, and you can change n for yourself and see the resulting distribution of r.

Excel

(e) Advanced procedure: establishing confidence limits for the value of *r*

Advanced

You may be still a bit frustrated with the fact that the hypothesis test described above could lead to a conclusion that is not intuitive to you, for instance, when we said that it is possible to have a ***strong*** *correlation that is* ***not significant***.

Now the question is: *Can we make a hypothesis test for a value of* ρ *that is different from zero*? Can we test with one of the values proposed in the rules of thumb for r, which classifies correlations as weak, moderate, or strong? Can we establish confidence limits for the values of r? The answer to all of these questions is 'yes, we can, but it requires a more advanced procedure'. Nevertheless, it is not difficult to implement, given the knowledge base we have already developed on the topic of hypothesis testing, and we will describe how to conduct these hypothesis tests below.

Procedure for samples with $n > 50$

The hypothesis test shown in item 'd' applies only when the null hypothesis is that $\rho = 0$, so it cannot be applied to test a null hypothesis for ρ equal to another value, other than zero.

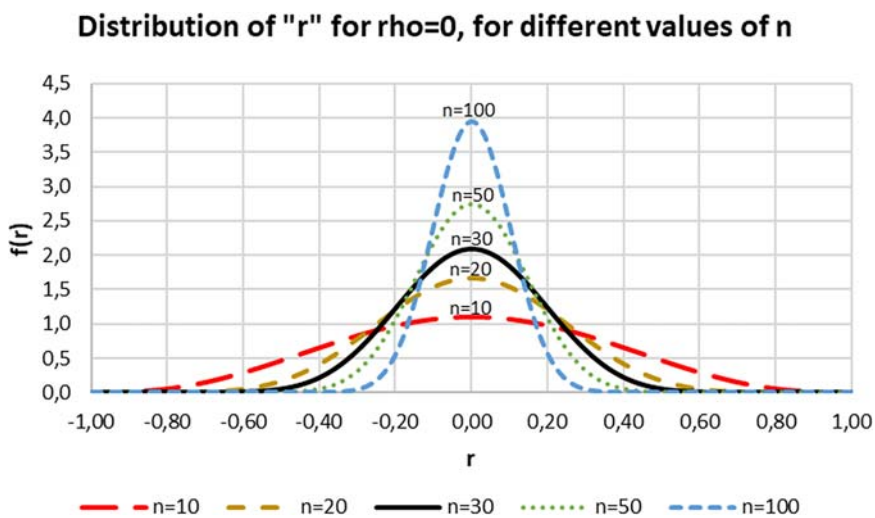


Figure 11.5 Distribution of the correlation coefficient 'r' for $\rho = 0$ (null value of the population correlation coefficient) for different values of the sample size (n).

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

For ρ close to ± 1.0, the distribution of sample values of r is markedly asymmetrical, and the equation should not be applied unless the sample is very large ($n > 500$). To overcome this difficulty, we transform r to a function z, developed by Fisher (1915). The formula for z is as follows:

$$z = 0.5 \ln\left(\frac{1+r}{1-r}\right) \tag{11.3}$$

As discussed in detail in the study by Sokal and Rohlf (1995), this expression is recognized as $z = \tanh^{-1} r$, the formula for the *inverse hyperbolic tangent of r*. You can use the Excel function **TANH**, which returns the hyperbolic tangent of a number. In this case, the *hyperbolic tangent of z will return the value of r*.

Inspection of Equation 11.3 shows that when $r = 0$, z also equals zero, since $0.5 \times \ln(1)$ equals zero. As r approaches 1, the quotient $(1 + r)/(1 - r)$ approaches infinity and, consequently, z approaches infinity. Therefore, substantial differences between r and z occur at the higher values of r. Thus, when $r = 0.1$, $z = 0.100$; when $r = -0.5$, $z = -0.549$; and when $r = 0.9$, $z = 1.472$. From these examples, you can see that the closer r is to 1, the more z departs from the value of r.

For values of r between 0 and 1, the corresponding values of Fisher's z will lie between 0 and ∞; and for values of r from 0 to -1, the corresponding z values will fall between 0 and $-\infty$.

The advantage of the z-transformation is that, while correlation coefficient values are distributed in a skewed shape for values of $\rho \neq 0$, the values of z are approximately normally distributed for any value of its parameter, which is called ζ (zeta), following the usual convention. The expected variance of z is calculated as follows:

$$\sigma_z^2 = \frac{1}{n-3} \tag{11.4}$$

Therefore, the standard deviation of z is as follows:

$$\sigma_z = \frac{1}{\sqrt{n-3}} \tag{11.5}$$

An interesting aspect about the variance of z, based on Equation 11.4, is that it is independent of the value of r and it is simply a function of the sample size n.

The **critical value of *t*** (t_{crit}) can be calculated for a two-tailed t test using the following Excel function, adopting infinity as the number of degrees of freedom:

T.INV.2T (probability; degrees of freedom) = T.INV.2T (α; ∞). For practical purposes, infinite can be replaced by a very large number in the Excel function (e.g., 10,000,000,000).

Having an infinite number of degrees of freedom in the inverse t function is equivalent to adopting the absolute value of the inverse of the standard normal variable for $\alpha/2$:

ABS(NORM.S.INV(α/2)). For the typical α values used in hypothesis tests, we have: for $\alpha = 0.05 \rightarrow t_{crit} = 1.960$; for $\alpha = 0.10 \rightarrow t_{crit} = 1.645$; and for $\alpha = 0.01 \rightarrow t_{crit} = 2.576$.

Therefore, for $\alpha = 0.05$, $t_{crit} = 1.960$, or $t_{0.05;\,\infty} = 1.960$ (which is the same as $Z_{0.05/2} = 1.960$).

To obtain the confidence limits of r, we convert the sample r to z (Equation 11.3), calculate the confidence limits for z, and then transform these limits back to the r-scale.

The **confidence limits for *z*** (for $\alpha = 0.05$) are calculated as follows:

$$95\%\ \text{confidence limits}: z \pm t_{0.05;\infty} \times \sigma_z \tag{11.6}$$

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

With z obtained from Equation 11.3, σ_z obtained from Equation 11.4, and the critical value of t obtained above, we calculate the lower and upper confidence limits (UCLs) for z using Equations 11.7 and 11.8.

$$\text{Lower limit for } Z\text{: } \mathrm{LL}_Z = z - t_{0.05,\infty}\frac{1}{\sqrt{n-3}} \tag{11.7}$$

$$\text{Upper limit for } Z\text{: } \mathrm{UL}_Z = z + t_{0.05,\infty}\frac{1}{\sqrt{n-3}} \tag{11.8}$$

Now, we retransform these z-values (obtained from Equations 11.7 and 11.8) to the r-scale by means of the hyperbolic tangent function:

$$\text{Lower limit for } r\text{: } \mathrm{LL}_r = \tanh \mathrm{LL}_Z = (e^{\mathrm{LL}_Z} - e^{-\mathrm{LL}_Z})/(e^{\mathrm{LL}_Z} + e^{-\mathrm{LL}_Z}) \tag{11.9}$$

$$\text{Upper limit for } r\text{: } \mathrm{UL}_r = \tanh \mathrm{UL}_Z = (e^{\mathrm{UL}_Z} - e^{-\mathrm{UL}_Z})/(e^{\mathrm{UL}_Z} + e^{-\mathrm{UL}_Z}) \tag{11.10}$$

You can also use the Excel function **TANH**, which returns the hyperbolic tangent of a number. You apply TANH() to the values of Z for LL_z and UL_z and obtain directly the values of LL_r and UL_r.

Thus, the 95% confidence limits around r are LL_r and UL_r.

In item 'd' above, you carried out a hypothesis test to verify whether the population correlation coefficient (ρ) was significantly different from zero. Now let us suppose that **you want to test whether ρ is equal to a different value ρ_0**, say, one of the values that compose the *rules of thumb* also presented in item 'd'. The interpretation is as follows (see also Figure 11.6):

- You have a **$1 - \alpha$ confidence interval** that the value of ρ_0 should be situated **between the lower limit LL_r and the upper limit UL_r**. You can compare these values with those proposed in the *rules of thumb*.
- If you are testing a value of ρ_0 that is **below the lower limit LL_r**, this will indicate that there is a **significant difference between ρ_0 and the sample correlation coefficient r**.
- If you are testing a value of ρ_0 that is **above the upper limit UL_r**, this will indicate that there is a **significant difference between ρ_0 and the sample correlation coefficient r**.
- If you are testing a value of ρ_0 that is **between the lower limit LL_r and the upper limit UL_r**, this will indicate that there is **no significant difference between ρ_0 and the sample correlation coefficient r**.
- If your **confidence interval includes the value of $r = 0$**, this will indicate that the **population correlation coefficient ρ is not significantly different from zero**. This should have been already detected in the traditional hypothesis test (item 'd'), which used the null hypothesis H_0: $\rho = 0$.

Procedure for samples with n between 10 and 50

The procedure shown above is for samples with $n > 50$. For samples with $n \leq 50$ (but greater than 10), we need to introduce some adaptations. However, we expect that the test outcomes from both approaches are not likely to be very different unless the critical values of the test statistic are near the threshold for significance.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

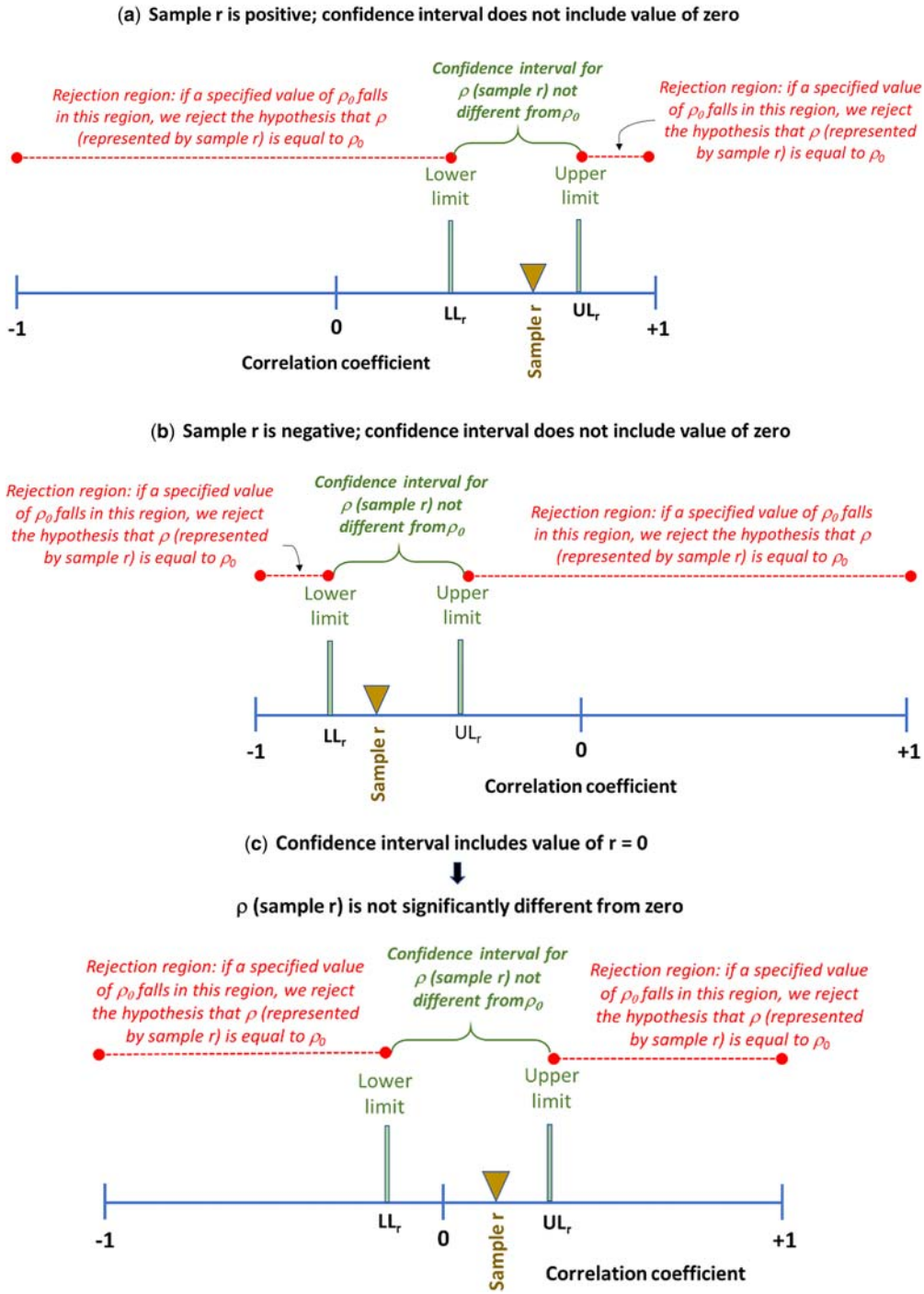


Figure 11.6 Interpretation of confidence intervals and rejection regions for the sample correlation coefficient *r*.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

For small samples, calculating exact probabilities is difficult. The following modified z-transformation has been suggested by Hotelling (1953) for use in small samples, as cited by Sokal and Rohlf (1995). We calculate modified z^* and σ^* according to the following equations:

$$z^* = z - \frac{3z + r}{4(n-1)} \tag{11.11}$$

$$\sigma^* = \frac{1}{\sqrt{(n-1)}} \tag{11.12}$$

The distribution of z^* is closer to a normal distribution than z. However, **this transformation should not be used for small sample sizes ($n < 10$)**.

To obtain the confidence limits for r, we convert the sample r to z (Equation 11.3); calculate z^* (Equation 11.11), σ^* (Equation 11.12), and the confidence limits for z^* (Equation 11.13); and then transform these values back to the r-scale.

The **confidence limits for z^*** (for $\alpha = 0.05$) are calculated as follows:

$$95\%\,\text{confidence limits}: z^* \pm t_{0.05;\infty} \times \sigma^* \tag{11.13}$$

The lower and UCLs for z^* are calculated using Equations 11.14 and 11.15.

$$\text{Lower limit for } Z^*: \text{LL}_{Z^*} = z^* - t_{0.05,\infty}\frac{1}{\sqrt{n-1}} \tag{11.14}$$

$$\text{Upper limit for } Z^*: \text{UL}_{Z^*} = z^* + t_{0.05,\infty}\frac{1}{\sqrt{n-1}} \tag{11.15}$$

Now, we retransform these z^*-values (obtained from Equations 11.14 and 11.15) to the r-scale by means of the hyperbolic tangent function (TANH Excel function):

$$\boxed{\text{Lower limit for } r: \text{LL}_r = \tanh \text{LL}_{Z^*} = (e^{\text{LL}_{Z^*}} - e^{-\text{LL}_{Z^*}})/(e^{\text{LL}_{Z^*}} + e^{-\text{LL}_{Z^*}})} \tag{11.16}$$

$$\boxed{\text{Upper limit for } r: \text{UL}_r = \tanh \text{UL}_{Z^*} = (e^{\text{UL}_{Z^*}} - e^{-\text{UL}_{Z^*}})/(e^{\text{UL}_{Z^*}} + e^{-\text{UL}_{Z^*}})} \tag{11.17}$$

The interpretation is the same as the one shown above (e.g., in Figure 11.6).

Example 11.1 illustrates the whole calculation sequence, for sample sizes greater than 50 and for sample sizes between 10 and 50. The Excel spreadsheet associated with the example performs the calculations automatically.

(f) Advanced procedure: hypothesis test $\rho = \rho_0$

Advanced

Following the determination of the confidence limits for r, we may continue with this analysis and formulate hypotheses that are different from the traditional one (H_0: $\rho = 0$). Now, using some concepts described in item 'e', we may test whether the correlation coefficient (ρ) is equal to any other value we specify (ρ_0). The test hypotheses are as follows:

- **Null hypothesis H_0**: $\rho = \rho_0$
- **Alternative hypothesis H_a**: $\rho \neq \rho_0$

To test the hypothesis $\rho = \rho_0$, where $\rho_0 \neq 0$, we cannot use the t-test, but must make use of the z-transformation and then use the following expressions for obtaining the value of the test statistic t_s (t_{calc}).

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Procedure for $n > 50$

The test statistic t_s (t_{calc}) is given by Equations 11.18 and 11.19:

$$t_s = \frac{z - \zeta}{\frac{1}{\sqrt{n-3}}} = (z - \zeta)\sqrt{n-3} \tag{11.18}$$

$$\zeta = 0.5 \ln\left(\frac{1+\rho}{1-\rho}\right) \tag{11.19}$$

where z and ζ (zeta) are the transformations of r and ρ, respectively. In Equation 11.19, in the place of ρ, we use the value of ρ_0 we want to test for the null hypothesis. Then, we compare the t_s value with the critical value of the t-distribution, $t_{\alpha,\,\infty}$ (t_{crit}). Note that the appropriate degrees of freedom for the z-transformation is infinity.

Procedure for n between 10 and 50

We calculate the test statistic t_s (t_{calc}) based on the transformation of z to z^* and ζ to ζ^*:

$$t_s = (z^* - \zeta^*)\sqrt{n-1} \tag{11.20}$$

$$\zeta^* = \zeta - \frac{(3\zeta + \rho)}{4n} \tag{11.21}$$

Comparison of t_{calc} and t_{crit} (for $n > 50$ and for n between 10 and 50)

In both cases ($n > 50$ and n between 10 and 50), we obtain the value of t_{crit} as already demonstrated in item 'e'. The **critical value of t** (t_{crit}) can then be calculated for a two-tailed t test using the following Excel function, adopting infinity as the number of degrees of freedom:

T.INV.2T (probability; degrees of freedom) = T.INV.2T (α; ∞). For practical purposes, infinity can be replaced by a very large number in the Excel function (e.g., 10,000,000,000).

As in all hypothesis tests, we compare t_{calc} with t_{crit}, or, in this case, t_s with $t_{\alpha,\,\infty}$.

- If $t_{calc} > t_{crit}$: *reject null hypothesis* and conclude that **ρ is significantly different from the specified value of ρ_0** (at a confidence level of $1 - \alpha$).
- If $t_{calc} \leq t_{crit}$: do not reject null hypothesis and accept that **ρ is not significantly different from the specified value of ρ_0** (at a confidence level of $1 - \alpha$).

***p*-value**

The p-value for the test may be obtained using the following Excel function for a two-tailed t distribution:

$$p\text{-value} = \mathbf{TDIST}(x, \text{deg_freedom, tails}) = \text{TDIST}(\text{ABS}(t_{calc}); \text{infinity}; 2)$$
$$\cong \text{TDIST}(\text{ABS}(t_{calc}); 10000000000; 2)$$

This is equivalent to using the standard normal variable distribution:

$$p\text{-value} = 2 \times (1 - \text{NORMSDIST}(\text{ABS}(t_{calc})))$$

As in the other hypothesis tests, you reject the null hypothesis H_0 if the p-value is less than the significance level α.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Example 11.1 illustrates the whole calculation sequence, for sample sizes greater than 50 and between 10 and 50. The Excel spreadsheet associated with the example performs the calculations automatically.

Example

EXAMPLE 11.1 EXAMPLE OF THE CALCULATION OF PEARSON'S COEFFICIENT OF CORRELATION (*R*)

Suppose you collected data for two water constituents in a river and you want to test whether the two concentrations are linearly correlated. You obtained 20 paired values of constituent X and constituent Y ($n = 20$). Calculate the Pearson coefficient of correlation r and perform hypothesis tests on this coefficient to determine whether it is a significant correlation.

The data are shown in the table below.

Measured values of constituents *X* and *Y*

Sample Number	Constituent *X* (mg/L)	Constituent *Y* (mg/L)	Sample Number	Constituent *X* (mg/L)	Constituent *Y* (mg/L)
1	4.7	6.9	11	6.9	7.4
2	5.2	7.7	12	7.5	7.6
3	5.1	7.4	13	7.7	7.8
4	4.7	6.8	14	7.1	8.3
5	3.5	6.3	15	7.5	8.6
6	3.3	5.2	16	7.3	8.7
7	3.8	5.4	17	6.8	7.7
8	4.0	6.0	18	5.2	7.0
9	5.9	6.6	19	4.9	6.8
10	7.3	7.3	20	4.3	6.6

Excel

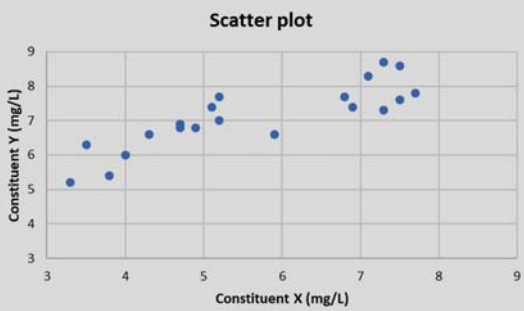
Note: This example is also available as an Excel spreadsheet.

Solution:

In most cases, you will do the analyses only related to items 'a' and 'b' below, but it is worthwhile also including item 'c'. For a more advanced evaluation, you might also incorporate the procedures described in items 'd' and 'e'.

(a) Scatterplot of the data

The first step in examining the relation between *y* and *x* is to build a scatterplot. The plot is shown below.



Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

The plot indicates that there is an imperfect but generally increasing relation between x and y. A linear (straight-line) relation appears plausible, and there is no evidence of the need to make transformations in the data. Also, there is no detection of an outlier falling far from the general pattern of the data. As a result, we continue with the study of the linear correlation between the two variables.

(b) Data and calculations for computing r

The required calculations for obtaining the value of r are shown in the table below.

Computational table to calculate r

Sample Number	Constituent X (mg/L)	Constituent Y (mg/L)	$x \cdot y$	x^2	y^2
1	4.7	6.9	32.4	22.1	47.6
2	5.2	7.7	40.0	27.0	59.3
3	5.1	7.4	37.7	26.0	54.8
4	4.7	6.8	32.0	22.1	46.2
5	3.5	6.3	22.1	12.3	39.7
6	3.3	5.2	17.2	10.9	27.0
7	3.8	5.4	20.5	14.4	29.2
8	4.0	6.0	24.0	16.0	36.0
9	5.9	6.6	38.9	34.8	43.6
10	7.3	7.3	53.3	53.3	53.3
11	6.9	7.4	51.1	47.6	54.8
12	7.5	7.6	57.0	56.3	57.8
13	7.7	7.8	60.1	59.3	60.8
14	7.1	8.3	58.9	50.4	68.9
15	7.5	8.6	64.5	56.3	74.0
16	7.3	8.7	63.5	53.3	75.7
17	6.8	7.7	52.4	46.2	59.3
18	5.2	7.0	36.4	27.0	49.0
19	4.9	6.8	33.3	24.0	46.2
20	4.3	6.6	28.4	18.5	43.6
Σ	**112.7**	**142.1**	**823.7**	**677.8**	**1026.6**
Mean	**5.6**	**7.1**			

The correlation between x and y is computed using Equation 11.1:

$$r = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\left[\sum x^2 - \frac{(\sum x)^2}{n}\right]\left[\sum y^2 - \frac{(\sum y)^2}{n}\right]}} = \frac{823.7 - (112.7 \times 142.1)/20}{\sqrt{\left[677.8 - \frac{(112.7)^2}{20}\right]\left[1026.6 - \frac{(142.1)^2}{20}\right]}} = \mathbf{0.850}$$

If we use the Excel function CORREL(array1, array2), in which array1 is the 20 data points for constituent X and array2 is the 20 data points for constituent Y, we also obtain $r = 0.850$.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

(c) Hypothesis test for linear correlation ($\rho = 0$)

Our test hypotheses are as follows:

- Null hypothesis H_0: $\rho = 0$
- Alternative hypothesis H_a: $\rho \neq 0$

The test statistic (t_{calc}) is given by Equation 11.2:

$$t_{calc} = \frac{r - \rho}{\sqrt{\frac{1 - r^2}{n - 2}}} = \frac{0.850 - 0}{\sqrt{\frac{1 - 0.850^2}{20 - 2}}} = \frac{0.850}{0.124} = \mathbf{6.848}$$

The critical value, for a significance level $\alpha = 0.05$, and degrees of freedom df $= n - 2 = 20 - 2 = 18$, is obtained from the Excel function T.INV.2T(probability; deg_freedom) = T.INV.2T(0.05; 18) = **2.101**.

Since $t_{calc} > t_{crit}$, $6.848 > 2.101$, we reject the null hypothesis H_0 that $\rho = 0$, thus rejecting the hypothesis that there is no linear correlation between the two variables and accepting the alternative hypothesis that **there is a significant linear correlation between the constituents *X* and *Y*.**

The conclusion about the hypothesis test can be also obtained by using the concept of the *p*-value. For the *t* test, Excel has a function that allows direct calculation of the *p*-value: T.DIST.2T (*x*; deg_freedom) = T.DIST.2T(ABS(t_{calc}); $n - 2$), where ABS(t_{calc}) is the absolute value of t_{calc}. In our example, we have:

***p*-value** = T.DIST.2T(ABS(6.848); 18) = $\mathbf{2.081 \times 10^{-6}}$

Since this *p*-value is lower than our significance level ($\alpha = 0.05$), we reject the null hypothesis H_0. Again, we are able to accept the alternative hypothesis that **there is a significant linear correlation between the constituents *X* and *Y*.**

(d) Advanced approach for establishing confidence limits for the value of the sample correlation coefficient *r*

S. 11.2.1

We will follow here the procedure outlined in Section 11.2.1(e). There are procedures when your sample size is large ($n > 50$) and when it has an intermediate size (n between 10 and 50). Even though our sample size in this example is $n = 20$, we will carry out both calculations, in order to demonstrate the procedure for both methods. The Excel spreadsheet associated with this example performs all calculations automatically and includes graphs to facilitate the interpretation.

We will establish the confidence limits for a 95% confidence level. Therefore our significance level is 5%, that is, $\alpha = 0.05$ (as previously adopted in this example).

Procedure for $n > 50$

Initially, we calculate *z*, using Equation 11.3, and knowing that $r = 0.850$ (calculated above):

$$z = 0.5 \ln\left(\frac{1 + r}{1 - r}\right) = 0.5 \ln\left(\frac{1 + 0.85}{1 - 0.85}\right) = 1.256$$

We then calculate the critical value of the *t* statistic. The **critical value of *t*** (t_{crit}) can be calculated for a two-tailed *t* test using the following Excel function, adopting infinity as the number of degrees of freedom:

T.INV.2T (probability; degrees of freedom) = T.INV.2T (α; ∞). For practical purposes, infinity can be replaced by a very large number in the Excel function (e.g., 10,000,000,000). Therefore, for $\alpha = 0.05$, $t_{crit} = 1.960$, or $t_{0.05;\,\infty} = \mathbf{1.960}$.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

To calculate the confidence limits, we first convert the sample r to z, set confidence limits to z, and then transform these limits back to the r-scale. The information we need has been already calculated or established: $r = 0.850$, $n = 20$, $z = 1.256$, $\alpha = 0.05$ (for a 95% confidence interval), $t_{0.05;\,\infty} = 1.960$.

The confidence limits for z are calculated from Equations 11.7 and 11.8:

$$\text{Lower limit for } Z\text{: } LL_Z = z - t_{0.05,\infty}\frac{1}{\sqrt{n-3}}$$

$$= 1.256 - \frac{1.960}{\sqrt{20-3}} = 0.781$$

$$\text{Upper limit for } Z\text{: } UL_Z = z + t_{0.05,\infty}\frac{1}{\sqrt{n-3}} = 1.256 + \frac{1.960}{\sqrt{20-3}}$$

$$= 1.732$$

Now, we retransform these z-values to the r-scale by means of the hyperbolic tangent function:

$$\textbf{Lower limit for } r\textbf{: } LL_r = \tanh LL_Z = (e^{LL_Z} - e^{-LL_Z})/(e^{LL_Z} + e^{-LL_Z})$$

$$= (e^{0.781} - e^{-0.781})/(e^{0.781} + e^{-0.781}) = \mathbf{0.653}$$

$$\textbf{Upper limit for } r\textbf{: } UL_r = \tanh UL_Z = (e^{UL_Z} - e^{-UL_Z})/(e^{UL_Z} + e^{-UL_Z})$$

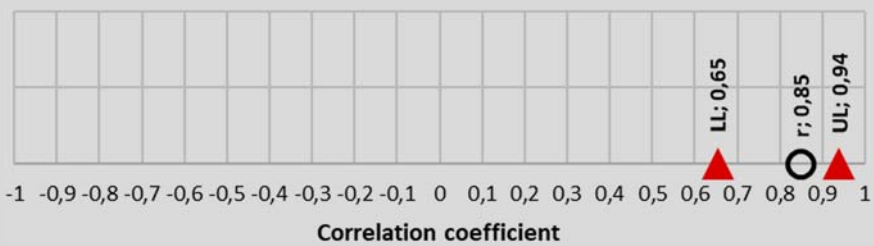
$$= (e^{1.732} - e^{-1.732})/(e^{1.732} + e^{-1.732}) = \mathbf{0.939}$$

You can also use the Excel function **TANH**, which returns the hyperbolic tangent of a number. You apply TANH() to the values of Z for LL_z and UL_z and obtain directly the values of LL_r and UL_r. TANH(0.781) = 0.653; TANH(1.732) = 0.939.

Thus, the 95% confidence limits around $r = 0.850$ are **0.653** and **0.939**

The following figure shows these results. You can interpret it using Figure 11.6a (positive correlation). Since the 95% confidence interval ranges from 0.65 to 0.94, this means that the true correlation of these two constituents in the *population* will be within this range, with 95% confidence. You can consult the rules of thumb available in informal statistical tests (such as those exemplified in Section 11.2.1(d)) and see whether the proposed values for a weak, intermediate, or strong correlation are inside or outside these limits. Review Section 4.5.3 for more details about the meaning of a confidence interval. The same concepts apply for our confidence interval around the estimated correlation coefficient.

S. 11.2.1



To increase your understanding about these concepts, try changing the sample size (value of n) in the spreadsheet. It is equal to 20 in this example. Put, for instance, a value of 10, and you will see that the confidence limits will become wider. After that, put a value of 100, and see how the width of the confidence interval decreases.

Also note that the lower and upper limits of r (0.653 and 0.939) are not equidistant around the value of r (0.850). The upper and lower values of z are equidistant around z, but when we transform them into the r-scale, they may be not equidistant anymore.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Procedure for *n* between 10 and 50

Based on the values of r (0.850), n (20), and z (1.256), we estimate the values of z^* and σ^* using Equations 11.11 and 11.12:

$$z^* = z - \frac{3z + r}{4(n-1)} = 1.256 - \frac{3(1.256) + 0.850}{4(20-1)} = 1.1957$$

$$\sigma_{z^*} = \frac{1}{\sqrt{(n-1)}} = \frac{1}{\sqrt{(20-1)}} = 0.2294$$

We calculate the lower and upper limits for z^* using Equation 11.13, or their developed versions, Equations 11.14 and 11.15. To use them, we need the value of t_{crit}, or $t_{0.05;\,\infty}$, which was determined above as **1.960**.

$$\text{Lower limit for } Z^* : \text{LL}_{Z^*} = z^* - t_{0.05,\infty}\frac{1}{\sqrt{n-1}}$$

$$= 1.1957 - 1.960 \times 0.2294$$

$$= 0.746$$

$$\text{Upper limit for } Z^* : \text{UL}_{Z^*} = z^* + t_{0.05,\infty}\frac{1}{\sqrt{n-1}}$$

$$= 1.1957 + 1.960 \times 0.2294$$

$$= 1.645$$

Similarly to what we did before, we retransform these z-values to the r-scale by means of the hyperbolic tangent function (tanh function):

$$\textbf{Lower limit for } r\textbf{: LL}_r = \tanh \text{LL}_{Z^*} = (e^{\text{LL}_{Z^*}} - e^{-\text{LL}_{Z^*}})/(e^{\text{LL}_{Z^*}} + e^{-\text{LL}_{Z^*}})$$

$$= (e^{0.746} - e^{-0.746})/(e^{0.746} + e^{-0.746}) = \textbf{0.633}$$

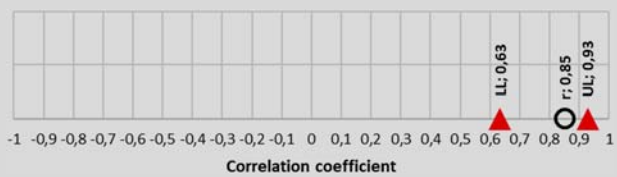
$$\textbf{Upper limit for } r\textbf{: UL}_r = \tanh \text{UL}_{Z^*} = (e^{\text{UL}_{Z^*}} - e^{-\text{UL}_{Z^*}})/(e^{\text{UL}_{Z^*}} + e^{-\text{UL}_{Z^*}})$$

$$= (e^{1.645} - e^{-1.645})/(e^{1.645} + e^{-1.645}) = \textbf{0.928}$$

You can also use the Excel function **TANH**, which returns the hyperbolic tangent of a number. You apply TANH() to the values of Z for LL_z and UL_z and obtain directly the values of LL_r and UL_r. TANH(0.746) = 0.633; TANH(1.645) = 0.928.

Thus, the 95% confidence limits around $r = 0.850$ are **0.633** and **0.928**. Notice that these values are very similar to those obtained for samples larger than 50 ($n > 50$), reinforcing the comment that both approaches are likely to lead to similar overall results and conclusions, unless you are testing for a threshold value close to the limits.

The figure below shows the values obtained for the confidence limits for r. The interpretation is similar to the one made above (for $n > 50$).



(e) Advanced approach for testing a null hypothesis that the correlation coefficient (ρ), or the sample r, is equal to any value we specify (ρ_0)

In item 'c' of this example, we tested whether our correlation coefficient was significantly different from zero, or, in other words, whether our linear correlation could be considered

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

significant. Now, suppose we want to test a different value other than zero, say, one of the values found in the 'rules of thumb'. Let us suppose that we want to test whether our correlation coefficient is significantly different from **0.70**, which is the boundary value between an intermediate and a strong correlation, as suggested by one of the available rules of thumb for *r*.

As a matter of fact, we would not need to do any further test. If we observe the confidence limits calculated above (item 'd' of this example), we see that 0.70 is inside the limits of the confidence interval. In other words, we could have already concluded that our sample correlation coefficient *r* (0.85) is **not** significantly different from 0.70. A similar conclusion could be obtained for the value of 0.90, which is also inside the confidence limits. However, if we wanted to compare with the value of 0.95, we would see that it is outside the limits, and therefore, we would say that our correlation coefficient *r* (0.85) is significantly different from 0.95.

Nevertheless, we will carry out a hypothesis test to deepen our knowledge about the correlation between constituents *X* and *Y*. For this, we will go back to the value of **0.70** as the threshold against which we want to test our correlation coefficient *r* (0.85).

We need to establish our null and alternative hypotheses. In general, they are as follows:

- Null hypothesis H_0: $\rho = \rho_0$
- Alternative hypothesis H_a: $\rho \neq \rho_0$

In our case, we make ρ_0 equal to 0.70, and our hypotheses become

- Null hypothesis H_0: $\rho = 0.70$
- Alternative hypothesis H_a: $\rho \neq 0.70$

S. 11.2.1

We will follow the procedure described in Section 11.2.1(f). We will split the calculations into the two possibilities ($n > 50$ and n between 10 and 50).

Procedure for $n > 50$

We use Equations 11.18 and 11.19 to estimate the **test statistic t_s (t_{calc})**. The value of *z* had already been calculated as 1.256 in item 'd' of this example.

$$\zeta = 0.5 \ln\left(\frac{1+\rho}{1-\rho}\right) = 0.5 \ln\left(\frac{1+0.70}{1-0.70}\right) = 0.867$$

$$t_s = (z - \zeta)\sqrt{n-3} = (1.256 - 0.867) \times \sqrt{20-3} = \mathbf{1.604}$$

We now compare the absolute value of t_{calc} with the **critical value t_{crit}** (which was calculated in item 'd' of this example as $t_{crit} = 1.960$, or $t_{0.05;\,\infty} = \mathbf{1.960}$).

Since $|t_{calc}| < t_{crit}$, or $|t_s| = 1.604 < t_{0.05;\,\infty} = 1.960$, **we do not reject H_0**. In other words, we do not reject the hypothesis that our sample correlation coefficient $r = 0.850$ (representing the population correlation coefficient ρ) is not significantly different from the specified value of $\rho_0 = 0.70$.

As we had mentioned above, we already knew this conclusion, simply by inspection of the confidence limits calculated in item 'd'. We saw that 0.70 was inside the confidence interval for $r = 0.850$, indicating that they were not significantly different.

The ***p*-value** for the test may be obtained using the following Excel function for a two-tailed *t* distribution:

***p*-value** = **TDIST**(x, deg_freedom, tails) = TDIST(ABS(t_{calc}); infinity; 2)

$\cong$ TDIST(ABS(1.604); 10000000000; 2) = **0.109**

Since this *p*-value is greater than the significance level adopted ($\alpha = 0.05$), we **do not reject the null hypothesis**.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Procedure for *n* between 10 and 50

We calculate the test statistic t_s (t_{calc}) based on the transformation of z to z^* and ζ to ζ^*, using Equations 11.20 and 11.21. The value of z^* has already been calculated as 1.1957 in item 'd' and ζ was estimated as 0.867 just above. The value of ρ_0 adopted for the test is 0.70.

$$\zeta^* = \zeta - \frac{(3\zeta + \rho)}{4n} = 0.867 - \frac{(3 \times 0.867 + 0.70)}{4 \times 20} = 0.826$$

$$t_s = (z^* - \zeta^*)\sqrt{n-1} = (1.1957 - 0.826) \times \sqrt{20 - 1} = \mathbf{1.611}$$

This value is very similar to the one calculated for samples with $n > 50$. We now compare the absolute value of t_{calc} with the **critical value t_{crit}** (which was calculated in item 'd' of this example as $t_{crit} = 1.960$, or $t_{0.05;\,\infty} = \mathbf{1.960}$).

Since $|\boldsymbol{t_{calc}}| < \boldsymbol{t_{crit}}$, or $|t_s| = 1.611 < t_{0.05;\,\infty} = 1.960$, **we do not reject H_0**. In other words, we do not reject the hypothesis that our sample correlation coefficient $r = 0.850$ (representing the population correlation coefficient ρ) is equal to the specified threshold value of $\rho_0 = 0.70$.

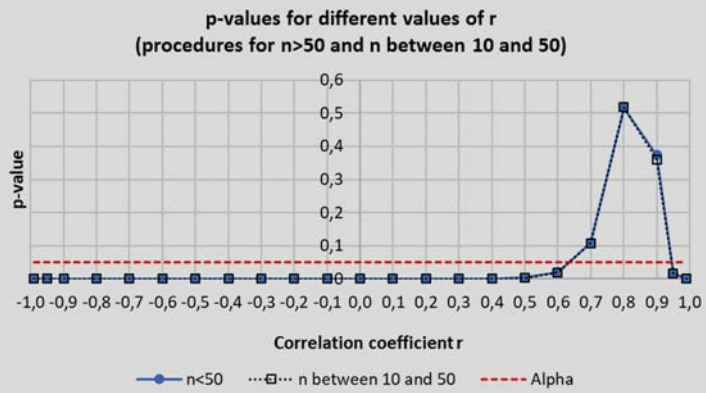
The ***p*-value** for the test may be obtained using the following Excel function for a two-tailed t distribution:

$$\textbf{\textit{p}-value} = \textbf{TDIST}(\text{x, deg_freedom, tails}) = \text{TDIST}(\text{ABS}(t_{calc}); \text{infinity}; 2)$$
$$\cong \text{TDIST}(\text{ABS}(1.611); 10000000000; 2) = \mathbf{0.107}$$

Since this *p-value is greater than the significance level adopted ($\alpha = 0.05$)*, again we **do not reject the null hypothesis** (this conclusion was already reached in the calculations above; calculating the *p*-value is just a different method to arrive at the same conclusion).

Comparison between the procedures for *n* > 50 and *n* between 10 and 50

In the Excel spreadsheet associated with this example, we have also prepared a graph comparing the *p*-values calculated for values of ρ_0 ranging from -0.99 to $+0.99$, according to the two procedures. We can see that, for this particular example, the two methods yield virtually the same results, since both lines overlap. You can also see, from the chart, the values of r that lead to *p*-values greater than $\alpha = 0.05$, indicating the non-rejection region. As expected, the boundaries of this region coincide with the confidence limits calculated and plotted above.



Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Final comment: We showed you different ways of interpreting the value of the correlation coefficient r obtained from your experimental data. Despite the breadth of the statistical methods we presented, it is still up to you to use your best judgment to interpret the results obtained, based on the knowledge you have about the system you are studying.

11.2.2 Spearman rank correlation coefficient (non-parametric)

Advanced

S. 11.2.1

S. 10.2.2

S. 10.4.3

If we have data obtained from a population that shows substantial departures from the normal distribution, then the correlation procedures we saw in Section 11.2.1, using the Pearson coefficient, are not applicable. In this case, we need to use non-parametric methods to assess correlation based on the ranks (order) of the measurements for each variable, instead of their original values. In this section, we demonstrate the utilization of the non-parametric **Spearman rank correlation coefficient**, r_s. A *discussion of* ***parametric*** *and* ***non-parametric*** *tests was done in* Section 10.2.2(c), and you should consult this section again to reinforce your decision about whether to use a parametric or a non-parametric test.

The ranking of a variable was illustrated in Example 10.6, Section 10.4.3. The ranking is internal for each variable – it is not done by grouping the two variables together. Ranking can also be performed using the following Excel function:

RANK.AVG(number; ref; [order])

- **Number**. The number whose rank you want to find.
- **Ref**. Array of all values in the particular sample.
- **Order**. A number specifying how to rank the number (0 or omitted: descending order; any non-zero value: ascending order)

After the measurements of each variable have been ranked, Equation 11.22 is applied to the rank to obtain the **Spearman correlation coefficient** r_s (when there are no ties in the rankings).

$$r_s = 1 - \frac{6\sum d_i^2}{n^3 - n} \tag{11.22}$$

where d_i is the difference between x and y ranks: $d_i = (\text{rank of } x_i) - (\text{rank of } y_i)$.

The value of r_s, representing an estimate of the population rank correlation coefficient, ρ_s, may **range from −1 to +1**, and it is *dimensionless*. Its value will not be the same as the value of the Pearson correlation coefficient r that you may have calculated using the original data instead of their ranks.

If there are **tied data**, then they are assigned average ranks, as done in Example 10.6. The Excel function RANK.AVG already computes the averages of tied data. There are procedures for correcting r_s for the effect of the ties. However, they are more laborious and are necessary only when you have a large number of ties relative to the total sample size n. In our book, for the sake of simplicity, we will not introduce this correction. If you desire to incorporate this factor, please use a specialized statistical software that accounts for this correction factor.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

(a) Assessing the significance of r_s

We can assess the significance of r_s using a critical values 'look-up' table, which is available in most statistical textbooks. In our book, we present a version of this look-up table only for the significance level $\alpha = 0.05$ (see Table 11.2). If you want to conduct your test with other significance levels, please consult additional references.

With your test statistic r_{calc} (r_s) and the critical value r_{crit} ($r_{s\ \alpha;n}$) obtained from Table 11.2, you can test the following hypotheses:

- Null hypothesis H_0: $\rho_s = 0$.
- Alternative hypothesis H_a: $\rho_s \neq 0$.

If r_{calc} is greater than r_{crit} ($r_s > r_{s\ \alpha;n}$), we *reject the null hypothesis*.

Table 11.2 Critical values for the r_s statistic (Spearman correlation coefficient) for a two-tailed test with significance level $\alpha = 0.05$ and number of data points n varying from 5 to 100.

n	r_s crit	n	r_s crit	n	r_s crit	n	r_s crit
1	n/a	26	0.390	51	0.276	76	0.226
2	n/a	27	0.382	52	0.274	77	0.224
3	n/a	28	0.375	53	0.271	78	0.223
4	n/a	29	0.368	54	0.268	79	0.221
5	1.000	30	0.362	55	0.266	80	0.220
6	0.886	31	0.356	56	0.264	81	0.219
7	0.786	32	0.350	57	0.261	82	0.217
8	0.738	33	0.345	58	0.259	83	0.216
9	0.700	34	0.340	59	0.257	84	0.215
10	0.648	35	0.335	60	0.255	85	0.213
11	0.618	36	0.330	61	0.252	86	0.212
12	0.587	37	0.325	62	0.250	87	0.211
13	0.560	38	0.321	63	0.248	88	0.210
14	0.538	39	0.317	64	0.246	89	0.209
15	0.521	40	0.313	65	0.244	90	0.207
16	0.503	41	0.309	66	0.243	91	0.206
17	0.485	42	0.305	67	0.241	92	0.205
18	0.472	43	0.301	68	0.239	93	0.204
19	0.460	44	0.298	69	0.237	94	0.203
20	0.447	45	0.294	70	0.235	95	0.202
21	0.435	46	0.291	71	0.234	96	0.201
22	0.425	47	0.288	72	0.232	97	0.200
23	0.415	48	0.285	73	0.230	98	0.199
24	0.406	49	0.282	74	0.229	99	0.198
25	0.398	50	0.279	75	0.227	100	0.197

Source: Zar (1999), modified.

Note: The test is not possible for sample sizes $n < 5$.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

(b) Assessing significance of r_s using the t test

There is also a second way to compute an approximate value of t without using the critical values table. To get an approximate value for the test statistic t (t_{calc}), we use the following equation:

$$t_{calc} \approx \sqrt{\frac{r_s^2 \times df}{1 - r_s^2}} \tag{11.23}$$

where degrees of freedom $df = n - 2$.

The critical value of t (t_{crit}) is obtained for a two-tailed test as a function of the significance level α (usually 0.05) using the Excel function:

T.INV.2T(probability; deg_freedom) = **T.INV.2T**(α; $n - 2$).

If t_{calc} is greater than t_{crit} ($t > t_{\alpha;n-2}$), we *reject the null hypothesis*.

The p-value for the test may be obtained using the following Excel function for a two-tailed t distribution:

***p*-value** = **T.DIST.2T**(x, deg_freedom) = T.DIST.2T(ABS(t_{calc}); $n - 2$)

As in the other hypothesis tests, you reject the null hypothesis H_0 if p-value is less than the significance level α.

(c) Assessing the significance of r_s using the ranked data with the same procedure for Pearson correlation coefficient

You can also use a third and even more simplified approach, applying the *Pearson correlation procedure* (as in Example 11.1) to your data rank values. For this, the procedures detailed in the Pearson worksheet, including the use of the CORREL Excel function, will be also applicable to the ranked values in the Spearman worksheet, which are both part of the spreadsheet associated with Example 11.1. The advantage with this approximate method is that you can complete all of the more advanced calculations we presented for the Pearson correlation coefficient (e.g., the establishment of critical values and hypothesis testing for different values of ρ).

These calculations are demonstrated in Example 11.2, using the same data as Example 11.1. In the associated Excel spreadsheet, the calculations are performed automatically.

Example

EXAMPLE 11.2 EXAMPLE OF THE CALCULATION OF THE SPEARMAN RANK CORRELATION COEFFICIENT (R_S)

Suppose you collected data from two water constituents in a river, and you want to test whether they are linearly correlated. You obtained 20 paired values of constituent X and constituent Y ($n = 20$). You decided to use a non-parametric test to calculate the Spearman rank coefficient of correlation r_s.

The data are the same as those used in Example 11.1.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Excel

Note: This example is also available as an Excel spreadsheet.

Solution:

From our data, we set up the following computational table. The ranking is done for each variable (one separate ranking is done for X and another separate ranking is done for Y). You may use the Excel function RANK.AVG, as explained above.

Computational table to calculate r_s

Sample Number	Constituent X (mg/L)	Constituent Y (mg/L)	Rank of X	Rank of Y	d	d^2
1	4.7	6.9	6.5	9	−2.5	6.3
2	5.2	7.7	10.5	15.5	−5.0	25.0
3	5.1	7.4	9	12.5	−3.5	12.3
4	4.7	6.8	6.5	7.5	−1.0	1.0
5	3.5	6.3	2	4	−2.0	4.0
6	3.3	5.2	1	1	0.0	0.0
7	3.8	5.4	3	2	1.0	1.0
8	4.0	6.0	4	3	1.0	1.0
9	5.9	6.6	12	5.5	6.5	42.3
10	7.3	7.3	16.5	11	5.5	30.3
11	6.9	7.4	14	12.5	1.5	2.3
12	7.5	7.6	18.5	14	4.5	20.3
13	7.7	7.8	20	17	3.0	9.0
14	7.1	8.3	15	18	−3.0	9.0
15	7.5	8.6	18.5	19	−0.5	0.3
16	7.3	8.7	16.5	20	−3.5	12.3
17	6.8	7.7	13	15.5	−2.5	6.3
18	5.2	7	10.5	10	0.5	0.3
19	4.9	6.8	8	7.5	0.5	0.3
20	4.3	6.6	5	5.5	−0.5	0.3
Σ					0.0	183.0

Note: d is the difference: rank X – rank Y

The Spearman rank correlation coefficient (in the absence of ties) is obtained using Equation 11.22, knowing that $n = 20$ and the sum of d^2 is 183.0 (calculated above). Normally, if there are ties, the rank correlation coefficient is corrected by a correction factor; however, we will not perform the correction for ties in this example. Unless the number of ties is very large relative to the total sample size, the correction factor will not have that great of an influence on the value of r_s.

$$r_s = 1 - \frac{6\sum d_i^2}{n^3 - n} = 1 - \frac{6 \times 183.0}{20^3 - 20} = \mathbf{0.862}$$

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

(a) Assessing the significance of r_s using a look-up table

To test whether this value of r_s is significantly different from zero, we formulate the hypotheses:

- Null hypothesis H_0: $\rho_s = 0$.
- Alternative hypothesis H_a: $\rho_s \neq 0$.

The critical value of r_s is obtained from Table 11.2, for $n = 20$ and $\alpha = 0.05$, as $r_{crit} = 0.447$.

Since $r_s = 0.862 > r_{s\ \alpha;n} = r_{s\ 0.05;20} = 0.447$, $r_{calc} > r_{crit}$, and we reject H_0, concluding that the correlation coefficient r_s is significantly different from zero.

(b) Assessing the significance of r_s using the *t* test

The calculated value of the *t* statistic (t_{calc}) can be estimated using Equation 11.23:

$$t_{calc} \approx \sqrt{\frac{r^2 \times df}{1 - r^2}} = \sqrt{\frac{0.862^2 \times 18}{1 - 0.862^2}} = \mathbf{7.214}$$

The critical value of *t* (t_{crit}) is obtained for a two-tailed test as a function of the significance level α (usually 0.05) using the Excel function:

$$\textbf{T.INV.2T}(\text{probability}; \text{deg_freedom}) = \text{T.INV.2T}(\alpha; n-2) = \text{T.INV.2T}(0.05; 20-2) = \mathbf{2.101}.$$

Since $t_{calc} > t_{crit}$, or $t = 7.214 > t_{crit} = 2.101$, we reject H_0.

The *p*-value for the test may be obtained using the following Excel function for a two-tailed *t* distribution:

$$\textbf{\textit{p}-value} = \textbf{T.DIST.2T}(x, \text{deg_freedom}) = \text{T.DIST.2T}(\text{ABS}(t_{calc}); n-2)$$
$$= \text{T.DIST.2T}(\text{ABS}(7.214; 20-2) = \mathbf{1.034 \times 10^{-6}}$$

Since $p < \alpha$, or $1.034 \times 10^{-6} < 0.05$, we reject H_0.

There may be small differences due to rounding errors. The values presented here have been obtained using calculations in Excel.

(c) Assessing the significance of r_s using the data rank values with Pearson correlation coefficient

Example 11.1 presented in detail the calculation of the Pearson correlation coefficient *r* using the original data and the Excel function CORREL. A series of additional calculations have also been done, and the results are discussed.

We can employ the same systematic approach here by simply using the ranked data (columns 4 and 5 of the computational table) instead of the original data (columns 2 and 3), employing the Excel function CORREL to the ranked data, and then performing all the complementary calculations. You will obtain the results using the Excel spreadsheet associated with this example, on the worksheet labelled 'Spearman'. All calculations are done automatically, using the same procedures from the worksheet 'Pearson', but with the ranked data. The main results obtained are as follows:

Spearman rank correlation coefficient: $\boldsymbol{r_s = 0.8620}$

$t_{calc} = \mathbf{7.214}$
$t_{crit} = \mathbf{2.101}$ (already calculated above)
p-value $= \mathbf{1.03395 \times 10^{-6}}$

These values are the same as those obtained in item 'b', and the interpretation and conclusions are also the same.

You can also follow the additional calculations performed in the worksheet to calculate the confidence limits for r_s and test hypotheses for ρ_0 values equal to and different from zero. These have been extensively detailed in Example 11.1.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

11.3 CORRELATION MATRIX

11.3.1 Pearson correlation matrix

Basic

In Section 11.2.1, we saw in detail how to estimate and interpret the Pearson linear correlation coefficient between two variables. In your study, you most likely have more than two variables, and you may be interested to investigate the correlation between all of them. You can do the calculations for each pair of variables and report them one by one. However, the presentation of the results will be more organized if you arrange your results in a matrix format, such as the one illustrated in Table 11.3 (with the correlation coefficients r), complemented with Table 11.4, which shows the associated p-values for whether the correlations are significant (based on tests of the null hypothesis that $\rho = 0$).

S. 11.2.1

After you interpret the values from the correlation matrix, you may decide to go in more depth in the study of the correlation between two or more specific variables and carry out the additional tests that explained in Section 11.2.1 (confidence limits for ρ and hypothesis tests for values of ρ different from zero).

S. 11.2.1

Table 11.3 Example of a correlation matrix for four variables, showing the values of the correlation coefficients r.

Variables	Variable A	Variable B	Variable C	Variable D
Variable A	1	0.0703	-0.0022	0.0897
Variable B	0.0703	1	-0.2571	-0.2556
Variable C	-0.0022	-0.2571	1	0.9022
Variable D	0.0897	-0.2556	0.9022	1

Notes:

- The table presents the values of the Pearson correlation coefficient between each pair of variables.
- Pay attention to the positive and the negative signs of the correlation coefficients.
- The value of r for the correlation between, say, variables B and C is the same as the one for variables C and B (they are presented twice, in this version of the correlation matrix).
- The diagonal has values of $r = 1$, since they represent the correlation between each variable and itself.

Table 11.4 Example of a complement to the correlation matrix for four variables, showing the p-values for the null hypothesis that $\rho = 0$.

Variables	Variable *A*	Variable *B*	Variable *C*	Variable *D*
Variable *A*	—	0.6837	0.9899	0.6029
Variable *B*	0.6837	—	0.1301	0.1324
Variable *C*	0.9899	0.1301	—	$5.7314 \times 10^{-14*}$
Variable *D*	0.6029	0.1324	$5.7314 \times 10^{-14*}$	—

Notes:

- The table presents the p-values for the t tests performed for the Pearson correlation coefficient between each pair of variables.
- If the p-value is lower than the significance level α you adopted for the test, the correlation between the two variables is significant (we reject the null hypothesis that $\rho = 0$).
- In this example, for $\alpha = 0.05$, we see that the correlation between C and D is significant (we highlighted this with an *). None of the other correlations are significant.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Example

EXAMPLE 11.3 CONSTRUCTION OF A CORRELATION MATRIX (PEARSON LINEAR CORRELATION COEFFICIENT *R*)

Suppose you obtained the monitoring data from a facultative pond treating wastewater. You obtained monthly data, comprising 24 data points for each variable. You then decided to investigate the existence of possible linear correlations between the variables. For this study, you selected the following four variables:

- *Effluent biochemical oxygen demand (BOD) concentration (mg/L)*
- *Mass loading rate (MLR), or surface organic loading rate [(kgBOD/d)/ha]*
- *Air temperature (°C)*
- *Solar insolation [(kWh/d)/m^2]*

Data:

BOD Effluent (mg/L)	MLR (kgBOD/ha . d)	Temperature (°C)	Insolation (kWh/m^2 . d)
136	80	22.5	5.23
48	167	23.5	5.84
33	163	22.7	5.31
73	29	21.7	4.98
150	199	19.7	4.47
105	215	18.1	4.41
60	82	22.7	5.59
40	283	22.4	5.31
36	198	23.2	5.61
90	85	23.3	5.84
78	181	21.7	4.98
60	221	19.5	4.47
96	96	18.3	4.57
50	51	19.7	5.17
54	147	21.8	5.42
110	357	22.8	5.59
80	391	22.2	5.31
36	153	22.6	5.23
66	164	23.4	5.61
110	62	23.5	5.84
108	136	22.7	5.31
120	385	21.3	4.98
30	170	19.7	4.47
20	238	18.2	4.41

Construct a Pearson correlation matrix and test the significance of the correlation coefficients.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Excel

Note: This example is also available as an Excel spreadsheet. Our spreadsheet allows for a maximum of nine variables. If you have more than this, you can use a statistical software or the Excel add-in 'Correlation'.

Solution:

S. 11.2.1

Here, we will not show you how to calculate the correlation coefficient *r* and do the hypothesis tests again, since we already demonstrated this in previous sections. Please refer to Section 11.2.1 and Example 11.1 for a review on these methods. The difference here is the presentation of the results in a matrix format, using the values of *r* calculated for each pair of variables. We will use the values calculated on an automatic basis in the associated Excel spreadsheet. In the spreadsheet, we use the Excel function CORREL to obtain the values of the correlation coefficients.

The correlation matrix obtained is shown in the table below.

Correlation matrix with Pearson correlation coefficients

Variables	BOD Effluent (mg/L)	MLR (kgBOD/ha . d)	Temperature (°C)	Insolation (kWh/m^2 . d)
BOD effluent (mg/L)	1	0.0525	−0.00999	−0.0491
MLR (kgBOD/ha d)	0.0525	1	−0.0652	−0.148
Temperature (°C)	−0.00999	−0.0652	1	0.917
Insolation (kWh/m^2 d)	−0.0491	−0.148	0.917	1

Note: We are presenting the results with several decimal cases obtained in the calculation with Excel, for you to be able to check your own calculations.

The *p*-values, for testing the null hypothesis that $\rho = 0$, are also shown in a matrix format in the table below.

The *p*-values for the correlation matrix with Pearson correlation coefficients

Variables	BOD Effluent (mg/L)	MLR (kgBOD/ha . d)	Temperature (°C)	Insolation (kWh/m^2 . d)
BOD effluent (mg/L)	—	0.8076	0.9630	0.8198
MLR (kgBOD/ha d)	0.8076	—	0.7622	0.4888
Temperature (°C)	0.9630	0.7622	—	$3.087 \times 10^{-10*}$
Insolation (kWh/m^2 d)	0.8198	0.4888	$3.087 \times 10^{-10*}$	—

*Significant *p*-value (at $\alpha = 0.05$ significance level).

We see that most of the correlation coefficients *r* are low, suggesting a weak linear relationship between most variables. This is endorsed by the *p*-values, which are almost all greater than $\alpha = 0.05$. The only exception is the correlation between temperature and insolation, which has a high value of *r* (0.902) and a very low *p*-value (3.09×10^{-10}), substantially lower than $\alpha = 0.05$, indicating a significant linear correlation between these two variables.

We analyse these values and agree that there is a good physical basis for having air temperature correlated with insolation. In particular, we analyse the correlations between effluent BOD and the other three variables and, based on the very small and non-significant correlation coefficients, decide to analyse the treatment system using different methods, such as process-based evaluation methods.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

11.3.2 Spearman rank correlation matrix (non-parametric)

Advanced

S. 11.2.2

The concept of the **Spearman** rank correlation matrix is similar to the Pearson correlation matrix, with the only difference that the correlation coefficient r_s is calculated based on the **ranked data**. In Section 11.2.2, we presented the concept, calculation, and interpretation of the Spearman rank correlation coefficient. We can apply the same concepts for building the correlation matrix.

As in Section 11.2.2, the ranking is done for each variable separately. You may use the Excel function RANK.AVG, as explained previously in this section.

Example

EXAMPLE 11.4 CONSTRUCTION OF A CORRELATION MATRIX (SPEARMAN RANK CORRELATION COEFFICIENT R_S)

After you built the Pearson correlation matrix in Example 11.3, you decided to make a similar matrix, but now using the non-parametric version of the Spearman rank correlation matrix. The data are the same as those shown in Example 11.3.

Excel

Note: This example is also available as an Excel spreadsheet. Our spreadsheet allows for a maximum of nine variables. If you have more than this, you can use a statistical software or the Excel add-in 'Correlation'.

Solution:

S. 11.2.1

S. 11.2.2

We will not show you how to calculate the Spearman rank correlation coefficient r_s and to do the hypothesis tests again since these methods were covered in previous sections. Please consult Sections 11.2.1 and 11.2.2 and also Examples 11.1 and 11.2 for a review on these methods. The only difference here is that we will present the results in a matrix format, using the values of r_s calculated for each pair of the ranked variables. We will use the values calculated on an automated basis in the associated Excel spreadsheet. In the spreadsheet, we use the Excel function CORREL applied to the **ranked data** to obtain the values of the correlation coefficients. Ranking was done using the Excel function RANK.AVG, as explained in Section 11.2.2.

The following table presents the values of the rankings of each variable. Tied ranks are reported as the average of their values. As given in Example 11.3, $n = 24$ for each variable.

Ranking of data of the variables presented in Example 11.3

BOD Effluent	MLR	Temperature	Insolation
23	4	14	11.5
7	13	23.5	23
3	11	17	14.5
13	1	9.5	8
24	17	6	4
18	18	1	1.5
10.5	5	17	18.5
6	21	13	14.5
4.5	16	20	20.5

(*Continued*)

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

BOD Effluent	MLR	Temperature	Insolation
16	6	21	23
14	15	9.5	8
10.5	19	4	4
17	7	3	6
8	2	6	10
9	9	11	17
20.5	22	19	18.5
15	24	12	14.5
4.5	10	15	11.5
12	12	22	20.5
20.5	3	23.5	23
19	8	17	14.5
22	23	8	8
2	14	6	4
1	20	2	1.5

From the Excel spreadsheet, using the CORREL function for the ranking values, we obtain the Spearman correlation matrix.

Correlation matrix with Spearman correlation coefficients

Variables	BOD Effluent (mg/L)	MLR (kgBOD/ha . d)	Temperature (°C)	Insolation (kWh/m^2 . d)
BOD effluent (mg/L)	1	−0.04307	0.02115	0.003063
MLR (kgBOD/ha d)	−0.04307	1	−0.2466	−0.2602
Temperature (°C)	0.02115	−0.2466	1	0.9461
Insolation (kWh/m^2 d)	0.003063	−0.2602	0.9461	1

Note: We are presenting the results with several decimal cases obtained in the calculation with Excel, for you to be able to check your own calculations.

The p-values, for testing the null hypothesis that $\rho_s = 0$, are also shown in a matrix format.

The p-values for the correlation matrix with Spearman correlation coefficients

Variables	BOD Effluent (mg/L)	MLR (kgBOD/ha . d)	*T* (°C)	Insolation (kWh/m^2 . d)
BOD effluent (mg/L)	–	0.8416	0.92187	0.9887
MLR (kgBOD/ha d)	0.8416	–	0.2453	0.2195
Temperature (°C)	0.92187	0.2453	–	2.989×10^{-12}*
Insolation (kWh/m^2 d)	0.9887	0.2195	2.989×10^{-12}*	–

*Significant p-value (at $\alpha = 0.05$ significance level).

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

In this case, the interpretation is very similar to the one we made in Example 11.3 for the Pearson correlation matrix. There were some changes in the signs of some of the correlation coefficients, when comparing Spearman with Pearson, but these correlation coefficients were very low, close to zero, and non-significant anyway. In general, we see that most of the Spearman correlation coefficients r_s are quite low, suggesting a weak or non-existent monotonic relationship between most variables. Spearman's correlation coefficient is a statistical measure of the strength of a monotonic relationship between the paired data. In a monotonic relationship, variables tend to move in the same relative direction, but not necessarily at a constant rate.

The weak relationships are endorsed by the *p*-values, which are almost all greater than $\alpha = 0.05$. The only exception is the correlation between temperature and insolation, which has a high value of r_s (0.946) and a very low *p*-value (2.99×10^{-12}), substantially lower than $\alpha = 0.05$, indicating a significant linear correlation between these two variables. Regarding the interpretation of the physical meaning of these correlations, we draw similar conclusions to those presented in Example 11.3.

11.4 CROSS-CORRELATION AND AUTOCORRELATION

11.4.1 Cross-correlation

Advanced

S. 13.2
S. 14.4
S. 14.5
S. 11.2.1
S. 11.2.2

So far, for variables arranged as time series, we have considered the correlation between variables whose measurements are related to the same data sequence. For instance, in a monitoring programme, we are relating only pairs of samples collected at week 1, and then we relate pairs of samples collected at week 2, and so on. However, what if the treatment unit or water body you are studying introduces a lag on the output variable – for instance, associated with the travelling time, retention time (Section 13.2), or hydraulic behaviour (sections 14.4 and 14.5) of your reactor or water body? The analysis of this type of correlation with a lag is called an analysis of **cross-correlation**: *correlated data that are shifted in the data sequence*.

From the two data sets we want to analyse (X and Y), we calculate the correlation coefficient (Pearson r and Spearman r_s) in the same way we performed in Sections 11.2.1 (Pearson) and 11.2.2 (Spearman). We then shift one of the variables (say, X) one step in the data sequence and calculate the resulting correlation coefficients between X with 1 lag and Y. After that, we shift it another step and calculate the coefficient of correlation between X with 2 lags and Y. We repeat this procedure as many times we want and interpret the various correlation coefficients obtained. We call this step '**lag**' because we introduce a *lag in the data sequence*. It can also be called a '**lead**', depending on which direction of your data sequence you are looking at. Most commonly, the *data sequence is related to a time variable*, and Data 1 could be Day 1, Data 2 could be Day 2, and so on. Therefore, a lag of 1 would correspond to shifting one day, a lag of 2 would correspond to shifting two days, etc. A similar comment could be made if the time unit of your time series is months, hours, minutes etc.

Table 11.5 presents a simplified example of how to arrange your data. For the sake of simplicity, we showed only three lags, but typically we carry out this analysis for a larger number of lags (12, 24, etc). We see that, for each lag we introduce, we lose data (and degrees of freedom) to perform the calculations of the correlation coefficients between the pairs of variables (X lag 0 and Y), (X lag 1 and Y), (X lag 2 and Y), and (X lag 3 and Y).

The sequence of correlation coefficients for lag 0 up to lag k is plotted on a column chart known as **cross-correlogram**. This graph will be illustrated in Example 11.5. The correlation coefficients may be calculated using the Excel function **CORREL**, as shown previously.

The cross-correlogram also includes testing for the **significance of *r*** (null hypothesis H_0: $\rho = 0$; alternative hypothesis H_a: $\rho \neq 0$) and the **confidence limits** for the estimates of r. These are calculated as

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Table 11.5 Example of the preparation of monitoring data for a cross-correlation, showing the shifting of data for each lag introduced.

Data sequence	Y	X			
		Lag 0	Lag 1	Lag 2	Lag 3
1	1.4	3.8			
2	1.2	4.1	3.8		
3	3	3.6	4.1	3.8	
4	3.5	5.4	3.6	4.1	3.8
...	...	...	5.4	3.6	4.1
...	...	...	...	...	...
n	0.7	5.9	...	...	...
Number of data (n) for correlation with Y		***n***	***n-1***	***n-2***	***n-3***
Correlation coefficient		***r for X lag 0 and Y***	***r for X lag 1 and Y***	***r for X lag 2 and Y***	***r for X lag 3 and Y***

S. 11.2.1

described in Section 11.2.1, using the t test for the distribution of r (for $\rho = 0$), and the calculations are summarized here.

$$t = \frac{r}{\sqrt{\frac{1 - r^2}{(n - \text{lag}) - 2)}}} \tag{11.24}$$

where

t = test statistic (t_{calc})
r = correlation coefficient between X and Y for each lag
n = number of data points
lag = number of lags introduced in the analysis.

The value of $\boldsymbol{t_{\text{crit}}}$ can be obtained from the Excel function **T.INV.2T**(probability; deg_freedom). The probability is the significance level for the test (e.g., $\alpha = 0.05$), and the degrees of freedom are (n-lag-2).

The associated ***p*-value** is obtained using the Excel function **T.DIST.2T**(x; deg_freedom) = T.DIST.2T (t_{calc}; $n - \text{lag} - 2$).

The **confidence limits for *r*** (*lower confidence limit* **LCL** and *upper confidence limit* **UCL**) are calculated assuming that the ρ distribution has a mean of $\mu = 0$ and a standard deviation of $\sigma = 1$. These confidence limits are also included in the cross-correlogram.

$$\begin{aligned} \text{LCL} &= \text{Mean} - t_{\text{crit}} \times \frac{\text{Standard deviation}}{\sqrt{n - \text{lag}}} \\ &= 0 - t_{\text{crit}} \times \frac{1}{\sqrt{n - \text{lag}}} \end{aligned} \tag{11.25}$$

$$\begin{aligned} \text{UCL} &= \text{Mean} + t_{\text{crit}} \times \frac{\text{Standard deviation}}{\sqrt{n - \text{lag}}} \\ &= 0 + t_{\text{crit}} \times \frac{1}{\sqrt{n - \text{lag}}} \end{aligned} \tag{11.26}$$

Example 11.5 illustrates the construction and interpretation of a cross-correlogram.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Example

EXAMPLE 11.5 CONSTRUCTION OF A CROSS-CORRELOGRAM (PEARSON AND SPEARMAN)

Suppose you obtained the monitoring data from a pollutant in a river at two separate points. For the first sample point (upstream), water quality samples were collected in the river at a location immediately after the discharge of an industrial effluent. Samples were collected at a second sample point (downstream), located in the same river, but further downstream. Assume that the industrial effluent discharge is not constant, but varies during the day, causing a diurnal variation in the river's water quality. The upstream and the downstream samples were collected at approximately the same time; however, there was a lag period for the water to travel from the upstream location to the downstream location.

Assess the correlation in the pollutant concentration at both sampling points, to analyse the possible decay in the river between the upstream and downstream locations.

Data:

You obtained 48 samples, arranged in a time sequence. The samples were collected at approximate intervals of 6 hours (4 samples/day). Therefore, you had data covering a period of 48 samples ÷ 4 samples/d ≅ 12 days. You labelled the upstream samples X and the downstream samples Y.

Data Sequence	Downstream (*Y*)	Upstream (*X*)	Data Sequence	Downstream (*Y*)	Upstream (*X*)
1	1.4	3.8	25	4.5	1.1
2	1.2	4.1	26	4.2	4.3
3	3.0	3.6	27	2.5	5.4
4	3.5	5.4	28	1.4	4.1
5	0.7	5.9	29	1.7	4.9
6	1.4	5.0	30	4.2	0.9
7	2.4	4.1	31	3.6	2.5
8	4.2	1.4	32	2.3	3.6
9	3.3	0.5	33	2.8	5.2
10	3.9	1.8	34	0.5	6.1
11	2.8	3.2	35	1.7	4.5
12	1.1	5.4	36	2.4	2.9
13	0.3	6.5	37	3.6	0.4
14	1.4	5.8	38	3.3	1.4
15	2.2	4.3	39	2.5	4.1
16	4.1	2.2	40	2.2	5.8
17	5.0	1.4	41	0.3	6.1
18	3.8	2.3	42	0.8	5.6
19	3.8	3.8	43	3.7	4.5
20	1.4	5.6	44	3.5	2.2
21	0.8	5.8	45	4.1	1.3
22	1.8	4.7	46	3.2	2.2
23	2.9	3.2	47	3.5	3.4
24	3.7	1.6	48	1.4	5.2

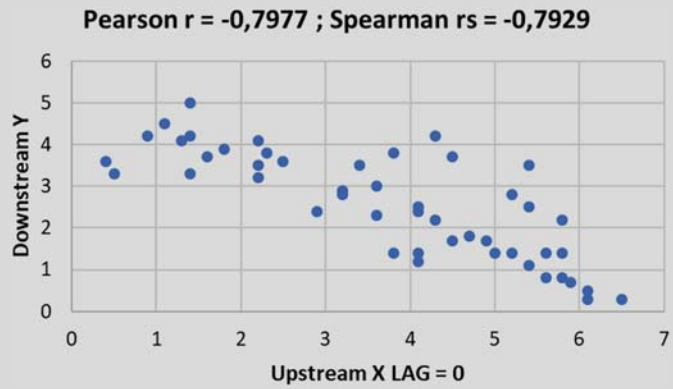
Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Excel

Note: This example is also available as an Excel spreadsheet. Our spreadsheet allows for a maximum of 24 lags. If you want more than this, either adapt the spreadsheet accordingly or use a statistical software.

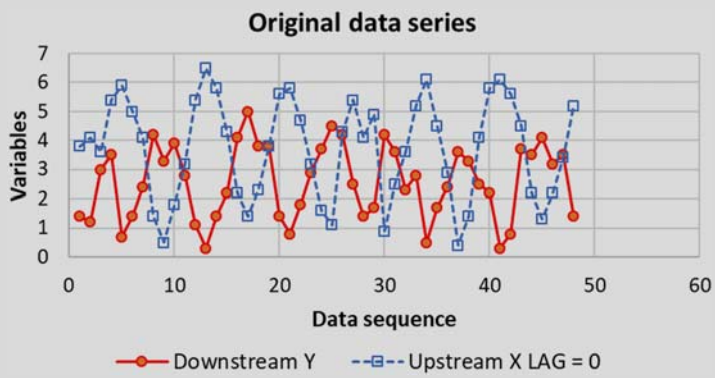
Solution:

As in all correlation studies, it is advisable to build a scatter plot with the data. The scatter plot you obtained is shown below.



At first sight, the results you obtained appear different from what you initially imagined: higher values of the upstream concentration were associated with low values in the downstream monitoring point, and lower upstream concentrations were associated with high downstream concentrations. This is supported by the negative correlation coefficients (both Pearson and Spearman in this case). Therefore, you could not draw specific conclusions about the decay of the pollutant.

However, if we look at a time series plot with the results from both sampling stations, we can see that the two series have opposite behaviour in cyclical patterns: peaks in the upstream concentration are paired with valleys in the downstream concentration, and vice versa.



Based on this finding, you should analyse the results in more detail, specifically building a cross-correlogram between both data series. The structure of the table is as follows.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Data Sequence	(Downstream) *Y*	(Upstream) *X* lag 0	*X* lag 1	*X* lag 2	*X* lag 3	*X* lag 4	*X* lag 5	*X* lag …
1	1.4	3.8						
2	1.2	4.1	3.8					
3	3	3.6	4.1	3.8				
4	3.5	5.4	3.6	4.1	3.8			
5	0.7	5.9	5.4	3.6	4.1	3.8		
6	1.4	5.0	5.9	5.4	3.6	4.1	3.8	
7	2.4	4.1	5.0	5.9	5.4	3.6	4.1	…
8	4.2	1.4	4.1	5.0	5.9	5.4	3.6	…
9	3.3	0.5	1.4	4.1	5.0	5.9	5.4	…
10	3.9	1.8	0.5	1.4	4.1	5.0	5.9	…
11	2.8	3.2	1.8	0.5	1.4	4.1	5.0	…
12	1.1	5.4	3.2	1.8	0.5	1.4	4.1	…
13	0.3	6.5	5.4	3.2	1.8	0.5	1.4	…
14	1.4	5.8	6.5	5.4	3.2	1.8	0.5	…
15	2.2	4.3	5.8	6.5	5.4	3.2	1.8	…
16	4.1	2.2	4.3	5.8	6.5	5.4	3.2	…
…	…	…	…	…	…	…	…	…
48	…	…	…	…	…	…	…	…

We showed only up to 5 lag periods, but our spreadsheet calculates up to 24 lag periods. Recall that in this study, one lag period corresponds to approximately 6 hours, since that was the frequency with which samples were collected. Therefore, five lag frequencies would be a lag period of $5 \times 6 = 30$ hours.

We will show how to construct the cross-correlogram for lag period 1. The calculations are similar for lag 0 and for lag periods 2 through 24.

The correlation coefficients may be calculated using the Excel function **CORREL**(array 1; array 2) = CORREL(column with Downstream *Y*; column with Upstream *X* lag 1) = **−0.6150**. This value will be plotted in the cross-correlogram, for lag period = 1.

The test statistic t_{calc} is calculated from Equation 11.24:

$$t_{calc} = \frac{r}{\sqrt{\frac{1-r^2}{(n-\text{lag})-2}}} = \frac{-0.6150}{\sqrt{\frac{1-(-0.6150)^2}{(48-1)-2}}} = \mathbf{5.232}$$

The value of $\boldsymbol{t_{crit}}$ can be obtained from the Excel function T.INV.2T(probability; deg_freedom) = T.INV.2T(α; *n*-lag-2) = T.INV.2T(0.05; 48 − 1 − 2) = **2.014**. Since $t_{calc} > t_{crit}$ (5.232 > 2.014), we conclude, at the 5% significance level, that the correlation coefficient for lag period 1, $r = -0.6150$, is *significantly different from zero*.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

The associated ***p*-value** is obtained from the Excel function T.DIST.2T(x; deg_freedom) = T.DIST.2T (t_{calc}; n-lag-2) = T.DIST.2T(5.232; $48-1-2$) = $\mathbf{4.23552 \times 10^{-6}}$. Since the p-value $<$ significance level ($4.23552 \times 10^{-6} < 0.05$), we can conclude again that the *correlation coefficient r for the lag period 1 is significant*.

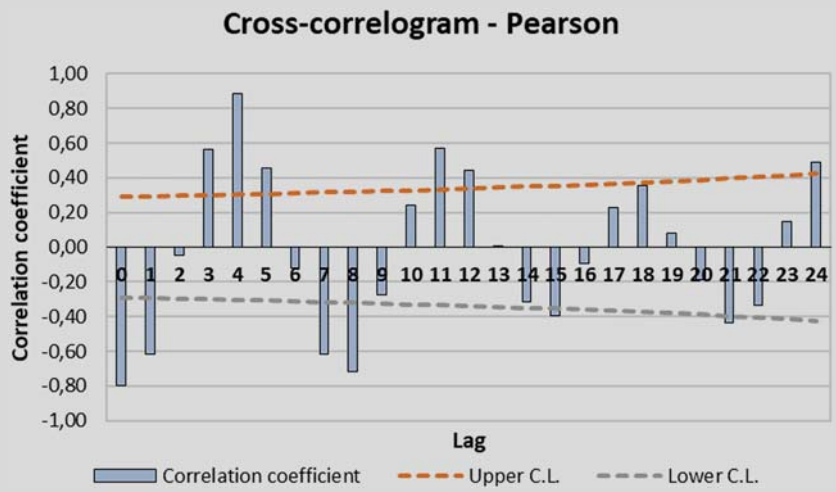
In the cross-correlogram, we also plot the **confidence limits** (in this case, for $\alpha = 0.05$, we have limits for a 95% confidence level). The confidence limits for lag period 1 are calculated using Equations 11.25 and 11.26:

$$\text{LCL} = -t_{crit} \times \frac{1}{\sqrt{n - \text{lag}}} = -2.014 \times \frac{1}{\sqrt{48-1}} = \mathbf{-0.294}$$

$$\text{UCL} = t_{crit} \times \frac{1}{\sqrt{n - \text{lag}}} = 2.014 \times \frac{1}{\sqrt{48-1}} = \mathbf{0.294}$$

Since the value of r for lag period 1 ($r = -0.6150$) is less than the LCL, we can conclude, once more, that the *correlation for lag period 1 is significant*.

If we carry out the same calculations for all lag periods, we end up with the cross-correlogram, which is displayed in the following figure.

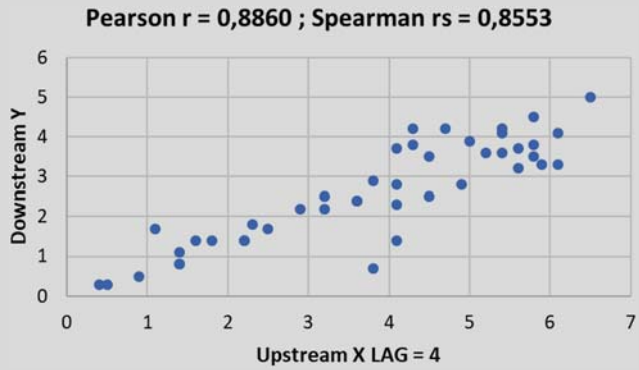


Analysing this correlogram, we see some points:

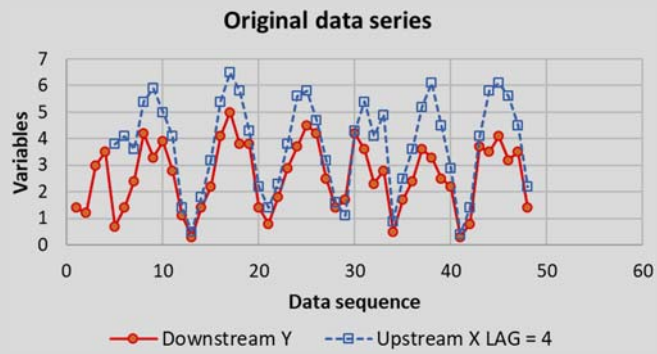
- The correlation coefficients are the vertical bars, and the confidence limits are the dashed lines.
- The correlation coefficient for lag 0 (no lag) is negative, as already shown before.
- The correlation coefficient for lag period 1 ($r = -0.6150$), already calculated above, is shown in the cross-correlogram.
- The highest correlation coefficient is positive and is found at lag period 4.
- The correlation coefficients display a wave shape, alternating sequences of positive values with negative values, indicating the cyclical nature of the data.
- The significant correlation coefficients are those that project themselves beyond the bounds of the confidence limits.
- As the lag periods increase, the confidence limits get wider (as a result of the loss of the degrees of freedom with each subsequent lag period), and the correlation coefficients tend to become less significant.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Based on the important conclusion that the highest correlation was found for lag 4, we make the scatter plot for lag 4 (simply alter the value of the number of lags in the tab 'Graphs' of the Excel spreadsheet).



We now see a clear pattern, with a positive correlation between the upstream sample *X* (with a lag of 4) and the downstream sample *Y*. We also decide to plot the time series with the lagged data.



Now we can see clearly the ups and downs from both samples coinciding. How can we interpret these results? The pollution at the vicinity of the discharge point (monitoring station called 'upstream') took some time to reach the 'downstream' sampling point. This is due to the travelling time of the water between the two sampling stations. A peak load at the upstream point travels in the river until reaching the downstream point. We see that there is some decrease in the concentrations along this stretch in the river since the downstream values are lower than the upstream. Periods of low concentration in the upstream station (probably at night, if the industry does not operate in the night shift) are reflected at the downstream station, after some time.

What is this time that is associated with four lag periods? We mentioned that we have approximately one sample every 6 hours. Therefore, four lag periods is associated with a time of 4×6 hours per data point $= 24$ hours of lag.

You then decided to search for a physical explanation for this. Based on the distance between the two sampling points and the flow velocity of the river, you concluded that the water, as a matter of fact, takes approximately 24 hours to flow downstream from the upstream sampling point to the downstream sampling point. You are then satisfied that you were able to understand better the behaviour of the river you are studying, integrating statistical and process-based calculations.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Notes:

- We showed here the results for the Pearson correlation coefficient (r). The Excel spreadsheet also computes the **Spearman** rank correlation coefficient (r_s). The calculations are basically the same, with the difference that the correlogram is constructed with the **ranks** of the data, instead of the original monitored values. The spreadsheet does all calculations automatically.
- Cross-correlograms are frequently plotted with positive lags (as we did here) and negative lags (called leads). To simplify our calculations, we presented only positive lags. If you want to analyse the 'leads', change the order of the variables X and Y in the spreadsheet and introduce the lags for Y.

11.4.2 Autocorrelation

Advanced

S. 11.4.1

In Section 11.4.1, we saw the meaning, calculation, and interpretation of cross-correlations. Now we will see a similar concept, but with the difference that we analyse **only one variable**, and the correlation is analysed in terms of the lags introduced in the variable itself. This procedure is called **autocorrelation**, and the related graph is called an **autocorrelogram**.

You may have a variable whose current measurement is related to some previous measurement (e.g., the previous measurement, lag period of 1; or to the measurement taken 24 hours prior, or even 48 or 72 hours prior, etc.). This is true if the variable has a **cyclical pattern**, and the measurements are taken every hour (for instance, in this particular example, the data sequence is organized on an hourly basis, and then each lag corresponds to 1 hour of shifting). This could be the case, for instance, for the time series of inflow to a wastewater treatment plant, with its diurnal variations following a cyclical pattern.

C. 15

Another important utilization of the analysis of autocorrelation is the investigation of the **properties of the residuals** from a mathematical model (either a regression-based model, such as those analysed in this chapter, or a process-based model, as discussed in Chapter 15). As seen in these two chapters, a *residual is the difference between the observed and the estimated values*. When we complete a residual analysis as a part of our assessment of the model performance, one of the properties that the residuals need to possess is that they should be **independent**, that is, they *should not be autocorrelated*. You can assess this by completing an autocorrelation study of the residuals.

S. 11.4.1

Autocorrelation is analysed in a similar way to that described for the cross-correlations (Section 11.4.1). From the data set we want to analyse (variable X), we shift it one step in the data sequence (lag 1) and calculate the resulting correlation coefficient between the two variables, now represented by X and X lag 1. After that, we shift it another step and calculate the coefficient of correlation between X and X lag 2. We repeat this procedure as many times we want and interpret the various correlation coefficients obtained.

Table 11.6 presents a simplified example of how to arrange the data. For the sake of simplicity, we show only three lags, but typically we carry out this analysis for a larger number of lags (12, 24, etc.). We see that, for each lag we introduce, we lose data (and degrees of freedom) to perform the calculations of the correlation coefficients between the pairs of variables: X and X lag 1, we lose one degree of freedom; X and X lag 2, we lose two degrees of freedom; and X and X lag 3, we lose three degrees of freedom; and so on.

The sequence of correlation coefficients for lag period 1 up to lag period k is plotted on a column chart known as **autocorrelogram**. The correlation coefficients may be calculated using the Excel function **CORREL** (array 1; array 2). All the elements that make up the autocorrelogram (testing of the significance of r and establishment of the upper and lower confidence limits) are calculated as explained in Section 11.4.1 for cross-correlations (Equations 11.24–11.26).

S. 11.4.1

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Table 11.6 Example of the preparation of data for an autocorrelation study, showing the shifting of data for each lag introduced.

Data sequence	X			
	X	X lag 1	X lag 2	X lag 3
1	3.8			
2	4.1	3.8		
3	3.6	4.1	3.8	
4	5.4	3.6	4.1	3.8
...	...	5.4	3.6	4.1
...	...	...	...	...
n	5.9	...	...	...
Number of data for correlation (n)	*n*	*n-1*	*n-2*	*n-3*
Correlation coefficient	*1*	*r for X and X lag 1*	*r for X and X lag 2*	*r for X and X lag 3*

Figure 11.7 shows the autocorrelogram of the time series of X for the 'upstream sampling point' in Example 11.5, with the Pearson correlation coefficient. In that example we saw, from the time-series plot, that the data showed a cyclical pattern. This is endorsed by the autocorrelogram, which indicates a significant correlation at lag 1, followed by successive periods with positive and negative correlations, clearly emphasizing the cyclical nature of the data.

To perform a more advanced study of autocorrelation and develop models based on it, some additional steps may be necessary. For example, you may need to *remove trends* in the time series by processes

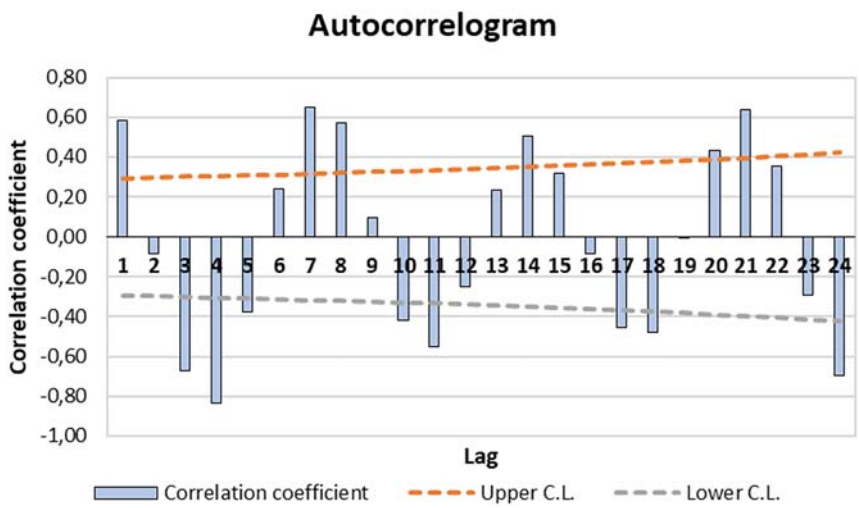


Figure 11.7 Example of an autocorrelogram showing a cyclical pattern of the data. The data used are the same from Example 11.5 (upstream X variable).

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

of *non-seasonal decomposition*, aiming to make the new series stationary. One such process is called *first-order differencing*, which is where we subtract the series by the same series with a lag of one period. The environmental data are also subject to seasonality (daily cycles of hourly variations, or annual cycles of monthly variations), as discussed here and as illustrated in Figure 11.7. Seasonality also influences the analysis of autocorrelation, which may require that we complete some procedures of *seasonal decomposition* to remove the cyclical pattern. If we remove trend and seasonality, we can perform more advanced analyses based on the so-called *autocorrelation function*. Statistical softwares that have a time-series component are capable of completing this type of analysis. If you would like to structure a model based on autocorrelation, you may want to study the so-called **ARIMA** (autoregressive integrated moving average) models. Most references cite the classical Box-Jenkins texts (see Box et al., 2015).

S. 15.3.5

Example 11.6 presents the typical application of autocorrelation analysis for the study of **model residuals**, as covered in Section 15.3.5. We would like our model residuals to follow a random pattern, in which there are no autocorrelations. We can check the compliance with this requirement by constructing an autocorrelogram.

Example

C. 15

EXAMPLE 11.6 AUTOCORRELATION ANALYSIS OF MODEL RESIDUALS

In Example 15.2 (Chapter 15), we carry out a full analysis of model residuals. One of the elements of a residuals analysis is testing whether the residuals are autocorrelated. Here, we complement this analysis by building an autocorrelogram with the residuals generated by the model.

The data are presented below. We show only the model residuals, which will be used here (see Example 15.2 to see the data and methods used to calculate these residuals).

Data Sequence	Residual	Data Sequence	Residual	Data Sequence	Residual
1	0.0	13	0.0	25	−0.2
2	−0.4	14	0.1	26	−0.8
3	−0.6	15	0.4	27	−0.3
4	−0.3	16	0.4	28	0.3
5	0.6	17	−1.1	29	0.0
6	0.3	18	−0.2	30	0.2
7	0.4	19	−0.1	31	0.2
8	0.7	20	0.3	32	−0.5
9	0.3	21	−0.5	33	0.6
10	0.5	22	−1.0	34	0.1
11	−1.0	23	0.3	35	−0.3
12	0.1	24	−1.0	36	1.0

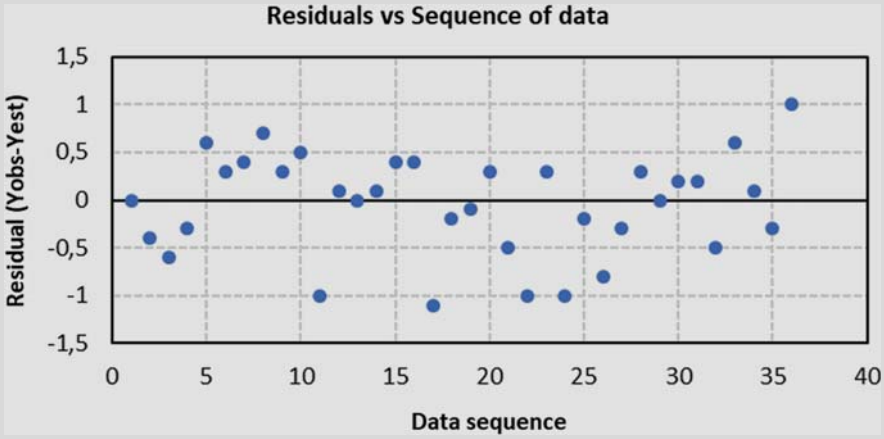
Excel

Note: This example is also available as an Excel spreadsheet. Our spreadsheet allows for a maximum of 24 lags. If you want more than this, either adapt the spreadsheet accordingly or use a statistical software.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Solution:

As in all correlation studies, we start by visually analysing our original data. A time series plot of the residuals is shown below. It is not clear, from this plot, whether the series will present autocorrelation. We will perform the autocorrelation analysis to be able to draw an appropriate conclusion.



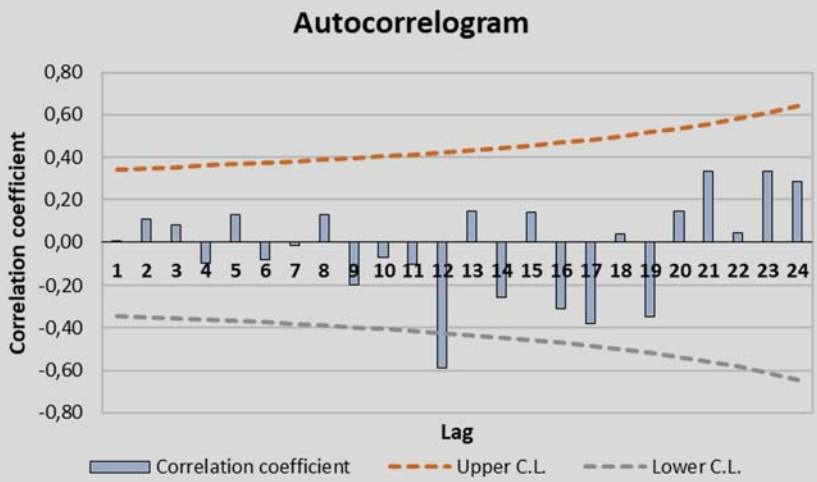
The structure of the table for performing the autocorrelation analysis is as follows.

Data Sequence	*X*	*X* lag 1	*X* lag 2	*X* lag 3	*X* lag 4	*X* lag 5	*X* lag …
1	0.0						
2	−0.4	0.0					
3	−0.6	−0.4	0.0				
4	−0.3	−0.6	−0.4	0.0			
5	0.6	−0.3	−0.6	−0.4	0.0		
6	0.3	0.6	−0.3	−0.6	−0.4	0.0	
7	0.4	0.3	0.6	−0.3	−0.6	−0.4	…
8	0.7	0.4	0.3	0.6	−0.3	−0.6	…
9	0.3	0.7	0.4	0.3	0.6	−0.3	…
10	0.5	0.3	0.7	0.4	0.3	0.6	…
11	−1.0	0.5	0.3	0.7	0.4	0.3	…
12	0.1	−1.0	0.5	0.3	0.7	0.4	…
…	…	…	…	…	…	…	…
36	…	…	…	…	…	…	…

We showed only up to lag period 5, but our spreadsheet calculates up to 24 lag periods.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

The calculation of the statistics follows the same procedure shown for the cross-correlation, illustrated in Example 11.5, and will not be repeated here. We will go directly to the autocorrelogram that results from these calculations.



We observe that almost all correlations are non-significant, since they are between the upper and the lower confidence intervals. The only exception is the negative correlation at lag 12. This may not be sufficient autocorrelation to invalidate the required property of independence (i.e., the absence of autocorrelation), but it would be worthwhile to carefully check your data and the model. At lag period 1, the correlation is very small, as endorsed by the Durbin–Watson test performed in Example 15.2, which shows no evidence of a first-order (lag period 1) correlation. To get a more complete view, you need to analyse all other results from the residuals analysis, as shown in Example 15.2 and as described in Section 15.3.5.

S. 15.3.5

11.5 SIMPLE LINEAR REGRESSION

11.5.1 The simple linear regression equation

(a) Approaches for performing a regression analysis

Basic

In Section 11.1, we explained the basic concepts of correlation and regression. Sections 11.2–11.4 were devoted for investigating the linear relationship between variables, as measured by **correlation coefficients**. In your studies, if you detect a significant linear relationship between variables, you may decide to expand upon your investigation with a regression analysis, for which you will fit a model to represent a relationship between the variables. The simplest and the most widely-used model is the **linear regression** model. Understanding its scope and limitations is fundamental for understanding other types of regression models.

S. 11.1
S. 11.2
S. 11.3
S. 11.4

The first step in linear regression analysis is to make a **scatter plot** of the variables x and y and to *observe if the two variables are actually linearly related*, as we have explained in the previous sections of this chapter. If the relationship does not appear to be linear, you should either try data transformations to make the relationship approach linearity or perhaps look into a different type of non-linear model.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

In this book, we will show you several different approaches to perform a regression analysis between X and Y variables:

- Adding a **trendline** in the scatter plot
- Using Excel **functions** to calculate regression coefficients
- Using the Excel add-in **Analysis ToolPak** regression tool
- Perform the calculations associated with the regression analysis **formulas**

✓ **Trendline**

Using the trendline option is very easy in Excel. In your scatter plot of the $X - Y$ data, simply click on the data points of the series you want to analyse (click left-button of the mouse, and then select '+', or right-click on the data points, and then select 'Add trendline'). Since we are now dealing with linear regression, you should select 'Linear' and allow Excel to plot the line of best fit.

Figure 11.8 shows a scatter plot with and without the fitted line. Note that we have the following options:

– Set the intercept to a specific value (e.g., set intercept to zero, if you want your straight line to include the origin with $X = 0$ and $Y = 0$; see more about the concept of intercept in item 'd' below).
– Display the equation of the line on the chart (very useful; recommended).
– Display the R^2 value on the chart (recall that R^2 is the Coefficient of Determination, and it is a measure of the goodness of fit of your model; see Section 11.5.2(f); this option is very useful and is recommended).
– You can also plot forecast values (forward or backward) for a specified number of periods.

S. 11.5.2

This is a very simple and useful procedure, and it will probably satisfy many of your needs if you want to do a simple regression analysis. The graph, model equation, and R^2 value are dynamic. If you change your data, the graph, equation, and R^2 will be updated automatically.

✓ **Excel functions**

Excel has plenty of very useful functions that allow you to do direct calculations related to the regression analysis. We will make use of several of them in this chapter. However, please note that new functions are added with new Excel versions, so keep updated on this regard and consult Excel's 'Help' resources.

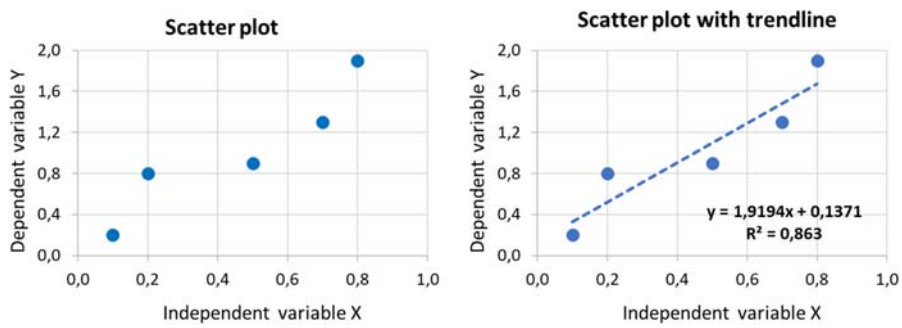


Figure 11.8 Example of the utilization of the Excel function 'add trendline' to a scatter plot.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

✓ **Excel add-in 'Analysis ToolPak's Regression tool'**

'Regression' is one of the several statistics add-in tools available in Excel. It allows you to perform the full calculations related to the regression analysis, including an analysis of variance and the confidence intervals of the coefficients. It is particularly useful for multiple linear regression analysis (see Section 11.7), which is more complex to perform. But note that this option is static: if you change your data, you will need to repeat the test. Also, it is less transparent than other options, because the add-in does not show you how the calculations are done (however, we will show the relevant equations in this chapter).

S. 11.7

✓ **Calculations using formulas**

You can do the regression analysis using the formulas provided in the statistics literature. These calculations are more laborious than the automated built-in regression options in Excel, but once you set up a template with the equations into an Excel spreadsheet, subsequent regression analyses will be straightforward because you can use the same template. We will demonstrate the main calculations in this chapter and in the full example provided here (Example 11.7), which also includes an associated Excel spreadsheet with a template that is ready to use. The advantage of using this approach over the add-in tool is that you will be able to see how the calculations are done and compare the procedures with statistical textbooks.

(b) The linear regression model

Basic

The linear model for the **population** of values is given by the general equation:

$$Y = \alpha + \beta X + \varepsilon \tag{11.27}$$

where

Y = dependent variable
X = independent variable
α = linear coefficient (intercept of the line with the Y-axis)
β = angular coefficient (slope of the line)
ε = error or residual; difference between the actual (observed or measured) Y and the Y predicted (estimated) by Equation 11.27. The sum of the errors ε is zero.

Figure 11.9 shows a typical scatter plot and a candidate for the estimated regression line. Estimates of α and β should result in a line that is the 'best fit' for the data. Karl Gauss (1777–1855) proposed estimating the parameters α and β in Equation 11.27 in a way that minimized the sum of the squares of the vertical deviations in the model.

As such, the criterion for 'best fit' that is most commonly employed in regression analyses utilizes the concept of **least squares**, i.e., a minimization of the sum of the squares. This criterion considers the vertical deviation of each observed Y value from the line with the estimated value of $\hat{Y}$ (i.e., the difference $Y_i - \hat{Y}_i$). In the example in Figure 11.9, we have five data points, and thus, we have five pairs of observed Y_i and estimated $\hat{Y}_i$ (i.e., $i = 1, 2, 3, 4, 5$). You can see in the example that some errors (or residuals) are positive, while others are negative. However, if we simply summed the five errors (without squaring them), the negative errors would cancel the positive values and there could be several different ways to minimize the sum of the errors. Thus, we need to use the squared values of the errors, or residuals, to find the best fit. By squaring all errors, they become positive values, and our error function to be minimized is defined as the sum of the squared errors (SSEs). The best-fit line is that which results in the smallest SSEs, that is, the minimum value for the summation $\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$, where n is the number of data points in the statistical sample (the

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

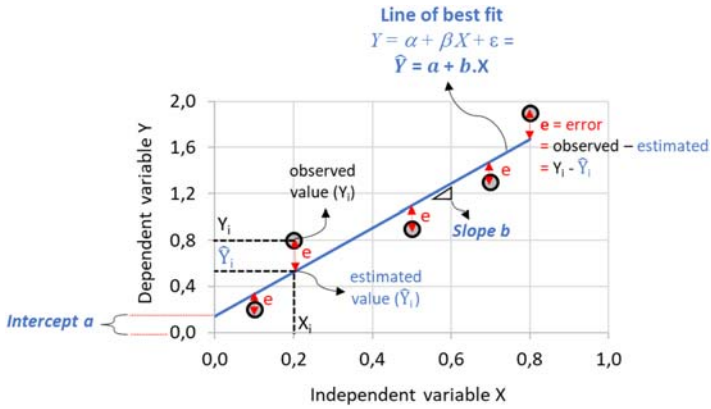


Figure 11.9 Observed values (Y_i), estimated values ($\hat{Y}_i$), errors ($e = Y_i - \hat{Y}_i$), and line of best fit in a typical regression analysis.

number of paired X and Y values). This is why the method is called 'least squares'. The sum of squares (SS) of the deviations is called the ***residual sum of squares*** (or, sometimes, the *error sum of squares* (***ESS***)).

The only way to determine the population parameters α and β with confidence and accuracy would be to know all values from the population. Since this is impossible, we have to estimate these parameters from a statistical **sample** of n data points. For example, to obtain a statistical sample of $n = 30$ data points for concentrations of some constituent with respect to the concentration of some other constituent, we would need to collect and analyse 30 water samples for the constituents of interest.

Then, the calculations involved in regression analysis require a variety of important concepts, involving the computation of sums of squared deviations from the means. The implementation of the least squares method requires the use of some fairly *long computations for finding the slope and intercept*. These computations will be shown in this chapter for the sake of completion. However, fortunately, **Excel** has several very **useful functions** that allow us to directly obtain the values of these model parameters and other important information for our regression analysis. These relevant Excel functions will also be described and utilized here.

(c) The regression coefficient '*b*' (slope)

Basic

The parameter β is termed the regression coefficient or the **slope** of the best-fit regression line, and the best estimate, from your sample data, is given by '***b***':

$$b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}} \tag{11.28}$$

The denominator in this calculation is always positive, but the numerator may be positive, negative, or zero, and thus, the value of the slope 'b' theoretically can range from $-\infty$ to $+\infty$, including zero.

In Excel, we can also obtain the value of the slope b directly using the function **SLOPE**:

SLOPE(known_y's, known_x's)

- **Known_y's** Required. An array or cell range of numeric dependent data points.
- **Known_x's** Required. The set of independent data points.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

(d) The *Y* intercept '*a*'

Basic

The value of Y in the population when $X = 0$ is called the Y **intercept** and is represented by the parameter α.

The best estimate of the intercept α is given by the sample intercept '***a***':

$$a = \bar{y} - b\bar{x} \tag{11.29}$$

In Excel, the intercept can also be obtained directly using the function **INTERCEPT**:

INTERCEPT(known_*y*'s, known_*x*'s)

- **Known_y's** Required. The dependent set of observations or data.
- **Known_x's** Required. The independent set of observations or data.

(e) Linear regression equation

Basic

The **sample** regression equation is given by:

$$\boxed{\bar{Y} = a + b.x + e} \tag{11.30}$$

where $e = Y - \hat{Y}$ (error or residual = observed Y – estimated Y)

Now, knowing the parameters a and b for the linear regression equation, we can predict the expected value of the dependent variable for a given value of X_i.

In Excel, we can estimate the values of Y for given values of X using the functions: **FORECAST** (or **FORECAST.LINEAR**, in newer Excel versions) and **TREND**:

- The **FORECAST** function calculates (predicts) a future value using existing values. The predicted value returned from the function is a y-value for a given x-value input. The known values are existing x-values and y-values, and the new value is predicted using linear regression. The following syntax is used:
  FORECAST(x, known_y's, known_x's)
  - **X** Required. The data point for which you want to predict a value.
  - **known_y's** Required. The dependent array or range of data.
  - **known_x's** Required. The independent array or range of data.
- The **TREND** function returns values along a linear trend. It fits a straight line (using the method of least squares) to the array's known_y's and known_x's. TREND returns the y-values along that line for the array of new_x's that you specify. Syntax:
  TREND(known_y's, [known_x's], [new_x's], [const])
  - **known_y's.** Required. The set of y-values you already know in the relationship $y = a + b.x$
  - **known_x's**. Optional. The set of x-values that you may already know in the relationship $y = a + b.x$
  - **new_x's**. Optional. New x-values for which you want TREND to return corresponding y-values
  - **const**. Optional. A logical value specifying whether to force the constant a to equal 0. TRUE for normal equation $Y = a + bX$. FALSE to force intercept to be zero, with regression equation $Y = bX$.

It is important to note that it is not safe to estimate $\hat{Y}_i$ (predicted values) for X_i values ***outside the observed range of X_i*** from the data set used to fit the model. This is called extrapolation. If you must do extrapolations, be sure to critically analyse the results using your knowledge of the system and to make it clear in your report that the estimation is outside the boundaries from which the equation has been derived.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Basic

(f) Assumptions of the regression analysis

When using regression analysis, we have to comply with the following assumptions:

- We must assume that for any value of X, the population contains a **normal distribution of Y values**. Also, for each value of X, the population has a **normal distribution of the error (e)**.
- We must assume **homogeneity of variances**: the variances of the population distribution of Y values (and errors e) must all be equal.
- In the population, the **mean of the Y values** for a given X lies on a **straight line** with all other mean Y values for the other X values. The actual relationship in the population is linear.
- The values of Y should have been obtained **randomly** from the sampled population and should be **independent** of one another.
- The measurements of X are obtained without error. Since this requirement is almost impossible to be fulfilled, we assume that the **error in the X data is negligible** or at least small compared with the **measurement error in the Y data**.

11.5.2 Testing the significance of a regression

Advanced

(a) Hypothesis test for the slope

Since the simple linear regression equation was obtained based on the data of a sample, its validity must be checked using a significance (hypothesis) test.

The dependence of Y on X is represented by the *slope coefficient b*, almost always obtained from a data sample (estimate of the true population coefficient, β). To determine the existence of a significant linear relationship between the variables X and Y, we must test whether the coefficient β (the slope in the population) is equal to 0. However, we should remember that the detection of a dependence of Y on X in the sample ($b \neq 0$) does not necessarily mean that there is dependence in the population ($\beta \neq 0$). In order to gain insight about the potential dependence in the population based on the evidence in our sample, we use a hypothesis test on the slope.

The null hypothesis and the alternative hypothesis for testing the **significance of the slope** is given below.

- **Null hypothesis H_0**: $\beta = 0$ (there is no linear relationship between X and Y).
- **Alternative hypothesis H_a**: $\beta \neq 0$ (there is a linear relationship between X and Y).

S. 11.2.1

Note that in Section 11.2.1(d), we analysed the linear correlation between two variables using the correlation coefficient 'r', and we *performed hypothesis tests to investigate whether the correlation was significant*. This is equivalent for testing the slope 'b', as we are doing here.

Let us understand this better by analysing the scatter plot and the best-fit line in Figure 11.10. We can see that the slope of the line is equal to 0.00. Therefore, the equation of the best-fit line is simply $Y = 0.25$, indicating that it is simply a line that is equal to the average of the Y values (=0.25). We can see that there is no dependence or relationship between Y and X. As a result, the correlation coefficient r is equal to 0.00 and, of course, $R^2 = 0.00$.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

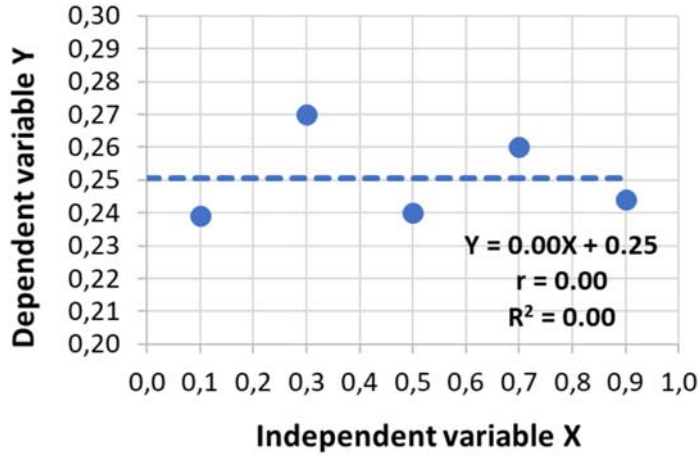


Figure 11.10 Example of a scatter plot between two variables that are not correlated.
Note: The slope of the line of best fit is 0.00 and so are the correlation coefficient *r* and the Coefficient of Determination R^2. The line is equal to the average of the *Y* data (0.25).

(b) Analysis of variance (test for the slope)

Advanced

The null hypothesis H_0 may be tested by the analysis of variance (ANOVA) procedure. The overall variability of the dependent variable is calculated by computing the sum of squares (SS) of the deviations of the *observed values* (Y_i) from the *mean of the observed values* ($\bar{Y}$), called ***total sum of squares (TSS)***:

$$\text{Total SS (TSS)} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} \tag{11.31}$$

The total variability of the data, expressed by the **TSS** is divided into:

- variation explained by the model or *explained variation* (**regression sum of squares (RSS)**)
- *non-explained variation* (**residual sum of squares,** or Error Sum of Squares (ESS), or sum of the squares for error (SSE), or sum of the squares of the residuals (SSR))

The regression sum of squares (**RSS**) is given by Equations 11.32 and 11.33:

$$\text{Regression SS (RSS)} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \tag{11.32}$$

$$\text{Regression SS (RSS)} = a\sum_{i=1}^{n} y_i + b\sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} \tag{11.33}$$

The **residual (or error) sum of squares** is obtained by Equations 11.34 and 11.35:

$$\text{Residual SS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{11.34}$$

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

$$\text{Residual SS} = \text{Total SS} - \text{Regression SS} = \text{TSS} - \text{RSS} \tag{11.35}$$

The following **Excel functions** may be used for obtaining the SS of interest:

- Total sum of squares (**total SS, TSS**): **DEVSQ**(array of observed values y_i).
- Regression sum of squares (**regression SS, RSS**): **SUMXMY2**(array of predicted values $\hat{y}_i$; array repeating the mean of the observed values $\hat{y}$) = Total SS – Residual SS.
- Residual sum of squares (**residual SS**): **SUMXMY2**(array of observed values y_i; array of predicted values $\hat{y}_i$).

Table 11.7 presents a typical format for a summary of the **ANOVA table** for testing the hypothesis H_0: $\beta = 0$ against H_a: $\beta \neq 0$. The **degrees of freedom (df)** are as follows:

- df associated with the total variability of Y_i values: **Total df** $= \boldsymbol{n - 1}$.
- df associated with the variability of Y_i's due to regression: **Regression df** $= \mathbf{1}$ in a simple linear regression.
- df associated with residuals: **Residual df** = Total df − Regression df $= (n - 1) - 1 = \boldsymbol{n - 2}$.

The *regression mean squares* (**regression MS**) and the *residual mean squares* (**residual MS**) are calculated as follows:

$$\text{Regression mean squares (Regression MS)} = \frac{\text{Regression SS}}{\text{Regression df}} \tag{11.36}$$

$$\text{Residual mean squares (Residual MS)} = \frac{\text{Residual SS}}{\text{Residual df}} \tag{11.37}$$

We then calculate the $\boldsymbol{F}$ **statistic** ($\boldsymbol{F_{calc}}$), using the values calculated in Equations 11.36 and 11.37:

$$\boxed{F = \frac{\text{Regression MS}}{\text{Residual MS}}} \tag{11.38}$$

The **critical value of** $\boldsymbol{F}$ ($\boldsymbol{F_{crit}}$) is obtained using look-up tables for the right-tailed inverse F distribution or using the Excel function **F.INV.RT**(probability, deg_freedom1, deg_freedom2). The probability is the test significance level (α). The degrees of freedom are df_1 = regression df = 1 and df_2 = residual df = $n - 2$. Thus, for a simple linear regression, $F_{\alpha,df1,df2}$ = F.INV.RT (α; 1; $n - 2$).

The test F statistic (F_{calc}) is then compared with the critical value (F_{crit}), $F_{\alpha,df1,df2}$. If $\boldsymbol{F_{calc} > F_{crit}}$, then we **reject the null hypothesis** $\mathbf{H_0}$: $\boldsymbol{\beta = 0}$ and accept the alternative hypothesis H_a: $\beta \neq 0$,

Table 11.7 ANOVA—Summary output for a regression analysis.

Source	df	Sum of squares (SS)	Mean squares (MS)	*F*
Regression	1	$\text{Regression SS} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$	$\frac{\text{Regression SS}}{\text{Regression df}}$	$F = \frac{\text{Regression MS}}{\text{Residual MS}}$
Residual	$n - 2$	Total SS − Regression SS	$\frac{\text{Residual SS}}{\text{Residual df}}$	
Total	$n - 1$	$\text{Total SS} = \sum_{i=1}^{n} (y_i - \bar{y})^2$		

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

thus concluding that the slope is significant or, in other words, that there is a significant linear relationship between X and Y.

We can also calculate the associated ***p*-value** for the F statistic. For this, we use the Excel function **F.DIST.RT**(x,deg_freedom1,deg_freedom2) = F.DIST.RT (F_{calc}; 1; $n-2$). If ***p*-value** $<$ $\boldsymbol{\alpha}$, we **reject the null hypothesis H_0: $\beta = 0$.**

(c) Standard error of the estimate, S_{yx}

Advanced

The residual MS (see Equation 11.37) is also often written as S_{yx}^2, a representation denoting that it is the variance of Y after taking into account the dependence of Y on X. The square root of this value, S_{yx}, is called the ***standard error of estimate*** (or the *standard error of the regression*):

$$S_{yx} = \sqrt{\frac{\text{Residual SS}}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}} \tag{11.39}$$

You can also obtain this value by using the Excel function **STEYX**:

STEYX. Returns the SE of the predicted y-value for each x in the regression. The SE is a measure of the amount of error in the prediction of y for an individual x. Syntax:

STEYX(known_y's, known_x's)

- Known_y's Required. An array or range of dependent data points.
- Known_x's Required. An array or range of independent data points.

The *standard error of the estimate* is an overall indication of the accuracy with which the fitted regression function predicts the dependence of Y on X. The magnitude of S_{yx} is proportional to the magnitude of the dependent variable Y.

(d) *t* test for the slope *b*

Advanced

In item 'b', we performed ANOVA, using the F statistic, to test whether β was significantly different from zero. This can also be tested by using Student's t statistic. The t statistic ($\boldsymbol{t_{calc}}$) for the testing of two-tailed hypotheses, H_0: $\beta = 0$ and H_A: $\beta \neq 0$, is calculated as follows:

$$\boxed{t = \frac{b-\beta}{S_b}} \tag{11.40}$$

where:

$$S_b = \frac{S_{yx}}{\sqrt{\text{SQX}}} \tag{11.41}$$

and

$$\text{SQX} = \sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{11.42}$$

***S_{yx}* is the standard error of estimate**, calculated in Equation 11.39. **SQX** can be calculated using the Excel function **DEVSQ**(array of observed values x_i). In Equation 11.40, $\beta = 0$, since we are testing for $\beta = 0$ (null hypothesis). S_b is the standard error of the slope b.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

After we calculate t_{calc}, we need to calculate t_{crit}, and compare both. The value of $\boldsymbol{t_{crit}}$ can be obtained from look-up tables or from the Excel function **T.INV.2T**(probability; deg_freedom). The probability is the significance level α for the test, and the degrees of freedom are $n - 2$.

If $t_{calc} > t_{crit}$, we reject H_0 and conclude that the slope is significant (i.e., there is a linear relationship between X and Y).

We can also calculate the associated ***p*-value**. For this, we use the Excel function **T.DIST.2T**(x; deg_freedom) = T.DIST.2T(ABS(t_{calc}); $n - 2$). If p-value $< \alpha$, of course, we reject the null hypothesis.

As a complement to our analysis of the significance of the slope of the equation, we may also estimate the **confidence interval for the slope (β)**. The confidence interval for the slope of the regression can be calculated for the $(1 - \alpha)$ confidence level as follows:

$$\text{Confidence interval for } \beta = b \pm t_{\alpha,n-2}.S_b \tag{11.43}$$

where

b = slope
$t_{\alpha,\ n-2}$ = t_{crit}, as calculated above
S_b = standard error of the slope b, given by Equation 11.41.

Therefore, the **lower confidence limit for b (LCL_b)** and the **upper confidence limit for b (UCL_b)** are given by:

$$\textbf{LCL for the slope } b = \text{LCL}_b = b - t_{\alpha,n-2}.S_b \tag{11.44}$$
$$\textbf{UCL for the slope } b = \text{UCL}_b = b + t_{\alpha,n-2}.S_b \tag{11.45}$$

(e) The coefficient of correlation *r*

Basic

As we saw in detail in Section 11.2.1, the **correlation coefficient *r*** is a measure of the strength of the linear relationship between the two variables x and y. In that section, we mentioned the simplified approach of obtaining r directly from the Excel function **CORREL** (array1, array2).

S. 11.2.1

The calculation of the correlation coefficient r is by using ANOVA (see Equations 11.32 and 11.31):

$$r = \sqrt{\frac{\text{Regression SS}}{\text{Total SS}}} \tag{11.46}$$

Note that r is computed using the same quantities used in fitting the least squares line. We already saw that a value of ***r* near or equal to 0 implies little or no linear relationship between *Y* and *X***. In contrast, the **closer *r* is to 1 or −1, the stronger is the linear relationship between *Y* and *X***. If $r = 1$ or $r = -1$, all the points fall exactly on the least squares line. *Positive values of* r *imply that* Y *increases as* X *increases, and negative values of* r *imply that* Y *decreases as* X *increases*.

S. 11.2.1

In Section 11.2.1(d), we presented the hypothesis test used to test the significance of a correlation, where the **null hypothesis, H_0, is $\rho = 0$, and the alternative hypothesis, H_a, is $\rho \neq 0$** (and ρ is the population correlation coefficient). In that case, we tested the hypothesis that X contributes no information for the prediction of Y, using the linear model, against the alternative that the two variables are at least linearly related. Note that this is *equivalent to the test we performed for the slope*, when we tested **H_0: $\beta = 0$ against H_a: $\beta \neq 0$** (item 'd' above).

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Thus, **$\beta = 0$ implies that $r = 0$, and vice versa** (see Figure 11.10 again). Consequently, **the null hypothesis H_0: $\rho = 0$ is equivalent to the hypothesis H_0: $\beta = 0$**, and the information provided by both tests about the utility of the least squares model is to some extent redundant. Furthermore, the slope β gives us additional information on the amount of increase (or decrease) in Y for every 1-unit increase in X. However, remember that in Section 11.2.1, items 'e' and 'f', we performed advanced calculations for setting up confidence limits for the correlation coefficient and for *testing whether it could be equal to any value other than zero*. Therefore, the usefulness of testing for ρ is clear, if we consider these broader goals.

(f) The Coefficient of Determination r^2 or R^2

Basic

Another way to measure the utility of the regression model is to quantify the contribution of x in predicting y. To do that, we compute how much the errors of the prediction of y were reduced by using the information provided by x.

The **Coefficient of Determination r^2** (or also **R^2**) is the *proportion (or percentage) of the total variation in* y *that is explained by the fitted regression model*. Therefore, if we have a value of r^2 equal to, say, 0.79, this means that 0.79 (or 79%) of the variance of y has been explained by our model.

The calculation of the r^2 value is made by using ANOVA (see Equations 11.32 and 11.31):

$$r^2 = \frac{\text{Regression SS}}{\text{Total SS}} = 1 - \frac{\text{Residual SS}}{\text{Total SS}} \tag{11.47}$$

From the notations and Equations 11.46 and 11.47, we see that r^2 is simply the correlation coefficient r raised to the power two. Therefore, we may conclude that **r^2 varies from 0 to +1**.

S. 15.2.3

C. 11

In Section 15.2.3(b), we will further discuss the concept of the Coefficient of Determination (CoD) from a broader perspective, showing its calculation and also its interpretation for regression-based models (as seen here in this chapter) and non-regression-based models (or process-based models). In this chapter, you will see that, for regression-based models, CoD is the same as r^2, and thus, it varies from 0 to +1. However, for non-regression-based models, CoD may vary from $-\infty$ to +1.

In Excel, we can calculate r^2 directly using one of the following functions:

- **RSQ**(known_y's, known_x's)
- **PEARSON** raised to power 2, where PEARSON(array1, array2)
- **CORREL** raised to power 2, where CORREL(array1, array2)
- **SUMXMY2** (numerator of right-hand side Equation 11.47) and **DEVSQ** (denominator of Equation 11.47)

Excel uses the notation R^2 in the graphs that make use of the 'add trendline' feature.

11.5.3 Confidence intervals and prediction intervals

Advanced

Our linear regression equation allows us to estimate a value of Y based on a value of X, given the best fit values of the intercept a and slope b. But this equation has been developed based on our **sample**, which comprises measured values. If we think in terms of a **population**, there is likely to be a variation in our prediction of Y, and it would be important for us to set the limits for our prediction based on a certain confidence level.

S. 4.5.4

In Section 4.5.4, we presented the concepts of confidence intervals and prediction intervals, and you should go there to review these concepts. For our application here, we can describe these concepts as follows:

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

- A **confidence interval** tells you the interval within which the **true mean value** of the population will fall, with a given probability (e.g., 95%).
- A **prediction interval** tells you the interval within which a **single value** of $\hat{Y}_i$ taken from the population will fall, with a given probability (e.g., 95%).

Figure 11.11 illustrates these concepts for the application in a regression analysis. You may see that, as expected, the *width of the prediction interval for a single value of* $\hat{Y}_i$ *is broader than the width of the confidence interval for the mean value of* $\hat{Y}$. From the entire population, for a given value of X, the *true mean of* $\hat{Y}$ *is expected to be inside the boundaries of the confidence interval*, while the *estimate for a single value of* $\hat{Y}_i$ *is expected to be within the limits of the prediction interval*, for a certain confidence level (equal to $1 - \alpha$).

In general, the estimation of an interval based on the t statistic and the standard error (SE) of the statistic is given by the following equation:

$$\text{Confidence interval} = \text{statistic} \pm (t)\ (\text{SE of statistic}) \tag{11.48}$$

(a) Confidence interval for the prediction of the mean value of $\hat{Y}$

Advanced

The regression equation allows us to predict the **mean value of $\hat{Y}$** for a given value of X. The *SE of the prediction of the mean value of* $\hat{Y}$ *is* given by:

$$S_{\hat{y}i} = S_{YX}\sqrt{\left[\frac{1}{n} + \frac{(X_i - \bar{X})^2}{\text{SQX}}\right]} \tag{11.49}$$

where

$S_{\hat{y}i}$ = standard error of the prediction of the mean value of Y
S_{YX} = standard error of estimate (Equation 11.39)
X_i = value of X for which estimation of Y will be made
SQX = $\sum(x_i - \bar{x})^2$ (Equation 11.42).

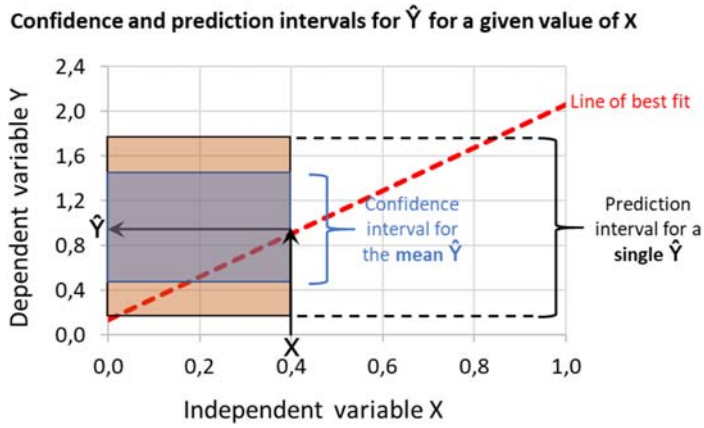


Figure 11.11 Concept of confidence intervals and prediction intervals in a regression analysis.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

We can see from Equation 11.49 that the standard error has a minimum of $X_i = \bar{x}$ and that it increases as the estimates are made at values of X_i farther away from the mean.

$$\text{Confidence interval for the mean } \hat{Y}_i = \hat{Y}_i \pm t_{\alpha,n-2}.S_{\hat{Y}i} \tag{11.50}$$

where

$\hat{Y}_i$ = predicted value of Y_i, for a given value of X_i, using the linear regression equation
$t_{\alpha,n-2} = t_{\text{crit}}$, for significance level α and $n-2$ degrees of freedom. It can be calculated using Excel function **T.INV.2T**(probability; deg_freedom) = **T.INV.2T**(α; $n-2$).

Therefore, the **lower confidence limit for the mean of $\hat{Y}_i$ (LCL)** and the **upper confidence limit for the mean of $\hat{Y}_i$ (UCL)** are given by:

$$\textbf{Lower confidence limit (LCL)} = \hat{Y}_i - t_{\alpha,n-2}.S_{\hat{Y}i} \tag{11.51}$$

$$\textbf{Upper confidence limit (UCL)} = \hat{Y}_i + t_{\alpha,n-2}.S_{\hat{Y}i} \tag{11.52}$$

(b) Prediction interval for a single value of $\hat{Y}$

Advanced

If we wish to estimate the prediction interval for the value of a **single observation** taken from the population for a specified value of X_i, Equation 11.53 may be used. This equation estimates the *standard error of the prediction of a single value of* $\hat{Y}_i$:

$$(S_{\hat{Y}i})_1 = S_{YX}\sqrt{\left[1 + \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\text{SQX}}\right]} \tag{11.53}$$

The prediction interval for the single value of $\hat{Y}_i$ is calculated using the same procedure illustrated above, only by using $(S_{\hat{Y}i})_1$ (Equation 11.53) instead of $S_{\hat{Y}i}$.

$$\text{Prediction interval for } \hat{Y}_i = \hat{Y}_i \pm t_{\alpha,n-2}.\left(S_{\hat{Y}i}\right)_1 \tag{11.54}$$

Therefore, the **lower prediction limit for a single value of $\hat{Y}_i$ (LPL)** and the **upper prediction limit for a single value of $\hat{Y}_i$ (UPL)** are given by:

$$\textbf{Lower prediction limit (LPL)} = \hat{Y}_i - t_{\alpha,n-2}.\left(S_{\hat{Y}i}\right)_1 \tag{11.55}$$

$$\textbf{Upper prediction limit (UPL)} = \hat{Y}_i + t_{\alpha,n-2}.\left(S_{\hat{Y}i}\right)_1 \tag{11.56}$$

11.5.4 Residual analysis

Advanced

S. 15.3

S. 15.3

We have been using the term 'error' or 'residual' here to express the difference between an observed (Y_i) and a predicted ($\hat{Y}_i$) value of the dependent variable Y. The analysis of the residuals is an integral part of the development and assessment of our model's performance. This topic is discussed in detail in Section 15.3, for models in general (regression-based and non-regression-based models). The principles are the same.

For the sake of completeness on the topic of regression analysis, we cover this subject here. However, you should consult Section 15.3 for a broader view of residuals analysis. The discussion below will give you an introduction to the subject.

Basically, the **residuals** generated from our regression model need to comply with the following **assumptions**:

- **Linearity**: The residuals ε_i have mean of 0.
- **Independence**: The residuals ε_i are independent.
- **Normality**: The residuals ε_i are normally distributed.
- **Homogeneity of variances**: The residuals ε_i have the same variance σ^2.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

(a) Linearity

Advanced

Violations of the *linearity* assumption are very serious. If we fit a linear model to the data that are non-linearly related, our predictions are likely to be severely wrong, *especially when we extrapolate beyond the range of the sample data*. The nonlinearity is usually most evident in a plot of *observed versus predicted values* or a plot of *residuals versus predicted values*, which are a part of a standard regression output. The points should be symmetrically distributed around a diagonal line in the former plot or around a horizontal line in the latter plot, with a roughly constant variance. The **residual-versus-predicted plot** is better than the observed-versus-predicted plot for this purpose, because it eliminates the visual distraction of a sloping pattern. Look carefully for evidence of a 'bowed' pattern, indicating that the model makes systematic errors whenever it is making unusually large or small predictions.

(b) Independence

Advanced

Violations of *independence* are potentially very serious, particularly for time series regression models: serial correlation in the errors (i.e., correlation between consecutive errors or errors separated by some other number of periods) means that there is room for improvement in the model, and extreme serial correlation is often a symptom of a badly mispecified model. Serial correlation (also known as **autocorrelation** – see Section 11.4.2) is sometimes a by-product of a violation of the linearity assumption, as in the case of a simple (i.e., straight) trend line fitted to the data that are growing exponentially over time.

S. 11.4.2

Independence can also be violated in non-time-series models if errors tend to always have the **same sign** under particular conditions, i.e., if the model systematically **underpredicts** or **overpredicts** the dependent variable when the independent variables have a particular configuration.

S. 11.4.2

You can diagnose this by interpreting the **autocorrelogram** of the residuals – see Section 11.4.2 and Example 11.6, and by analysing autocorrelations in comparison with confidence intervals (autocorrelations should be inside the envelope of the confidence limits). Pay special attention to significant correlations in the first lag periods and in the vicinity of the seasonal period, because these are probably not due to mere chance and they are also fixable. Also, you can calculate the Durbin–Watson statistic, as described in Section 15.3.5, to test for significant residual autocorrelation at lag period 1.

S. 15.3.5

(c) Homogeneity of variances

Advanced

Violations of homogeneity of variances (which are called '**heteroscedasticity**') make it difficult to derive the true standard deviation of the forecast errors, usually resulting in *confidence intervals that are too wide or too narrow*. In particular, if the variance of the errors is increasing over time, confidence intervals for out-of-sample predictions will tend to be unrealistically narrow. Heteroscedasticity may also have the effect of giving too much weight to a small subset of the data (namely the subset where the error variance was largest) when estimating coefficients.

We should generate **plots of residuals versus independent variables** to look for consistency. Because of imprecision in the coefficient estimates, the errors may tend to be slightly larger for forecasts associated with predictions or values of independent variables that are extreme in both directions. We hope not to see errors that systematically get larger in one direction by a significant amount.

(d) Normality

Advanced

Violations of normality create problems for determining whether **model coefficients are significantly different from zero** and for **calculating confidence intervals for forecasts**.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Sometimes the error distribution is 'skewed' by the presence of a few large outliers. Since parameter estimation is based on the minimization of the sum of squared errors, a few extreme observations can exert a disproportionate influence on parameter estimates. The calculation of confidence intervals and significance tests for coefficients are all based on assumptions of normally distributed errors. If the error distribution is significantly non-normal, *confidence intervals may be too wide or too narrow*.

Technically, the normal distribution assumption is less serious if you are willing to assume that the model equation is correct and your only goal is to estimate its coefficients using minimized mean squared error and to generate point estimate predictions. The formulas for estimating coefficients require no more than that, and some references on regression analysis do not list normally distributed errors among the key assumptions. But generally, we are interested in making inferences about the model and/or estimating the probability that a given forecast error will exceed some threshold in a particular direction, in which case distributional assumptions are important. Also, a significant violation of the normal distribution assumption is often a warning, indicating that there is some other problem with the model assumptions or that there are a few unusual data points that should be studied more closely.

S. 8.2.8
S. 15.3.2

Verification of normality can be done following the procedures described in Sections 8.2.8 and 15.3.2, involving the interpretation of **normal probability and Q-Q plots** and, if necessary, performing statistical **tests for normality** (such as the *Shapiro–Wilk test*).

11.5.5 The effect of influential observations and outliers in the regression analysis

Advanced

In any sample of observations, there is the possibility of having one or more unexpectedly high or low values. These values are so far away in magnitude from the other observations that they do not seem to be representative of the sample – that is, they do not apparently have the same distribution. Such unexpectedly high or low values are called **outliers** and can unduly influence the estimation of the parameters of a probability model, unless we identify and deal with them appropriately. We have to examine whether these points are **influential** in the model results, which means that if we remove them, there will be a significant change in the estimates of the model parameters. The detection of outliers is discussed in Section 5.5.

S. 5.5

As mentioned in Section 5.5, the prior detection of outliers is often based on probability plots or box plots and depends on the type of data and how they are presented. After the initial detection, more formal identification is possible through appropriate statistical tests. Under the normal null hypothesis, there are several methods, which can be used, such as the Shapiro-Wilk's test and the skewness and kurtosis statistics.

Now, let us analyse the possible influence of outliers in the regression analysis by a simple example, depicted in Figure 11.12. The points inside the marked area are clearly different from the main cloud of the data points and, because they are more distant, they substantially change (in this case, increase) the value of the Coefficient of Determination. When these outliers are removed from the analysis, the data behaviour seems to be more realistic, and a smaller r^2 value is obtained. You should also analyse whether the slope of the line of best fit changed, because this would imply an important modification to your model. But, note that *removing outliers is a complex issue*, and you should reflect considerably on the pertinence of your decision. See again Section 5.5 to review these concepts.

S. 5.5

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

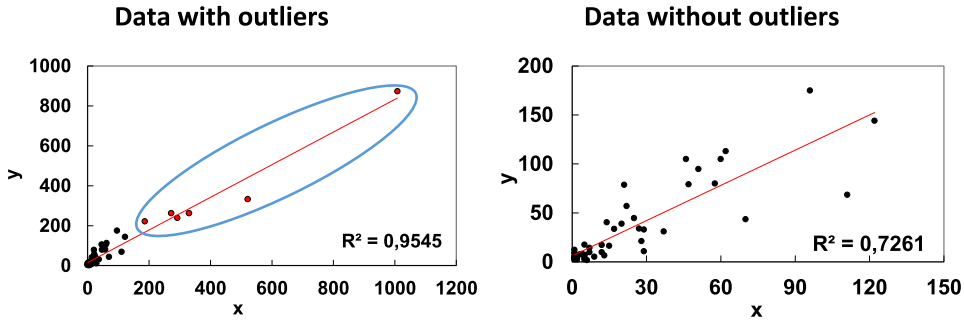


Figure 11.12 Example of the influence of outliers in a linear regression analysis.

11.5.6 Data transformation

Advanced

The analysis of the residuals can provide important information on the performance of our model and of the possible need of data transformations to be introduced in order to improve its explanatory capacity.

As discussed in the previous sections, the testing of regression hypotheses and the computation of confidence intervals depend on the assumptions of normality and homoscedasticity, with regard to the values of y, the dependent variable. There are several options to transform the data to achieve closer approximations to these assumptions.

A transformation of the sample data is defined as a process in which the measurements on the original scale are systematically converted to a new scale. For example, if the original variable is y and the variances associated with the variable across the range of x values are not equal (i.e., they are heterogeneous), it may be necessary to work with a new variable such as $\sqrt{y}$, log y, or some other transformation of the variable y.

Finding the appropriate transformation is no easy task and often takes a great deal of experience. However, a good statistical computer package or a spreadsheet program such as Excel will be able to compute several transformations and work with the transformed data, so that you can analyse the results obtained and see whether you are satisfied with the final outcome.

For further information, we recommend consulting statistical textbooks by Sokal and Rohlf (1995), Zar (1999), Ott and Longnecker (2010), and Mendenhall and Sincich (2012).

11.5.7 Complete example of a simple linear regression

Example 11.7 presents a complete application of simple linear regression analysis, showing all calculations and relevant statistics, tables, and graphs.

Example

EXAMPLE 11.7 EXAMPLE OF A COMPLETE SIMPLE LINEAR REGRESSION ANALYSIS

In Example 11.1, we performed a full correlation analysis with the data from two water constituents in a river. Now we will use the same data and perform a complete simple linear regression analysis.

The data include 20 values of constituent X and 20 values of constituent Y ($n = 20$) that have been collected simultaneously in the river.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

The data from Example 11.1 are repeated below.

Measured values of constituents *X* and *Y*

Sample Number	Constituent *X* (mg/L)	Constituent *Y* (mg/L)	Sample Number	Constituent *X* (mg/L)	Constituent *Y* (mg/L)
1	4.7	6.9	11	6.9	7.4
2	5.2	7.7	12	7.5	7.6
3	5.1	7.4	13	7.7	7.8
4	4.7	6.8	14	7.1	8.3
5	3.5	6.3	15	7.5	8.6
6	3.3	5.2	16	7.3	8.7
7	3.8	5.4	17	6.8	7.7
8	4.0	6.0	18	5.2	7.0
9	5.9	6.6	19	4.9	6.8
10	7.3	7.3	20	4.3	6.6

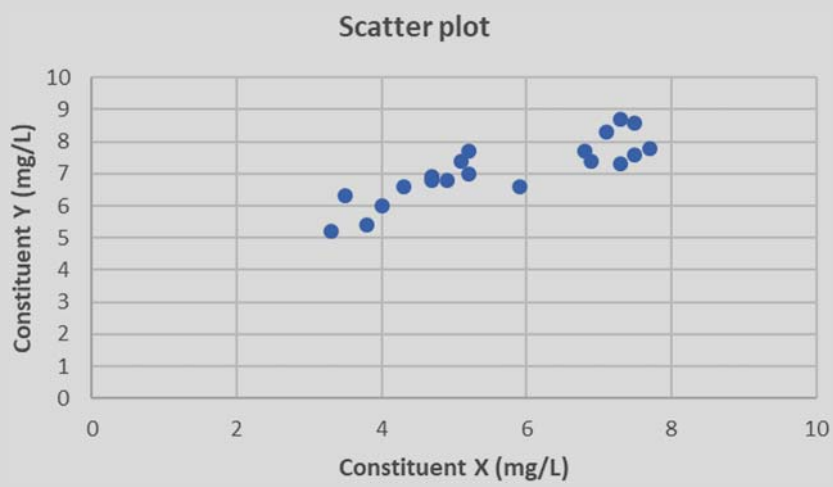
Excel

Note: This example is also available as an Excel spreadsheet.

Solution:

We will present here a complete example of a full regression analysis. In the **first part**, we will show how to interpret the results from the Summary Output table for the regression analysis undertaken using the **Excel add-in** '**Analysis ToolPak's Regression tool**' (or any statistical software). In the **second part** of the example, we will show you how to perform **all calculations step by step**. In the **third part**, we present the **Residuals Analysis**.

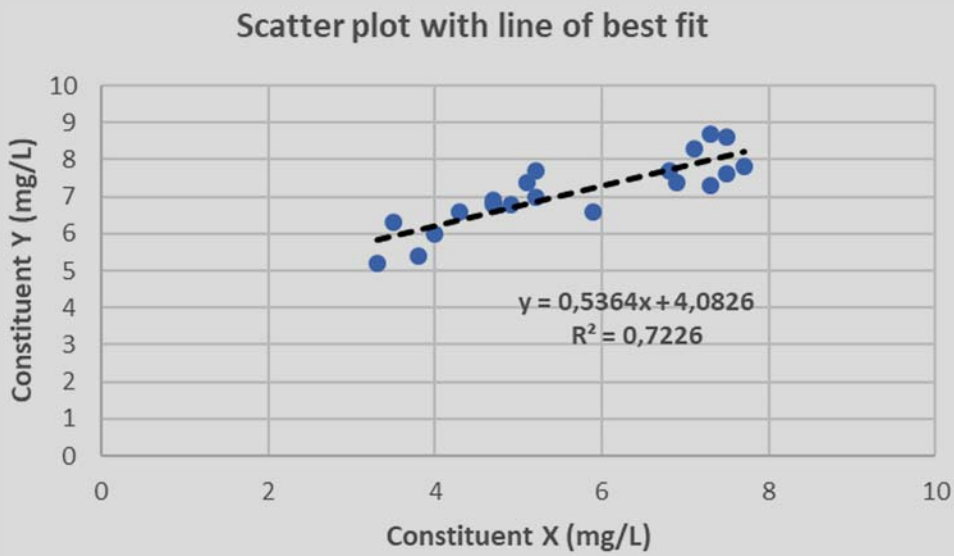
As always, our first step is to analyse the data visually. In this case, the graph we plot is the traditional scatter plot, the same way we did in Example 11.1 for the correlation analysis. The chart is shown below.



Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

The plot indicates that there is an imperfect but generally increasing relation between *x* and *y*. A linear (straight-line) relation appears plausible, and there is no evidence of the need to make transformations in the data. Also, there is no detection of any outlier falling far from the general pattern of the data. As a result, we continue with the study of the linear regression analysis between the two variables.

After viewing this scatter plot, we can use Excel to fit a line to the data points. This is accomplished by using the Excel feature '**Add a trendline**', selecting '**Linear**', and marking the selection for including the 'equation' and the value of 'R^2'. The resulting plot is shown below. Many users will go only as far as obtaining this chart because it includes the most important information we need. However, in this example, we will show you to go beyond this chart and the information associated with it.



From the graph, we have already obtained important information:

- Intercept: $\boldsymbol{a} = \mathbf{4.083}$
- Slope: $\boldsymbol{b} = \mathbf{0.536}$
- Coefficient of Determination: $\boldsymbol{R^2} = \mathbf{0.723}$

PART 1. INTERPRETATION OF THE SUMMARY OUTPUT TABLE FROM THE REGRESSION ANALYSIS

We can also use the Excel add-in '**Analysis ToolPak's Regression tool**'. As we mentioned, it is not dynamic, and you need to use it every time you want to do a regression analysis, if you change your input data, you will have to rerun the analysis tool. Also, you do not know how the calculations have been performed, because we do not see the functions being implemented. The analysis is complete and produces important information related to our regression analysis (regression statistics table, ANOVA table, and residuals results). The regression results are shown below.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Summary output from Excel add-in "Analysis ToolPak's Regression tool

	A	B	C	D	E	F	
1	**SUMMARY OUTPUT**						
2							
3	*Regression Statistics*						
4	Multiple R	0.850					
5	R Square	0.723					
6	Adjusted R Square	0.707					
7	Standard Error	0.512					
8	Observations	20					
9							
10	**ANOVA**						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	1	12.292	12.292	46.896	2.081E-06	
13	Residual	18	4.718	0.262			
14	Total	19	17.010				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*p-value*	*Lower 95%*	*Upper 95%*
17	Intercept	4.083	0.456	8.954	4,75E-08	3.125	5.041
18	X Variable 1	0.536	0.078	6.848	2.08E-06	0.372	0.701
19							
20							
21							
22	**RESIDUALS RESULTS**						
23							
24	*Observation*	*Y predicted*	*Residuals*				
25	1	6.603	0.297				
26	2	6.872	0.828				
27	3	6.818	0.582				
28	4	6.603	0.197				
29	5	5.960	0.340				
30	6	5.853	-0.653				
31	7	6.121	-0.721				
32	8	6.228	-0.228				
33	9	7.247	-0.647				
34	10	7.998	-0.698				
35	11	7.784	-0.384				
36	12	8.105	-0.505				
37	13	8.213	-0.413				
38	14	7.891	0.409				
39	15	8.105	0.495				
40	16	7.998	0.702				
41	17	7.730	-0.030				
42	18	6.872	0.128				
43	19	6.711	0.089				
44	20	6.389	0.211				

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Now, we will show how to interpret this **Summary Output** table. This will be discussed in the next six steps. We will then show you how to perform all the calculations after that (Part 2).

(a) ***Step 1*. Hypothesizing a straight-line model**

First, we hypothesize a straight-line model to relate the constituent concentrations:

$$y = a + bx$$

(b) ***Step 2*. Organizing the input data in a table format**

We obtain the x and y values for each of the $n = 20$ data points and organize them in a table format (see Computational table later in this example).

(c) ***Step 3*. Obtaining the model parameters (Summary Output table)**

From the Summary Output table, we obtain the unknown parameters in the component of the simple linear regression analysis.

The least-squares estimates of a and b are as follows:

- Intercept: $a = 4.083$
- Slope: $b = 0.536$

Thus, the simple linear regression equation is (after rounding) as follows:

$$\hat{\boldsymbol{y}} = \mathbf{4.083} + \mathbf{0.536}\boldsymbol{x}$$

This equation was displayed in the scatter plot with the best-fit line we showed above. The model equations were included in this plot and also in the summary regression output presented above.

The least squares estimate of the slope, $b = 0.536$, implies that the estimated mean value of constituent Y increases by 0.536 for each additional 1 mg/L of constituent X. The interpretation of the estimated intercept, $a = 4.083$, is that the mean concentration of constituent Y will be 4.083 when the concentration of constituent X is equal to zero.

(d) ***Step 4*. Performing a residuals analysis (not shown in the Summary Output table)**

S. 15.3
S. 11.5.4

We should perform a residuals analysis ($Y_i - \hat{Y}_i$) following the procedures outlined in Section 15.3 and supported by the additional discussion in Section 11.5.4. The values of the residuals are shown in the Summary Output table, but the analysis of the residuals in terms of the compliance with the required assumptions is not performed there. In Part 3 of this example, we will show you the outcome of the residuals analysis using the associated Excel spreadsheet for residual analysis.

(e) ***Step 5*. Testing the significance of the slope *b* and assessing the goodness of fit of the model (Summary Output table)**

We can check the utility of the hypothesized model, that is, whether x really contributes information for the prediction of y, using the straight-line model.

(i) **Significance of the slope *b*:** First, we test the null hypothesis that the population slope $\beta = 0$, that is, that there is no linear relationship between constituent Y and constituent X. We test this hypothesis against an alternative hypothesis that x and y are linearly related at a significance level of $\alpha = 0.05$. In mathematical notation, the null and alternative hypotheses are thus given as follows:

$$H_0 : \beta = 0$$
$$H_a : \beta \neq 0$$

The value of the test statistic highlighted on the Summary Output table is $\boldsymbol{t} = \mathbf{6.848}$, and the two-tailed p-value of the test also highlighted is $\mathbf{2.08 \times 10^{-6}}$. Since p-value $< \alpha = 0.05$, there is sufficient evidence to reject H_0 and conclude that the constituent X concentration does

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

indeed contribute information for the prediction of constituent Y concentration in the river, and that the mean Y concentration increases as the concentration of X increases.

p-value = **T.DIST.2T**(x; deg_freedom) = T.DIST.2T(ABS(t_{calc}); $n-2$) = T.DIST.2T (6.848; $20-2$) = $\mathbf{2.08 \times 10^{-6}}$

(ii) Confidence interval for the slope β: Confidence interval for slope: a 95% confidence interval for β is highlighted on the Summary Output table. The values are as follows: LCL = 0.372; UCL = 0.701. Thus, we are 95% confident that the interval from 0.372 to 0.701 includes the true mean increase in constituent Y concentration per each additional 1 mg/L of the constituent X (i.e., slope β).

(iii) Coefficient of Determination r^2 and coefficient of correlation r: The numerical descriptive measure of model adequacy (highlighted in the Summary Output table) is the Coefficient of Determination $\mathbf{r^2 = 0.723}$. This value implies that about 72% of the sample variation in constituent Y concentration is explained by the constituent X concentration in a linear model.

The coefficient of correlation, $\mathbf{r = 0.850}$, which measures the strength of the linear relationship between x and y, is also shown and highlighted in the Summary Output table. The good correlation confirms the conclusion that b differs from 0 and that constituents Y and X are linearly correlated.

(f) *Step 6.* Use of the linear regression model

We can now use the least squares model. Suppose we want to predict the concentration of constituent Y for a constituent X concentration = **4.7 mg/L** (first value in the X sample).

$$\hat{y} = 4.083 + 0.536x$$
$$\hat{y} = 4.083 + 0.536 \times (4.7) = \mathbf{6.60\ mg/L}$$

The **95% confidence interval** for the **mean value of the prediction of *y*** is as follows:

$$\hat{y}_i \pm t_{\alpha,n-2}.S_{\hat{Y}i} \text{ and } S_{\hat{Y}i} = S_{YX}\sqrt{\left[\frac{1}{n} + \frac{(X_i - \bar{X})^2}{SQX}\right]}$$

where S_{yx} = standard error of estimate = **0.512**, highlighted in the Summary Output table (see Equation 11.39). **SQX** can be calculated using the Excel function **DEVSQ**(array of observed values x_i; see Equation 11.42). The mean value of X is 5.64 mg/L, SQX is 42.73, $n = 20$, $t_{\alpha,\ n-2} = t_{0.05,20-2} =$ 2.101. The X value for which we want to do the calculation is $X_i = 4.7$ mg/L.

$$S_{\hat{Y}i} = 0.512\sqrt{\left[\frac{1}{20} + \frac{(4.7 - 5.64)^2}{42.73}\right]} = 0.512 \times 0.266 = 0.136$$

$$\hat{y}_i \pm t_{\alpha,n-2}.S_{\hat{y}i} = 6.60 \pm 2.101 \times 0.136 = 6.60 \pm 0.286$$

Therefore, we predict that the true population **mean** value of constituent Y for a given value of constituent $X = 4.7$ mg/L will fall between **6.32** and **6.89 mg/L**, with 95% confidence. Our best estimate of the mean value for constituent Y is 6.60 mg/L.

PART 2. FULL CALCULATION OF THE SIMPLE LINEAR REGRESSION STEP BY STEP

The following table presents the calculations required for the regression analysis. You may find some differences from the calculations here and those in the associated Excel spreadsheet due to rounding errors.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Computational table for the regression analysis

Sample	Constituent X (mg/L)	Constituent Y (mg/L)	xy	x^2	y^2	$\hat{y}$	$(x-\bar{x})^2$	$(y-\bar{y})^2$	$(\hat{y}-\bar{y})^2$	$y-\hat{y}$	$(y-\hat{y})^2$	$e=y-\hat{y}$	$(e)^2$
1	4.7	6.9	32.4	22.1	47.6	6.60	0.874	0.042	0.252	0.297	0.088	0.297	0.09
2	5.2	7.7	40.0	27.0	59.3	6.87	0.189	0.354	0.054	0.828	0.686	0.828	0.69
3	5.1	7.4	37.7	26.0	54.8	6.82	0.286	0.087	0.082	0.582	0.339	0.582	0.34
4	4.7	6.8	32.0	22.1	46.2	6.60	0.874	0.093	0.252	0.197	0.039	0.197	0.04
5	3.5	6.3	22.1	12.3	39.7	5.96	4.558	0.648	1.311	0.340	0.116	0.340	0.12
6	3.3	5.2	17.2	10.9	27.0	5.85	5.452	3.629	1.569	−0.653	0.426	−0.653	0.43
7	3.8	5.4	20.5	14.4	29.2	6.12	3.367	2.907	0.969	−0.721	0.520	−0.721	0.52
8	4	6	24.0	16.0	36.0	6.23	2.673	1.221	0.769	−0.228	0.052	−0.228	0.05
9	5.9	6.6	38.9	34.8	43.6	7.25	0.070	0.255	0.020	−0.647	0.419	−0.647	0.42
10	7.3	7.3	53.3	53.3	53.3	8.00	2.772	0.038	0.798	−0.698	0.487	−0.698	0.49
11	6.9	7.4	51.1	47.6	54.8	7.78	1.600	0.087	0.460	−0.384	0.147	−0.384	0.15
12	7.5	7.6	57.0	56.3	57.8	8.11	3.478	0.245	1.001	−0.505	0.255	−0.505	0.26
13	7.7	7.8	60.1	59.3	60.8	8.21	4.264	0.483	1.227	−0.413	0.170	−0.413	0.17
14	7.1	8.3	58.9	50.4	68.9	7.89	2.146	1.428	0.617	0.409	0.167	0.409	0.17
15	7.5	8.6	64.5	56.3	74.0	8.11	3.478	2.235	1.001	0.495	0.245	0.495	0.24
16	7.3	8.7	63.5	53.3	75.7	8.00	2.772	2.544	0.798	0.702	0.493	0.702	0.49
17	6.8	7.7	52.4	46.2	59.3	7.73	1.357	0.354	0.390	−0.030	0.001	−0.030	0.00
18	5.2	7	36.4	27.0	49.0	6.87	0.189	0.011	0.054	0.128	0.016	0.128	0.02
1	4.9	6.8	33.3	24.0	46.2	6.71	0.540	0.093	0.155	0.089	0.008	0.089	0.01
20	4.3	6.6	28.4	18.5	43.6	6.39	1.782	0.255	0.513	0.211	0.045	0.211	0.04
Σ	**112.7**	**142.1**	**823.7**	**677.8**	**1026.6**	**142.10**	**42.726**	**17.010**	**12.292**	**0.000**	**4.718**	**0.000**	**4.72**
Mean	**5.6**	**7.1**											

(a) Regression equation (model parameters)

The slope is calculated from Equation 11.28, which uses results from the computational table.

$$b = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}} = \frac{823.7 - \frac{112.7 \times 142.1}{20}}{677.8 - \frac{112.7^2}{20}} = 0.536$$

The intercept is calculated from Equation 11.29:

$$a = \bar{y} - b\,\bar{x} = 7.1 - 0.536 \times 5.6 = 4.083$$

The slope and the intercept can be also determined using the Excel functions **SLOPE** (known_*y*'s, known_*x*'s) and **INTERCEPT**(known_*y*'s, known_*x*'s).

Therefore, the simple linear regression equation is as follows:

$$\hat{y} = 4.083 + 0.536x$$

$\hat{y}$: estimated, predicted, or expected value of constituent Y as a function of constituent X.

(b) Testing the significance of the regression

The test hypotheses for the significance of the slope β of the equation are as follows:

Null hypothesis H_0: $\beta = 0$.

Alternative hypothesis H_a: $\beta \neq 0$.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

H_0 may be tested using the ANOVA procedure.

Source	df	Sum of Squares (SS)	Mean Squares (MS)	*F*
Regression	1	$\text{Regression SS} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$	$\frac{\text{Regression SS}}{\text{Regression df}}$	$F = \frac{\text{Regression MS}}{\text{Residual MS}}$
Residual	$20 - 2 = 18$	Total SS − Regression SS	$\frac{\text{Residual SS}}{\text{Residual df}}$	
Total	$20 - 1 = 19$	$\text{Total SS} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$		

df, degrees of freedom; n, number of data (n = 20).

Sum of squares SS

Total SS (Equation 11.31):

$$\text{Total SS} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = 17.010$$

Regression SS (Equation 11.32):

$$\text{Regression SS} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = 12.292$$

Residual SS (Equation 11.34):

$$\text{Residual SS} = \text{Total SS} - \text{Regression SS} = 17.010 - 12.292 = 4.718$$

The following **Excel functions** may be used for obtaining the sum of squares of interest:

- Total sum of squares (**total SS**): **DEVSQ**(array of observed values y_i)
- Regression sum of squares (**regression SS**): **SUMXMY2**(array of predicted values $\hat{y}_i$; array repeating the mean of the observed values $\hat{y}$) = Total SS – Residual SS
- Residual sum of squares (**residual SS**): **SUMXMY2**(array of observed values y_i; array of predicted values $\hat{y}_i$)

Mean squares MS

Regression MS (Equation 11.36):

$$\text{Regression MS} = \frac{\text{Regression SS}}{\text{Regression df}} = \frac{12.292}{1} = 12.292$$

Residual MS (Equation 11.37):

$$\text{Residual MS} = \frac{\text{Residual SS}}{\text{Residual df}} = \frac{4.718}{18} = 0.262$$

***F* test for the slope β**

The *F* statistic (F_{calc}) is given by Equation 11.38:

$$F = \frac{\text{Regression MS}}{\text{Residual MS}} = \frac{12.292}{0.262} = 46.896$$

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

The critical value of F ($\boldsymbol{F_{crit}}$) is obtained using look-up tables for the right-tailed inverse F distribution or the Excel function F.INV.RT(probability,deg_freedom1,deg_freedom2) = F.INV.RT (α, 1, $n-2$) = F.INV.RT(0.05, 1, 20 – 2) = **4.414**.

Since $F_{calc} > F_{crit}$, or 46.896 > 4.414, we **reject the null hypothesis H_0** that the slope is equal to zero and thus conclude that the slope is significant (at $\alpha = 0.05$).

We can also calculate the associated ***p*-value** for the F statistic. For this, we use the Excel function **F.DIST.RT**(x,deg_freedom1,deg_freedom2) = F.DIST.RT (F_{calc}; 1; n – 2) = F.DIST.RT (46.896; 1; 20 – 2) = $\mathbf{2.081 \times 10^{-6}}$.

Since the p-value $< \alpha$, we reject the null hypothesis H_0: $\beta = 0$.

***t* test for the slope β**

Using t statistic for the testing of two-tailed hypotheses, H_0: $\beta = 0$ and H_a: $\beta \neq 0$ (using Equations 11.40 to 11.42), we have:

$$S_b = \frac{S_{yx}}{\sqrt{SQX}} = \frac{S_{yx}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

The standard error of estimate S_{yx} is calculated from Equation 11.39:

$$S_{yx} = \sqrt{\frac{\text{Residual SS}}{n-2}} = \sqrt{\frac{4.718}{20-2}} = 0.512$$

You can calculate S_{yx} directly using the Excel function STEYX(known_*y*'s, known_*x*'s).

Then,

$$S_b = \frac{S_{yx}}{\sqrt{SQX}} = \frac{S_{yx}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} = \frac{0.512}{\sqrt{42.726}} = \mathbf{0.078}$$

and

$$t = \frac{b - \beta}{S_b} = \frac{0.536 - 0}{0.078} = \mathbf{6.848}$$

The value of $\boldsymbol{t_{crit}}$ can be obtained from the Excel function **T.INV.2T**(probability; deg_freedom). The probability is the significance level α for the test and the degrees of freedom are n – 2. The resulting critical value of t, at 0.05 significance level, is $t_{0.05,n-2} = t_{0.05,\ 18} = \mathbf{2.101}$.

Since $t = 6.848 > 2.101$, we reject H_0.

We can also calculate the associated ***p*-value**. For this, we use the Excel function **T.DIST.2T**(*x*; deg_freedom) = T.DIST.2T(ABS(t_{calc}); $n-2$) = T.DIST.2T(ABS(6.848); 20 – 2) = $\mathbf{2.081 \times 10^{-6}}$. Note that this has the same value of the p-value calculated using the F statistic.

Since p-value $< \alpha$, we **reject the null hypothesis H_0: $\boldsymbol{\beta = 0}$.**

Confidence interval for the regression coefficient (slope β)

For the 95% level of confidence, the limits of β are as follows (see Equations 11.43 to 11.45):

$b \pm t_{\alpha,\ n-2}.S_b = 0.536 \pm 2.101 \times 0.078 = 0.536 \pm 0.164$
$0.536 - 0.164 = 0.372$
$0.536 + 0.164 = 0.700$
$\mathbf{0.372 \leq b \leq 0.700}$

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Thus, we can state, with 95% confidence, that 0.372 and 0.700 form an interval that includes the population regression coefficient β. The true slope is estimated, with 95% confidence, to be between 0.372 and 0.700. Since these values are above zero, it can be concluded that there is a significant linear relationship between x and y.

(c) Confidence interval for a mean value of $\hat{y}_i$

The standard error of the prediction of the ***mean value of*** $\hat{Y}_i$ for a given X_i value is obtained from Equation 11.49:

$$S_{\hat{Y}i} = S_{YX}\sqrt{\left[\frac{1}{n} + \frac{(X_i - \bar{X})^2}{\text{SQX}}\right]} = 0.512\sqrt{\left[\frac{1}{20} + \frac{(X_i - 5.635)^2}{42.726}\right]}$$

If we want to obtain the lines with the lower and upper confidence limits to be included in the scatter plot with the regression line, we need to do this calculation for all values of constituent X (all X_i values).

To illustrate this procedure, let us show the calculations for the first value in the sample of constituent X ($X_i = \mathbf{4.7\ mg/L}$).

From Equation 11.49, we obtain:

$$S_{\hat{Y}i} = S_{YX}\sqrt{\left[\frac{1}{20} + \frac{(4.7 - 5.635)^2}{42.726}\right]} = 0.512 \times 0.266 = \mathbf{0.136}$$

From the regression equation we have:

$\hat{y} = 4.083 + 0.536x$

$\hat{y} = 4.083 + 0.536 \times (4.7) = \mathbf{6.60\ mg/L}$

The **95% confidence interval** for the **mean value of the prediction of *y*** is as follows (see Equations 11.50–11.52):

$$\hat{y}_i \pm t_{\alpha,n-2}.S_{\hat{y}_i} = 6.60 \pm 2.101 \times 0.136 = 6.60 \pm 0.286$$

Therefore, we estimate (with 95% confidence) that the **mean** value of the predicted $\hat{y}$ for constituent Y, for a given value of constituent $X = 4.7$ mg/L, will fall between the lower confidence limit of LCL = **6.32 mg/L** and upper confidence limit of UCL = **6.89 mg/L**.

After that, you do a similar calculation for all other X_i values and plot these confidence limits in the scatter plot with the regression line.

(d) Prediction interval for a $\hat{y}$ value for a single observation

If we wish to predict the $\hat{y}$ value of a single observation taken from the population at a specified x value, Equation 11.53 may be used for the calculation of the *standard error of the prediction of a single value of* $\hat{Y}_i$:

$$(S_{\hat{Y}i})_1 = S_{YX}\sqrt{\left[1 + \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\text{SQX}}\right]}$$

$$= 0.512\sqrt{\left[1 + \frac{1}{20} + \frac{(X_i - 5.635)^2}{42.726}\right]}$$

If we want to obtain the lines for the lower and upper prediction limits to be included in the scatter plot with the regression line, we need to do this calculation for all values of constituent X (all X_i values).

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

To illustrate this procedure, we will perform the calculations for the first value in the sample of constituent X ($X_i = \mathbf{4.7\ mg/L}$). From Equation 11.53, we obtain:

$$(S_{\hat{Y}i})_1 = 0.512\sqrt{\left[1 + \frac{1}{20} + \frac{(4.7 - 5.635)^2}{42.726}\right]} = 0.512 \times 1.035 = \mathbf{0.530}$$

The **95% prediction interval** for a **single value of the prediction of *y*** is (see Equations 11.54–11.56) as follows:

$$\hat{y}_i \pm t_{\alpha,n-2}.(S_{\hat{Y}i})_1 = 6.60 \pm 2.101 \times 0.530 = 6.60 \pm 1.114$$

Therefore, we estimate (with 95% confidence) that a **single** value of constituent Y for a given value of constituent $X = 4.7$ mg/L will fall between the lower prediction limit of **LPL = 5.49 mg/L** and upper prediction limit of **UPL = 7.72 mg/L**.

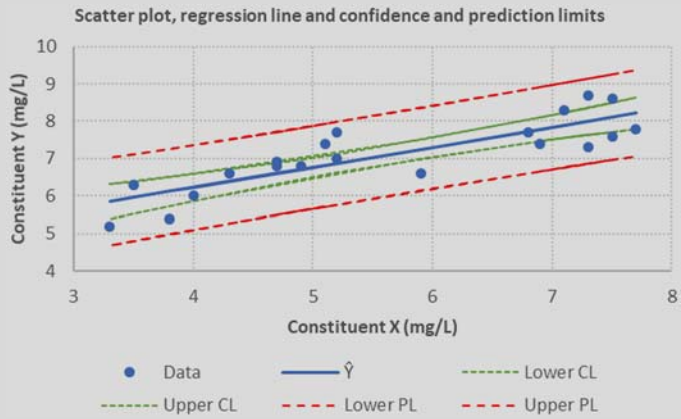
After that, you do a similar calculation for all other X_i values and plot these prediction limits in the scatter plot with the regression line.

The following table presents the confidence and prediction intervals for all values of X_i and Y_i at the 95% level.

Constituent *X* Concentration (mg/L)	Constituent *Y* Concentration (mg/L)	Predicted Value of $\hat{y}$ (from Regression Equation)	95% Confidence Limits for Mean $\hat{y}_i$		95% Prediction Limits for a Single $\hat{y}_i$	
			Lower CL	Upper CL	Lower PL	Upper PL
4.7	6.9	6.60	6.32	6.89	5.49	7.72
5.2	7.7	6.87	6.62	7.12	5.77	7.98
5.1	7.4	6.82	6.56	7.07	5.71	7.92
4.7	6.8	6.60	6.32	6.89	5.49	7.72
3.5	6.3	5.96	5.53	6.39	4.80	7.12
3.3	5.2	5.85	5.40	6.31	4.69	7.02
3.8	5.4	6.12	5.73	6.51	4.98	7.26
4	6	6.23	5.87	6.59	5.09	7.36
5.9	6.6	7.25	7.00	7.49	6.14	8.35
7.3	7.3	8.00	7.63	8.36	6.86	9.13
6.9	7.4	7.78	7.47	8.10	6.66	8.91
7.5	7.6	8.11	7.72	8.50	6.96	9.25
7.7	7.8	8.21	7.80	8.63	7.06	9.37
7.1	8.3	7.89	7.55	8.23	6.76	9.02
7.5	8.6	8.11	7.72	8.50	6.96	9.25
7.3	8.7	8.00	7.63	8.36	6.86	9.13
6.8	7.7	7.73	7.42	8.04	6.61	8.85
5.2	7	6.87	6.62	7.12	5.77	7.98
4.9	6.8	6.71	6.44	6.98	5.60	7.82
4.3	6.6	6.39	6.06	6.71	5.27	7.51

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

The following figure shows the scatter plot with the data points, the adjusted regression line, and the 95% confidence and prediction limits for *y* values.



Note: We should be very careful in using this model to make predictions for *X*-values less than 3.3 mg/L (minimum value in the sample) or more than 7.7 mg/L (maximum value in the sample). It is **always risky to do extrapolations**, that is, to use the model to make predictions outside the range of the sample data used to fit the model. You should always take into account the knowledge you have from your system and whether it is acceptable to assume that the same linear relationship between the two variables is expected to occur outside the boundaries of the experimental data.

(e) Assessing the strength of correlation between variables and the goodness of fit of the model

The **Coefficient of Determination** (r^2) is given by Equation 11.47. It indicates the proportion of the variability of the dependent variable *Y*, which is explained by the explanatory variable *X*. The closer to 1, the better the fit of the model. See Sections 11.5.2(f) and 15.2.3(b) for a detailed discussion on the interpretation of this important goodness-of-fit indicator.

S. 11.5.2

S. 15.2.3

$$r^2 = \frac{\text{Regression SS}}{\text{Total SS}} = \frac{12.292}{17.010} = \mathbf{0.723}$$

The correlation coefficient *r* is given by Equation 11.46. See Section 11.5.2(e) for a discussion about its interpretation.

S. 11.5.2

$$r = \sqrt{0.723} = \mathbf{0.850}$$

These coefficients can be calculated directly using Excel functions (**RSQ** and **CORREL**). Please refer to Sections 11.5.2(e) and 11.5.2(f) for further information.

S. 11.5.2

PART 3 RESIDUALS ANALYSIS

Here, we present some of the results of the residuals analysis, without going into a comprehensive interpretation, since the detailed background for this is provided in Sections 11.5.4 and 15.3. The calculations here make use of the Excel spreadsheets Residuals Analysis (Chapter 15) and Autocorrelation (this chapter). Simply input the values of Y_i and $\hat{Y}_i$ into the spreadsheet Residuals Analysis and the values of the residuals in the spreadsheet Autocorrelation. The statistics and graphs shown below have been obtained from these spreadsheets (we will not show the calculations here).

S. 11.5.4

S. 15.3

Excel

C. 15

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

(a) Testing for mean of residuals = 0

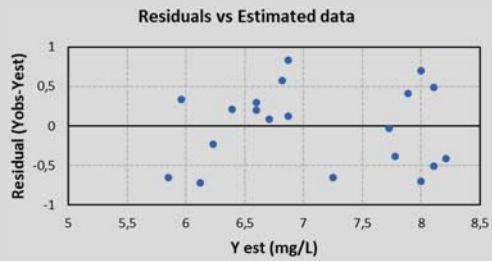
The mean of the 20 residuals values is 8.88×10^{-17} (in practical terms, it is 0.000).

Null hypothesis H_0: mean of residuals = 0
Alternative hypothesis H_a: mean of residuals ≠ 0

Student *t* test: *p*-value = 1.000 > α = 0.05. Conclusion: we cannot reject the null hypothesis. Therefore, we cannot say that the mean of residuals is different than zero. This is the conclusion we would expect.

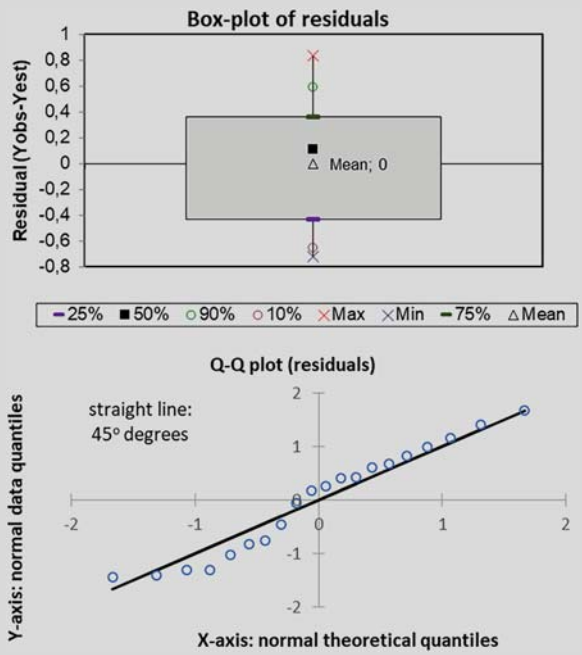
(b) Testing for linearity

See below a plot of residuals versus predicted values of *Y*. The residuals seem to be well distributed on both sides of the zero line, and there are no indications that the variance is not constant.



(c) Testing for normality

The box-plot and the Q–Q plot are shown below. Although the distribution does not seem to be perfectly normal (some asymmetry in the box plot and data points not entirely on top of the straight line in the Q–Q plot), there are no indications of strong departures from normality.



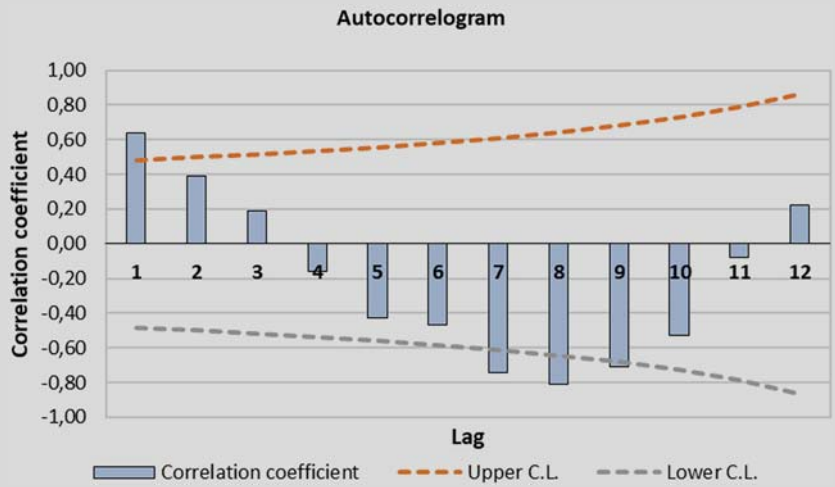
Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

In order to assess this more formally, we carried out the **Shapiro–Wilk** test using a statistical software (calculations not shown, neither here nor in the Excel spreadsheet) and obtained the *p*-value of **0.2093**. Since this *p*-value is greater than the significance level of $\alpha = 0.05$, we can conclude that the distribution of the residuals is *not significantly different from a normal distribution.*

(d) Testing for independence

The autocorrelogram is plotted below. We see that there is a significant autocorrelation at some lags (lag 1, and then lags 7 to 9), suggesting the existence of some dependence in the data. This is endorsed by the Durbin–Watson test, which gave **DW = 0.710**. This value indicates significant autocorrelation at lag 1 (see Section 15.3.5).

S. 15.3.5



(e) Conclusion from the Residuals Analysis

From our residuals analysis, we can see that most of the assumptions required for the residuals have been satisfied. The only concern lies in terms of the independence of the data since we identified the existence of some significant autocorrelations in the residuals. Take a look at Section 11.5.4(b) to analyse the possible implications of this.

S. 11.5.4

11.5.8 Conceptual problems of a linear regression equation traditionally used in wastewater treatment design and evaluation

Advanced

This item is based on a problem identified by Dr. Jeremy Lumbers, further developed by von Sperling (1999) and endorsed by Kadlec and Wallace (2009). It could have been covered in Chapter 13, which deals with the loading rates, but it is presented here, because it is mainly concerned with regression analysis. It is a specific topic, but is within the context of our book, since it deals with a typical equation used for design and evaluation of treatment plants.

We analyse here the applicability of classical equations used for the design and evaluation of some wastewater treatment systems, based on the structure of a simple linear regression such as '*removed BOD load* $= a + b$. *applied BOD load*', or:

$$L_r = a + b.L_a \tag{11.57}$$

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

where

L_r = removed BOD load [(kgBOD/d)/ha, (gBOD/d)/m^2 or other similar units]
L_a = applied BOD load [(kgBOD/d)/ha, (gBOD/d)/m^2 or other similar units]
a, b = regression coefficients.

The removed load L_r is simply the applied load L_a minus the effluent load L_e.

In spite of the broad utilization of the above equation, its structure contains some problems of statistical determination, which should be taken into account by its potential users. The equation is biased, because the right-hand side variable (applied load L_a), called **independent variable, is not actually independent**. Even though not directly perceptible, *L_a is also present on the left-hand side, since the removed load L_r contains in itself L_a. This comes from the fact that* $L_r = L_a - L_e$, as stated above. The non-recognition of this interdependence is responsible for the problem, which is not translated by the correlation coefficient, which is usually high in this equation. *This limitation is also encountered in the design equations, which use the removal efficiency as one of its variables*.

In order to allow a simple demonstration, a very small number of data are included in this example, representing only five hypothetical treatment plants (biological reactors). Table 11.8 shows values of applied load (L_a), removed load (L_r), and effluent load (L_e), which could be used to develop an empirical equation based on the classical structure. It is assumed that L_a and L_e have been determined experimentally for each reactor (average values of the historical data series), allowing the calculation of L_r ($L_r = L_a - L_e$).

Analysing the relationship between L_a and L_e, it can be seen that extreme values of the applied load lead to the same values of the effluent load. For instance, reactors 1 and 5 have completely different values of L_a, but the same value of L_e. A possible reason for this could be, for instance, the fact that reactor 1 is situated in a cold location, whereas reactor 5 is situated in a warm place.

A linear regression analysis of L_e on L_a leads to the following results (see also Figure 11.13):

$$L_e = 70 + 0.0 \times L_a$$

Coefficient of Determination $R^2 = 0.000$

Besides the frustrating value of R^2 (0.000), the equation above simply states that the effluent load is equal to 70 $kgBOD_5$/ha d, irrespective of the applied load. The value of 70 $kgBOD_5$/ha d is nothing more than the average of L_e.

However, a different and misleading picture can be obtained by the utilization of the classical structure (Equation 11.57) for the regression analysis. The results obtained are as follows (see also Figure 11.14):

$$L_r = -70 + 1.0 \times L_a$$

Coefficient of Determination $R^2 = 0.982$

Table 11.8 Experimental values of L_a and L_e and calculated values of L_r for five treatment reactors

Plant (reactor)	L_a ($kgBOD_5$/ha · d)	L_e ($kgBOD_5$/ha · d)	L_r ($kgBOD_5$/ha · d)
1	100	50	50
2	200	75	125
3	300	100	200
4	400	75	325
5	500	50	450

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

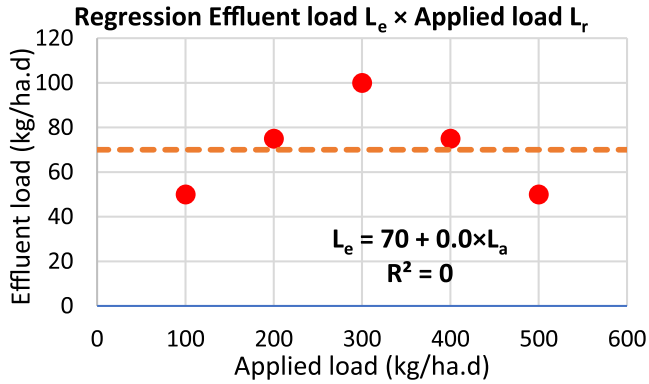


Figure 11.13 Scatter plot and best-fit line of the regression $L_e \times L_a$.

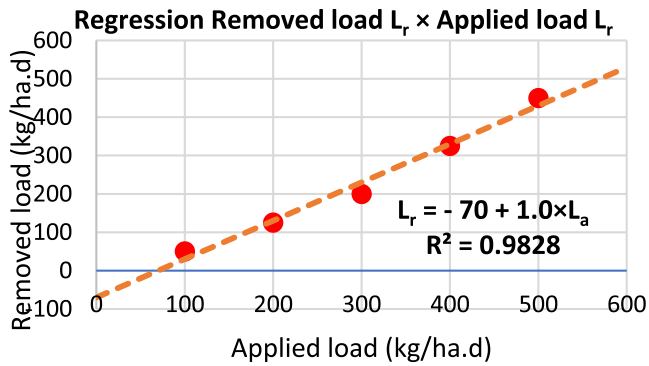


Figure 11.14 Scatter plot and best-fit line of the regression with the traditional structure $L_r \times L_a$.

The Coefficient of Determination R^2 is very high, indicating an excellent fit of the model to the experimental data. Figure 11.14 also appears to support the same conclusions. Similarly, good correlations have been obtained by the various authors, which derived empirical equations based on this classical structure and reported them in the literature.

However, if one uses the equation $L_r = -70 + 1.0 \times L_a$ to predict the removed load and, consequently, the effluent load L_e (and, as a result, the effluent BOD concentration, which is the main objective of the equation), a value of L_e always equal to 70 ($kgBOD_5/d$)/ha is obtained, irrespective of the value of L_a. The usefulness of this linear regression in this case is, therefore, highly questionable.

We have illustrated here a typical and widely used application of regression analysis, emphasizing that you should always search for a thorough interpretation of the results and of the inherent limitations of the model you select. Always check whether you are complying with the underlying assumptions for applying your model – in this case, the assumptions related to linear regression analysis.

11.6 MULTIPLE LINEAR REGRESSION

11.6.1 Basics of multiple linear regression

Advanced

S. 11.5

Simple linear regression, which we covered in Section 11.5, is a particular case of multiple linear regression. In multiple linear regression, we have more than one independent variable, which can often

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

help us to obtain a model with a greater explanatory capacity. Models used often in multiple regression are of the type:

$$\boxed{y = a + b_1x_1 + b_2x_2 + \cdots + b_nx_n + \text{error}} \tag{11.58}$$

where

y = dependent variable
$x_1, x_2, \ldots, x_n$ = independent variables
a = intercept
$b_1, b_2, \ldots, b_n$ = slope for each independent variable
error = difference between the observed and the predicted y.

Most of the concepts described for simple linear regression are also applicable to multiple linear regression, and they will not be repeated here.

The determination of the model parameters will not be shown here. Excel provides the add-in '**Analysis ToolPak's Regression tool**'. This add-in was illustrated in Example 11.7, but it is particularly useful for multiple regression analysis. On the basis of interpretation we provided in Example 11.7, you should be able to understand the Summary Output table, which shows the statistics for *all independent variables*. You can also use statistical software to perform the calculations and obtain the model outputs.

S. 11.5

Model fitting is evaluated by the **Coefficient of Determination R^2**, as demonstrated in Section 11.5. The interpretation of R^2 is also linked to the number of independent variables included in the model (k) and the number of data (n), that is, the degrees of freedom of the model. The R^2 value increases as more variables are introduced into the model and can reach values very close to 1, without the model contributing any more to the prediction of Y. If our model has the same number of independent variables as the number of data points used to fit the model, then R^2 will be equal to 1. Because of this, we can calculate a corrected R^2 value, known as the '**adjusted R^2**' (see Equation 11.59). To assist us with this analysis, the F test of ANOVA should also be performed. This is particularly important in research studies that work with small samples.

$$\boxed{R^2 \text{ adjusted} = 1 - (1 - R^2)\frac{n-1}{n-k-1}} \tag{11.59}$$

where

R^2 = Coefficient of determination determined in the usual way
n = number of data
k = number of independent variables.

The overall utility of the model can be evaluated by the F test of ANOVA (included in the Excel add-in and in all basic statistical software programs). With this test, you can evaluate whether each of the coefficients of the model (a, b_1, b_2, …, b_n) are significantly different from 0. A coefficient equal to 0 implies that the variable associated with it does not contribute significantly to the model. You can also perform the F test for each model parameter (also available in most statistical software programs), allowing the exclusion of those variables that do not contribute to the model. According to the principle of parsimony, the simplest possible models should be adopted, so long as they have the desired accuracy for estimation.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

11.6.2 Potential problems or difficulties with multiple linear regression

In multiple regression analysis, you may encounter the following **problems** or **difficulties**:

- Difficulty to estimate the parameters (when there are few X values)
- Difficulty to interpret all the parameters (impossible to infer cause and effect relationships)
- Non-linear relationship between the variables (especially relevant for environmental variables, which frequently exhibit non-linear behaviour)
- Multicollinearity (correlation between independent variables)
- Prediction outside the experimental region (extrapolation)
- Correlated errors or residuals (violation of the assumption of independent errors)

The **multicollinearity** problem is quite frequent in multiple regression, since it is often difficult for our designated independent variables to really be independent of each other, that is, not correlated. Multicollinearity can cause rounding errors in the estimates of parameters and statistics, as well as confusing results, with possible inversions of the signs of the coefficients (e.g., positive versus negative coefficients). The following findings might indicate multicollinearity: (a) significant correlations between pairs of independent variables in the model; (b) non-significant results to tests for the contribution of each parameter (t test), even when the global F test yields a significant result; and (c) coefficients with opposite signs from what you would have expected (for instance, you would expect a positive regression coefficient for one independent variable, but because it is correlated with another independent variable, it may become negative).

11.6.3 Graphical outputs for multiple linear regression

In terms of **graphical outputs** from a multiple regression analysis, since there are many variables involved, there is no single graph that can allow for the visualization of all of them together, including the predicted Y values. Therefore, the following are typical graphs we use to plot the results of a multiple regression analysis:

- if data are arranged as time series or data sequence, you can plot a **time series graph**, with the *X-axis for time* (or data sequence) and the *Y-axis for both observed Y (Y_i) and predicted Y ($\hat{Y}$)*
- **scatter plot** of *observed Y (Y_i) versus predicted Y ($\hat{Y}$)*
- **residuals analysis** (usual graphs, including all the statistics for residual analysis – see Sections 11.5.4 and 15.3).

S. 11.5.4
S. 15.3

11.6.4 Data transformations to linearize a model for using in a multiple regression model

For some model structures, you can apply **transformations** to the original equation, so that it is **linearized**, and then use the computation for a multiple linear regression. For instance, if you want to test the following **multiple power model**:

$$y = a.x_1^{b_1} x_2^{b_2} \ldots x_n^{b_n} \tag{11.60}$$

You can linearize Equation 11.60 by applying logarithm to the terms of the original equation:

$$\ln(y) = \ln(a) + b_1 \ln(x_1) + b_2 \ln(x_2) + \cdots + b_n \ln(x_n) \tag{11.61}$$

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

You then perform a multiple linear regression having as dependent variable $\ln(y)$ and independent variables $\ln(x_1)$, $\ln(x_2)$, … $\ln(x_3)$. After you obtain the intercept and the regression coefficients (slopes), you calculate 'a' as $e^{intercept}$. The slopes b_1, b_2, …, b_n will be the same as those calculated in the multiple linear regression.

You can also use the multiple linear regression for a **polynomial model**, which has the following format:

$$y = a + x^1 + x^2 + x^3 + \cdots + x^n \tag{11.62}$$

where

n = order of the polynomial equation.

Simply create columns for each independent variable x, taking into account that the first variable will be x^1 (x raised to the power 1), the second variable will be x^2 (x raised to the power 2), the third variable will be x^3 (x raised to the power 3), and so on. Perform the multiple linear regression as usual and obtain the model coefficients (intercept and slopes) directly.

11.7 NON-LINEAR REGRESSION

Advanced

Non-linear regression is also very useful for the treatment plant and the water quality modelling, given the frequent occurrence of non-linear biological and biochemical phenomena in environmental systems. Non-linear regression allows the fitting of models with different forms from those seen previously, introducing greater flexibility in the regression analysis.

The **estimation of regression coefficients** can be made using the following approaches:

- Using Excel's '**Add trendline**' feature, available in the scatter plots.
- **Linearization** of the equation (when possible, see Section 11.6.3).
- **Numerical methods** for minimizing the error function (using Excel Solver or a specific algorithm).

(a) Using Excel 'Add trendline'

S. 11.5

In Excel's scatterplots, you have the option of adding a trendline. Besides the linear model, which we already saw in Section 11.5, we have the possibility of adding the following alternative models to our data:

- exponential
- logarithmic
- polynomial
- power

This is a very convenient and easy-to-use feature. In the scatter plot, you obtain the fitted line, and you have the possibility of including and viewing the equation, as well as the Coefficient of Determination R^2. Furthermore, you can forecast forward and backward as desired. Nevertheless, this feature should be used judiciously.

A special **word of caution must be made for polynomial models**. They are very powerful in the sense that they can easily provide an apparently good fit to your experimental data, especially if you select a high-order polynomial. However, you need to pay attention to the following aspects. Take the case exemplified in Figure 11.15, suppose the data represent concentrations of some constituent in a water body. The fourth-order polynomial was able to pass exactly through the experimental data (and the R^2 value was equal to 1.0). However, its **interpolation** led to negative results, which have no physical meaning, since they are concentrations. Now let us take

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

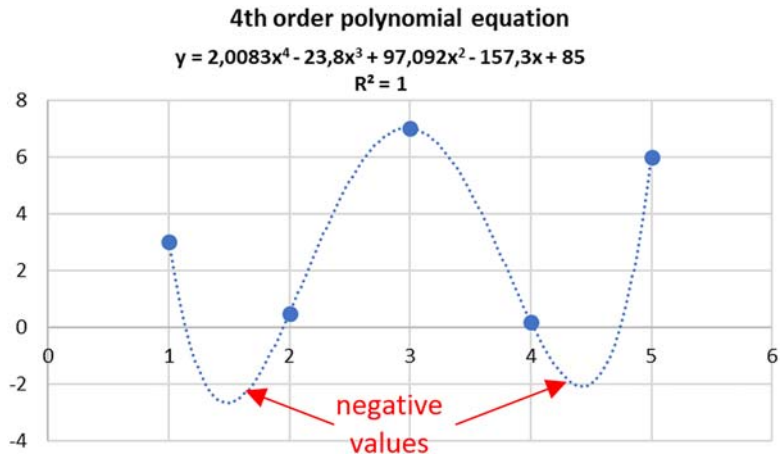


Figure 11.15 High-order polynomial, with perfect fit to the data, but producing results without physical meaning (negative values), even for interpolation.

the case of **extrapolation**. In Figure 11.16, a second-order polynomial gave very good fit ($R^2 = 0.9896$) to the experimental data, which was increasing along the data sequence, but seemed to reach a maximum (saturation) point. However, when we use the equation to extrapolate forward, we might be surprised by the outcome, which indicates an unexpected decrease after reaching the maximum. This is normal for a second-order polynomial model, but it might not be the best model to describe the phenomenon we are seeing in our data. Therefore, we recommend that you *use polynomial equations* ***only in very specific scenarios****, for which you have full control*. After all, *you are not only searching for a good fitting curve, you should be dedicating your efforts to obtain a model that helps elucidate the possible behaviour of the system you are studying*.

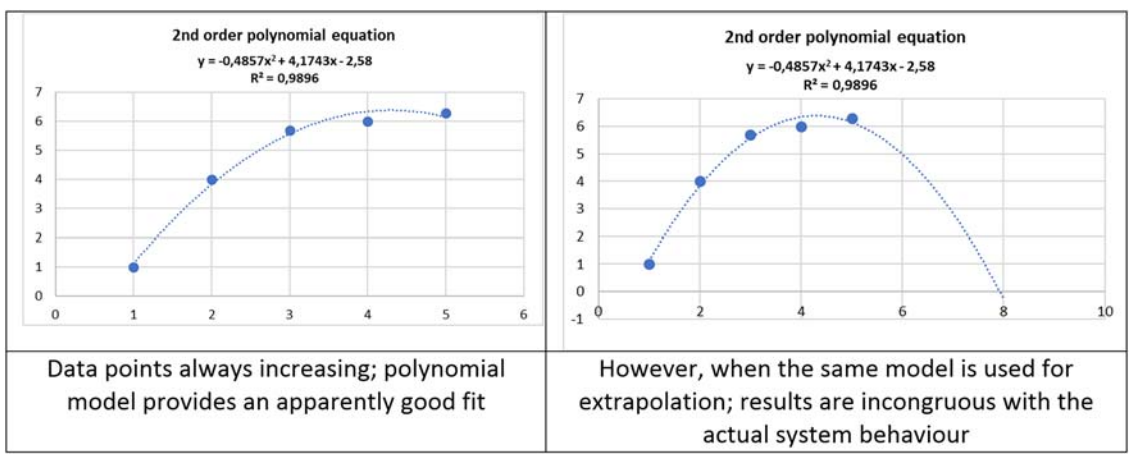


Figure 11.16 Polynomial model, with an apparently good fit (left figure), but with very poor extrapolation capacity (right figure).

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

(b) Linearization of equations and use of linear regression for the transformed data

Depending on the model structure, we can apply transformations to linearize it, so that we can apply a linear regression model. We gave an example in Equation 11.60, which was linearized in Equation 11.61.

For the regression models presented in Excel, we can apply the transformations shown in Table 11.9.

For instance, if you want to fit an exponential model to your data, you can take the natural logarithms values of your data x and y ($\ln x$ and $\ln y$) and perform a linear regression with the transformed data. You then obtain the values for the intercept and the slope of the straight line in the usual way. Then, in order to obtain the values of the coefficients a and b, you need to transform them back to the original base. From Table 11.9, you see that this transformation is as follows: $a = e^{\text{intercept}}$ and $b = \text{slope}$.

S. 14.3

In Section 14.3, we also perform a linearization in order to be able to calculate the coefficients of a kinetic model. Have a look at that section to see a practical application of the concept, including Example 14.2. As we show in Example 14.2, we should interpret the **values of R^2** taking into account that they are based on the *transformed data*, in order to obtain a linearized plot. The sums of the squares are calculated from the transformed data and not the original ones. Therefore, by transforming the data, we also modify the capability of the R^2 coefficient of being a true indicator of the goodness of fit of our original (untransformed) data.

(c) Numerical methods for minimizing the error function

For a regression-based model with any structure, you can obtain the regression coefficients using an iterative numerical procedure that **minimizes the error function** (sum of the squared errors) or **maximizes the Coefficient of Determination R^2**. There are several numerical algorithms, and you should adopt the one indicated by your statistical software. We recommend that you always try to understand how the algorithm works.

S. 14.3

In Excel, you can use the **Solver tool**, which we have already used in several parts of our book. In Section 14.3, we exemplify the utilization of Solver in the case of the determination of coefficients for a kinetic model (see Example 14.3). In Section 15.2.2, we discuss model calibration (regression-based and non-regression-based models) by performing the minimization of the residuals. Visit this section to learn about the applicability and constraints of this method. One interesting fact to consider is that these methods do not necessarily guarantee that we have found

S. 15.2.2

Table 11.9 Transformations in some models to obtain a linear structure.

Model	Equation	Transformation	Variable X in regression	Variable Y in regression	a	b
Exponential	$\hat{y} = a.e^{bx}$	$\ln \hat{y} = \ln a + bx$	x	$\ln y$	$e^{\text{intercept}}$	slope
Logarithm	$\hat{y} = a + b.\ln x$	$\hat{y} = a + b \ln x$	$\ln x$	y	intercept	slope
Power	$\hat{y} = a.x^b$	$\ln \hat{y} = \ln a + b \ln x$	$\ln x$	$\ln y$	$e^{\text{intercept}}$	slope

Source: Adapted from Lapponi (2005).

Notes:

- Intercept: intercept of the regression equation, after transformation.
- Slope: slope of the regression equation, after transformation.
- a, b = coefficients of the original equation.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

the global minimum, that is, the values of the coefficients that give the smallest global error. Many times, the algorithm will stop after finding a local minimum, not knowing that an even smaller global minimum exists. Therefore, we need to run the algorithm several times, using different starting values to see if it produces the same results.

11.8 CHECK-LIST FOR YOUR REPORT

- ✓ Start by plotting the independent or explanatory variable(s) x and the dependent or response variable y of your data set to visualise the possible correlation or relationship between them.
- ✓ If your data set has more than one independent variable (e.g., x_1, x_2, …, x_n), then make a scatter plot of each combination of independent variables, so that you can start to understand if there is evidence of correlation between them.
- ✓ Assess the data set for outliers and state clearly in your report which method(s) and justification(s) were used to assess and remove outlier(s) from the data set.
- ✓ Calculate a correlation coefficient, using a parametric method (such as the Pearson correlation coefficient) if the relationship appears to be linear and using a non-parametric method (such as the Spearman rank correlation coefficient) if the relationship appears to be non-linear.
- ✓ Use a hypothesis test (e.g., where the null hypothesis is that the correlation coefficient equals zero) to determine if the correlation you found is significant. If desired, use hypothesis testing to assess whether the correlation coefficient is significantly different from some threshold (e.g. 0.4, 0.7, etc.). Report the p-value and ideally, the confidence interval of the sample correlation coefficient. Use appropriate methods to determine the confidence limits, depending on the sample size (e.g., $n > 50$ versus n between 10 and 50).
- ✓ If you have multiple independent variables (e.g., x_1, x_2, …, x_n), then construct a correlation matrix to determine which combinations have significant correlation coefficients.
- ✓ Report whether you assessed cross-correlation (e.g., for time-series data with a lag) and whether you assessed autocorrelation as one way to test for the independence assumption. Plot your data as a time series if applicable, to help visualise temporal trends in the data. Report any lag periods that produce significant correlation coefficients, using a cross-correlogram and an auto-correlogram. Some of this information might go into supporting information document or the appendix to your report.
- ✓ Report the method used to conduct any linear regression analysis: Did you perform the regression by adding a trendline to your scatter plot in Excel? Did you use Excel functions to calculate regression coefficients? Did you use the Excel add-in Analysis ToolPak regression tool? Did you manually perform the calculations associated with the regression analysis formulas?
- ✓ Calculate and report the Coefficient of Determination (i.e., the R^2 value in the case of linear regression), which is an indication of the goodness of fit for the model.
- ✓ Perform a residuals analysis, perform an autocorrelation analysis, and construct plots of the residuals versus predicted values, etc., to ensure that the model is satisfying the assumptions of linearity, independence, normality of residuals, and homogeneity of variances. Most of these plots will go into the appendix of your report, or into a supporting information document, but in the body of your report or paper, you should mention that you checked the assumptions and state whether the assumptions were satisfied.
- ✓ Test the significance of your regression and its coefficients using a hypothesis test, where the null hypothesis is that there is no linear relation between X and Y variables. Report the p-value for this significance test and interpret it appropriately.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

✓ Calculate the values for the confidence interval and prediction interval for your regression curve and plot them (as appropriate) along with the line of best fit, also showing the data points used to fit the line. Report the confidence or prediction interval limits when interpolating with the model. Do not extrapolate values outside of the range of the data used to fit the model, unless absolutely necessary. If you do extrapolate, be very clear about this in your report and warn readers of the limitations of extrapolating. Make sure you have a very good understanding about the behaviour of your system and that your model is correctly representing this behaviour, even under the extrapolated conditions.

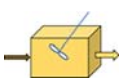
Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Chapter 12

Water and mass balances

This chapter highlights the importance of the concept of water and mass balances, which are key elements for understanding the behaviour of a treatment plant. The concepts of steady state and dynamic state, which influence the structuring of mass balances and the interpretation of data, will be presented. From this chapter, you are expected to be able to construct simple water and mass balances around a treatment unit.



The contents in this chapter, in the way they have been structured, are mainly applicable to treatment plant monitoring. However, the overall concepts of steady and dynamic states, water balance, and mass balance are also applicable to water bodies.

CHAPTER CONTENTS

12.1 Steady State and Dynamic State **480**
12.2 Water Balance **481**
12.3 Mass Balance **487**
12.4 Check-List for Your Report **497**

© 2020 The Authors. This is an Open Access book chapter distributed under the terms of the Creative Commons Attribution Licence (CC BY-NC-ND 4.0), which permits copying and redistribution for non-commercial purposes with no derivatives, provided the original work is properly cited (https://creativecommons.org/licenses/by-nc-nd/4.0/). This does not affect the rights licensed or assigned from any third party in this book. The chapter is from the book Assessment of Treatment Plant Performance and Water Quality Data: A Guide for Students, Researchers and Practitioners, Marcos von Sperling, Matthew E. Verbyla and Sílvia M. A. C. Oliveira (Authors).

doi: 10.2166/9781780409320_0479

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

12.1 STEADY STATE AND DYNAMIC STATE

Basic

You have already learned about the analyses and elements that comprise performance evaluation of a treatment plant. Now you will advance your understanding of some influencing factors that impact the performance. This chapter will start to help you learn preparing for some of the concepts that will be addressed in subsequent chapters.

If you want to mathematically represent your system using a model with any degree of complexity (both simple and complex models) or understand the influence of loading rates and environmental conditions on the performance of your treatment plant, you should be aware that the analysis can be done assuming one of the following two distinct conditions:

- *Steady-state conditions*
- *Dynamic conditions*

Figure 12.1 illustrates the concept of steady and dynamic states as a representation of the variation in the concentration of a constituent in the influent or effluent of a treatment unit with respect to time.

Steady state is when there is no accumulation of the constituent in the system (or in the volume being analysed). As a result, *the concentration of the constituent inside the tank or in the effluent is constant*. During steady-state conditions, the input and output flows and concentrations are constant and so are the environmental conditions to which your treatment plant is subjected. There is a perfect equilibrium between the positive and the negative terms of the water and mass balances around your tank (as will be further discussed in this chapter), which, when summed, lead to no accumulation of constituents in the system. Of course, this represents a strong simplification of reality, especially for real, full-scale plants, but mathematically, this assumption is, in many cases, worthwhile. For instance, it is common to assume simplified steady-state conditions for the **design** of wastewater treatment plants. It may also be advantageous to use this simplified assumption when analysing the performance of the treatment plant you are studying. Additionally, if you are performing controlled experiments in the laboratory, your reactors may actually be very close to steady-state conditions.

The **dynamic state** is the one in which there are accumulations or losses in the mass of the constituent in the system. *The concentration of the constituent inside the reactor or in the effluent is therefore variable with respect to time* and can increase or decrease, depending on the balance between the positive and negative terms. In a treatment plant, the input flow and/or the input concentration are usually variable, and other external stimuli to the system (e.g., temperature changes) may also cause changing concentrations of the constituent with respect to time. For this reason, dynamic conditions are the ones that are usually

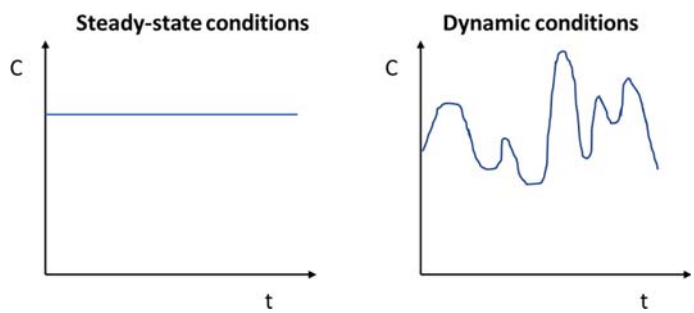


Figure 12.1 Profile of the concentration of a constituent in the influent, effluent, or inside a tank or reactor with respect to time, for steady-state and dynamic-state conditions.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

prevailing in actual treatment plants. Dynamic models are usually more adequate for the **operational control** of a treatment plant, due to the frequent variation of internal and external conditions of the system. Dynamic models can also be used for design, principally for evaluating the impact of variable influent loads on the performance of the plant. In general, it can be said that dynamic models have been used less frequently due to the greater complexity associated with solving their equations. Compared to simple steady-state models, dynamic models also have greater data requirements for fitting model coefficients and variables. However, the increasing availability of highly powerful, commercially available software with numerical integration capabilities has contributed to the increased use of dynamic models. It should be emphasized that the steady state is a particular case of the dynamic state.

In summary, we have

- **Steady-state assumption**
  - is simpler to represent mathematically and requires fewer input data
  - may be used with confidence when loading and environmental conditions are controlled and fixed, such as in lab or bench-scale experiments
  - may be used with relative confidence in pilot-plant studies, when operating conditions are somewhat stable
  - is often used for design purposes
  - is used for the sake of simplicity in the representation of average long-term conditions in a full-scale real treatment plant
- **Dynamic-state assumption**
  - is a better representation of reality but involves more complex mathematical representations and greater data requirements
  - is used to represent variations over time in plants subjected to variations in influent flow, influent concentrations, operating conditions, and environmental factors
  - is used to represent short-term variations, in the time scale of minutes, hours, or days (depending on the system dynamics)

12.2 WATER BALANCE

Basic

In most cases, you can assume that the outflow from a treatment unit is the same as the inflow. In many situations, this can be a relatively good approximation to reality, but in other situations, a refinement must be done. A water balance, or hydrologic budget, can be made around a treatment unit, taking into account the following terms (see Equation 12.1 and Figure 12.2):

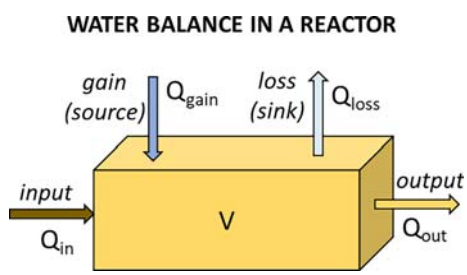


Figure 12.2 Water balance (hydrologic budget) in a treatment unit or reactor.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

- *flow that enters the unit from the influent line (influent flow or input flow)*
- *flow that leaves the unit in the effluent (effluent flow or output flow)*
- *flow that enters the unit from another route (gain or source; e.g., precipitation)*
- *flow that leaves the unit by another route (loss or sink; e.g., leakage, infiltration to the soil, evaporation, evapotranspiration)*
- *liquid volume that is accumulated in the unit*

Advanced

$$\frac{dV}{dt} = Q_{in} - Q_{out} + Q_{gain} - Q_{loss} \quad (12.1)$$

where

dV/dt = change in the liquid volume inside the reactor per day (m^3/d)
Q_{in} = input flow (m^3/d)
Q_{out} = output flow (m^3/d)
Q_{gain} = gain of liquid from other sources, such as precipitation (m^3/d)
Q_{loss} = loss of liquid from sinks (leakage, evaporation, acnd evapotranspiration) (m^3/d).

You may think that the term dV/dt, because it is expressed in the form of a differential equation, is complicated, but it is a very simple concept. It represents the accumulation term (volume per unit time, e.g., m^3/d), which can be positive or negative:

- **When the accumulation term dV/dt is positive**, it means that the volume of liquid (V) inside the treatment unit has increased by a certain volume (dV) during a certain amount of time (dt). In this case, the sum of the positive terms (input flow and gain of liquid) is greater than the sum of the negative terms (output flow and losses of liquid). For instance, if dV/dt is equal to $+10\ m^3/d$, it means that, after one day, the volume occupied by the liquid inside the tank has increased by $10\ m^3$. If the volume of liquid inside the reactor was $100\ m^3$, after this specific day it will be $100 + 10 = 110\ m^3$. Considering that the surface area of the existing tank is constant (fixed values of length and width), the increase in liquid volume means that the water level will rise in the tank.
- **When the accumulation term dV/dt is negative**, it means that the volume of liquid (V) inside the treatment unit has decreased by a certain volume (dV) during a certain amount of time (dt). In this case, the sum of the positive terms (input flow and gain of liquid) is less than the sum of the negative terms (output flow and losses of liquid).
- **When the accumulation term dV/dt is equal to zero**, it means that the volume of liquid (V) inside the treatment unit is stable. If this persists over a long time, we can imply that we have approached or reached steady state. As highlighted before, *this assumption is usually adopted in most studies of treatment plant evaluation*, unless advanced mathematical modelling of the unit is performed.

When there is no accumulation term ($dV/dt = 0$), Equation 12.1 can be simplified, leading to a simple estimation of the outflow:

Basic

$$0 = Q_{in} - Q_{out} + Q_{gain} - Q_{loss} \quad (12.2)$$

$$\boxed{Q_{out} = Q_{in} + Q_{gain} - Q_{loss}} \quad (12.3)$$

In some studies, in which the outflow is different from the inflow, and when the flow value is necessary for calculating other variables, such as hydraulic retention time, you may assume a practical position, for the sake of simplicity, and adopt the average of both flows, that is, $\boldsymbol{Q = (Q_{in} + Q_{out})/2}$.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Furthermore, if water gains and losses are negligible ($Q_{gain} = 0$ and $Q_{loss} = 0$), Equation 12.3 can be further simplified, stating that the output flow is equal to the input flow. *This assumption is usually adopted in most studies*.

$$Q_{out} = Q_{in} \tag{12.4}$$

In your reports, you should always make it clear which flow you are using. Important concepts to clarify are whether the flow is measured or calculated; inflow or outflow; original value or mean of different values; assuming steady state or dynamic state?

Advanced

Note that in Equations 12.1–12.4, **other time units can be used, such as hours, months, or years**, depending on the system. Also note that there may be **more than one input line to the tank** (e.g., influent line and return sludge line in activated sludge aeration tanks) and **more than one outflow line** (e.g., effluent line and sludge underflow line in sedimentation tanks), as illustrated in Figure 12.3, but the principles of the water balance remain the same.

For you to carry out a full water balance according to Equation 12.1 or 12.3, all flow components must be measured or estimated:2

C. 2

- **Influent** and **effluent** flows can be measured as described in Chapter 2.
- If the gain is by **precipitation** (rainfall), the volume of water that is incorporated into the tank during a certain time (e.g., one day) is given by the value of the daily precipitation [mm/d, converted into m/d, which is equivalent to $(m^3/d)/m^2$] multiplied by the area (m^2) of the open tank or reactor, resulting in m^3/d. This is a positive value, which is added to the water balance. Precipitation values are made available from weather stations – try to select the one that is closest to your treatment plant.
- Water losses by **evaporation** are estimated in a similar fashion. From a weather station, the daily evaporation [mm/d, converted into m/d, which is equivalent to $(m^3/d)/m^2$] is multiplied by the area (m^2) of the tank or reactor and results in m^3/d. This is a negative value, which is subtracted from the water balance.
- Water losses by **evapotranspiration** (evaporation + plant transpiration) are more difficult to measure or estimate, because of the transpiration component. The methods for estimating evapotranspiration are outside the scope of this book but are routinely practiced in hydrological or agricultural studies and the procedures can be applied here. The computation is similar to the one shown for evaporation (above).
- Water losses by **leakage** and the resulting infiltration of liquid to the soil are difficult to measure or estimate, but they may end up becoming important if the tank bottom is deteriorating in quality or is not properly sealed.

In the computation of a water balance, if you **estimate a component that has not been measured** by calculating it as difference from the other factors in the water balance, you may obtain some possible

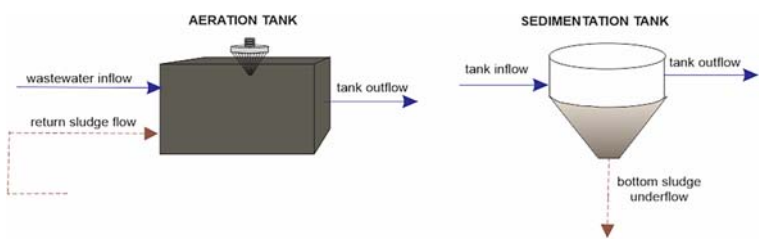


Figure 12.3 Left: example of one tank with two inlet flows and one outlet flow. Right: example of one tank with one inlet flow and two outlet flows.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

errors. For instance, suppose you want to estimate the liquid losses by leakage in a reactor. Suppose you measure the inflow and outflow, estimate the precipitation and evapotranspiration, and assume that dV/dt is equal to zero. If you then calculate the losses due to leakage by summing or subtracting all other components ($Q_{leakage} = Q_{in} - Q_{out} + Q_{gain} - Q_{evapotranspiration}$) from Equation 12.1, this approach may lead to incorrect values. This is because dV/dt may not be zero and there may be errors in the measurement or estimation of the other components (inflow, outflow, precipitation, and evapotranspiration). *In many cases, the water balance does not close entirely*, and you should make a critical analysis and the best possible interpretation based on the available data and on the knowledge of the expected system behaviour.

As mentioned before, in most studies, for the sake of simplicity, you may assume that the outflow is equal to the inflow (Equation 12.4), but this may not be the case in some specific cases. Natural treatment systems (ponds, overland flow, constructed wetlands), for instance, have a large, open surface area, which increases the relative contribution of precipitation, evaporation, and evapotranspiration (see Example 12.1). For instance, in treatment wetlands, which use large, open reactors, the hydrological budget is usually an important element in the assessment of the system behaviour. Even for compact treatment units (aeration tanks, anaerobic reactors, sedimentation tanks, and others) with small and well-sealed tanks, the input flow may be greater or lower than the output flow, unless steady-state conditions are prevailing (see Example 12.2).

Advanced

You should also pay attention to another aspect regarding the quality of effluent from a **system with substantial water losses** by evaporation and evapotranspiration, such as in treatment wetlands. If water is simply lost to the atmosphere, this lost water is totally pure (zero concentration of pollutants), and the constituents in the effluent become more concentrated simply due to water loss. In general, the effect of water losses to the atmosphere in the constituent's effluent concentrations can make the percent removal appear lower than it really is. In order to avoid this problem, you can calculate the percent removal based on the loading rates (see Chapter 7):

C. 7

$$E = \frac{C_{in} \cdot Q_{in} - C_{out} \cdot Q_{out}}{C_{in} \cdot Q_{in}} \tag{12.5}$$

For instance, suppose the effluent biochemical oxygen demand (BOD) from a treatment wetland was measured to be 50 mg/L and the influent was measured to be 200 mg/L, but it was noticed that the unit lost 30% of the water to the atmosphere (the effluent flow was 350 m^3/d, 30% lower than the measured influent flow of 500 m^3/d). In this case, if we calculated percent removal based on the concentrations, we would obtain a value of $(200 - 50)/200 = 75\%$. But if we calculate percent removal based on mass loadings, we would obtain a value of $((200 \times 500) - (50 \times 350))/(200 \times 500) = 82.5\%$. In this case, there were water losses but also no mass losses (only mass removal by treatment processes). This reasoning does not apply if the water losses are by leakage, because the liquid leaves the treatment unit together with the constituent, and so there are also mass losses from the system.

Simple calculations for water balances are shown in Examples 12.1 and 12.2.

Example

EXAMPLE 12.1. WATER BALANCE UNDER STEADY-STATE CONDITIONS

Based on several years of measurements of inflows and outflows from an extensive pond system (surface area of 2000 m^2), together with precipitation and evaporation values from a nearby weather station, the following yearly average flow rate values have been obtained:

- *inflow: 10,000 m^3/year*
- *outflow: 9000 m^3/year*

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

- *precipitation: 1000 mm/year*
- *evaporation: 1500 mm/year*

Complete a water balance around this unit and interpret the results.

Solution:

(a) Estimate the gain of water by precipitation
The precipitation of 1000 mm/year is the same as 1.0 m/year or 1.0 $(m^3/year)/m^2$.

$$\text{Gain of flow by precipitation} = \text{surface area} \times \text{precipitation} = 2000\,m^2 \times 1.0(m^3/year)/m^2$$
$$= 2000\,m^3/year$$

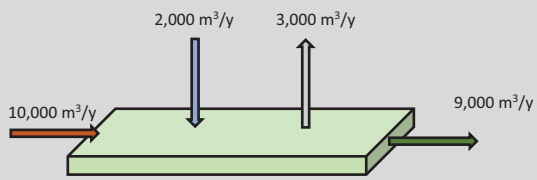
(b) Estimate the loss of water by evaporation
The evaporation of 1500 mm/year is the same as 1.5 m/year or 1.5 $(m^3/year)/m^2$.

$$\text{Loss of flow by evaporation} = \text{surface area} \times \text{evaporation} = 2000\,m^2 \times 1.5(m^3/year)/m^2$$
$$= 3000\,m^3/year$$

(c) Estimate the expected outlet flow
Using Equation 12.3, you have

$$Q_{out} = Q_{in} + Q_{gain} - Q_{loss} = 10{,}000 + 2000 - 3000 = 9000\,m^3/year$$

(d) Interpret the results
The schematics of the water balance are presented below.



The outflow value estimated by the water balance matches the average measured value, so we would say that the water balance has completely closed. However, this is not always the case, and differences may occur due to uncertainty in most measurements, possible leakages, problems in flow measurements, and utilization of measurements from weather stations which are not in the exact same location as the treatment plant. You are responsible for analysing the uncertainty behind all data and judging the adequacy of the water balance. Nevertheless, by doing these calculations and searching for explanations of possible deviations, you obtain more insight into the behaviour of this treatment system, and you acknowledge that exact calculations frequently do not match real data from real systems.

Example

EXAMPLE 12.2 WATER BALANCE UNDER DYNAMIC CONDITIONS

Complete a water balance analysis for a tank, based on measurements of the inflow and outflow over a period of 24 h. The tank does not have a large surface area, is well sealed, and any water losses or gains are negligible. The initial volume of water in the tank was 1000 m^3.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Measured data:

Hour of the Day	Inflow (m^3/h)	Outflow (m^3/h)	Hour of the Day	Inflow (m^3/h)	Outflow (m^3/h)
1	110	112	13	213	202
2	101	114	14	189	182
3	91	116	15	178	180
4	98	120	16	161	175
5	114	129	17	168	170
6	130	142	18	170	169
7	163	158	19	188	162
8	184	174	20	177	158
9	222	218	21	168	148
10	238	225	22	129	143
11	259	239	23	113	132
12	248	231	24	109	119

Excel

Note: This example is also available as an Excel spreadsheet.

Solution:

From Equation 12.1, you have

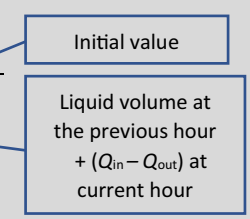
$$\frac{dV}{dt} = Q_{in} - Q_{out} + Q_{gain} - Q_{loss}$$

Since there are no gains or losses ($Q_{gain} = 0$ and $Q_{loss} = 0$), this equation is simplified to

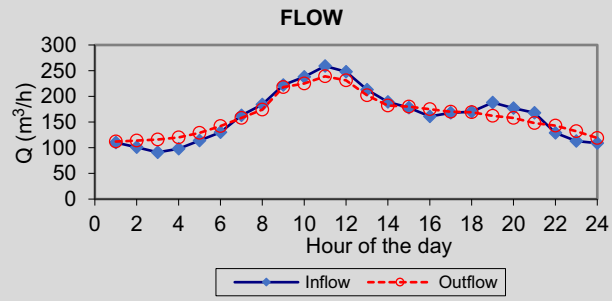
$$\frac{dV}{dt} = Q_{in} - Q_{out}$$

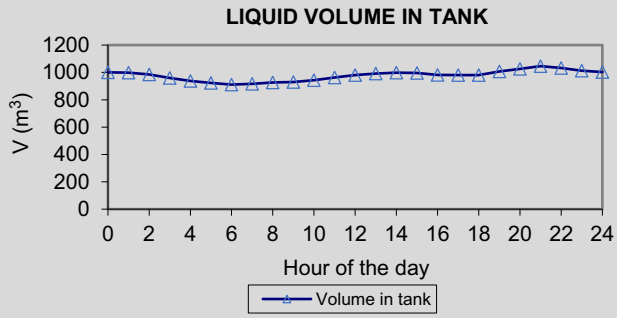
The computational table of the resulting volume of liquid in the tank is presented below, together with the associated graphs.

Hour of the Day	Inflow Q_{in} (m^3/h)	Outflow Q_{out} (m^3/h)	$Q_{in} - Q_{out}$ (m^3/h)	Volume in Tank ($V_{previous} + Q_{in} - Q_{out}$) ($m^3$)
0	-	-	-	1000
1	110	112	–2	998
2	101	114	–13	985
3	91	116	–25	960
4	98	120	–22	938
...	...	...	...	...
23	113	132	–19	1013
24	109	119	–10	1003



Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest





Summing up the 24-h values of inflow and outflow, the total daily flows to the tank were: inflow = 3921 m^3/d and outflow = 3918 m^3/d. Both values are very similar but, during the 24 h of this particular day, the inflow was slightly higher than the outflow (positive difference of only 3 m^3/d). The liquid volume in the tank varied little. It started at 1000 m^3 and ended with a volume of 1003 m^3, therefore reflecting the net increase of 3 m^3. The average liquid volume during this day was 976 m^3.

From the graphs, we can see that the outflow hydrograph was slightly smoother compared with the inflow hydrograph, indicating that a slight equalization took place in this tank.

Basic

12.3 MASS BALANCE

After understanding the concept of water balance, we can now move onto the generic concept of **mass balances**. For this, **mass loads (load = flow × concentration)** are used.

Structuring a mass balance around a treatment unit is a very useful way to understand its behaviour. The concentration of a certain constituent in a reactor (or in any place or control volume inside it) is a function of not only the physical, chemical, and biochemical reactions that take place inside it but also of the transport mechanisms (input and output) of the constituent. **Reactor** is a name we give to tanks or to generic volumes in which reactions occur. Part of this section is based on von Sperling and Chernicharo (2005) and von Sperling (2007).

A mass balance is a *quantitative description of all the materials that enter, leave, and accumulate in a system with defined physical boundaries*. The mass balance is based on the law of conservation of mass, that is, mass is neither created nor destroyed. The basic mass balance expression should be derived based on a chosen boundary volume, which can be either a tank or a reactor as a whole, or any volume

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

element within them. In the mass balance, there are terms for the following items (Tchobanoglous and Schroeder, 1985):

- *materials that enter*
- *materials that leave*
- *materials that are generated (produced)*
- *materials that are consumed (decayed and removed)*
- *materials that are accumulated in the selected volume*

When preparing a mass balance, you should consider the following steps (Tchobanoglous & Schroeder, 1985; Metcalf & Eddy, 2014):

- Prepare a simplified schematic or flowsheet of the system or process for which the mass balance will be prepared.
- Draw the system boundaries to define where the mass balance will be applied.
- List all the relevant data that will be used in the preparation of the mass balance in the schematic or flowsheet.
- List all the chemical or biological reaction equations that are judged to represent the process.
- Select a convenient basis on which the numerical calculations will be performed.

In any selected volume (see Figure 12.4), the quantity (mass per unit time, or load) of the accumulated material must be equal to the quantity of the material that enters, minus the quantity that leaves, plus the quantity that is generated, minus the quantity that is consumed. In linguistic terms, the mass balance can be expressed in the following general form:

$$\text{Accumulation} = \text{input} - \text{output} + \text{production} - \text{consumption} \tag{12.6}$$

S. 12.2

Note that this expression is similar to the one used for water balance (Section 12.2). As a matter of fact, a water balance is a particular case of a general mass balance.

Some authors prefer not to assume a negative sign for the term representing consumption, instead expressing it as a produced material with a negative reaction rate coefficient. The convention adopted in this text is the one of Equation 12.6, which leads to a clearer understanding of the four main components

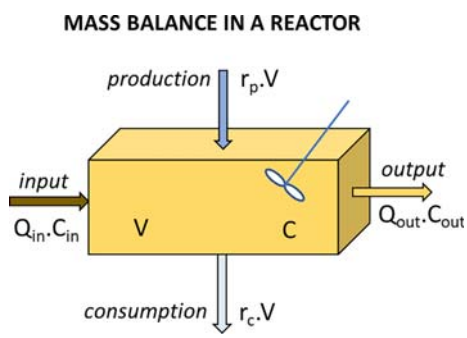


Figure 12.4 Mass balance in a reactor.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

involved in the mass balance. Therefore, you must exercise care and coherence with the signs of each term when adopting one convention or another.

From Figure 12.4, we see that we have the following important terms when structuring a mass balance:

Transport terms (shown as horizontal arrows in Figure 12.4***)***:

$$\text{Load}(\text{g/d}) = \text{flow } Q(\text{m}^3/\text{d}) \times \text{concentration } C(\text{g/m}^3) \tag{12.7}$$

Reaction terms (shown as vertical arrows in Figure 12.4***)***:

$$\text{Load}(\text{g/d}) = \text{rate } r[(\text{g/m}^3)/\text{d}] \times \text{volume of reactor } V(\text{m}^3) \tag{12.8}$$

Mass of constituent (inside the reactor):

$$\text{Mass}(\text{g}) = \text{volume of reactor } V(\text{m}^3) \times \text{concentration in reactor } C(\text{g/m}^3) \tag{12.9}$$

C. 2-4

The **transport terms** and the **mass of the constituent** are simple to obtain, since they are usually based on measurements of flows and concentrations (e.g., see Chapters 2–4). The **reaction terms** are more complex to obtain and can be inferred by the mass balance (but with caution, as will be detailed later in this chapter), by laboratory experiments (provided that they represent well the system under study), or by other suitable strategies.

Mathematically, the linguistic representation in Equation 12.6 can be expressed as

Advanced

$$\frac{\text{d}(C \cdot V)}{\text{d}t} = Q_{\text{in}} \cdot C_{\text{in}} - Q_{\text{out}} \cdot C_{\text{out}} + r_{\text{p}} \cdot V - r_{\text{c}} \cdot V \tag{12.10}$$

where

C = concentration of the constituent inside the reactor at a time t (g/m^3)
C_{in} = input concentration of the constituent (g/m^3)
C_{out} = output concentration of the constituent (g/m^3)
V = volume of the reactor or volume element of any reactor (m^3)
Q = flow (m^3/d)
t = time (d)
r_{p} = reaction rate of production of the constituent [(g/m^3)/d]
r_{c} = reaction rate of consumption of the constituent [(g/m^3)/d].

The units of time can be days, hours, minutes, or any other time unit, depending on how fast the dynamics of the constituent inside the reactor are. The units of mass can be g, mg, or kg, and the resulting units of load can be g/d, kg/d, or any other unit, provided consistency in all units in the equation are adopted.

The terms r_{p} (**rate of production**) and r_{c} (**rate of consumption**) can be expressed as (g/m^3)/d. If they are multiplied by the volume of the tank (V), expressed in m^3, the resulting product has the units of g/d or in other words, mass load. In some cases, there may be no production terms, such as a simple model for BOD decay, and in this case, r_{p} is set equal to zero. Some examples of constituents that have simultaneous production and consumption in the mass balance include oxygen (aeration and deoxygenation) and nitrite (ammonia conversion into nitrite and nitrite conversion into nitrate). Reaction rates will be

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

C. 14

discussed further in Chapter 14. Here, we are just presenting the basic concept of a mass balance in which there are reactions of mass production or consumption.

The representation in Figure 12.4 and in Equation 12.10 is for a generic fluid element. If we want to extrapolate it to a tank, we need to assume, for the sake of simplicity, that a completely stirred tank reactor, also called a complete mix reactor, is being represented. In a complete mix tank, the concentrations are assumed to be equal in all parts of the tank, and thus the output concentration (C_{out}) is the same as the prevailing concentration (C) inside the tank. The concept of **complete mix reactors** will be detailed in Chapter 14, which covers the hydraulic behaviour of tanks.

C. 14

Equation 12.10 can be expressed in the following alternative form, in which the left-hand term has been expanded:

$$C \cdot \frac{dV}{dt} + V \cdot \frac{dC}{dt} = Q_{in} \cdot C_{in} - Q_{out} \cdot C_{out} + r_p \cdot V - r_c.V \tag{12.11}$$

In most systems, the volume (V) in biological reactors can usually be considered to be fixed ($dV/dt = 0$), making the first term on the left-hand side of Equation 12.11 disappear. This leads to the simplified and more usual form of the mass balance, presented in Equation 12.12. Since in this equation the only dimension is time, it represents an *ordinary differential equation*, in which the analytical solution (or numeric computation) is much simpler. However, it must be emphasized that the mass balance in other systems, such as the sludge volume in secondary sedimentation tanks for activated sludge systems, also implies variations in volume (in addition to concentration variations). In this particular case, there are two dimensions (time and space), which lead to a *partial differential equation*. The solution of partial differential equations requires greater mathematical sophistication and is outside the scope of this book. However, for completely mixed reactors (with fixed volumes), the more usual mass balance, expressed in Equation 12.12, is used.

$$V \cdot \frac{dC}{dt} = Q_{in} \cdot C_{in} - Q_{out} \cdot C_{out} + r_p \cdot V - r_c \cdot V \tag{12.12}$$

This equation is very useful and is frequently adopted in most mathematical models of reactors in treatment plants. The left-hand side of the equation has the following units: V in m^3 and dC/dt in $(g/m^3)/d$. The resulting product leads to a mass load with units g/d, which is the same as in all terms in the right-hand side (g/d). On the right-hand side of Equation 12.12, the first two terms are **transport terms** and the third and fourth terms are **reaction terms**.

You may think that the term dC/dt is complicated, since it is expressed in the form of a differential equation, but it is a very simple concept to understand once you break it down (note, we made a similar comment for the water balance and the term dV/dt in Section 12.2). The term dC/dt, expressed as $(g/m^3)/d$ (or other equivalent units for concentration change over time), represents the accumulation term, which can be positive or negative:

S. 12.2

- **When the accumulation term dC/dt is positive**, it means that the concentration of the constituent (C) inside the treatment unit has increased by a certain concentration (dC) during a certain amount of time (dt). If the concentration increases, so does the mass, since mass is equal to the volume multiplied by the concentration ($V{\cdot}C$). In this case, the sum of the positive terms (input and production) is greater than the sum of the negative terms (output and consumption). For instance, if dC/dt is equal to $+2$ $(g/m^3)/d$, it means that, after one day, the concentration of the constituent inside the tank has increased by 2 g/m^3. If the previous concentration in the tank was 50 g/m^3, after this specific day it will be $50 + 2 = 52$ g/m^3. If the volume of liquid inside the reactor was 100 m^3, then the mass

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

increase of the constituent after this specific day will be 1 d × 2 (g/m^3)/d × 100 m^3 = 200 g. The mass of the constituent was previously 50 g/m^3 × 100 m^3 = 5000 g, and after this day it has increased to 5000 + 200 = 5200 g.

- **When the accumulation term dC/dt is negative**, it means that the concentration of the constituent (C) inside the treatment unit has decreased by a certain concentration (dC) during a certain amount of time (dt). In this case, the sum of the positive terms (input and production) is lower than the sum of the negative terms (output and consumption).
- **When the accumulation term dC/dt is equal to zero**, it means that the concentration of the constituent (C) inside the treatment unit is stable. If this persists over a long time, we can imply that we have approached or reached steady state. As highlighted before, *this assumption is usually adopted in most studies of treatment plant evaluation*, unless an advanced mathematical model is used.

S. 12.1

From the concepts of steady state and dynamic state presented in Section 12.1, we can clearly see that Equation 12.12 represents the **dynamic state**, because the input and output variables change over time, and dC/d$t \neq 0$. *The concentration of the constituent in the system is therefore variable with time* and can increase or decrease, depending on the balance between the positive and negative terms.

If you want to assume **steady-state** conditions in your study, you should remember that there are no accumulations of the constituent in the system (or in the volume being analysed). Thus, dC/d$t = 0$, that is, *the concentration of the constituent is constant*. Under these conditions, in which dC/d$t = 0$, the mass balance is given by a simplification of Equation 12.12, leading to

$$0 = Q_{in} \cdot C_{in} - Q_{out} \cdot C_{out} + r_p \cdot V - r_c \cdot V \tag{12.13}$$

If the objective of the model is to estimate the output concentration C_{out}, then Equation 12.13 can be rearranged to lead to an even simpler form. Furthermore, if the inflow is equal to the outflow ($Q_{in} = Q_{out}$), and both are generically represented by Q, and if V is constant, a simplified form is achieved:

$$Q \cdot C_{out} = Q \cdot C_{in} + V \cdot (r_p - r_c) \tag{12.14}$$

$$C_{out} = C_{in} + \frac{V}{Q} \cdot (r_p - r_c) \tag{12.15}$$

S. 13.2

Knowing that V/Q is equal to the hydraulic retention time (t) (see Section 13.2 for more details about retention time), you have now a simple form of a **basic steady-state equation** (Equation 12.16). You should know that *the rates of mass production and consumption (r_p and r_c) are usually not constant and may be variable over time. They also may be a function of the constituent concentration or influenced by limiting factors* (see Chapter 14).

$$\boxed{C_{out} = C_{in} + t \cdot (r_p - r_c)} \tag{12.16}$$

Similar to our comment made above for water balances, if you use a *mass balance* to **estimate a component that has not been measured** by calculating it from the sum and difference of the other factors, you may find that the estimated value has errors, especially if the mass balance is complex and involves several components. For instance, suppose you want to estimate the nitrite conversion rate into nitrate (rate of consumption r_c) in a reactor. Suppose you measure the input and output nitrite loads and estimate the conversion rate of ammonia into nitrite (production rate r_p) using laboratory experiments. Mathematically, the computation of the nitrite consumption rate can be made by assuming that dC/dt is equal to zero and that the consumption rate is equal to the difference from the other components (input

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

load – output load + production) from Equation 12.13. However, this strategy may lead to incorrect values. This is because dC/dt may not be zero and also because there may be errors in the measurement or estimation of the other components (input and output loads) and especially, in the rate of nitrite production (which was estimated based on lab experiments and may not be applicable to your reactor). *In many cases, the mass balance does not close entirely*, and it is necessary that you analyse all factors critically and extract the best possible conclusions.

Example

EXAMPLE 12.3 MASS BALANCE UNDER STEADY-STATE CONDITIONS

Consider the same system covered in Example 12.1, but now we will expand upon that example to complete a mass balance. The analysis is based on several years of measurement of inflows and outflows, together with influent and effluent chemical oxygen demand (COD) concentrations from an extensive system (surface area of 2000 m^2). Precipitation and evaporation values from a nearby weather station were also used. The following yearly average values have been obtained:

- *inflow: 10,000 m^3/year*
- *outflow: 9000 m^3/year*
- *precipitation: 2000 m^3/year*
- *evaporation: 3000 m^3/year*
- *influent COD concentration: 450 mg/L*
- *effluent COD concentration: 120 mg/L*

Complete a mass balance around this unit and interpret the results.

Solution:

(a) Calculate the COD loads for the four components of the mass balance:

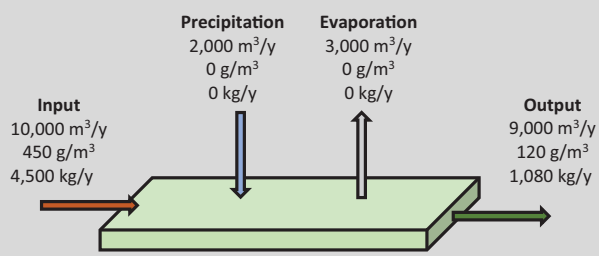
$$\text{Input load} = Q_{in} \times C_{in} = 10{,}000\ \text{m}^3/\text{year} \times 450\ \text{g/m}^3 = 4{,}500{,}000\ \text{g/year} = 4500\ \text{kg/year}$$

$$\text{Output load} = Q_{out} \times C_{out} = 9000\ \text{m}^3/\text{year} \times 120\ \text{g/m}^3 = 1{,}080{,}000\ \text{g/year} = 1080\ \text{kg/year}$$

$$\text{Load gain from precipitation} = 0\ \text{kg/year}\ \text{(no COD comes from rain water)}$$

$$\text{Load loss to evaporation} = 0\ \text{kg/year}\ \text{(no COD is lost with evaporated water)}$$

(b) Interpret the results
The schematics of the mass balance are presented below.



You can see that precipitation and evaporation contributed only to the water balance but not to the COD mass balance (because precipitation does not contain any COD, and no COD is lost due to evaporation).

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

If we calculate the removal efficiency in terms of COD concentration, we obtain

$$E_{\text{concentration}} = \frac{C_{\text{in}-C_{\text{out}}}}{C_{\text{in}}} = \frac{450 - 120}{450} = 0.73 = 73\%$$

S. 12.2

However, the water losses have affected the output concentrations, as discussed in Section 12.2. Therefore, we should calculate removal efficiencies based on influent and effluent loads. The result is

$$E_{\text{load}} = \frac{\text{load}_{\text{in}} - \text{load}_{\text{out}}}{\text{load}_{\text{in}}} = \frac{4500 - 1080}{4500} = 0.76 = 76\%$$

There is a slight difference in the values, but the calculation based on loads is a better representation of the actual removal efficiency and should be preferentially used. If the water losses were higher, the difference between both calculations would be greater, reinforcing even more the adequacy of reporting efficiencies based on loads in systems with substantial water losses.

Example

EXAMPLE 12.4 MASS BALANCE FOR ESTIMATING AN UNMEASURED COMPONENT

A secondary sedimentation tank receiving the effluent from an aeration tank has been monitored over a long time. Average values of input and output flows and suspended solids (SS) concentrations are given below. However, there are no measurements of the flow and SS concentrations that leave from the bottom of the sedimentation tank (underflow), and you would like to estimate their average values based on water and mass balances of the tank.

Data:

- *inflow to the sedimentation tank:* $Q_{in} = 1000\ m^3/d$
- *outflow from the tank (effluent):* $Q_{out} = 600\ m^3/d$
- *underflow from the tank:* $Q_{under} = ?$
- *SS concentration in the influent to the sedimentation tank:* $C_{in} = 3000\ g/m^3$
- *SS concentration in the effluent from the sedimentation tank:* $C_{out} = 30\ g/m^3$
- *SS concentration in the underflow:* $C_{under} = ?$

Solution:

(a) Water balance

Based on these flows, compute the water balance around the sedimentation tank.

The influent flow to the tank is

$$\text{Influent flow: } Q_{\text{in}} = 1000\ \text{m}^3/\text{d}$$

The effluent flow from the tank is

$$\text{Effluent flow: } Q_{\text{out}} = 600\ \text{m}^3/\text{d}$$

Since the sedimentation tank is water sealed, and influences of precipitation and evaporation are negligible given the tank's small surface and short hydraulic retention time of the liquid (as is usual in secondary sedimentation tanks), the water balance can be assumed to be close to steady-state

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

conditions, and the underflow can be estimated by difference between the inflow and the outflow:

$$Q_{\text{underflow}} = Q_{\text{in}} - Q_{\text{out}} = 1000 - 600 = 400\,\text{m}^3/\text{d}$$

(b) Mass balance
Based on these SS concentrations, compute the mass balance around the sedimentation tank. The influent load to the tank is

$$\text{Influent SS load} = Q_{\text{in}} \cdot C_{\text{in}} = 1000\,\text{m}^3/\text{d} \times 3000\,\text{g/m}^3 = 3{,}000{,}000\,\text{g/d} = 3000\,\text{kg/d}$$

The effluent load from the tank is

$$\text{Effluent SS load} = Q_{\text{out}} \cdot C_{\text{out}} = 600\,\text{m}^3/\text{d} \times 30\,\text{g/m}^3 = 18{,}000\,\text{g/d} = 18\,\text{kg/d}$$

You can see that the effluent SS load is much smaller than the influent load, highlighting the average good efficiency of the final clarifier in transferring solids to the bottom.
In secondary sedimentation tanks, it is usually assumed that there are no reactions taking place inside them, and thus the mass balance can be expected to be due only to the transport terms. Therefore, the SS load in the underflow can be computed as the difference between the input and the output loads:

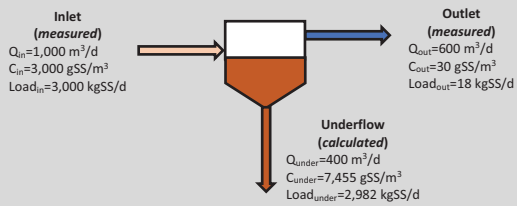
$$\text{Underflow SS load} = \text{influent SS load} - \text{effluent SS load} = 3000 - 18\,\text{kg/d} = 2982\,\text{kg/d}$$

Since we have the SS load and the flow in the underflow from the bottom of the sedimentation tank, the SS concentration can be computed knowing that concentration is equal to load divided by flow:

$$\text{SS concentration in underflow}\left(\frac{\text{g}}{\text{m}^3}\right) = \frac{\text{SS load in underflow(g/d)}}{\text{flow in underflow(m}^3\text{/d)}}$$

$$= \frac{2{,}982{,}000\,\text{g/d}}{400\,\text{m}^3/\text{d}} = 7455\,\text{g/m}^3$$

(c) Schematics of the water and mass balances around the sedimentation tank
The figure illustrates the main components of the water and mass balances (average values):



EXAMPLE 12.5 MASS BALANCE UNDER DYNAMIC CONDITIONS

Example

A dissolved refractory (non-biodegradable) constituent is present in the influent to a treatment plant and as such enters and leaves a tank without being removed (no sedimentation, no conversion, and no decay). Flow has been measured at the inlet and outlet of the tank over a period of 24 h. The flows are the same as those from Example 12.2, and so the water balance is the same. Only influent concentration has been measured, so effluent concentrations will need to be estimated. Complete a mass balance of the constituent in the tank and compute the accumulation terms and

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

resulting concentrations inside the tank. Assume that the tank is well mixed, and thus the effluent concentrations are equal to those prevailing inside the tank.

The initial liquid volume in the tank was 1000 m^3 and the initial concentration of the constituent inside the tank was 10.0 g/m^3. The measured data of the input and output flows and input concentrations in this particular day are shown as follows:

Hour of the Day	Inflow (m^3/h)	Outflow (m^3/h)	C_{in} (g/m^3)
1	110	112	9.30
2	101	114	8.40
3	91	116	8.70
4	98	120	9.20
5	114	129	9.10
6	130	142	8.60
7	163	158	8.50
8	184	174	9.80
9	222	218	11.70
10	238	225	13.50
11	259	239	15.20
12	248	231	13.10
13	213	202	11.70
14	189	182	11.80
15	178	180	12.10
16	161	175	11.60
17	168	170	11.20
18	170	169	10.10
19	188	162	9.70
20	177	158	9.30
21	168	148	8.70
22	129	143	8.40
23	113	132	7.90
24	109	119	7.70

Excel

Note: This example is also available as an Excel spreadsheet.

Solution:

The initial mass of the constituent inside the tank is equal to the product of the initial volume and the initial concentration:

$$\text{Initial mass} = 1000\,\text{m}^3 \times 10.0\,\text{g/m}^3 = 10{,}000\,\text{g}$$

Since there are variations over time, this problem involves the assumption of a *dynamic state*. From Equation 10.10, you have

$$\frac{\mathrm{d}(C \cdot V)}{\mathrm{d}t} = Q_{in} \cdot C_{in} - Q_{out} \cdot C_{out} + r_p \cdot V - r_c \cdot V$$

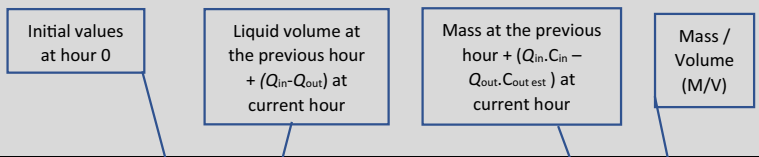
Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

The constituent is refractory, and there are no production and consumption terms (r_p and r_c are equal to zero). Therefore, the equation of mass change in the reactor may be simplified to

$$\frac{d(C \cdot V)}{dt} = Q_{in} \cdot C_{in} - Q_{out} \cdot C_{out}$$

Since we have no measurements of the effluent concentration C_{out}, we will use the mass balance to find the *estimated* value of C_{out}. Since we are dealing with a refractory constituent, the mass balance is less complex, involving only transport terms (no reactions terms), and we will assume that this would be our best choice.

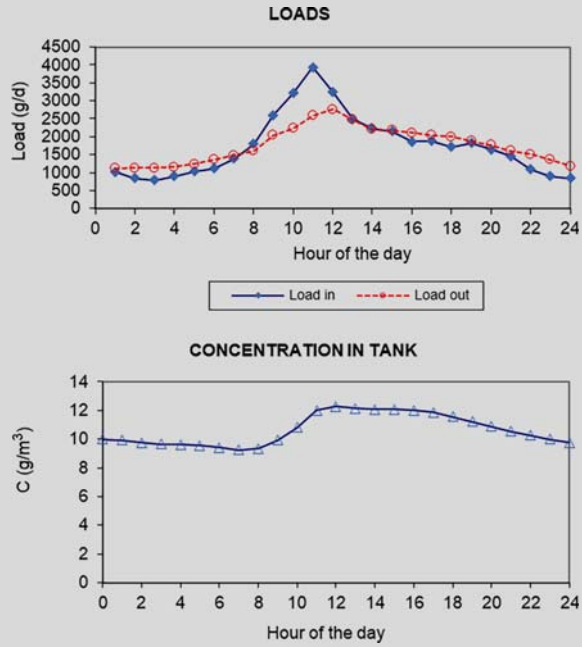
The computational table is presented below. The columns for flow and volume are the same as those from Example 10.2. The mass variations in the tank result from the balance of input and output loads, $load_{in} - load_{out}$ or $Q_{in} \cdot C_{in} - Q_{out} \cdot C_{out\ est}$, as indicated in the equation above. The resulting concentration in the tank comes from the division of mass by volume.



	Flows			Volume	Concentration	Loads			Mass	Concent
Hour	Q_{in}	Q_{out}	Q_{in}-Q_{out}	V	C_{in}	$Q_{in}.C_{in}$	$Q_{out}.C_{out}$	$Q_{in}.C_{in} - Q_{out}.C_{out\ est}$	M	$C_{out\ est}$
	m^3/h	m^3/h	m^3/h		g/m^3	g/h	g/h	g/h	g	g/m^3
0				1000					10000	10.00
1	110	112	-2	998	9.30	1023	1120	-97	9903	9.92
2	101	114	-13	985	8.40	848	1131	-283	9620	9.77
3	91	116	-25	960	8.70	792	1133	-341	9279	9.67
4	98	120	-22	938	9.20	902	1160	-258	9021	9.62
5	114	129	-15	923	9.10	1037	1241	-203	8818	9.55
6	130	142	-12	911	8.60	1118	1357	-239	8579	9.42
7	163	158	5	916	8.50	1386	1488	-102	8477	9.25
8	184	174	10	926	9.80	1803	1610	193	8670	9.36
9	222	218	4	930	11.70	2597	2041	556	9226	9.92
10	238	225	13	943	13.50	3213	2232	981	10207	10.82
11	259	239	20	963	15.20	3937	2587	1350	11557	12.00
12	248	231	17	980	13.10	3249	2772	477	12033	12.28
13	213	202	11	991	11.70	2492	2480	12	12045	12.15
14	189	182	7	998	11.80	2230	2212	18	12063	12.09
15	178	180	-2	996	12.10	2154	2176	-22	12041	12.09
16	161	175	-14	982	11.60	1868	2116	-248	11793	12.01
17	168	170	-2	980	11.20	1882	2042	-160	11633	11.87
18	170	169	1	981	10.10	1717	2006	-289	11344	11.56
19	188	162	26	1007	9.70	1824	1873	-50	11294	11.22
20	177	158	19	1026	9.30	1646	1772	-126	11168	10.89
21	168	148	20	1046	8.70	1462	1611	-149	11019	10.53
22	129	143	-14	1032	8.40	1084	1506	-423	10596	10.27
23	113	132	-19	1013	7.90	893	1355	-463	10133	10.00
24	109	119	-10	1003	7.70	839	1190	-351	9782	9.75

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

The 24-h time series graphs of input and output loads and concentrations of the constituent in the reactor are shown below. The concentration inside the tank increased from 8:00, when the input load was higher than the output load, and decreased mildly after 15:00, when output loads were higher than input loads.



From the computational table, the sum of the loads over the 24 h led to a value of 41,994 g/d for the input and 42,212 g/d for the output. The difference between them is -218 g/d. This is why the mass in the tank decreased, from the initial value of 10,000 g, to the final value of $10{,}000 - 218 = 9782$ g.

Note that these computations, although founded on the dynamic state, are based only on measured values at hourly intervals. They do not involve integration of ordinary differential equations, which are typical from dynamic models, and which would *require much shorter time steps* (small fractions of an hour) to give accuracy to the numerical calculations (see Section 14.2 for a discussion on numerical integration).

S. 14.2

If you were investigating a reactive constituent, the mass balance equation would involve production or consumption terms, the rate of which is not so simple to obtain. Good kinetic models are necessary to represent well the conversion process, and these need to be incorporated into a suitable hydraulic model for the reactor (see Chapter 14). In the current example, the simplified assumption that the entire reactor was completely mixed was adopted, which substantially simplified the calculations.

C. 14

12.4 CHECK-LIST FOR YOUR REPORT

- ✓ Specify clearly if you are adopting the steady-state or the dynamic-state assumption.
- ✓ Make clear all the components of your water and mass balances.
- ✓ Specify which components of the water and mass balances are measured and which are estimated.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

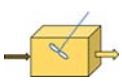
- ✓ If the components of water and mass balances are estimated, explain clearly how they were estimated.
- ✓ If a component of a water or mass balance has been estimated by simple difference from the sum of the other terms, assuming that the balance would close perfectly, make this assumption clear. Analyse critically the implications of this assumption, which may not be a good representation of reality.
- ✓ Check consistency of all units of the water and mass balances (units of time, mass, and volume) and make the units clear in your calculations and summary tables or figures.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Chapter 13

Loading rates applied to treatment units

This chapter presents the different types of hydraulic and mass loading rates, and how to calculate and interpret them. Loading rates are used for the design of treatment units and for experimental studies that aim to investigate treatment performance under different loading conditions. In existing plants, knowledge of the applied loading rates assists in the understanding of the behaviour of the treatment units. This chapter will help you to understand the concept and the physical meaning of the different types of hydraulic and mass loading rates.



The contents in this chapter are only applicable to treatment plant studies, and not to the evaluation of water bodies.

CHAPTER CONTENTS

13.1 The Different Types of Loading Rates **500**
13.2 Hydraulic Retention Time **507**
13.3 Volumetric Hydraulic Loading Rate **517**
13.4 Surface Hydraulic Loading Rate **518**
13.5 Volumetric Mass Loading Rate **520**
13.6 Surface Mass Loading Rate **521**
13.7 Specific Surface Mass Loading Rate **522**
13.8 Food-to-microrganism Ratio (F/M) **523**
13.9 Sludge Age **525**
13.10 Check-List for Your Report **529**

© 2020 The Authors. This is an Open Access book chapter distributed under the terms of the Creative Commons Attribution Licence (CC BY-NC-ND 4.0), which permits copying and redistribution for non-commercial purposes with no derivatives, provided the original work is properly cited (https://creativecommons.org/licenses/by-nc-nd/4.0/). This does not affect the rights licensed or assigned from any third party in this book. The chapter is from the book Assessment of Treatment Plant Performance and Water Quality Data: A Guide for Students, Researchers and Practitioners, Marcos von Sperling, Matthew E. Verbyla and Sílvia M. A. C. Oliveira (Authors).
doi: 10.2166/9781780409320_0499

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Basic

13.1 THE DIFFERENT TYPES OF LOADING RATES

Treatment performance may be highly influenced by the applied loading rates on the unit you are studying. This chapter will present the most common loading rates that are used for design purposes and to evaluate the performance of a treatment unit.

The loading rates covered here are:

- **surface** and **volumetric *hydraulic* loading rates**, which associate the unit surface (m^2) or volume (m^3) with the **flow** (m^3/d) it receives.
- **surface** and **volumetric *mass* loading rates**, which associate the unit surface (m^2) or volume (m^3) with the **mass load** (g/d or kg/d) it receives (the load may be of some constituent, such as chemical oxygen demand (COD), biochemical oxygen demand (BOD), suspended solids (SS), and ammonia-N).
- **other loading rates**, expressed in different forms (F/M ratio, sludge age).

All of these loading rates are used for **design purposes**. When a treatment plant is built, the operators will start by following the design specifications regarding loading rates, but after some time, with more experience gained from the specific plant, they may adapt the loading rate to optimize the plant's performance. In many cases, not much can be done to alter the loading rates, because the areas and volumes of the treatment units are fixed, and the plant needs to treat whatever flow rate is received. However, in some situations, flexibility may be incorporated in the design to operate using different configurations that alter the loading rates. Also, for some loading parameters, such as the F/M ratio or sludge age, **operational procedures** may be modified to find the loading rates that optimize treatment performance.

You should remember that treatment plants are designed for flows and loads that will occur in the **future**, after some years (the planning horizon for design is usually 20 or 30 years) (von Sperling & Chernicharo, 2005). Therefore, during the **initial years** of operation, the incoming flows and loads may be much lower than those adopted in the design, and the operational staff needs to understand these implications.

Also, remember that **designs** are usually made based on **estimates** of future flows, and concentrations and loads of the main constituents of interest, usually supported by the literature or by design guidelines or standards. It is not uncommon for the **monitoring** results from a new treatment plant to show different **actual values** than that were anticipated at the design stage, which may be a result of the fact that the units initially receive hydraulic and mass loads that are different from what was envisaged during design. The operational staff needs to understand these implications.

In general, a treatment unit in a real plant may find itself in one of the following situations regarding the applied loads:

- **Operation at the desired loading rate**. Loading rates are considered adequate according to the literature, design guidelines, regional experience, or successful adjustments made by the operational staff. *Performance is expected to be adequate*.
- **Operation at overloading conditions**. Overloading conditions indicate that the treatment unit is receiving a higher hydraulic or mass load than expected. Loading rates are considered above the recommended values according to the literature, design guidelines, or regional experience. *Performance may not necessarily be directly affected, but there are risks of deterioration in performance* if the system becomes too overloaded due to insufficient retention time, insufficient pumping, piping or hydraulic appurtenances capacity, pollutant loads in excess of the biomass capability, oxygen consumption higher than the supply capacity, etc.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

- **Operation at underloading conditions**. Underloading conditions occur if the treatment unit is receiving a lower hydraulic or mass loading than expected. Loading rates are considered below the recommended values according to the design, literature, guidelines, or regional experience. In principle, *performance should not be negatively affected and, in general, it could be expected to improve*, due to the application of loads that are below the system's capacity. However, there are situations when underloading may not be desired, such as when it leads to long retention times at unaerated locations where undesired anaerobic conditions can take place, leading to bad smell. Also, in some treatment units meant to operate at anaerobic conditions, very low loads may make the unit oscillate between aerobic and anaerobic, which is undesirable for the biological community.

Note that the considerations above are for volumetric and mass loadings, in which *high values* may tend to cause overloading. For hydraulic retention time (HRT) and solids retention time (see Table 13.1) it is the opposite: *low values* are those which may be associated with overloading. Also note that hydraulic underloading is not generally a problem for large systems designed with *redundancy*, for example, an activated sludge system with four aeration basins in parallel, which could operate at a maximum capacity using all four of the units, but if the hydraulic loading is very low, the plant may only use two or three of the units.

Apart from the operation of real full-scale plants, you may also be involved in studies of **laboratory or pilot scale systems**, in which consideration of loading conditions is also an integral part of your experimental design. Because it is easy to apply different loads under controlled conditions, the system's capacity and the resulting performance may be tested at different loading rates, and you may extract useful results from these experimental manipulations. Typical applications are as follows (see Chapter 10 for the statistical treatment of these types of study):

C. 10

- **Comparison of phases in sequence**. For instance, your experiment may involve one treatment unit that is operated at two, three, or more different phases in time, one after the other, each for a period of several months in sequence, with each phase having different loading conditions.
- **Comparison of units in parallel**. Alternatively, your treatment system may have two, three, or more equal units in parallel, all operating simultaneously, with each unit receiving a different load, with the experiment lasting a period of several months.

Table 13.1 summarizes the most commonly used **types of loading rates** in water and wastewater treatment practice, together with their **main applications**. The following notation is used in this book, but there are also other notations widely adopted in the literature:

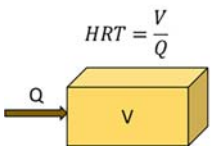
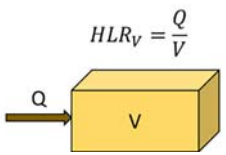
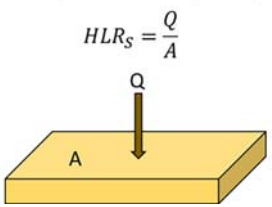
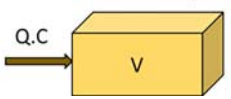
- ***Hydraulic* loading rate: HLR**
  - **Volumetric** hydraulic loading rate: use subscript 'V' for *volume*: $\mathbf{HLR_V}$
  - **Surface** hydraulic loading rate: use subscript 'S' for *surface* area: $\mathbf{HLR_S}$
- ***Mass* loading rate: MLR**
  - **Volumetric** mass loading rate: use subscript 'V' for *volume*: $\mathbf{MLR_V}$
  - **Surface** mass loading rate: use subscript 'S' for *surface* area: $\mathbf{MLR_S}$

In the specific but widely used concept of **organic loading rate (BOD or COD)**, the notation can be $\mathbf{OLR_V}$ and $\mathbf{OLR_S}$, standing for *volumetric* organic loading rate and *surface* organic loading rate, respectively. Therefore, the term 'mass' is substituted by 'organic,' but the concept is the same.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

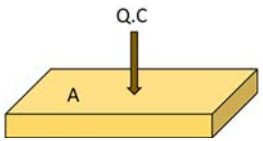
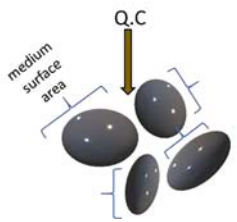
Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

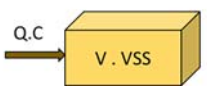
Table 13.1 Different types of loading rates used in water and wastewater treatment practice.

Loading Rate	Concept	Examples of Typical Applications
Hydraulic loading rates		
Hydraulic retention time (HRT) 	• Calculated by the quotient of the tank volume (V) and the applied flow (Q) • It is called 'theoretical' hydraulic retention time, because the actual average retention time of fluid elements is likely to be different in real tanks • It is not exactly a loading rate, but may be considered one, since it involves volume and flow • Typical units: d, h, min	• Most tanks and reactors in water and wastewater treatment practice.
Volumetric hydraulic loading rate (HLR_V) 	• Calculated by the quotient of the applied flow (Q) and the tank volume (V) • In tanks with continuous feeding and without support medium, it is equal to the reciprocal of the hydraulic retention time • Typical units: $(m^3/d)/m^3$, $(m^3/h)/m^3$	• Anaerobic reactors • Coarse filters
Surface hydraulic loading rate (HLR_S) 	• Calculated by the quotient of the applied flow (Q) and the tank surface area (A) • Typical units: $(m^3/d)/m^2$, $(m^3/h)/m^2$, $(L/d)/m^2$, $(L/h)/m^2$, m/d, mm/d	• Grit chambers • Sedimentation tanks • Filters • Horizontal wetlands • Vertical wetlands • Overland flow • Infiltration systems • Trickling filters • Aerated biofilters • Membranes • Sludge dewatering
Mass loading rates		
Volumetric mass loading rate (MLR_V) 	• Calculated as the applied mass load ($Q{\cdot}C$) divided by the tank volume (V) • The concept is general, but a common application is for organic loading rate, in which C is the input concentration of BOD or COD • Typical units: $(kg/d)/m^3$, $(kg/h)/m^3$, $(g/d)/m^3$, $(g/h)/m^3$, $(kg/year)/m^2$	• Anaerobic ponds • Anaerobic reactors • Trickling filters • Aerated biofilters • Sludge digesters

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Table 13.1 Different types of loading rates used in water and wastewater treatment practice. (*Continued*).

Loading Rate	Concept	Examples of Typical Applications
Surface mass loading rate (MLR_S) $MLR_S = \frac{Q.C}{A}$ 	• Calculated as the applied mass load ($Q{\cdot}C$) divided by the tank surface area (A) • Also called 'mass flux' • The concept is general, but a common application is for organic loading rate, in which C is the input concentration of BOD or COD • Typical units: (kg/d)/m^2, (kg/h)/m^2, (g/d)/m^2, (g/h)/m^2, (kg/year)/m^2, (kg/d)/ha	• Secondary clarifiers (activated sludge) • Sand filters • Facultative ponds • Horizontal wetlands • Vertical wetlands • Membranes • Sludge thickening • Sludge dewatering
Specific surface mass loading rate (MLR_S) $MLR_S = \frac{Q.C}{total\ media\ surface}$ 	• Calculated as the applied mass load ($Q{\cdot}C$) divided by the total available surface area of all elements comprising the medium (total media surface area) • For instance, if stones have a specific surface area of 50 m^2/m^3, then for each load unit (e.g., 1 g/d) applied to a m^3 of reactor volume, the specific MLR_S will be 1 g/d per 50 m^2 of medium surface, or 0.020 (g/d)/m^2 of medium surface • Typical units: (kg/d)/m^2 media, (kg/h)/m^2 media, (g/d)/m^2 media, (g/h)/m^2 media	• Trickling filters (e.g., ammonia-N load for nitrification) • Activated sludge with fixed film carriers • Biological activated carbon filters • Ion exchange systems
Other types of loading rates		

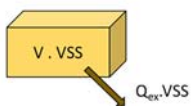
Food-to-microorganism ratio (F/M)
$F/M = \frac{Q.C}{V.VSS}$
 
	• Calculated as the applied substrate load ($Q{\cdot}C$) divided by the mass of biomass in the reactor ($V{\cdot}$VSS)
	• Substrate load is usually referring to the BOD or COD load, so the C in the equation is representing the input BOD or COD concentration
	• Mass of biomass in the reactor is usually VSS or SS mass, which is given by the product of the reactor volume and the VSS or SS concentration
	  ○ VSS: volatile suspended solids, also called mixed liquor volatile suspended solids or MLVSS
	  ○ SS: suspended solids, also called mixed liquor suspended solids or MLSS
	• In the F/M ratio, the numerator (substrate load) is understood as the 'food,' and the denominator (mass of biomass) represents the microorganisms that 'eat the food'
	• Typical units: (gBOD/d)/gVSS, (gBOD/d)/gSS, (gCOD/d)/gVSS, (gCOD/d)/gSS
		• Aeration tanks in the activated sludge process

(*Continued*)

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Table 13.1 Different types of loading rates used in water and wastewater treatment practice. (*Continued*).

Loading Rate	Concept	Examples of Typical Applications

Sludge age (solids retention time) (SRT)
$SRT = \frac{V.VSS}{Q_{ex}.VSS}$
 
 	
	• Calculated as the mass of biomass in the reactor (V·VSS) divided by the mass of biomass that is removed from the system per day (Q_{ex}·VSS)
	• The mass of biomass in the reactor is usually represented as the VSS or SS mass, which is given by the product of the reactor volume and the VSS or SS concentration
	• The mass of biomass that is removed from the system per day (or load removed) is usually the product of the biomass concentration (expressed as VSS or SS) and the flow of excess sludge (also called waste sludge, surplus sludge, biological sludge) removed from the system
	  ○ VSS: volatile suspended solids, also called mixed liquor volatile suspended solids or MLVSS
	  ○ SS: suspended solids, also called mixed liquor suspended solids or MLSS
	• It is not exactly a loading rate, but may be considered one, since it involves a mass and loads
	• Another name is "mean cell residence time" (MCRT)
	• Typical units: days
		• Aeration tanks in the activated sludge process

Let us make a note about the **units of expression** for loading rates. As you saw, the loading rates are expressed in terms of flow or load per unit surface or per unit volume. In Table 13.1, you see different units representing flows, loads, and areas. The question here is not the units we should use, but how we should report the combination of units in the numerator and denominator of the loading rates. Consider the following:

- **Expression of units using scientific notation**. Examples: $m^3 \cdot m^{-2} \cdot d^{-1}$ and $g \cdot m^{-3} \cdot d^{-1}$ or $m^3\ m^{-2}\ d^{-1}$ and $g\ m^{-3}\ d^{-1}$. This is the preferred choice for scientific publications because it makes it clear what goes in the numerator and what goes in the denominator.
- **Expression of units using casual notation**. Examples: m^3/m^2 d and g/m^3 d. Although this does not make it clear what goes in the denominator (is 'd' in the numerator or denominator?), it is informally used in many books, and readers seem to accept easily that 'd' is in the denominator and do not have much trouble to understand it.
- **Notation used in this book**. Examples: $(m^3/d)/m^2$ and $(g/d)/m^3$. Although this notation is not so widely applied, it does not bring confusion on what goes in the numerator and in the denominator, and it also emphasizes the concept of loading rates: the numerators are flows or loads, and the denominators are areas or volumes.

The loading rates presented in Table 13.1 are all used for **design purposes**. For design applications, you should rearrange the equations with the loading rates, putting the volume V or surface area A on the left-hand-side (in a design scenario, this is usually the unknown value that you want to calculate, in order to size the system). By adopting a value of the loading rate based on recommendations from the literature, design standards, design guidelines, or regional experience, you can calculate the necessary volumes or surface areas. The design of treatment plants is outside of the scope of this book, though there are many excellent texts that go into great detail about how to design a treatment system (e.g., Hammer & Hammer, 2014; Metcalf & Eddy, 2014; and several others). You can also consult the books in the Biological Treatment Series and the book *Biological Wastewater Treatment in Warm Climate Regions* (von Sperling & Chernicharo, 2005), which we believe are companions to our current book, because they are also available as open access sources by IWA Publishing.

For the **assessment of existing treatment plants**, which is the main objective of this book, in order to better understand the behaviour of your system, you should try to *calculate the applied loading rates, compare the values with the recommended ones in the literature, and analyse whether their variation has influenced system performance*.

The recommended loading rate values are specific to each process, and the presentation of their typical values is also outside the scope of this book. You should seek the recommended ranges in the pertinent literature and draw your conclusions based on the procedures listed in this chapter.

Advanced

If you are investigating different operational phases in your treatment system, and each phase is associated with different applied loading rates, your report should indicate a clear **description of each phase** and the associated loading rates. A typical **summary table** to be included in the methods section of your report could look like the one presented in Table 13.2.

We have the following suggestions for your work and your report, associated with Table 13.2:

- Give a number, letter, or an acronym to each operating phase, but *do not use acronyms in excess* – your reader will be confused. In the Conclusion section of your report, do not make conclusions mentioning Phase 1 or Phase 2, but rather describe the operational condition that each one covered and their influence on the results. For instance, instead of concluding that '*… the reactor in Phase 1 had a better removal efficiency than in Phase 2…*,' state that '*… the applied loading rates of xxx led to a better performance of the reactor than loading rates of yyy…*'

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Table 13.2 Example of a table showing a description of different operational phases, associated with the applied loading rates, to be included in the Methods section of your report.

Phase	Period	Duration (months)	Target Loading Rate (Unit)	Actual Loading Rate (Unit)	Liquid Temperature (°C)	Etc	Etc
1	Feb/2017 to Sep/2017	8	3.0	2.7	22.5	…	…
2	Oct/2017 to June/2018	9	6.0	6.1	20.4	…	…
3	July/2018 to Jan/2019	7	9.0	8.8	21.9	…	…

- Stating the period of each phase (*month/year to month/year* or date *dd/mm/yyyy* to date *dd/mm/yyyy*) helps the reader to understand whether one phase was predominantly in the winter or predominantly in the summer. This is important in locations where there are substantial seasonal variations along the year.
- The duration of each phase should be sufficient to arrive at stable operating conditions and generate results that support your conclusions. With biological systems, the duration should be long enough to accommodate the inherent variations that take place within the microbial communities. *Avoid adopting a large number of operational phases, each one with a short duration*, because they may not be representative. Remember that the beginning of a new phase may still reflect the operating conditions of the previous phase, and microorganisms need time to adapt to the new situation or the new loading conditions. If you are undertaking physical–chemical experiments, the response is likely to be faster and less complex than with biological experiments.
- Report your target loading rate for the experiment. *Make it clear which type of loading rate you are reporting* (see Table 13.2) and the appropriate units.
- Unless you are running completely controlled experiments, it is very likely that *your actual applied loading rate will be different from the target one*. Even if you try to control the input flow, the input concentrations may be different from the ones you imagined, and may also be variable with respect to time, thus affecting your mass loading rates. Report the *mean or median values of the actual loading rate* you obtained during each phase. If necessary, include *descriptive statistics of the actual loading conditions* in each phase, together with relevant graphs (boxplot, time series, or other).
- *Include other variables that may be influential in the results of your experiments*. For instance, if the liquid temperature was different in each phase, and you know that temperature has an influence on the associated biochemical, chemical, and physical reactions, you should present the mean or median in each phase and, if judged necessary, its descriptive statistics.
- Include *any other information you may find important in this summary table*. It will assist you when presenting and discussing the results and will also assist the reader to find this important information all together in an organized way.
- Experiments with water and wastewater treatment are not simple. Of course, one should understand that, in practice, *it is difficult to accomplish the exact conditions you aimed to obtain for your experimental design*. This is part of studies in real-life systems, and you should be flexible to adjust the experiments to your time and resources constraints. But you are also responsible for *including and discussing possible limitations in the completion of your experiments* – you should be the one to show this to the reader, and not wait for the reader to discover on their own or have to try to guess them. Even with difficulties associated with running your experiments, there are always lessons to learn, knowledge to transfer, and results to be discussed.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Now that we have made a general description of the main types of loading rates and how you should discuss them in your report, we can go into a more detailed analysis of each of them. Note that there are **basic concepts** that you should know, and also some more **advanced concepts** that you should also become acquainted with, depending on the circumstances of your study.

13.2 HYDRAULIC RETENTION TIME

13.2.1 The general concept of hydraulic retention time

Basic

Retention time, also called *detention time* or *residence time*, corresponds to the average period spent in a given volume, most frequently a tank or reactor. In the case of a liquid, the hydraulic retention time corresponds to the average time spent by the water molecules in the reactor, from the moment they enter to the moment they exit. Hydraulic retention time in a continuously fed tank is usually represented by '**HRT**' or simply '**t,**' and is given by

$$\text{HRT} = \frac{\text{volume of liquid in tank}}{\text{volume of liquid } \textit{entering} \text{ the tank per unit time}} \tag{13.1}$$

The denominator of Equation 13.1 is the input flow to the tank. If we consider that the output flow from the tank is equal to the input flow ($Q_{out} = Q_{in}$), then Equation 13.2 can also be expressed as follows:

$$\text{HRT} = \frac{\text{volume of liquid in tank}}{\text{volume of liquid } \textit{leaving} \text{ the tank per unit time}} \tag{13.2}$$

If Q_{in} and Q_{out} are equal, Equations 13.1 and 13.2 can be expressed as follows:

$$\boxed{t = \frac{V}{Q}} \tag{13.3}$$

where

t = HRT (min, h, d)
V = volume of liquid in the tank (m^3)
Q = inflow or outflow (m^3/min, m^3/h, m^3/d)

Equation 13.3 is of paramount importance in water and wastewater treatment practice. The retention time calculated from this equation is understood as the '**theoretical**' or '**nominal**' hydraulic retention time. *In real tanks, the retention time is not a single number, but rather a distribution of times.* This is known as the **residence time distribution**. If you consider water molecules passing through a tank, you can imagine that some pass through faster while others pass through slower. If we were to plot the distribution of these times, we might typically get a right-skewed bell curve with a long tail. The mean of this bell curve is known as *the actual mean retention time of the reactor, and it is frequently different from the theoretical one* due to imperfections in the flow behaviour inside the tank. Approaches used to characterize the real hydraulic behaviour in existing tanks are presented in Section 13.2.6.

S. 13.2.6

The schematic representation of HRT for a generic tank receiving flow continuously is shown in Figure 13.1. You should know that *the theoretical hydraulic retention time is not affected by the direction of flow*: in all cases, it is given by Equations 13.1–13.3.

As we saw in Section 12.2, which dealt with water balances, **inflows and outflows may be variable over time**, following diurnal variations, changes throughout the week, with respect to the intrusion of rainwater

S. 12.2

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

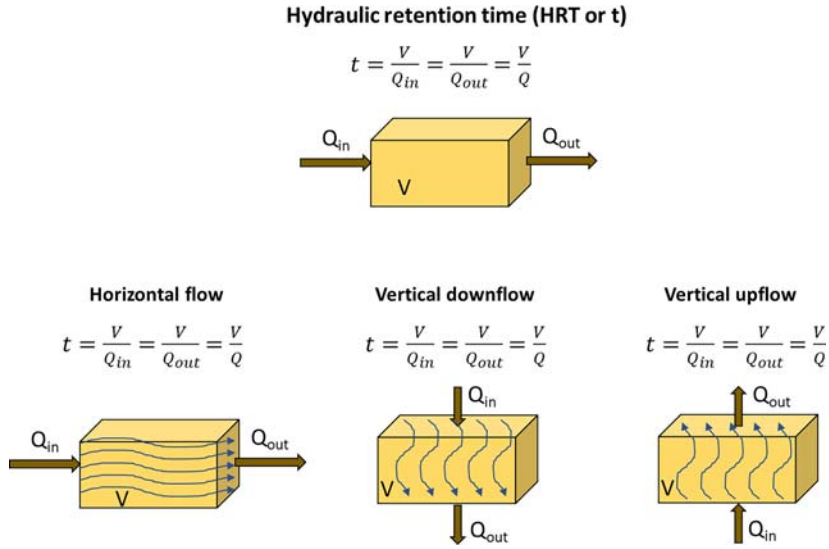


Figure 13.1 Schematic representation of HRT in a generic tank receiving continuous flow. Top: generic case. Bottom: horizontal flow, downflow, upflow. In all cases the theoretical HRT is equal to V/Q. All the reactor volume is occupied by liquid.

into the system, etc. *The concept of HRT being equal to the liquid volume V divided by the flow Q* (*even if it is variable*) *still holds, and thus HRT will be variable with respect to time*. A hydraulic surge (a sudden increase in the influent flow rate) will cause a drop in HRT (assuming that the liquid volume V in the tank remains approximately the same). The HRT may come back to its original value after the surge has passed. For instance, a flow increase over the weekend in a touristic town will cause a reduction on HRT during this period of flow increase. The relative impact of the flow variation on system performance will depend on the magnitude of the HRT. Large reactors, such as those used in natural treatment systems (ponds, wetlands) which have HRTs of several days or even weeks, will suffer little impact from a flow variation in a time scale of hours. However, compact reactors used in intensive systems, which have HRTs on the order of several hours, are likely to be affected by hourly variations in the inflow. When the flow variations are expected, because they are an integral part of the system (such as diurnal variations), the design usually should take them into account to ensure that the system will behave well during the hours of increased hydraulic loading.

Advanced

In systems in which there are **water losses or gains** (see water balance in Section 12.2), and in which the outflow may be different from the inflow, you may adopt, for the sake of simplicity, the average value between Q_{in} and Q_{out} for the computation of HRT. In this case, HRT may be calculated as follows:

S. 12.2

$$t = \frac{V}{(Q_{\text{in}} + Q_{\text{out}})/2} \tag{13.4}$$

where

Q_{in} = input flow (m^3/min, m^3/h, m^3/d)
Q_{out} = output flow (m^3/min, m^3/h, m^3/d)

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

13.2.2 Influence of the reactor dimensions on the theoretical hydraulic retention time

Basic

Note that the **theoretical HRT**, as given by the general Equation 13.3 is **not influenced by the tank dimensions** (length L, width W, height H), and is solely a **function of the liquid volume** in the tank and the applied flow. However, the **actual mean HRT**, in a real tank, may be influenced by the geometric dimensions of the reactor, and their relative position with respect to the flow direction, as we will discuss in Section 13.2.6.

S. 13.2.6

Advanced

Also note that the equation for predicting the ***theoretical*** mean HRT **does not consider the influence of baffles in a tank, although baffles clearly affect the actual mean HRT**. See Figure 13.2, in which there are two tanks with the same external dimensions, the same surface area, and the same volume. However, one tank has no baffles and the other tank has two baffles introduced in the longitudinal direction. In the tank without baffles, HRT is calculated in the usual way. In this idealized tank, if all flowlines moved in parallel from the inlet to outlet, the time (t) for crossing the length of the tank would be equal to the length (L) divided by the horizontal flow velocity (v). According to the continuity equation, this velocity (v) is equal to the flow (Q) divided by the cross-sectional area, which, in this case, is the width (W) times the liquid height (H). If we put both equations together, we end up with $t = (L \cdot W \cdot H)/Q$, which is basically the traditional equation of $t = V/Q$. In the tank with two baffles, the flow passes along the three resulting channels. Since each channel has $1/3$ of the original width, its cross-section will have $1/3$ of the area and, according to the continuity equation, the horizontal flow velocity will be multiplied by 3. However, also the length of the total flow path will be multiplied by 3, because the liquid needs to pass along the three channels. As a result, the flow velocity and the length will both be multiplied by 3, and the travelling time will remain the same. Again, we end up with $t = (L \cdot W \cdot H)/Q$, coming back to the traditional equation of $t = V/Q$. But recall that these calculations are for the *theoretical* mean HRT – the actual mean HRT will likely be affected (typically made longer) by the introduction of baffles.

13.2.3 Influence of internal recirculations on the theoretical hydraulic retention time

Advanced

Another important concept for you to understand is that **internal recirculations do not alter the calculation of the theoretical mean HRT, though they might affect the actual mean HRT**. If the recirculation is internal, that is, within the system boundaries, the HRT of a tank with volume V is only influenced by the flow (Q) that enters and leaves the system. The recirculation flow (Q_r), being internal

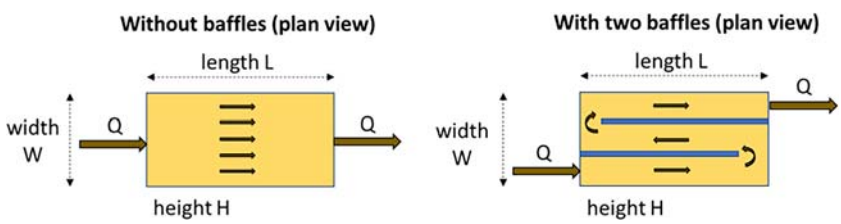


Figure 13.2 Schematic representation of the absence of influence of baffles on the theoretical hydraulic retention time. Left: tank with no baffles; right: tank with two longitudinal baffles and resulting three channels along the length. In both tanks the theoretical HRT is the same.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

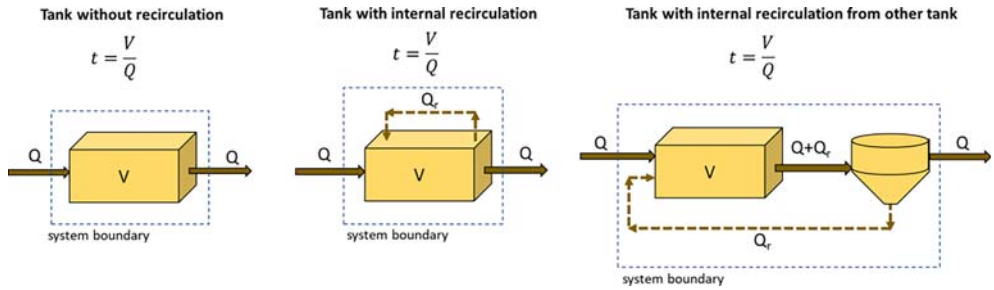


Figure 13.3 Illustration of the fact that internal recirculations do not alter the theoretical hydraulic retention time of a tank that has a volume V and receives a flow Q. The left figure shows the traditional concept, without recirculation, the middle figure shows a recirculation within the tank and the right figure shows a recirculation that comes from a separate tank.

to the system, has no influence on the calculation of the theoretical mean HRT. Figure 13.3 illustrates three possible situations, and in all of them the theoretical HRT is given by $t = V/Q$:

- The left-hand-side of the figure illustrates the **traditional case, in which there is no recirculation**, and the theoretical mean HRT is obviously V/Q.
- The middle figure shows a **tank that has an internal recirculation, from a region close to the outlet to a region close to the inlet**. In this case, the theoretical mean HRT continues to be equal to V/Q, irrespective of the value of Q_r. An easy way of understanding this is making a parallel with a tank that is completely mixed. In this type of tank, all contents are fully mixed in the internal volume, the concentrations are the same in all parts of the tank, and this mixing is equivalent to the existence of multiple internal recirculations, bringing fluid elements from one part of the tank to another. Note that even though this does not affect the way we calculate the theoretical HRT, it will likely affect the true mean HRT.
- The right-hand-side figure is more complex, but is typical for some treatment systems, notably the activated sludge process. The **recirculation (Q_r) comes from a second tank** (in the case of the activated sludge process, from the secondary sedimentation tank) **and is redirected to the first tank** (in the case of activated sludge, the aeration tank). Since the recirculation is internal to the system (see the system boundary, surrounding both tanks) it will not affect the theoretical HRT, which will remain equal to V/Q. One way of understanding this is by the following example: if the recirculation flow Q_r is equal to the flow Q (recirculation ratio $Q_r/Q = 1$), the influent to the aeration tank will be $Q + Q_r$, the total inflow will double, and so the retention time in this passage of the liquid will be halved. But because there is a recirculation with a ratio Q_r/Q equal to 1, there will be another chance of the liquid to pass again at the aeration tank, with half of the retention time. In total, we will have $(0.5 + 0.5) = 1.0$ HRT. If the recirculation ratio were $Q_r/Q = 2$, the retention time in each passing would be $1/3$ of the original HRT, but the number of passages would be 3, leading to $1/3 + 1/3 + 1/3 = 1.0$ HRT. Another way of understanding the fact that HRT in the tank will not be affected by the recirculation is to remember that the sedimentation tank is internal to the system, and sending part of the flow from one place of the system to another place will be similar to the example of the complete-mix tank, in the middle figure. Instead of thinking in terms of one tank, we can think in terms of the system. Therefore, HRT in the first tank will continue to be given by V/Q. For the sake of simplicity, in this example we did not include the component of sludge wasting, which usually represents only a minor fraction of the influent flow (1–2%).

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

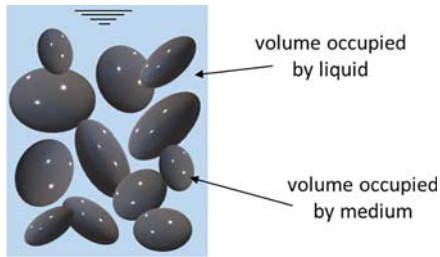


Figure 13.4 Schematic representation of part of a tank with medium, with an indication of the volume occupied by the liquid and by the medium. The tank is hydraulically saturated.

13.2.4 Influence of a support medium on the theoretical hydraulic retention time

Basic

Expanding our understanding of the concept of HRT, we need to analyse now the **case in which the tank is filled with a medium**, such as gravel, stones, sand, or plastic material, which is a common situation in several treatment units in water and wastewater treatment practice. This is the case, for instance, of horizontal subsurface flow constructed wetlands, coarse filters for effluent polishing, and other similar units that receive continuous flow which passes through a porous medium. The medium occupies part of the tank volume, and the volume occupied by the liquid, consequently, will be smaller (see Figure 13.4). In the figure, we assume that the tank is **hydraulically saturated**, that is, all pore (void) spaces are occupied by liquid.

For the calculation of the HRT, what counts is the volume occupied by the liquid. Since the medium is in a hydraulically saturated tank, the volume of liquid is equal to the volume of the pore spaces:

$$\begin{aligned}\text{Volume of liquid} &= \text{Volume of pore spaces} \\ &= \text{Volume of tank} \times \text{Medium porosity}\end{aligned} \tag{13.5}$$

The resulting theoretical HRT will be:

$$\boxed{\text{HRT} = \frac{\text{Volume of liquid}}{\text{Flow}} = \frac{\text{Volume of tank} \times \text{Porosity}}{\text{Flow}}} \tag{13.6}$$

Medium porosity varies with the type of material. For instance, clean sand and gravel commonly have porosity values between 0.30 and 0.45 (Kadlec & Wallace, 2009), while stones have porosities in the order of 0.50–0.60 and plastic support media have much higher values, in the order of 0.95–0.98 (Chernicharo & Bressani, 2019). You should find the porosity of the medium you are investigating – there are several experimental procedures for this which are outside the scope of this book. Also note that, with the passing of time in an operating unit, porosity may decrease because of accumulation of material around the medium.

For instance, a tank with 100 m^3 filled with a medium with porosity equal to 0.40 will have a liquid volume of $0.40 \times 100\ m^3 = 40\ m^3$, and the volume occupied by the medium will be $(1 - 0.40) \times 100\ m^3 = 60\ m^3$. If this tank receives a flow of 10 m^3/d, the theoretical HRT will be, according to Equation 13.6, $t = (40\ m^3)/(10\ m^3/d) = 4.0$ d.

In your report or scientific publication, you should clearly specify the characteristics of your support medium, including their **dimensions**, **porosity** and, if possible, their **specific surface area** (m^2 of medium surface area per m^3 of reactor volume). The dimensions of the support material should be

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

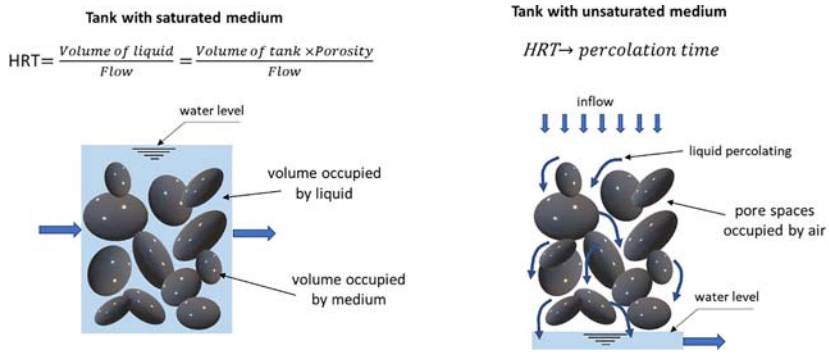


Figure 13.5 Schematic representation of a control volume in a saturated tank (left) and an unsaturated tank (right), both filled with support material. In the case of saturated media, HRT is equal to the volume occupied by liquid divided by flow. In the case of unsaturated media, there is no relationship between HRT and volume, and the concept used is that of percolation time or passage time.

specified in an unequivocal way. For instance, if you are dealing with sand or gravel, it is not sufficient to say, for instance, that '*the gravel had dimensions ranging from 10 to 20 mm*'. There are formal conventions for reporting this information, such as the diameters $\boldsymbol{d_{10}}$ and $\boldsymbol{d_{60}}$ and the ratio $\boldsymbol{d_{60}/d_{10}}$ (consult appropriate textbooks on water and wastewater treatment or material sciences for more information about these conventions).

Advanced

Now we need to analyse the specific case in which the **medium is hydraulically unsaturated**, that is, the void spaces are predominantly occupied by air. This is the case, for instance, of trickling filters, intermittently fed filters, and pulse-fed vertical flow wetlands. The inflow comes from the top surface, and the liquid simply percolates downwards, towards the bottom. Since the liquid simply percolates and does not fill the pore spaces, so the medium is unsaturated, meaning that the pores are not occupied by liquid, but rather by air. In this case, **HRT cannot be calculated as $\boldsymbol{V/Q}$**, because V is not the volume of the liquid, and we do not use the concept of HRT, but rather hydraulic **percolation time,** hydraulic **passage time,** or hydraulic **travelling time**. Typical passage times for these systems may only be on the order of minutes, but treatment still takes place because solids are retained by sorption onto the biofilm that grows on the support medium. Figure 13.5 illustrates the comparison between saturated and unsaturated media.

13.2.5 Hydraulic retention time in tanks operated in batch mode

Advanced

S. 13.2.1

S. 13.2.2

S. 13.2.3

S. 13.2.4

The situations described in Sections 13.2.1–13.2.4 are mainly for tanks or reactors that receive inflow on a continuous basis. However, in water and wastewater treatment practice, some reactors are operated in **batch mode**. One example is a sequencing batch reactor, an operational variant of the activated sludge process. At the beginning of the cycle, the tank is empty, the inlet is open, the outlet is closed, and the tank starts to fill. When it reaches the desired level, the inlet closes, and the tank enters into the reaction mode (no input, no output). Depending on the process, this can take hours or days. During this period, the concentration of the constituent of interest decreases from C_0 until it reaches the desired value (C_n), after a time of t_n (number of days or hours). After this, the outlet is opened and the treated effluent is emptied from the tank. In this case, the **HRT is the time dedicated to reaction** (filling and emptying can be optionally included), and thus, **HRT is not calculated by $\boldsymbol{V/Q}$**. For instance, if the removal of the constituent requires 10 days, the tank will be closed for 10 days, and so we can say that the liquid was retained for 10 days. After this cycle, the tank enters a new cycle (cycle 2, 3, …, n), all of them with similar characteristics. Figure 13.6 illustrates this operational mode.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

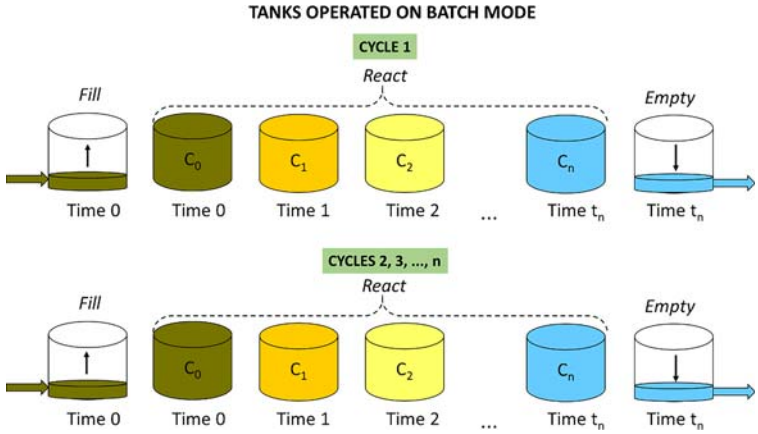


Figure 13.6 Hydraulic retention time in a tank operated in batch mode.

Now consider a different situation, distinct from the example of the batch operation described above. The purpose of the following example is to consider the **time required for filling an empty tank and for emptying a full tank**. Imagine that a tank with volume V is completely empty. If it receives a constant flow Q, the time taken to fill the tank will be equal to $t_{\text{filling}} = V/Q$. Conversely, if the tank is full, with a volume V, and the outlet withdraws a constant flow Q, the time for emptying the tank will also be $t_{\text{emptying}} = V/Q$. For instance, a tank with a volume of 100 m^3 is empty, and starts receiving a constant inflow of 20 m^3/h. The time taken to fill the tank will be $t_{\text{filling}} = (100\ m^3)/(20\ m^3/h) = 5.0$ h. If this same tank is now emptied with a constant outflow of 25 m^3/h, the time taken to empty will be $t_{\text{emptying}} = (100\ m^3)/(25\ m^3/h) = 4.0$ h. These concepts are illustrated in Figure 13.7.

13.2.6 Actual mean hydraulic retention time and departures from the theoretical behaviour

Advanced

In the previous sections, we presented formulations used to calculate the theoretical mean HRT. However, as mentioned in the previous sections, the actual mean HRT may be different from the theoretical mean HRT. Nevertheless, the **theoretical** HRT is a useful tool for **design** purposes.

But when it comes to real operational life, such as an existing treatment unit you are investigating, it is reasonable to assume that the flow behaviour will show departures from the theoretical behaviour. Therefore, you can imagine that the **actual** (real) HRT will be useful for your **diagnostic** studies.

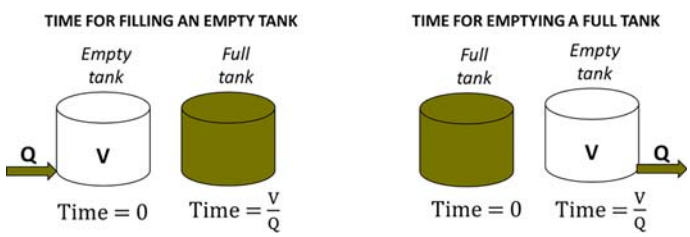


Figure 13.7 Time for filling an empty tank and time for emptying a full tank. The tank volume is equal to *V*. The flow is constant and is equal to *Q*.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

A complicating element is that it is not an easy task to estimate the actual HRT of a treatment unit. One of the tools we use to estimate the actual HRT is using **tracer tests**. Tracer tests involve the addition of an inert *tracer* (*chemical, radioactive, fluorescent, or another inert material*) at the inlet of the reactor and then measuring the distribution of concentrations with respect to time in the outlet. This task is laborious because it involves collecting and analysing samples or measuring effluent concentrations using sensors during a period of approximately three times the theoretical HRT. However, it is the best way to estimate the residence time distribution and the **mean HRT**, which is typically assumed to be the **actual** retention time. If you want to go in depth on the understanding of the behaviour of the treatment unit you are studying, you are *highly incentivized to complete a tracer test*. The main results that can be derived from tracer tests are, amongst others, the following:

- **Mean hydraulic retention time** (actual HRT)
- **Volumetric efficiency** (ratio of the mean HRT and the theoretical HRT, which is equivalent to the ratio between the useful volume and the total tank volume)
- **Dispersion number *d*** for the dispersed flow hydraulic model (see Chapter 14)
- Equivalent **number of apparent tanks in series (NTIS)** for the tanks-in-series hydraulic model (see Chapter 14)

C. 14

The description of tracer studies is outside the scope of our book. However, this topic is well covered in some treatment plant books (e.g., Teefy, 1996; Kadlec & Wallace, 2009; Metcalf & Eddy, 2014) and in chemical reaction engineering textbooks, with emphasis to the classic book by Levenspiel (1999).

We will now analyse two major factors that are responsible for departures from the theoretically expected hydraulic behaviour: **dead zones** and **short-circuiting**.

(a) **Dead zones**

Dead zones are zones in the reactor which are not flushed by flowing liquid. As a result, they do not participate in the main conversion and removal mechanisms that take place in the remaining parts of the tank. The opposite of a dead zone is an active region, also called useful (or net) volumes. Thus, the total tank volume is

$$\text{Total tank volume} = \text{Useful volume} + \text{Dead volume} \tag{13.7}$$

The actual HRT, instead of being a result of the total tank volume, is now mainly dictated by the useful volume, so that

$$\text{Mean HRT (d)} = \frac{\text{Useful volume (m}^3\text{)}}{\text{Flow (m}^3\text{/d)}} \tag{13.8}$$

Since the useful volume may be lower than the total volume due to the presence of dead zones, the actual mean HRT will be lower than the expected theoretical HRT, and treatment efficiency may be, consequently, lower. The ratio between both HRT values is called the **volumetric efficiency** of the tank:

$$\begin{aligned}\text{Volumetric efficiency} &= \frac{\text{Useful volume (m}^3\text{)}}{\text{Total volume (m}^3\text{)}} \\ &= \frac{\text{Actual HRT (d)}}{\text{Theoretical HRT (d)}}\end{aligned} \tag{13.9}$$

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Possible **causes for the occurrence of dead zones** are illustrated in Figure 13.8. The top illustration shows a frequent situation, in which parts of the tank or reactor (typically those situated in the corners) become dead zones due to *uneven inflow distribution* or *uneven outflow collection* at the inlet and outlet zones. The bottom-left figure shows a stabilization pond, in which a large portion of the total volume has been occupied by *sludge* (sediments), thus reducing the useful liquid volume. The bottom-right illustration depicts a horizontal subsurface-flow constructed wetland in which the zone close to the inlet suffers from *clogging* (pore spaces occupied by solids), and thus the liquid does not flow through it, turning it into a dead zone.

(b) **Hydraulic short-circuiting**

Hydraulic short-circuiting, or channelling, takes place when a fraction of the liquid follows a preferential flow path through the treatment unit, much faster than the ordinary flow paths.

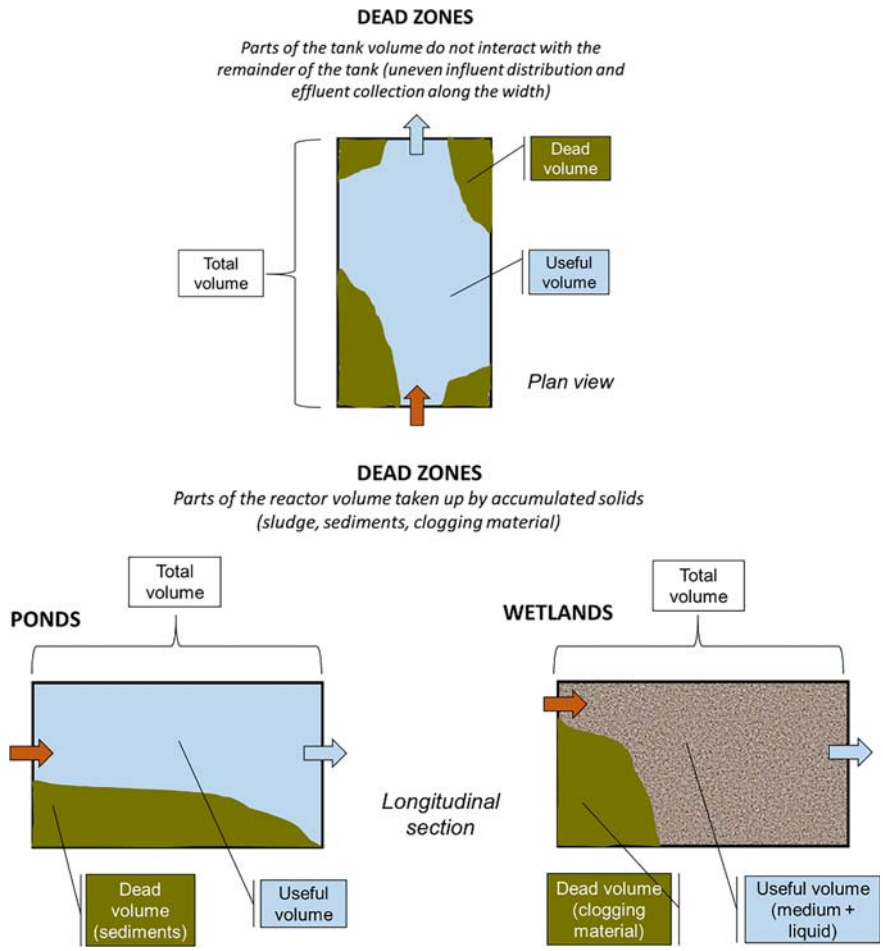


Figure 13.8 Possible causes for the presence of dead zones in tanks or reactors: in all cases, the useful volume is less than the total volume, and thus the actual HRT is lower than the theoretical HRT.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

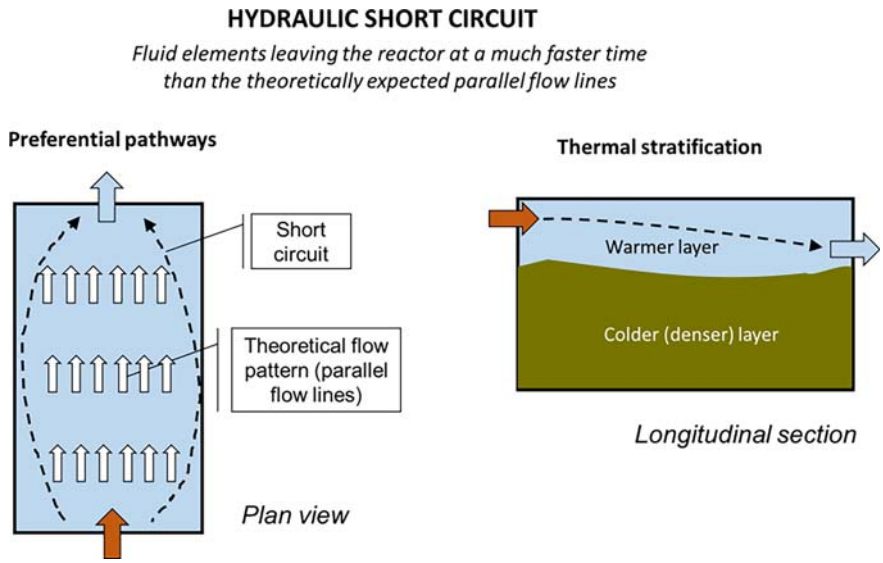


Figure 13.9 Possible causes for the occurrence of hydraulic short circuits in tanks or reactors: in all cases, the short-circuited liquid flows much faster that the remainder of the liquid, and its HRT may be much smaller than the overall mean HRT of the tank.

Figure 13.9 shows the possible occurrences of short-circuiting. The left side of the figure shows internal currents flowing close to the tank walls at a much faster velocity compared with the theoretical parallel streamlines. This may be a result of *inadequate flow distribution* or even the influence of *winds* (in this case, the fast flow would be occurring only on one side of the tank). The right side of the figure illustrates the occurrence of *thermal stratification* in the tank. As part of this phenomenon, the bottom layer (colder and denser) does not interact with the liquid, which flows quickly through the upper layer (warm and less dense).

Let us analyse the influence of hydraulic short-circuiting with two hypothetical examples. In the first one, we have a reactor that is expected to remove 90% of the influent **BOD** ($C_{in} = 300$ mg/L). As a result, the effluent concentration would then be expected to be $C_{out} = (1 - 0.90) \times 300 = \mathbf{30\ mg/L}$. A hydraulic short-circuit is identified, responsible for diverting 1% of the flow. In the short-circuited portion, due to the lower hydraulic retention time, the removal efficiency was only 50% (the remainder 99% still keep the 90% removal). The final effluent concentration will now be:

$$C_{out} = 300\left[\frac{1 \times (1 - 0.50) + 99 \times (1 - 0.90)}{1 + 99}\right] = 300 \times 0.104 = \mathbf{31.2\ mg/L}$$

Therefore, we can conclude that not much deterioration in terms of BOD removal resulted from this 1% of short-circuiting (effluent concentration increased from 30 to 31 mg/L).

Now let us consider the impact in terms of ***E. coli*** removal, in a similar example. We have a disinfection unit, designed specifically for coliform removal, that has an expected efficiency of 99.999% (5 log-units reduction). The influent *E. coli* concentration is $C_{in} = 1.00 \times 10^7$ MPN/100 mL. As a result, the effluent concentration would then be expected to be $C_{out} = (1 - 0.99999) \times 1.00 \times 10^7 = 1.00 \times 10^2 = \mathbf{100\ MPN/100\ mL}$, which is a low value. However, a

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

hydraulic short-circuit is identified, responsible for diverting 1% of the flow. In the short-circuited portion, due to the lower hydraulic retention time, the *E. coli* removal efficiency was only 90% (the other 99% of the liquid still provides the 99.999% removal). The final effluent concentration will now be:

$$C_{out} = 1.00 \times 10^7 \left[\frac{1 \times (1 - 0.90) + 99 \times (1 - 0.99999)}{1 + 99}\right]$$

$$= 1.00 \times 10^7 \times 0.0010099 = 1.0099 \times 10^4 = \mathbf{10{,}099\ MPN/100\ mL}$$

Now, the difference is very large. Just 1% of a poorly treated fraction made the effluent *E. coli* concentration raise from 100 to >10,000 MPN/100 mL, that is, a value more than 100 times greater than the expected one. This discrepancy occurred because, when we study microorganisms in water systems, we typically deal with log-scales, having very high concentrations and needing very high reduction efficiencies. Thus, small imperfections in the hydraulics of our reactor may lead to much less efficient results in terms of performance.

13.3 VOLUMETRIC HYDRAULIC LOADING RATE

Basic

Volumetric hydraulic loading rate (HLR_V) corresponds to the influent flow (m^3/min, m^3/h, m^3/d) that is applied to a unit tank volume (1 m^3). It is calculated by the inflow (Q) divided by the tank volume (V), as given by Equation 13.10 and illustrated in Figure 13.10.

$$HLR_V = \frac{Q}{V} \tag{13.10}$$

where

HLR_V = volumetric hydraulic loading rate [(m^3/min)/m^3, (m^3/h)/m^3, (m^3/d)/m^3]
V = volume of liquid in the tank (m^3)
Q = inflow (m^3/min, m^3/h, m^3/d)

Similar to what was mentioned for the hydraulic retention time (Figure 13.1), the concept holds true *regardless of the flow direction* (*horizontal, vertical downflow, vertical upflow*). This loading rate is important when the *removal processes are dependent on flow and are more influenced by volume than the surface area of the tank*.

Note, from Equation 13.10, that **HLR_V is the inverse of the mean theoretical hydraulic retention time (HRT)**, as given by Equation 13.3. Therefore, **$HLR_V = 1/HRT$**. The units of HLR_V are (m^3/min)/m^3,

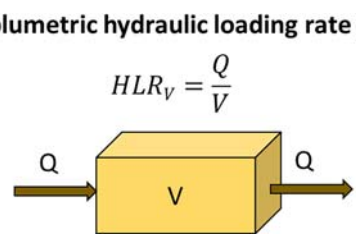


Figure 13.10 The concept of volumetric HLR.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

$(m^3/h)/m^3$, and $(m^3/d)/m^3$. If we cut out m^3 from the numerator and denominator, we are left with min^{-1}, h^{-1}, and d^{-1}, showing again the inverse relationship with time. But this inverse relationship can be only applied in the case with tanks **without support media**, in which the full tank volume is occupied by liquid.

HLR_V can also be interpreted as the **number of volume renewals per unit time**. For instance, for a tank with an average HLR_V of 4 $(m^3/d)/m^3$ or 4 d^{-1}, on average, the entire liquid contents are renewed 4 times per day. In other words, each day the tank receives an input volume that corresponds to 4 times the tank volume. Considering the concept that $HLR_V = 1/HRT$, the associated HRT is equal to $1/(4\ d^{-1}) = 0.25\ d = 6\ h$. Of course, the higher the renewal rate, the lower the hydraulic retention time.

In the case of **tanks with support medium**, the volumetric HLR is not the reciprocal of the hydraulic retention time ($HLR_V \neq 1/HRT$), because part of the tank volume is taken up by the medium. But, still in this case, **you can use the concept of HLR_V** for design purposes or for performance evaluation, because the literature reports values of HLR_V in terms of the total tank volume, and not the liquid volume.

13.4 SURFACE HYDRAULIC LOADING RATE

Basic

The surface hydraulic loading rate (HLR_S) corresponds to the influent flow (m^3/min, m^3/h, m^3/d, L/d) that is applied in a unit tank surface area (1 m^2). It is calculated by the inflow (Q) divided by the tank surface area (A), as given by Equation 13.11 and illustrated in Figure 13.11.

$$HLR_S = \frac{Q}{A} \tag{13.11}$$

where

HLR_S = surface hydraulic loading rate $[(m^3/d)/m^2, (m^3/h)/m^2, (L/d)/m^2, (L/h)/m^2, m/d, mm/d]$
A = surface area of the tank (m^2)
Q = inflow $(m^3/min, m^3/h, m^3/d, L/d)$

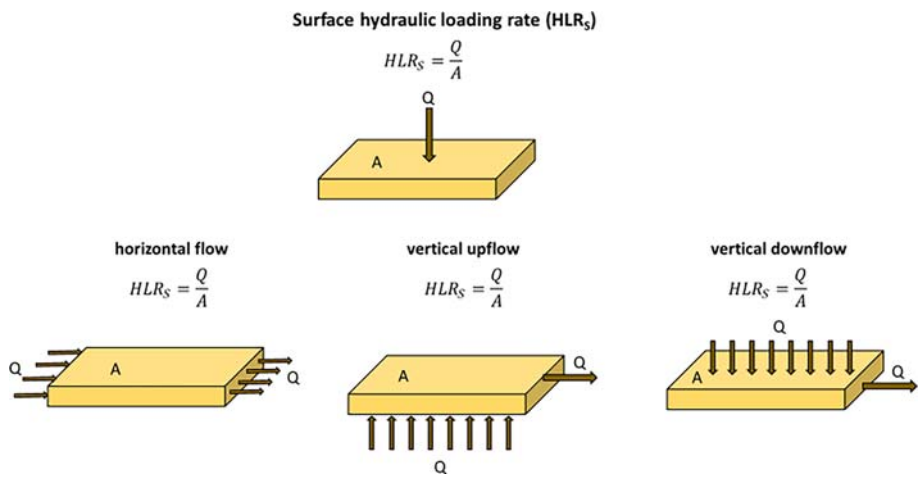


Figure 13.11 The concept of surface hydraulic loading rate (HLR_S). HLR_S can be used for tanks receiving horizontal flow, vertical upflow and vertical downflow, and calculations are the same.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

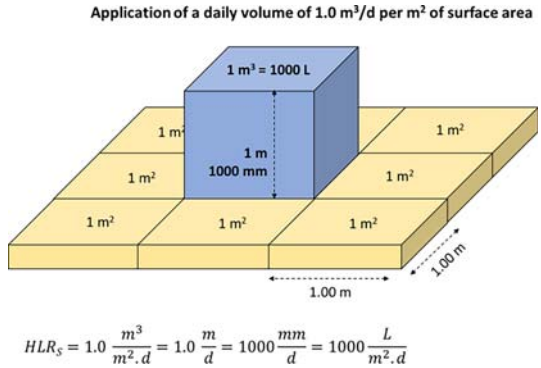


Figure 13.12 Schematic visualization of the application of a surface HLR of 1.0 $(m^3/d)/m^2$ and the different ways in which it can be physically interpreted.

You can see from Figure 13.11 that the same concept holds true *regardless of the flow direction* (*horizontal, vertical downflow, vertical upflow*). This loading rate is important when the *removal processes are dependent on flow and are more influenced by surface area than volume of the tank.*

When analysing the values of the surface HLR, you should try to conceptualize their physical meaning. Figure 13.12 gives an example, for a HLR_S of $(1.0\ m^3/d)/m^2$. The physical meaning is that each 1.0 m^2 of the surface area will receive 1.0 m^3/d, which is equivalent to 1000 L/d. This can also be understood as the application of a liquid height of 1.0 m each day or 1000 mm/d.

Advanced

The concept of surface hydraulic loading rate is used for several different types of treatment units, but it is the main design parameter for **sedimentation tanks**. In this case, *HLR_S has a direct equivalence with the settling velocity* of the particles or solids to be removed in the sedimentation tank. Settling velocity has a dimension of distance (height) over time (m/min, m/h, m/d), which corresponds to the same dimensions of HLR_S. For instance, a grit chamber, which is designed to remove sand particles, is designed according to this principle. If we are seeking to remove sand particles with a settling velocity greater than 1000 m/d, the design of the grit chamber can be done on the basis of a HLR_S of 1000 $(m^3/d)/m^2$. Knowing the flow and adopting this value of HLR_S, you can calculate the required surface area. Sedimentation tanks in water and wastewater treatment practice deal with solids or suspensions with much lower settling velocities, in the order of 0.5–1.0 m/h or 12–24 m/d. This means that the HLR_S values for the calculation of the required area in the design of clarifiers and sedimentation tanks are around 12–24 $(m^3/d)/m^2$. For existing sedimentation tanks, you can divide the inflow by the surface area, obtain the value of HLR_S and compare it with the design value or with recommendations from the literature.

Recognizing the importance of the surface area for the performance of some treatment units, such as sedimentation tanks, we must consider the *hydraulic behaviour of the tank*, which will influence its ability to make the most use out of the available surface area. Here, the occurrence of dead zones, with a reduction of the useful area, may lead to a deterioration of performance. See Section 13.2.6 for the concepts of total volume, dead volume, and useful volume. In the case here, we are interested in **total area**, **dead area,** and **useful area**, but the concept is the same, since we can convert volume into area dividing by the tank depth.

S. 13.2.6

You must take caution when applying the concept of **HLR_S for treatment systems that have units simultaneously in operation and resting**. For instance, the first-stage of the French vertical flow wetlands system typically has three units in parallel. They alternate with each other, so that there is always one unit in operation (feeding) and two units resting. You must specify clearly if the HLR_S you

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

are calculating uses the area of only one unit (HLR_S for the unit in operation) or the total area of the three units (HLR_S for the whole system). For instance, the system has three units in parallel, each with 30 m^2 (total surface area of $3 \times 30\ m^2 = 90\ m^2$) and receives an average flow of 10 m^3/d. The HLR_S for the unit in operation is $(10\ m^3/d)/30\ m^2 = 0.33\ (m^3/d)/m^2$ and for the total system is $(10\ m^3/d)/90\ m^2 = 0.11\ (m^3/d)/m^2$. Check with the literature what is the usual way of reporting these loading rates – for instance, for the French system of wetlands, the traditional way is reporting HLR_S for the unit in operation (Dotro *et al.*, 2017).

You can still use another way of reporting loading rates in units that operate on an alternating basis on the feed and rest mode by **expressing the loads per year**. This avoids confusion. In the example shown above, the inflow is 10 m^3/d. Since, on the long run, each unit operates for 1/3 of the time and rests for 2/3 of the time (there is always one unit on the feed mode and two units on the rest mode), the total yearly flow per unit is $(10\ m^3/d) \times (365\ d/year) \div (3\ units) = 1{,}216.7\ m^3/year$ per unit. The total system, with its three units in parallel, receives, per year, a total flow of $(10\ m^3/d) \times (365\ d/year) = 3650\ m^3/year$, which is exactly three times the flow received in each unit. Each unit receives, per year, a HLR_S of $(1{,}216.7\ m^3/year)/30\ m^2 = 40.6\ (m^3/year)/m^2$. If you make the calculation for the whole system, with its three units, and express it on a yearly basis, you arrive at exactly the same value of $HLR_S = (3{,}650\ m^3/year)/90\ m^2 = 40.6\ (m^3/year)/m^2$. This shows the convenience of reporting loading rates per year, when the units operate on an alternated basis.

13.5 VOLUMETRIC MASS LOADING RATE

Basic

Volumetric mass loading rate (MLR_V) corresponds to the influent load (kg/d, g/d, kg/h, g/h, kg/year) that is applied to a unit tank volume (1 m^3). It is calculated by the inflow (Q) times the input concentration (C) divided by the tank volume (V), as given by Equation 13.12 and illustrated in Figure 13.13.

$$MLR_V = \frac{Q \cdot C}{V} \tag{13.12}$$

where

MLR_V = volumetric mass loading rate [$(kg/d)/m^3$, $(kg/h)/m^3$, $(g/d)/m^3$, $(g/h)/m^3$, $(kg/year)/m^3$]
V = volume of liquid in the tank (m^3)
Q = input flow (m^3/d, m^3/h, $m^3/year$)
C = input concentration (g/m^3)

As with the volumetric hydraulic loading rate (HLR_V), the concept of volumetric mass loading rate (MLR_V) is also *independent of the flow direction* (*horizontal, vertical downflow, vertical upflow*). This

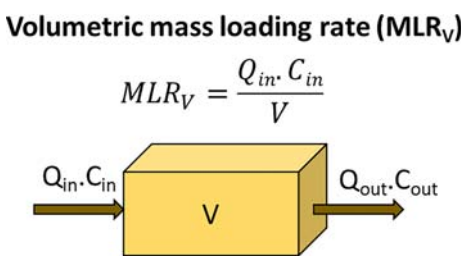


Figure 13.13 The concept of applied volumetric MLR.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

loading rate is important when the *removal processes are dependent on mass load and are more influenced by volume than the surface area of the tank or reactor.*

In the case of **tanks with a support medium**, you should calculate the volumetric MLR based on the *total tank volume*, and not on the volume occupied by liquid (pore spaces).

A special case of MLR_V is for the application of a load of **organic matter**, expressed as **BOD** or **COD**. In this case, the equivalent expression, frequently used, is $\mathbf{OLR_V}$ (**volumetric organic loading rate**). Influent BOD and COD concentrations are expressed in the usual form of g/m^3 (equal to mg/L) and the load is also calculated, as usual, by the product of flow × concentration.

13.6 SURFACE MASS LOADING RATE

Basic

Surface mass loading rate (MLR_S) corresponds to the influent load (kg/d, g/d, kg/h, g/h, kg/year) that is applied in a unit tank surface area (1 m^2 or ha). It is calculated by the inflow (Q) times the input concentration (C) divided by the tank surface area (A), as given by Equation 13.13 and illustrated in Figure 13.14. Load per unit area is also called **mass flux**.

$$MLR_S = \frac{Q \cdot C}{A} \tag{13.13}$$

where

MLR_S = surface mass loading rate [$(kg/d)/m^2$, (kg/d)/ha, $(kg/h)/m^2$, $(g/d)/m^2$, $(g/h)/m^2$, $(kg/year)/m^2$]
A = surface area of the tank (m^2, ha)
Q = input flow (m^3/d, m^3/h, $m^3/year$)
C = input concentration (g/m^3)

As shown in Figure 13.11 for the surface hydraulic loading rate, the same concept holds true here, for the surface mass loading rate, in that it is *independent of the flow direction* (*horizontal, vertical downflow, vertical upflow*). This loading rate is important when the *removal processes are dependent on mass load and are more influenced by the surface area than the volume of the tank.*

Loading rates for most treatment units are frequently reported as $\mathbf{(kg/d)/m^2}$. However, some systems that receive low values of loading rates are sometimes described in terms of $\mathbf{(g/d)/m^2}$, as is frequently done, for instance, for treatment wetlands. Another frequently used unit, in this case, for stabilization

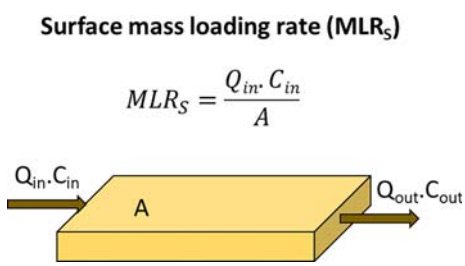


Figure 13.14 The concept of applied surface MLR.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

ponds, is **(kg/d)/ha**, considering that the required surface areas are large, and hectare is considered a suitable unit for area. For the conversion between units, we have:

$$1\,\frac{\text{g}}{\text{m}^2\cdot\text{d}} = 10\,\frac{\text{kg}}{\text{ha}\cdot\text{d}} \tag{13.14}$$

A special case of MLR_S is for **organic matter**, expressed as **BOD** or **COD**. In this case, the equivalent expression, frequently used, is $\textbf{OLR}_\textbf{S}$ (**surface organic loading rate**). Influent BOD and COD concentrations are expressed in the usual form of g/m^3 (equal to mg/L), and the load is also calculated, as usual, by the product of flow × concentration.

The concept of surface mass loading rate, or flux, is also used with **SS** in the case of secondary sedimentation tanks (activated sludge) and other treatment units, and ammonia-nitrogen or **TKN** (total Kjeldahl nitrogen) for units that aim at nitrification.

S. 13.4

S. 13.2.6

Similar to the comment we made in Section 13.4, we should aim to have *good hydraulic behaviour* in our tank, using as much as possible the available surface area. The occurrence of dead zones, with a reduction of the useful area, may lead to a deterioration of the tank performance. See Section 13.2.6 for the concepts of total volume, dead volume, and useful volume. In the case of surface mass loading rates, we are interested in **total area**, **dead area,** and **useful area**, but the concept remains the same as when we previously dealt with volumes, since we can convert volume into area dividing by the tank depth.

Advanced

S. 13.4

You must take care when applying the concept of $\textbf{MLR}_\textbf{S}$ **for treatment systems that have units simultaneously in operation (feeding) and units resting**. The concept here is similar to the one described for the surface HLR (see Section 13.4 for a detailed discussion on this matter).

13.7 SPECIFIC SURFACE MASS LOADING RATE

Advanced

The specific surface MLR (specific MLR_S) corresponds to the influent load (kg/d, g/d, kg/h, g/h, kg/year) that is applied over the entire surface area of all the elements composing the medium (m^2) in a tank with support material. It is calculated by the inflow (Q) times the input concentration (C) divided by the total media surface area (A_m), as given by Equation 13.15 and illustrated in Figure 13.15.

$$\boxed{\text{Specific MLR}_\text{S} = \frac{Q\cdot C}{A_\text{m}}} \tag{13.15}$$

where

$\text{Specific MLR}_\text{V}$ = specific surface mass loading rate [$(\text{kg/d})/\text{m}^2$ of medium, $(\text{kg/h})/\text{m}^2$ of medium, $(\text{g/d})/\text{m}^2$ of medium, $(\text{g/h})/\text{m}^2$ of medium

A_m = entire surface area of all the elements composing the medium (m^2)

Q = input flow (m^3/d, m^3/h)

C = input concentration (g/m^3)

This concept has been applied to **trickling filters** with respect to organic matter conversion and nitrification (Metcalf & Eddy, 2014). In this application, for instance, stones have a specific surface area between 50 and 70 m^2/m^3 (Chernicharo & Bressani, 2019). For each load unit (e.g., 1 g/d) applied to each m^3 of the reactor volume, the specific MLR_S will be 1 g/d distributed between 50 and 70 m^2 of medium surface, leading to values between $1/50 = 0.020$ $(\text{g/d})/\text{m}^2$ of medium and $1/70 = 0.014$ $(\text{g/d})/\text{m}^2$

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

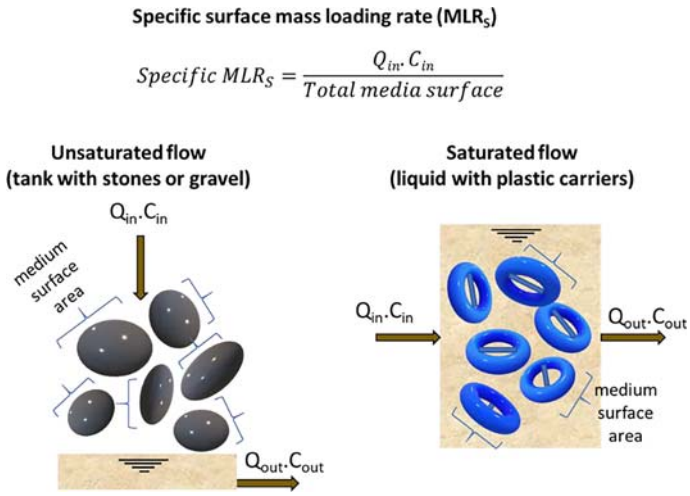


Figure 13.15 The concept of applied specific surface MLR. Left: example of a trickling filter with stones and unsaturated medium. Right: example of IFAS or MBBR activated sludge with plastic carriers inside the liquid in the reactor.

of medium. If we consider plastic media, which have a much larger surface area (between 80 and 98 m^2/m^3) (Chernicharo & Bressani, 2019), for each load unit (e.g., 1 g/d) applied to each m^3 of reactor volume, the specific MLR_S will be between $1/80 = 0.013$ and $1/98 = 0.010$ (g/d) per m^2 of medium surface. Therefore, for the same applied load, plastic media, due to their much higher specific surface area, will require a much smaller reactor volume. For existing trickling filters, plastic media units will be able to receive much higher loads compared with coarse material, such as stones.

Other applications are for variants of the activated sludge process, such as Integrated Fixed Film Activated Sludge (**IFAS**) and for Moving Bed Biofilm Reactor (**MBBR**), in which plastic carriers are used inside the biological reactor. Plastic carriers may have specific surface areas between 500 and 1000 m^2/m^3 or higher, depending on the medium type and manufacturer.

13.8 FOOD-TO-MICRORGANISM RATIO (*F/M*)

Advanced

The concept of the **F**ood-to-**M**icroorganism ratio, or F/M ratio, also represents a type of volumetric mass loading rate (MLR_V). In MLR_V, the input load is applied to a unit tank volume, irrespective of its biomass concentration. In the case of the F/M ratio, the **input load is applied to a unit mass of biomass**. The input load is the '*food*,' while the mass of biomass represents the microorganisms (that 'eat the food'). The F/M ratio is also commonly called the **sludge load**. The F/M ratio is discussed in virtually all books on the activated sludge process, and details can be found in the open-access companion books to this text (von Sperling & Chernicharo, 2005; von Sperling, 2007).

For biological wastewater treatment, the mass of biomass is more important than reactor volume, because it is the biomass which will be responsible for the conversion processes. In many biological reactors, it is very difficult to compute the biomass present in the reactor. However, in processes such as activated sludge, biomass may be computed by assuming that volatile suspended solids (VSS or X_v), also called mixed liquor volatile suspended solids (MLVSS) are a good representation of the biological mass present in the reactor, and that the product of the **biomass concentration** (**VSS, X_V, MLVSS**) times the **reactor**

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

volume (***V***) will give an estimate of the **mass of biomass**. In some cases, SS, X, or MLSS (mixed liquor suspended solids) are used for the representation of biomass – they are easier to measure, but are not such a good representation of the biological solids responsible for the treatment as the volatile suspended solids.

With regard to what is meant by 'food,' it usually is the load of organic matter applied to the reactor, expressed in terms of BOD or COD.

An illustration of the concept of the F/M ratio is presented in Figure 13.16, and its calculation is shown in Equation 13.16.

$$\boxed{\frac{F}{M} = \frac{Q_{in} \cdot C_{in}}{V \cdot \text{VSS}}} \tag{13.16}$$

where

F/M = food-to-microorganism ratio [gBOD/d)/gVSS or (gCOD/d)/gVSS)]
Q_{in} = influent flow to the reactor (m^3/d)
C_{in} = influent substrate concentration to the reactor; usually BOD or COD (g/m^3)
V = reactor volume (m^3)
VSS = volatile suspended solids concentration in the reactor, also called mixed liquor volatile suspended solids (g/m^3)

In Equation 13.16, Q_{in}/V corresponds to the reciprocal of the hydraulic retention time ($1/t$). Thus, F/M can also be represented by the simplified form in Equation 13.17, as a direct function of the hydraulic retention time (t).

$$\frac{F}{M} = \frac{C_{in}}{t \cdot \text{VSS}} \tag{13.17}$$

High F/M values are usually representative of **high loaded systems**, while low F/M values are associated with **low loaded systems**. There are several implications of the F/M ratio on the activated sludge process, associated not only with the required reactor volume, but also with the removal efficiency, biomass production, sludge digestion, oxygen consumption, and others.

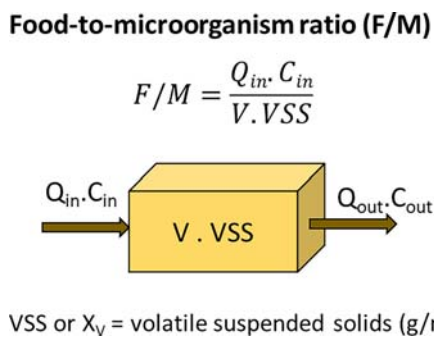


Figure 13.16 Concept of F/M ratio or sludge load. The numerator is usually the BOD or COD load applied to the reactor, and the denominator is the mass of biomass in the reactor, usually expressed in terms of volatile suspended solids.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

13.9 SLUDGE AGE

Advanced

Sludge age, also called the **Solids Retention Time (SRT)** or the **Mean Cell Residence Time (MCRT)**, is frequently represented using the symbol θ_c (theta representing time and 'c' representing 'cell'). Although the calculations are not so straightforward as others presented in this chapter, the concept is relatively simple. The **sludge age represents the average time that the biological solids (biomass) remain in the system**. In most cases, what is meant by the system is the biological reactor. Sludge age is the most important design and operational variable of the activated sludge process, but its concept also applies to other treatment systems, such as upflow anaerobic sludge blanket (UASB) reactors. There are several implications of the sludge age on the activated sludge process, associated with the required reactor volume, removal efficiency, biomass production, sludge digestion, oxygen consumption, and others. Sludge age is not exactly a loading rate, but is presented here because it associates loads with the reactor volume. Sludge age is covered in virtually all books on the activated sludge process, and details can be found in the open-access companion books to this text (von Sperling & Chernicharo, 2005; von Sperling, 2007).

In order to start understanding sludge age, or the average time spent by the biological solids in the system, you could remember the concept of HRT, or the average time spent by the liquid in the system. From Equations 13.1 and 13.2, we have:

$$\text{HRT} = \frac{\text{volume of liquid in the tank}}{\text{volume of liquid } \textit{entering} \text{ the tank per unit time}} = \frac{\text{volume of liquid in the tank}}{\text{volume of liquid } \textit{leaving} \text{ the tank per unit time}} \tag{13.18}$$

In Equation 13.18 it is assumed that the volumes of liquid entering and leaving the system are the same, in other words, the inflow is equal to the outflow (denominators of Equation 13.18) in the steady state. A similar relationship can be made for the biological solids (biomass) in the system:

$$\text{Sludge age} = \frac{\text{mass of solids in the system}}{\text{mass of solids } \textit{produced} \text{ in the system per unit time}} = \frac{\text{mass of solids in the system}}{\text{mass of solids } \textit{removed} \text{ from the system per unit time}} \tag{13.19}$$

In Equation 13.19, instead of liquid (as in HRT), we mention biological solids. These are produced in the system by the reproduction of the microorganisms as a result of the conversion of the organic matter supplied by the influent wastewater. We do not measure the microorganisms as such, but we use other proxy measures to represent them, such as VSS (see Section 13.8). In the *steady state*, we assume that the biological solids being produced are compensated by their removal at an equal rate, so that the *load of solids production is equal to the load of solids removal* (denominators of Equation 13.19).

S. 13.8

We will now analyse two situations:

- **System without solids retention** (no sludge recycle, such as aerated lagoons)
- **System with solids retention** (with sludge recycle, such as in the activated sludge process)

(a) **System without solids retention**

This is the case in which the system is comprised of a reactor and has no means of retaining the biomass, such as a sludge recirculation line (Figure 13.17). An example could be an aerated lagoon.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

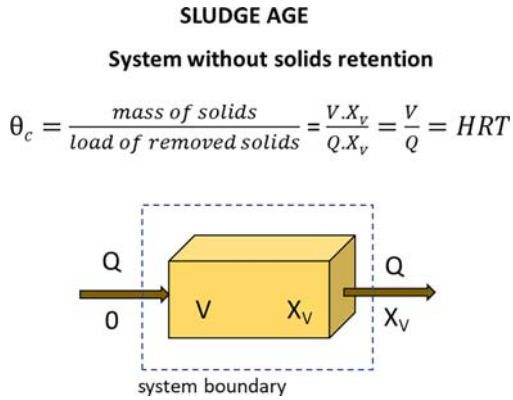


Figure 13.17 Schematic representation of the concept of sludge age in a system without solids retention.

Since it is difficult to calculate or measure solids production, the estimation of the sludge age is made based on the solids removed from the system. According to Equation 13.19, the mass of solids in the system (numerator of the equation) is given by

$$\text{Mass of solids in system} = V \cdot X_V \tag{13.20}$$

The denominator of Equation 13.19 is the load of solids removed from the system (reactor), which is given by Equation 13.21, since this is the only route for the solids to leave the system:

$$\text{Load of solids removed from system} = Q \cdot X_V \tag{13.21}$$

where
V = volume of reactor (m^3)
Q = output flow from the system (m^3/d)
X_V = VSS concentration in the reactor, representing biomass concentration (g/m^3)

Incorporating Equations 13.20 and 13.21 into Equation 13.19, we have the resulting equation for a **system without solids retention**:

$$\theta_C = \frac{V \cdot X_V}{Q \cdot X_V} = \frac{V}{Q} \tag{13.22}$$

This is an important conclusion for **systems without solids retention**: in this case, the **sludge age is equal to V/Q, and so it is the same as the HRT**. See Figure 13.17 for the schematic representation of this system. Note that the influent solids concentration was assumed as equal to zero, since the biological solids are produced inside the reactor.

(b) **System with solids retention**

Now let us analyse the case in which there are **means of retaining solids in the system**, for example, by using a **recirculation line** of sludge from the secondary clarifier to the biological reactor, as is typical for the activated sludge process. The flowsheet is now more complex, and so is the mass balance, which is represented in Figure 13.18. Without going into much detail,

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

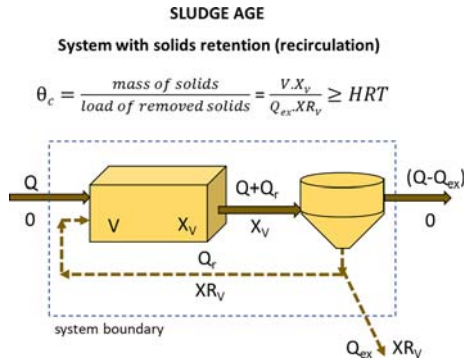


Figure 13.18 Schematic representation of the concept of sludge age in a system with solids retention (recirculation of solids from the bottom of the secondary sedimentation tank to the reactor).

the denominator of Equation 13.19 is the solids load that leaves in the line of excess sludge (waste sludge, surplus sludge):

$$\text{Load of solids removed from system} = Q_{ex} \cdot XR_V \tag{13.23}$$

As a result, the sludge age can be computed by Equation 13.24 in a system with solids retention. Again, for simplicity, biological solids in the plant influent and effluent are considered negligible, and the mass of solids inside the secondary sedimentation tank is also neglected.

$$\theta_C = \frac{V \cdot X_V}{Q_{ex} \cdot XR_V} \tag{13.24}$$

where

XR_V = volatile suspended solids concentration in the sludge return line (g/m^3)

The denominator in Equation 13.24 (system with solids retention) is much smaller than the denominator in Equation 13.22 (system without solids retention), and so $\theta_c > HRT$.

Decoupling the solids retention time from the hydraulic retention time is a very important characteristic of **high-rate systems**, as it leads to lower reactor volumes and higher removal efficiencies. The example shown above is for the **activated sludge** process, in which typical sludge age values are in the order of **days**, depending on the activated sludge variant. However, there are other treatment systems that have this capability. For instance, **UASB** reactors retain biological solids due to the fact that they settle in the upper compartment (sedimentation tank) and return by gravity to the reaction compartment, forming the sludge blanket, and leading to sludge ages that are on the order of **weeks**.

Example

EXAMPLE 13.1 CALCULATION OF APPLIED LOADING RATES

Calculate the applied loading rates on a reactor from a wastewater treatment plant, using the following data:

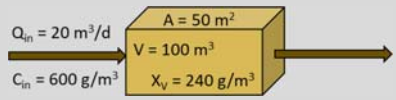
- *Volume of the reactor: V = 100 m^3*
- *Surface area of the reactor: A = 50 m^2*
- *Input flow: Q_{in} = 20 m^3/d*

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

- *Input COD concentration:* $C_{in} = 600\ g/m^3$
- *Volatile suspended solids concentration in the reactor:* $X_v = 240\ g/m^3$

Solution:

(a) Prepare the schematics of the reactor and input data



(b) Hydraulic retention time (HRT)
From Equation 13.3:

$$t = \frac{V}{Q} = \frac{100\,\text{m}^3}{20\,\text{m}^3/\text{d}} = 5.0\,\text{d}$$

(c) Volumetric hydraulic loading rate (HLR_V)
From Equation 13.10:

$$\text{HLR}_V = \frac{Q}{V} = \frac{20\,\text{m}^3/\text{d}}{100\,\text{m}^3} = 0.20\,\frac{\text{m}^3/\text{d}}{\text{m}^3} = 0.20\,\text{d}^{-1}$$

(d) Surface hydraulic loading rate (HLR_S)
From Equation 13.11:

$$\text{HLR}_S = \frac{Q}{A} = \frac{20\,\text{m}^3/\text{d}}{50\,\text{m}^2} = 0.40\,\frac{\text{m}^3}{\text{m}^2\cdot\text{d}} = 0.40\,\frac{\text{m}}{\text{d}} = 400\,\frac{\text{mm}}{\text{d}} = 400\,\frac{\text{L}}{\text{m}^2\cdot\text{d}}$$

(e) Volumetric mass loading rate (MLR_V)
From Equation 13.12:

$$\text{MLR}_V = \frac{Q\cdot C}{V} = \frac{\left(20\,\frac{\text{m}^3}{\text{d}}\right)\times\left(600\,\frac{\text{g}}{\text{m}^3}\right)}{100\,\text{m}^3} = \frac{12{,}000\,\text{g/d}}{100\,\text{m}^3} = 120\,\frac{\text{g}}{\text{m}^3\cdot\text{d}} = 0.120\,\frac{\text{kg}}{\text{m}^3\cdot\text{d}}$$

In this case, since the input constituent is COD, the MLR_V can also be called volumetric organic loading rate (OLR_V).

(f) Surface mass loading rate (MLR_S)
From Equation 13.13:

$$\text{MLR}_s = \frac{Q\cdot C}{A} = \frac{\left(20\,\frac{\text{m}^3}{\text{d}}\right)\times\left(600\,\frac{\text{g}}{\text{m}^3}\right)}{50\,\text{m}^2} = \frac{12{,}000\,\text{g/d}}{50\,\text{m}^2} = 240\,\frac{\text{g}}{\text{m}^2\cdot\text{d}} = 0.240\,\frac{\text{kg}}{\text{m}^2\cdot\text{d}}$$

In this case, since the input constituent is COD, the MLR_S can also be called surface organic loading rate (OLR_S).

(g) Food-to-microorganism ratio (F/M)
From Equation 13.16:

$$\frac{F}{M} = \frac{Q\cdot C}{V\cdot X_V} = \frac{20\,\frac{\text{m}^3}{\text{d}}\times 600\,\frac{\text{gCOD}}{\text{m}^3}}{100\,\text{m}^3\times 240\,\frac{\text{gVSS}}{\text{m}^3}} = \frac{12{,}000\,\text{gCOD/d}}{24{,}000\,\text{gVSS}} = 0.50\,\frac{\text{gCOD}}{\text{gVSS}\cdot\text{d}}$$

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

13.10 CHECK-LIST FOR YOUR REPORT

- ✓ Include only the loading rates which are typically used for the treatment system you are investigating. There is no need to calculate and express all of them if they are not relevant to your system.
- ✓ Check the consistency of all units for the loading rates (units of time, mass, volume, concentration) and make them clear in your calculations and summary tables or figures.
- ✓ In systems with support medium and **saturated flow**, make sure you use only the liquid occupied by liquid (pore spaces) when calculating the hydraulic retention time.
- ✓ In systems with support medium and **unsaturated flow**, make sure you have not used the traditional concept of HRT $= V/Q$, since, in this case, the pore spaces are occupied by air and not by liquid. In this case, you would use the concept of percolation time.
- ✓ Make it clear whether your reported HRT is the theoretical mean HRT, or the actual mean HRT, as derived from tracer tests. If the latter is the case, describe the main methods used for completing the tracer tests.
- ✓ In systems with units in parallel that alternate periods of feeding and resting, make sure you report clearly whether the loading rates you calculated apply only to the unit in operation (feeding) or to all units (feeding + resting).
- ✓ Analyse the performance of your treatment system with consideration of the applied loading rates.
- ✓ Interpret the physical meaning of the calculated loading rates and compare them with the values recommended in the literature or design guidelines for your treatment system.
- ✓ If you are comparing the performance of your treatment system with the performance of other systems, make sure you also report the loading rates used in the other systems.
- ✓ If you are analysing different experimental phases, each of them with a different applied loading rate, organize your results in summary tables and check whether you have followed the recommendations listed at the final paragraphs of Section 13.1 (Table 13.2 and comments).

S. 13.1

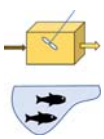
Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Chapter 14

Reaction kinetics and reactor hydraulics

In this chapter, we present an argument for why you should determine kinetic coefficients for the decay or removal of various constituents of concern in a water or wastewater treatment reactor, as these coefficients may be useful to other researchers or practitioners designing new systems or undertaking similar studies. We describe the main reaction orders (0, 1, and 2), show you how to derive them, and give an emphasis on the use of first-order reactions. We cover the determination of reaction coefficients based on batch experiments, under steady-state and dynamic conditions, and we provide some precautions about the estimation of reaction rate coefficients for continuous-flow reactors, due to the impact of hydraulic efficiency on these estimates. We demonstrate the need to characterize the hydraulics of continuous-flow reactors and provide some examples using the idealized plug-flow model, the idealized complete-mix model, the plug-flow model with dispersion, and the apparent tanks-in-series model for representing reactor hydraulics.



The contents in this chapter are applicable to both treatment plant monitoring and water quality monitoring. As the chapter is structured, most of the applications are for treatment plants reactors. However, we can also consider that water bodies are reactors, and several concepts presented here will also be applicable.

CHAPTER CONTENTS

14.1 Introduction **531**
14.2 Reaction Order **533**
14.3 Experimental Determination of the Reaction Order and Kinetic Coefficient in Batch Reactors . . . **541**
14.4 Idealized Flow Regimens in Continuous-Flow Reactors **556**
14.5 Plug-Flow with Dispersion and Apparent Tanks-In-Series Models **569**
14.6 Check-List for Your Report **592**

© 2020 The Authors. This is an Open Access book chapter distributed under the terms of the Creative Commons Attribution Licence (CC BY-NC-ND 4.0), which permits copying and redistribution for non-commercial purposes with no derivatives, provided the original work is properly cited (https://creativecommons.org/licenses/by-nc-nd/4.0/). This does not affect the rights licensed or assigned from any third party in this book. The chapter is from the book Assessment of Treatment Plant Performance and Water Quality Data: A Guide for Students, Researchers and Practitioners, Marcos von Sperling, Matthew E. Verbyla and Sílvia M. A. C. Oliveira (Authors).
doi: 10.2166/9781780409320_0531

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

14.1 INTRODUCTION

Advanced

C. 7

C. 11

C. 12

C. 14

In Chapter 7, we analysed how to compute *removal efficiencies* in one single reactor and in several reactors in series, leading to the overall removal obtained in the system. In Chapter 11, we gave you an incentive to search for *relationships between variables*, in order to obtain a better understanding of the system you are studying and possibly estimate effluent concentrations and removal efficiencies using regression analysis. In Chapter 12, we introduced the concepts of *water and mass balances*, which are essential for predicting effluent concentrations. Finally, in Chapter 13, we showed you the different types of *hydraulic and mass loading rates*, which are equally important factors for understanding the behaviour of treatment unit processes.

In this chapter, we will shift our discussion to the topics of **reaction kinetics** and **reactor hydraulics**. By incorporating these two elements into your study on the performance of treatment plants or the quality of ambient waters, you can broaden the impact of your findings, *making them potentially useful to others*. Rather than making inferences about your specific system, you can generalize the results from your research and estimate removal coefficients that may apply not only in your system but also in other treatment plants working in a similar fashion (either existing plants or plants that could be designed using the coefficients derived in your research). But this type of **generalization** requires **a lot of care**, and we will discuss in this chapter some basic methods and associated precautions to be taken.

Our objective of study here is to *describe mathematically how and at what rate a constituent is removed or converted over time in a reactor*. Simply stated, **reaction kinetics** deals with the rate of reaction of a constituent in a reactor, that is, its transformation (production or consumption).

Reactor hydraulics represents the pattern of fluid flow and the transport of constituents through the reactor, which will influence the effluent concentration of the constituent. *Reaction kinetics and reactor hydraulics go hand-in-hand, and both are equally important for the prediction of effluent concentrations and removal efficiencies*. If you want your results to be **applicable at full-scale plants**, you will need to take both components into consideration.

If we treat *rivers and lakes also as reactors* in which constituents are transported and converted, then several of the approaches used in this chapter will also be applicable for **water quality modelling**.

S. 13.2

As we saw in Section 13.2, reactors can be operated in **batch** or **continuous-flow** regimes. For continuous-flow regimes, when you have both the *kinetics* of the constituent removal and the *transport* of the constituent through the reactor, the model is called a **process model**. The following are two **idealized** types of reactors frequently used in process modelling, encompassing different tank configurations, hydraulic behaviours, and mixing conditions:

- **Complete-mix reactor**, also called completely stirred tank reactor (CSTR), continuous-flow stirred tank reactor (CFSTR), and completely mixed flow reactor (CMFR)
- **Plug-flow reactor** (PFR)

These two hydraulic conditions are at the opposite ends of the spectrum of mixing and dispersion. **Complete mixing** assumes *infinite dispersion* and **plug-flow** assumes *zero dispersion* of fluid elements as they travel from the inlet to the outlet of the reactor. We should realize that *these two idealized reactors do not occur in real life*, and the hydraulic behaviour and mixing conditions in actual reactors lie somewhere between these two idealized extremes. We can *approach* plug-flow conditions in a river that flows with minimal dispersion, from upstream to downstream. We can also *approach* complete-mix conditions in a squarish aeration tank subjected to intensive mixing or in mechanized coagulation rapid-mixing units. Our reactor can *approach* these idealized behaviours but will not fully comply with their conditions. In some cases (as in the reactors mentioned in this paragraph), it may be useful for us to assume these idealized conditions, but in most other reactors, we will need to elaborate more about their hydraulic behaviour. We will discuss about these two idealized flow regimens in Section 14.4.

S. 14.4

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Other models that represent more realistic hydraulics and mixing have been developed, such as the '**plug-flow with dispersion**' and the '**apparent tanks-in-series**' models. The use of these important process models together with the relevant reaction kinetics will be discussed in Section 14.5.

S. 14.5

We want to stress that this subject is covered in many textbooks on chemical engineering and wastewater treatment and associated texts, such as Levenspiel (1999), Arceivala (1981), von Sperling and Chernicharo (2005), von Sperling (2007), Kadlec and Wallace (2009), Metcalf and Eddy (2014), Mihelcic and Zimmerman (2014), and von Sperling *et al.* (2018).

The theory of reaction kinetics and reactor hydraulics has been presented in sufficient detail in the companion books of this series – von Sperling and Chernicharo (2005) and von Sperling (2007) – which are also available as 'open access' sources in the International Water Association (IWA) Publishing website. In these two books, the coverage of these topics has been mainly for their *utilization for* ***design*** *purposes of new reactors*. The coverage was detailed, and we will not repeat it here. Since these two references are 'open access', you are advised to consult them to obtain more background information about the subject.

Here, we will address mainly the application that is associated with the objectives of our book, which are *performance assessment of existing treatment plants*, and, to some extent, of *water bodies*.

In this chapter, we will cover the following items:

- Reaction order
- Analytical and numerical integration
- Determination of kinetic coefficients in batch experiments (mainly bench or pilot scales)
- Reactor hydraulics (complete-mix, plug-flow, dispersed flow, and tanks-in-series)
- Determination of first-order kinetic coefficients in continuous-flow reactors
- Estimation of effluent concentrations under steady-state conditions
- Estimation of effluent concentrations under dynamic conditions

Our coverage here will be **simple**, aimed at enticing you to the idea of incorporating a modelling component to your research, and to open your mind to the possibility of using simple and more advanced models. Mathematical modelling for representing the behaviour of treatment units is an extremely vast subject, so if you want to deepen your knowledge on this topic, you should consult specialized references. In the field of wastewater treatment, there are the classical family of models based on the activated sludge model (IWA, 2000) that were further adapted to include other processes, such as anaerobic digestion, constructed wetlands, biofilm reactors, and even river water quality. You can find references for these publications on the IWA Publishing website. Additionally, you should consult textbooks such as Henze *et al.* (2008), Van Haandel and Van der Lubbe (2012), Meijer and Brdjanovic (2012), Van Loosdrecht *et al.* (2016), and in addition, there are many other excellent references.

14.2 REACTION ORDER

14.2.1 Reaction orders – 0, 1, and 2

Advanced

The following section is partially based on von Sperling (2007). The *reaction rate* ***r*** is the term used to represent the removal or formation of a constituent or chemical species. Note that in Section 12.3, where we covered mass balances, we presented this concept of the reaction rate *r*.

S. 12.3

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

The relation between the *reaction rate* (r), the *concentration of the reagent* (C), and the *order of reaction* (n) is given by the expression

$$r = kC^n \tag{14.1}$$

where

r = reaction rate [typically $(g/m^3)/d$ or $(mg/L)/d$]
k = reaction coefficient (unit is variable, depending on the reaction type)
C = reagent concentration (typically g/m^3 or mg/L)
n = reaction order (note that n here represents reaction order; in other parts of the book, it represents number of data in the sample).

For different values of n, there are the following types of reactions:

- $n = 0$ *zero-order reaction*
- $n = 1$ *first-order reaction*
- $n = 2$ *second-order reaction*

If the logarithm is applied on both sides of Equation 14.1, the following equation is obtained:

$$\log r = \log k + n \log C \tag{14.2}$$

The visualization of the above relation for different values of n is presented in Figure 14.1.
The interpretation of Figure 14.1 is

- The *zero-order reaction* results in a horizontal line. The reaction rate is independent of the reagent concentration, that is, it is the same, independent of the reagent concentration.
- The *first-order reaction* has a reaction rate directly proportional to the reagent concentration.
- The *second-order reaction* has a reaction rate proportional to the square of the reagent concentration.

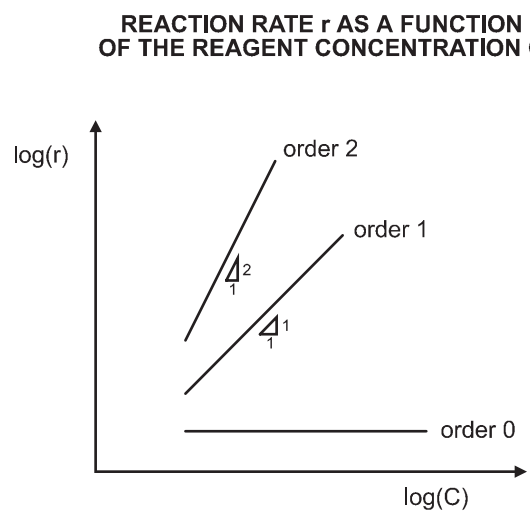


Figure 14.1 Determination of the reaction order on a logarithmic scale. Source: von Sperling (2007), adapted from Benefield and Randall (1980).

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

The **most frequent reaction order** used in treatment plant modelling is **first order**, but *you should check whether it is really applicable to the constituent you are analysing*.

Besides these reactions with constant order, there is another type of reaction, which is widely used in the area of wastewater treatment, called **saturation reaction, Michaelis-Menten reaction** or **Monod-type reaction**. The structure of this reaction is very useful for representing reaction rates dependent on a limiting substrate, but they will not be covered here. You should consult the references cited in the preceding paragraphs for obtaining the concept and utilization of this important reaction type.

14.2.2 Zero-order reactions

Advanced

Zero-order reactions are those in which the *reaction rate is independent of the reagent concentration*. In these conditions, the **rate of change of the reagent concentration (C) is constant**. This comment assumes that the reaction occurs in a batch reactor, in which there is no addition or withdrawal of the reagent during the reaction period. In the case of a reagent that is disappearing in the reactor (for example, through decomposition mechanisms), the rate of change is given by Equation 14.3. The derivative $\mathbf{dC/dt}$ represents the **rate of change**, that is, the change in concentration (dC, in g/m^3 or mg/L) per unit time (typically days, but can also be minutes, hours, or years, depending on how fast the reaction proceeds). The *minus* sign in the term on the right-hand side of the equation indicates *removal* of the constituent, whereas a *plus* sign would indicate *production* of the constituent.

$$\frac{dC}{dt} = -K \cdot C^0 \tag{14.3}$$

or

$$\frac{dC}{dt} = -K \tag{14.4}$$

The development of the rate of change (dC/dt) with time according to Equation 14.4 can be seen in Figure 14.2. You can note that the rate is constant with time.

The integration of Equation 14.4 with $C = C_0$ at $t = 0$ leads to

$$\boxed{C = C_0 - K \cdot t} \tag{14.5}$$

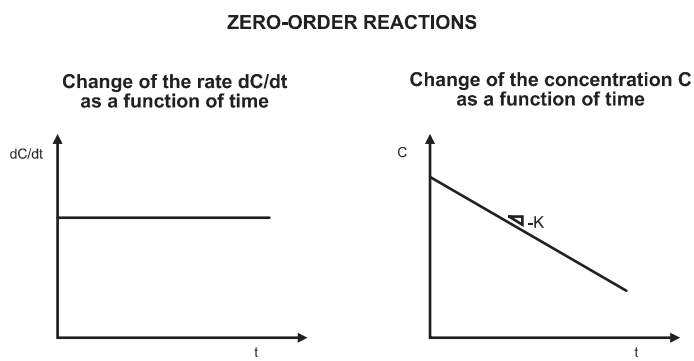


Figure 14.2 Zero-order reactions. (a) Change of the reaction rate dC/dt with time. (b) Change of the concentration C with time. Source: von Sperling (2007).

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

where

C = concentration of the constituent that is being removed (g/m^3 or mg/L)
C_0 = concentration of the constituent that is being removed (g/m^3 or mg/L) at time $t = 0$
t = time (d)
K = reaction coefficient [$(g/m^3)/d$ or (mg/L)/d].

S. 14.3

This equation can be visualized in Figure 14.2. If you undertake measurements of the concentration over time in a batch reactor (see Section 14.3) and your constituent decays following a zero-order reaction, this is the concentration profile you will obtain.

The coefficient K in a zero-order reaction reflects the concentration that was removed or converted per unit time. For instance, if the initial concentration of the constituent is $C_0 = 100$ mg/L and the reaction coefficient is $K = 10$ (mg/L)/d, this means that, after one day the concentration will be $100 - 10 = 90$ mg/L; after two days it will be $90 - 10 = 80$ mg/L; after three days it will be $80 - 10 = 70$ mg/L, and so on (as long as the assumption of a zero-order reaction holds throughout the reaction period). Note that $-K$ is the slope of the curve shown in Figure 14.2.

14.2.3 First-order reactions

(a) Structure of a first-order reaction

Advanced

First-order reactions are those in which the **reaction rate is proportional to the concentration of the reagent**. Therefore, in a batch reactor, the rate of change of the reagent concentration C is proportional to the reagent concentration at a given time. Assuming a reaction in which the constituent is being removed, the associated equation is

$$\frac{dC}{dt} = -K \cdot C^1 \tag{14.6}$$

or

$$\frac{dC}{dt} = -K \cdot C \tag{14.7}$$

The development of the rate of change (dC/dt) with time according to Equation 14.7 is presented in Figure 14.3. It can be noted that the ***rate*** *decreases linearly with time*.

Integrating Equation 14.7 with $C = C_0$ at $t = 0$ leads to

$$\ln C = \ln C_0 - K \cdot t \tag{14.8}$$

or

$$\boxed{C = C_0 \cdot e^{-K \cdot t}} \tag{14.9}$$

where

C = concentration of the constituent that is being removed (g/m^3 or mg/L)
C_0 = concentration of the constituent that is being removed (g/m^3 or mg/L) at time $t = 0$
t = time (d)
K = reaction coefficient (d^{-1}).

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

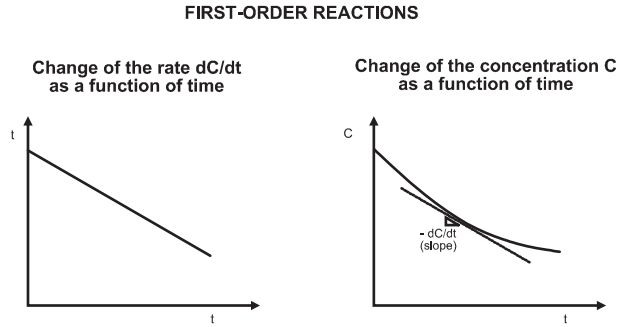


Figure 14.3 First-order reactions. (a) Change of the reaction rate dC/dt with time. (b) Change of the concentration C with time. Source: von Sperling (2007).

Note that the units of the reaction coefficient K are now $\mathbf{d^{-1}}$ (differently from the K units in the zero-order reaction, which were (mg/L)/d). Also note the **dimensionless product** $\boldsymbol{K \cdot t}$ ($d^{-1} \times$ d), because it will appear in several other equations shown in this chapter.

S. 14.3

Equation 14.9 is plotted in Figure 14.3. If you undertake measurements of the concentration over time in a batch reactor (see Section 14.3) and your constituent decays following a first-order reaction, this is the **concentration profile** you will obtain. Note that the slope of the curve shown in Figure 14.3 is $-K$, and in Figure 14.3, the derivative $\mathbf{dC/dt}$ is the **tangent** to the curve at any given time t.

If you are modelling *coliform* decay, you know that coliforms are usually plotted on a *log-scale* in the Y-axis. Therefore, you will not see a curve, as the one shown in Figure 14.3, but rather a straight line, because of the log-transformation of the coliform data.

When analyzing Equation 14.7 and the concentration profile in Figure 14.3, we see the important fact that, for first-order reactions, **the higher the concentration (C) at a given time, the higher the decay rate ($\mathbf{dC/dt}$)**. We need to understand this statement. *When we have a high concentration, we have a high value of removed* ***concentration*** *during a specified time period, but the removal* ***efficiency*** *during this time period is not influenced by the concentration*. For instance, if we start with a high concentration (300 mg/L) and it falls down to 100 mg/L in a period of five days, the reduction during this period is $300 - 100 = 200$ mg/L, and the removal efficiency is $(300 - 100) = 0.67 = 67\%$ after the period of five days. When the concentration becomes low, say, 30 mg/L, the concentration will decrease to 10 mg/L in a subsequent period of five days (provided that the reaction coefficient K remains the same). Because the concentration was low, it only decreased by $30 - 10 = 20$ mg/L during this period. However, the removal efficiency during these subsequent five days remained the same: $(30 - 10)/30 = 0.67 = 67\%$.

S. 14.3
S. 14.4.4
S. 14.5.4

First-order reactions are very common in water and wastewater treatment systems and also for modelling constituents in water bodies. Given the importance of first-order reactions in treatment plant performance assessment, we will provide more details about their interpretation and about the experimental determination of the reaction coefficient K (see also Sections 14.3, 14.4.4, and 14.5.4).

(b) Interpreting the first-order removal coefficient K

We will now interpret the **meaning of K** and its value. As shown in Equation 14.9, its dimension is time^{-1}, that is, min^{-1}, h^{-1}, d^{-1}, and year^{-1}. As already mentioned, these units are not so easy to

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

S. 14.2.2

interpret as those from the zero-order reactions, in which K units are $(g/m^3)/d$ or $(mg/L)/d$ (see Section 14.2.2).

Let us interpret the differential Equation 14.7. **If *K* values are small**, below, say, 0.4 d^{-1}, we can roughly interpret that ***K* represents the fraction of the constituent that decays per day**. For instance, if $K = 0.10\ d^{-1}$, we can say that, approximately 0.10 or 10% of the constituent is removed per day. If our initial concentration is 100 mg/L, after one day we would have 100–$0.10 \times 100 = 100–10 = 90$ mg/L.

The essence of a first-order reaction is that the rate dC/dt is directly proportional to the concentration C at any time t. Therefore, if we want to calculate the concentration after two days, we will use the value from day one, and obtain: $90 - 0.10 \times 90 = 90 - 9 = 81$ mg/L. After three days, we will use the value from day two, and get: $81 - 0.10 \times 81 = 81 - 8.1 = 72.9$ mg/L, and so forth for the other days. In Example 14.1 we present the sequence of calculations and compare it with the results from the analytical solution (Equation 14.9).

The **error** brought about by this simple numerical calculation with respect to the exact solution given by the analytical integration, expressed in Equation 14.9, **for one day**, is as presented in Table 14.1.

We can see that, if $K > 0.4\ d^{-1}$, the resulting deviation from the analytical (exact) solution will be *>10%*. Can we still use this numerical approach if we have K values greater than 0.4 d^{-1}? Yes, we can! We can simply solve this problem by adopting a suitable time unit, for instance, converting days into hours. For example, a coefficient $K = 0.72\ d^{-1}$ is the same as $K = (0.72\ d^{-1})/(24\ h/d) = 0.03\ h^{-1}$. Problem solved! Now, we have a low value of K, and we can roughly say that our constituent will decay approximately 3% *per hour*.

Looking at Table 14.1, you would think that we could not have K values that are greater than 1.0 d^{-1}. But we can, and several coefficients in water and wastewater modelling practice may be higher than 1.0 d^{-1} (for instance, coliform removal in maturation ponds or oxygen transfer by reaeration in river modelling). For example, we can have $K = 2.4\ d^{-1}$, and we will consider that it is equivalent to $K = (2.4\ d^{-1})/(24\ h/d) = 0.1\ h^{-1}$.

(c) Analytical integration of the equation of the first-order reaction

The **analytical solution** to the integration of the differential equation representing the first-order reaction (Equation 14.7) is **simple** and is given by Equation 14.9. To estimate the concentration (C) at any given time t, you need only the initial concentration (C_0) and the reaction coefficient K (see Example 14.1).

S. 14.3

This is the equation we will use in the experimental determination of the coefficient K, for given values of C and t obtained in a batch essay (see Section 14.3).

(d) Numerical integration of the equation of the first-order reaction

The numerical solution to the integration of the differential equation representing the first-order reaction (Equation 14.7) will build up from the discussion we just had about errors in the prediction when we have high values of K. The way out we found was to convert the coefficient to a smaller

Table 14.1 Percent difference, in one day, between the simple numerical calculation and the exact analytical solution (using Equation 14.9) for a first-order reaction, with different K values.

K (d^{-1})	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Error (%)	0.5	2.3	5.5	10.5	17.6	27.1	39.6	55.5	75.4	100.0

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

time dimension (for instance, from day to hour). Now, we will do something similar, but instead of converting the coefficient K, we will introduce the concept of a **time step (Δt)** in our calculations and move our calculations at *very small time steps*.

We will use the **simple numerical integration** applying the **method of Euler** (Swiss mathematician – search the internet for more details on this method), given by Equation 14.10.

$$C_t = C_{t-1} - K \times C_{t-1} \times \Delta t \tag{14.10}$$

where

C_t = concentration at time t (mg/L)
C_{t-1} = concentration at the previous time step $t - 1$ (mg/L)
t = time (d)
K = first-order reaction coefficient (d^{-1})
Δt = time step (fraction of one day).

The calculations we did in subsection 'b' above were essentially using this equation, but having a large time step ($\Delta t = 1$ d). Euler's method is *very simple and straightforward to be implemented in spreadsheets* (as we will show in some Excel sheets in this chapter). However, with simplicity comes a *potential inaccuracy*: the *errors from the numerical procedure are carried out from time step to time step*. One day may be a long time step, and the errors will be propagated along the days (see Example 14.1). However, if we divide one day into small time-intervals, say, into 100 time-intervals, each will have $1/100 = 0.01$ of a day ($\Delta t = 0.01$ d), and we will be able to make our calculation at small time steps and substantially reduce the error. *We do not need to convert our K coefficient into another time basis*, as we did before, when we changed days into hours. The integration procedure, with the insertion of the time step (Δt), will take care of this.

Of course there are other more efficient numerical integration procedures that produce a smaller error and that can work with larger integration steps, such as the Runge–Kutta methods, but these are outside the scope of this book and can be found in other literature that cover water quality modelling (Chapra *et al.*, 1997) or numerical methods.

Example

EXAMPLE 14.1 ANALYTICAL AND NUMERICAL INTEGRATIONS OF THE FIRST-ORDER REACTION

Undertake an analytical and a numerical integration of the differential equation that represents a first-order reaction. Calculate the resulting concentrations from days 1 to 10, considering the following input data:

$C_0 = 100$ *mg/L*
$K = 0.10$ d^{-1}

Excel

Note: This example is also available as an Excel spreadsheet.

Solution:

(a) Analytical soelution

From Equation 14.9, and given $C_0 = 100$ mg/L and $K = 0.10$ d^{-1}, we can estimate the concentrations at various days:

$$C = C_0 \cdot e^{-K \cdot t} = 100 \times e^{(-0.10 \times t)}$$

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

For t varying from 0 to 10, we obtain

t (d)	C (mg/L)	t (d)	C (mg/L)
0	100.0	6	54.9
1	90.5	7	49.7
2	81.9	8	44.9
3	74.1	9	40.7
4	67.0	10	36.8
5	60.7		

(b) Numerical solution

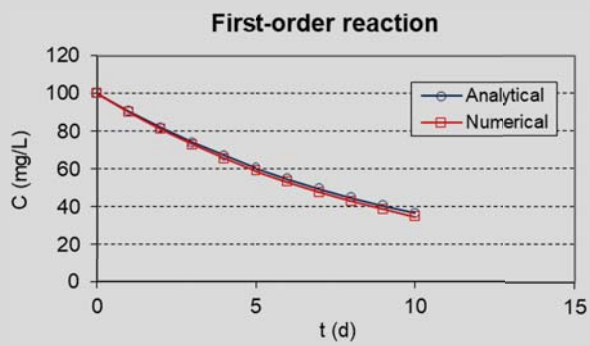
Using Euler's integration method and considering the time step of one day ($\Delta t = 1$ d), we can set up the following computational table:

t (d)	$C_t = C_{t-1} - K \times C_{t-1} \times \Delta t$	Calculations	C at the end of the day (mg/L)
0	–	–	100.0
1	$C_1 = C_0 - K \times C_0$	$C_1 = 100.0 - 0.10 \times 100.0$	90.0
2	$C_2 = C_1 - K \times C_1$	$C_2 = 90.0 - 0.10 \times 90.0$	81.0
3	$C_3 = C_2 - K \times C_2$	$C_3 = 81.0 - 0.10 \times 81.0$	72.9
4	$C_4 = C_3 - K \times C_3$	$C_4 = 72.9 - 0.10 \times 72.9$	65.6
5	$C_5 = C_4 - K \times C_4$	$C_5 = 65.6 - 0.10 \times 65.6$	59.0
6	$C_6 = C_5 - K \times C_5$	$C_6 = 59.0 - 0.10 \times 59.0$	53.1
7	$C_7 = C_6 - K \times C_6$	$C_7 = 53.1 - 0.10 \times 53.1$	47.8
8	$C_8 = C_7 - K \times C_7$	$C_8 = 47.8 - 0.10 \times 47.8$	43.0
9	$C_9 = C_8 - K \times C_8$	$C_9 = 43.0 - 0.10 \times 43.0$	38.7
10	$C_{10} = C_9 - K \times C_9$	$C_{10} = 38.7 - 0.10 \times 38.7$	34.9

(c) Graphical representation of both results

The chart below shows the calculated values of concentrations along 10 days. We can see that the analytical integration and numerical approximation led to very similar results. The reason is that the K coefficient is small ($K = 0.10\ \text{d}^{-1}$), as discussed in Section 14.2.3 and shown in Table 14.1.

S. 14.2.3

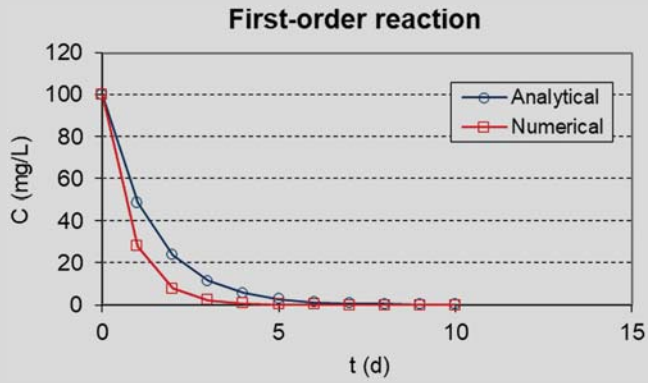


Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

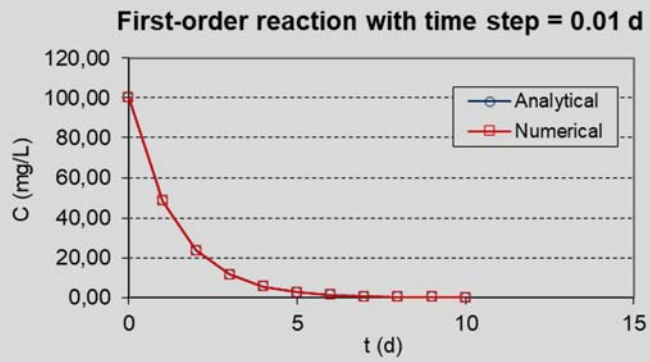
(d) Calculations with a higher *K* value

We will not show all the calculations again but will only explore the situation in which we have a high value of the first-order reaction coefficient. Now, let us use, for instance, $K = 0.72\ d^{-1}$. In the spreadsheet, simply change the *K* value in the corresponding cell.

If we keep the calculations with a long time step (one day), the errors will be large and will propagate along the days, as you can see in the graph below.



In order to circumvent this, we can make an integration with a small time step (0.01 d, that is, fractionating each day into 100 time-intervals). The calculations are shown in the second part of the Excel spreadsheet, and you can see that the errors are now very small. This is endorsed by the graphical output of the results, in which both curves overlap.



14.3 EXPERIMENTAL DETERMINATION OF THE REACTION ORDER AND KINETIC COEFFICIENT IN BATCH REACTORS

14.3.1 Estimation of the reaction order *n* and the reaction coefficient *K*

Advanced

The determination of the reaction order *n* and the associated kinetic coefficient *K* is frequently done at *bench* or *lab-scale* experiments, using a stirred-vessel operating on **batch mode**. The liquid medium is set to reproduce approximately the liquid (water, wastewater, and mixed liquor) that would be present in the reactor of the treatment system we want to represent. We fill the vessel and measure the concentration of

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

the constituent we are studying with respect to time. We tabulate the results, plot a time series graph, and try to fit a model that provides results that match well with the measured data. From the model that provides the best fit, we derive the reaction order n and/or the value of the kinetic coefficient K. See Figure 14.4 for the illustration of this typical sequence.

To decide whether our reaction is better represented by order 0, 1, or 2, we prepare graphs based on the **linearization of the models** and then undertake a **linear regression** to obtain the desired information. See Chapter 11 for the concept of linear regression.

C. 11

Table 14.2 presents a summary of the linearized plots and the information that can be obtained from the intercept and slope of the line of best fit.

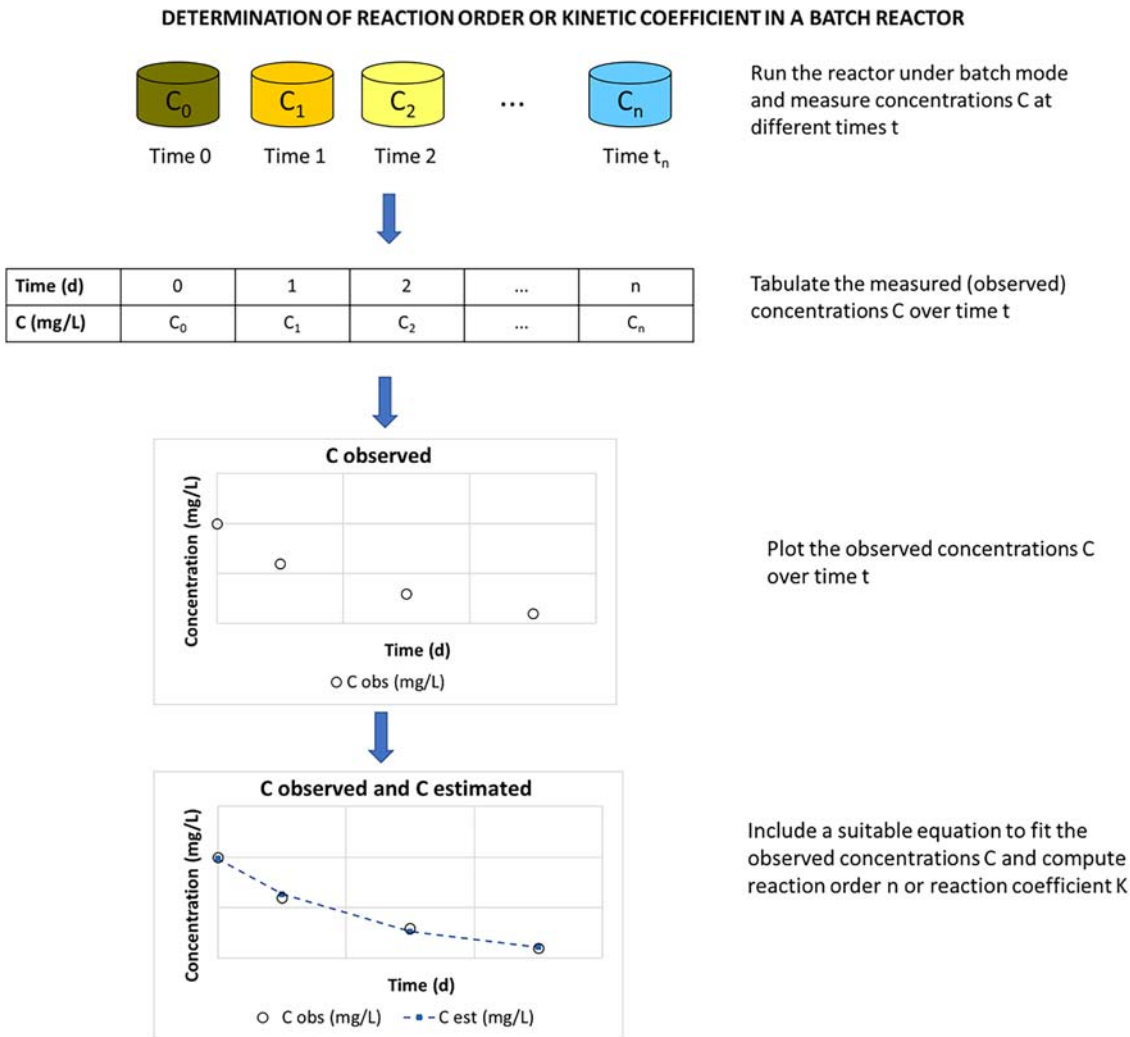


Figure 14.4 Experimental sequence for the determination of the reaction order and/or the value of the kinetic coefficient. Source: Inspired by an illustration provided in Chapra (1997).

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Table 14.2 Summary of the linearized plots of concentration versus time obtained in batch studies and the information associated with the intercept and slope of the line of best fit for reactions of orders 0, 1, and 2.

Reaction Order	Rate Units (Typical)	Dependent Variable (y)	Independent Variable (x)	Intercept	Slope
Zero ($n = 0$)	(mg/L)/d	C	t	C_0	$-K$
First ($n = 1$)	d^{-1}	$\ln C$	t	$\ln C_0$	$-K$
Second ($n = 2$)	(L/mg)/d	$1/C$	t	$1/C_0$	K
General ($n \neq 1$)	$[(L/mg)^{n-1}]/d$	C^{1-n}	t	C_0^{1-n}	$(n-1)K$

Source: Chapra (1997).

From Table 14.2, we can see that the **linearized forms** of the equations are

- **Zero-order:**

$$C = C_0 - K \cdot t \tag{14.11}$$

- **First-order:**

$$\ln C = \ln C_0 - K \cdot t \tag{14.12}$$

- **Second-order:**

$$\frac{1}{C} = \frac{1}{C_0} + K \cdot t \tag{14.13}$$

- **General** ($n \neq 1$, but a value very close to 1 can be used and lead to similar results, for instance, 0.999999999 or 1.000000001):

$$C = \frac{C_0}{[1 + (n-1) \cdot K \cdot C_0^{n-1} \cdot t]^{1/(n-1)}} \tag{14.14}$$

We will use these linearized equations in Example 14.2 to see which reaction order fits best the experimental data obtained in a batch experiment and calculate the coefficient K for $n = 0$, 1, and 2.

After that we will show in Example 14.3 how to simultaneously estimate the reaction order n and the reaction coefficient K without linearization, that is, using the original data, employing the Excel tool 'Solver' and applying the general Equation 14.14. Note, the Excel Solver tool is an 'add-in' which is not necessarily available by default after the installation of Microsoft Excel. It can be added by accessing the 'add-ins' and clicking on the check-box next to the Solver tool. On Windows, the add-ins are accessed under File > Options and on Macs, the add-in tools are accessed under Tools > Excel Add-ins.

Finally, in this subsection, we will use the Excel Solver tool to estimate the coefficient K for a first-order reaction ($n = 1$) using the analytical integration expressed in Equation 14.9.

In the analyses shown in the examples below, you can see that *the more data points you have in your experiments, the more reliable is your estimate of the parameter K* (von Sperling *et al.*, 2018). The least reliable estimate of K would be based on only two data points: one at the beginning of the experiment

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

($t = 0$; C equal to the initial concentration C_0) and one at the end of the experiment ($t = t_{\text{final}}$; C equal to the final concentration C_{final}). By inserting these values into the linearized equations, we could estimate the desired coefficient K. However, this approach is not recommended because it does not allow for the confirmation of what reaction order we have (0, 1, or 2), and we would not know which equation (14.11–14.13) to apply. We recommend that you use a minimum of five time points to construct a curve and determine the order and rate of a reaction.

Example

EXAMPLE 14.2 DERIVATION OF THE REACTION ORDER (N) AND THE REACTION COEFFICIENT (K) BY LINEARIZATION OF THE EQUATIONS

You completed a batch experiment and measured the concentration of a constituent with respect to time. You obtained the data shown below (C_{obs}) and want to know what is the closest reaction order (n) and the associated reaction coefficient (K).

Data:

Time (d)	*C_{obs} (mg/L)*
0	*100.0*
1	*86.0*
3	*77.0*
5	*57.0*
10	*40.0*
15	*19.0*
20	*16.0*

Excel

Note: This example is also available as an Excel spreadsheet.

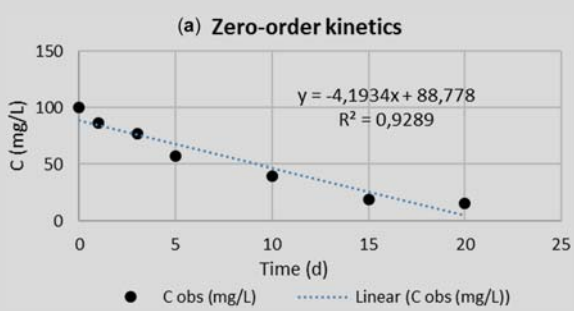
Solution:

(a) Zero-order reaction

For zero-order reactions, Equation 14.11 shows the progress of the concentration over time as

$$C = C_0 - K \cdot t$$

This equation is already linear. Thus, you plot the values and fit a straight line by regression analysis (automatically done in the accompanying Excel spreadsheet, which gives you the values of the intercept and slope of the line). The graph obtained is shown below.



Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

From the equation shown in the graph, we see that

- Intercept $= 88.78 \rightarrow C_0 = 88.78$ mg/L
- Slope $= -4.193 \rightarrow K = -(-4.193) =$ **4.193 (mg/L)/d**

(b) First-order reaction

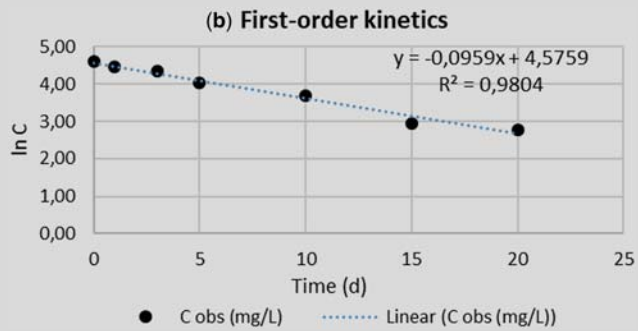
For first-order reactions, Equation 14.12 shows the linearized form of the progress of the concentration over time as

$$\ln C = \ln C_0 - K \cdot t$$

Thus, you calculate the natural logarithm of the observed concentration data (C_{obs}) and obtain the following values:

Time (d)	ln *C*
0	4.61
1	4.45
3	4.34
5	4.04
10	3.69
15	2.94
20	2.77

You plot these values and fit a straight line by regression analysis (automatically done in the accompanying Excel spreadsheet).



From the equation shown in the graph, we see that

- Intercept $= 4.576 \rightarrow C_0 = e^{4.576} = 97.12$ mg/L
- Slope $= -0.096 \rightarrow K = -(-0.096) =$ **0.096 d^{-1}**

(c) Second-order reaction

For second-order reactions, Equation 14.13 shows the linearized form of the progress of the concentration over time as

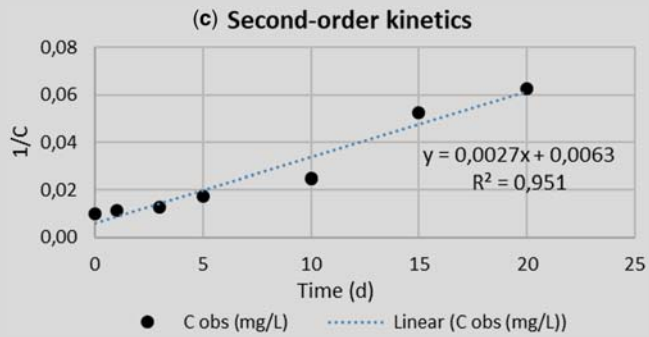
$$\frac{1}{C} = \frac{1}{C_0} + K \cdot t$$

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Thus, you need to calculate the inverse of the observed concentration data (C_{obs}):

Time (d)	1/*C*
0	0.010
1	0.012
3	0.013
5	0.018
10	0.025
15	0.053
20	0.063

You plot these values and fit a straight line by regression analysis (automatically done in the accompanying Excel spreadsheet).



From the equation shown in the graph, we see that

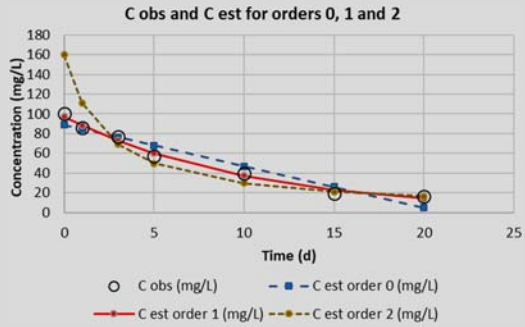
- Intercept $= 0.0063 \rightarrow C_0 = 1/0.0063 = 159.84$ mg/L
- Slope $= 0.0027 \rightarrow K =$ **0.0027 (L/mg)/d**

(d) Summary

The observed concentrations (C_{obs}) and estimated concentrations (C_{est}) for the three reaction rates are shown in the following table and graph:

Time (d)	C_{obs} (mg/L)	C_{est} order 0 (mg/L)	C_{est} order 1 (mg/L)	C_{est} order 2 (mg/L)
0	100	88.78	97.12	159.84
1	86	84.58	88.23	111.03
3	77	76.20	72.83	68.94
5	57	67.81	60.12	49.99
10	40	46.84	37.22	29.62
15	19	25.88	23.04	21.05
20	16	4.91	14.26	16.33

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest



Note that, visually speaking, the best fit to the observed data was obtained by the first-order reaction and that the worst fit was associated with the second-order reaction.

Besides the visual interpretation, we want to have a formal information of which of the reaction orders provided the best fit. Please observe that we cannot interpret directly the values of R^2 shown in the graphs (see Chapter 11 for the concept of R^2 in regression analysis). The R^2 values are all very good (above 0.90), but you should consider that these values are based on transformed data, in order to obtain a linearized plot. Therefore, by transforming the data, we also modify the capability of the R^2 coefficient of being a true indicator of the goodness-of-fit of our original (untransformed) data.

C. 15

The best approach is to use the Coefficient of Determination (CoD), which is explained in Chapter 15. The resulting values (calculated in the Excel spreadsheet) are

Reaction Order (n)	0	1	2
Coefficient of Determination CoD	0.9289	0.9896	0.3182

From the CoD values, we confirm that the best fit was provided by the first-order reaction, and that the second-order reaction led to a poor fit.

(e) Generalized equation

The Excel spreadsheet also shows you the utilization of the general Equation 14.14. The results obtained are naturally the same as those obtained above using the specific equations for each reaction order.

Example

EXAMPLE 14.3 SIMULTANEOUS ESTIMATION OF THE REACTION ORDER N AND THE REACTION COEFFICIENT K USING THE GENERALIZED EQUATION AND THE SOLVER TOOL

Based on the data you obtained from a batch experiment, estimate the reaction rate (n) and the associated reaction coefficient (K) using the generalized Equation 14.14 and the Excel 'Solver' tool. The observed data are the same as those reported in Example 14.2, and your estimation of n and K will be based on these original data, without any transformations.

Excel

Note: This example is also available as an Excel spreadsheet.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Solution:

Use the generalized Equation 14.14:

$$C = \frac{C_0}{[1 + (n-1) \cdot K \cdot C_0^{n-1} \cdot t]^{1/(n-1)}}$$

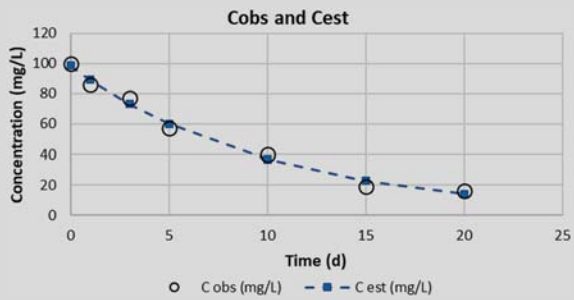
Provide initial 'guesses' for the values of C_0, K, and n, and run the Solver tool (maximize the value of the Coefficient of Determination by varying the values of C_0, K, and n). After convergence, you should obtain the values listed below. Note that there is no guarantee that you will arrive exactly at these values, because any optimization procedure may converge onto a local optimum, without necessarily arriving at the global optimum. This depends partly on the algorithm used for optimization and partly on the accuracy of the initial guesses. To increase your chances of finding the correct values, choose initial guesses that are reasonable, and try to solve again using different initial guesses to see if they produce the same optimized values of C_0, K, and n.

C_0	98.595
K	0.087
n	1.030

Note that n is very close to 1, indicating that this reaction approaches a **first-order reaction**. The resulting value of the reaction coefficient K is 0.087 d^{-1} (the unit of d^{-1} is specific for a first-order reaction).

The Coefficient of Determination (CoD) obtained is 0.9902, highlighting the excellent fitting.

The plot of observed (C_{obs}) and estimated (C_{est}) concentrations is shown below.



Example

EXAMPLE 14.4 ESTIMATION OF THE REACTION COEFFICIENT *K* FOR A FIRST-ORDER REACTION USING THE ANALYTICAL EQUATION AND THE SOLVER TOOL

Based on the data you obtained from the batch experiment, and based on the fact that you concluded that the reaction rate could be expressed using first-order kinetics, estimate the reaction rate coefficient (K) using the analytical solution represented by Equation 14.9 and the Excel 'Solver' tool. The observed data are the same as those reported in Example 14.2, and your estimation of K will be based on these original data, without transformations.

Excel

Note: This example is also available as an Excel spreadsheet.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Solution:

Use the generalized Equation 14.9:

$$C = C_0 \cdot e^{-K \cdot t}$$

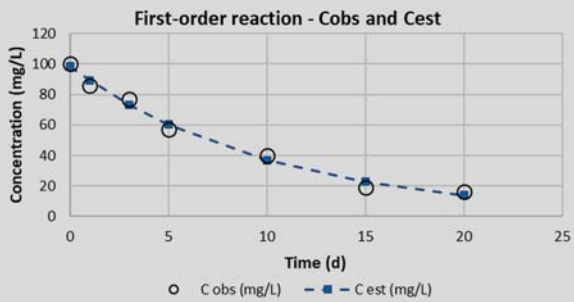
Provide initial 'guess' values for C_0 and K, and run the Solver tool (maximize the value of the Coefficient of Determination while changing the values of C_0 and K). After convergence, you should obtain the values listed below. Note that there is no guarantee that you will arrive exactly at these values, because any optimization procedure may converge onto a local optimum, without arriving at the global optimum. This depends partly on the algorithm used for optimization and partly on the accuracy of the initial guesses. To increase your chances of finding the correct values, choose initial guesses that are reasonable, and try to solve again using different initial guesses to see if they produce the same optimized values of C_0 and K.

C_0	98.400
K	0.098

The resulting value of the reaction coefficient K is 0.098 d^{-1}. This value is close to the values obtained in Examples 14.2 (small differences due to linearization) and 14.3 (small differences due to the fact that the reaction order was found to be not exactly 1 in that example).

The Coefficient of Determination (CoD) obtained is 0.9901, indicating an excellent fitting.

The plot of observed (C_{obs}) and estimated (C_{est}) concentrations is shown below.



Throughout Examples 14.2–14.4, we showed you complementary approaches for estimating the reaction order n and reaction coefficient K based on a batch experiment. They led to similar results, and it is up to you to adopt the approach with which you feel most comfortable and for which you really feel that you understand the entire calculation sequence.

14.3.2 Influence of a refractory fraction on the removal of a constituent (first-order reaction)

Advanced

S. 14.2

In Section 14.2, we saw Equation 14.9, which represented the temporal progression of a constituent that decays according to first-order kinetics. From that equation, if we extend the time t for a long duration or have a high reaction coefficient K, we will eventually end up with a concentration that approaches (but never reaches) zero.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

This may be the case for several constituents, but others, for instance, representing organic matter (biochemical oxygen demand or chemical oxygen demand) may not reach near-zero concentrations. This is because a fraction of the organic matter may consist of refractory (non-biodegradable, persistent) compounds that will not be decomposed by biological means. In order to accommodate this fact, we can use first-order kinetic models with residuals (Dotro *et al.*, 2017; Kadlec & Wallace, 2009):

- First-order model **without** residual (Equation 14.9):

$$C = C_0 \cdot e^{-K \cdot t} \tag{14.15}$$

- First-order model **with** residual C^*:

$$(C - C^*) = (C_0 - C^*) \cdot e^{-K \cdot t} \tag{14.16}$$

$$\boxed{C = C^* + (C_0 - C^*) \cdot e^{-K \cdot t}} \tag{14.17}$$

where

C^* = residual (refractory, persistent, or non-biodegradable) concentration (mg/L).

You can see that we introduced a simple modification into the traditional first-order equation: instead of having only C, we used $C - C^*$; instead of having only C_0, we used $(C_0 - C^*)$. The structure of the equation otherwise remains the same.

The value of C^* will vary with the composition of the constituent you are analysing. From your experiments, you can make the estimation incorporating it into the equation for first-order reaction kinetics and inserting C^* as an additional parameter to be estimated with the Solver tool (adapt the sequence shown in Example 14.4). Otherwise, if you have a long time series of measurements, you can tentatively adopt C^* as the minimum value found (or a low percentile, say, first percentile) and take it outside of the estimations by Solver. Alternatively, you can search for references in the literature that report the refractory fraction of organic matter in a system similar to your system.

Figure 14.5 shows an example of a plot that does and does not account for the residual concentration C^*. The values have been calculated using the following input data: $C_0 = 100$ mg/L, $K = 0.10$ d^{-1}, $C^* = 20$ mg/L. You can see that if the existence of a refractory concentration is not taken into account, the concentration will approach zero. However, if we incorporate a residual concentration C^* in the model, the concentration will approach (but never actually reach) the value of C^*, possibly providing a better representation of the reality for this particular constituent.

Excel

14.3.3 First-order reaction with lag phase

Advanced

In some cases, the reaction may not proceed immediately, and a 'lag' time is required for the reaction to be established. This case is called a first-order reaction with a lag phase, and the appropriate equations are (Manser *et al.*, 2015; Pecson *et al.*, 2007):

$$C = C_0[1 - (1 - e^{(-K \cdot t)})^m] \tag{14.18}$$

$$\text{Lag period} = \frac{\ln(m)}{K} \tag{14.19}$$

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

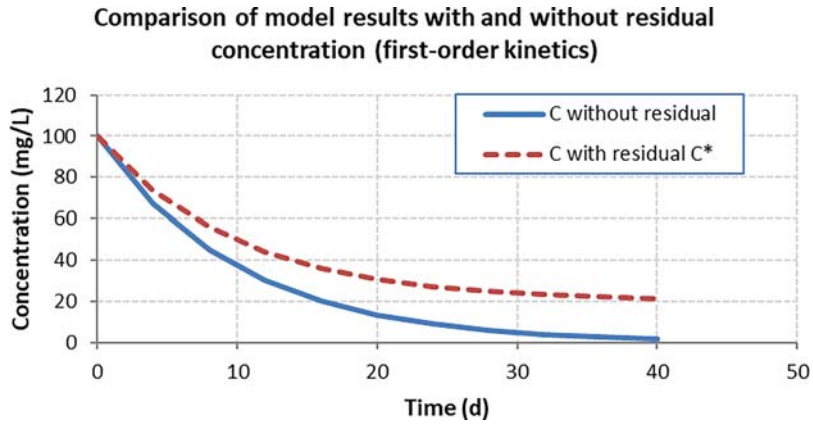


Figure 14.5 Plot of model results (first-order kinetics) with and without taking into account, the influence of a residual concentration. Calculations made with: $C_0 = 100$ mg/L, $K = 0.10$ d^{-1}, $C^* = 20$ mg/L.

where

m = lag coefficient

lag period = inflection point of the curve (d).

If there is no lag, $m = 1$, and the equation becomes the traditional first-order reaction equation (Equation 14.9).

Excel

In Figure 14.6, you can see the influence of the value of 'm' on the shape of the decay curve. We have used the same example as before, with $C_0 = 100$ mg/L and $K = 0.10$ d^{-1}. The figure shows the resulting curves (Equation 14.18) for different values of m (1, 2, 3, and 4), together with the associated lag periods, or inflection points (Equation 14.19) marked in the X-axis.

14.3.4 Influence of temperature on the reaction rate

Advanced

The rate of any chemical reaction increases with temperature, provided that this increase in temperature does not produce alterations in the reagents or in the catalyst. Biochemical reactions also have the same tendency to increase with respect to temperature (within certain ranges). However, there is an ideal temperature for the biological reactions, above which the rate decreases, possibly due to the destruction of enzymes at the higher temperatures (Benefield & Randall, 1980; Sawyer & McCarty, 1978).

Usually, the kinetic coefficients K are reported at a **standard temperature of 20°C**. However, their use will frequently be applied at different temperatures, leading to the following situations:

- If you plan to use a K coefficient from the literature, it is likely that it has been reported at 20°C. Therefore, you will probably want to modify it based on the liquid temperature in your reactor.
- If you obtained the K coefficient from experiments undertaken at a certain liquid temperature, you might wish to convert it to the K coefficient for a standard temperature of 20°C.

A usual form to estimate the variation of the reaction rate as a function of temperature is by means of the formulation based on the **van't Hoff-Arrhenius theory**, which can be expressed as

$$\frac{K_{T_2}}{K_{T_1}} = \theta^{T_2 - T_1} \tag{14.20}$$

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

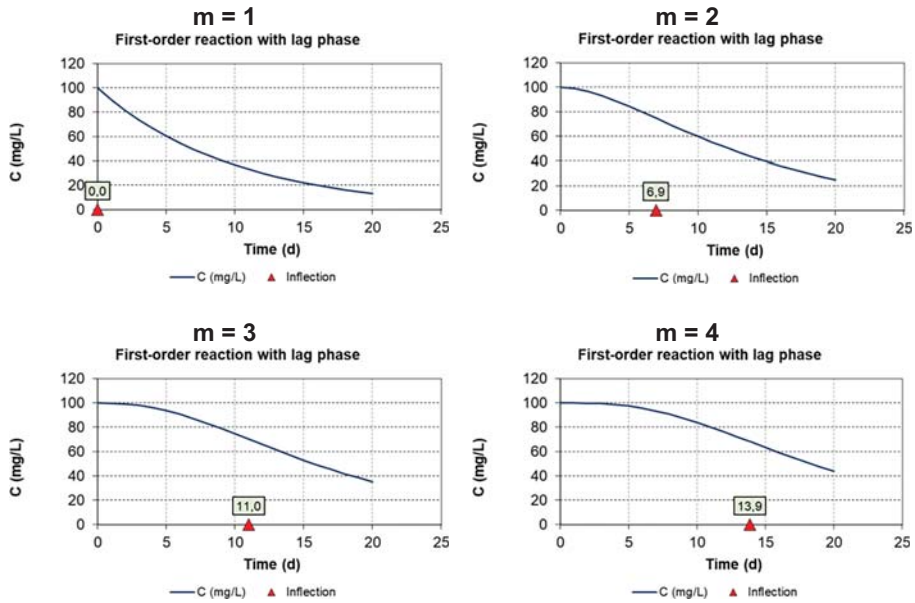


Figure 14.6 Plot of model results (first-order kinetics) without lag ($m = 1$) and with lag phase ($m = 2$, 3, and 4), with an indication of the inflection point (triangle in the *X*-axis). Calculations made with $C_0 = 100$ mg/L and $K = 0.10$ d^{-1}.

where

K_{T1} = reaction coefficient at the liquid temperature T_1
K_{T2} = reaction coefficient at the liquid temperature T_2
θ = temperature coefficient (dimensionless).

Or, algebraically rearranging Equation 14.20 to associate K with the standard temperature of 20°C:

$$K_T = K_{20}\theta^{T-20} \tag{14.21}$$

where

K_T = reaction coefficient at the liquid temperature T
K_{20} = reaction coefficient at the standard temperature of 20°C.

Note that Equation 14.21 has the form of a geometric growth, indicating that the *reaction coefficient K increases geometrically with the temperature*. For instance, if a certain reaction has a temperature coefficient $\theta = 1.05$, it means that there is a 5% increase in K for each increase of 1°C in T. For example, if T increases from 20°C to 21°C, K increases 5% with respect to K at 20°C. Furthermore, if T increases from 21°C to 22°C, K increases 5% with respect to K at 21°C, and so on.

Also note that the reference is for the temperature of the **liquid** and *not the air temperature*.

In the literature, the temperature coefficient θ *is usually reported together with the K coefficient at the standard temperature of 20°C* so that you can make the appropriate conversions. However, this is not always

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

the case, and in many situations, you will face problems when doing the conversions. You may encounter the following situations:

- If you obtained your coefficient K at experiments conducted at a different temperature from 20°C, you may try to find the corresponding **θ value from the literature** and calculate K for 20°C. In your report, you need to specify clearly: (a) your results: liquid temperature in your experiments, K coefficient obtained at the liquid temperature; (b) your assumption: temperature coefficient θ used, citing the reference; and (c) your calculated estimate of the K_{20} value.
- If you do not find any references for the temperature coefficient, **report only the K coefficient you obtained**, making it very clear that it has been for the liquid temperature of your experiments. Specify what the liquid temperature was during your experiments.
- Alternatively, you may wish to plan **controlled experiments at different temperatures** so that you obtain the K coefficients for each temperature and determine an estimate of the temperature coefficient θ. However, such a strictly controlled experiment is sometimes not simple to undertake. Example 14.5 shows you the estimation of K_{20} based on two K values derived at different temperatures (based on Chapra, 1997).
- Another possibility is that, based on **several results of uncontrolled experiments**, you simultaneously estimate K and θ. However, this is again complex, because both coefficients (K and θ) may be correlated and this may interfere with the convergence procedure in your regression analysis.

In summary, make everything clear in your report. There have been many experimental studies that obtained important estimates of K values, but did not specify the temperature of the experiments, or even whether the K coefficient was for the standard temperature of 20°C or some other temperature. As a result, the usefulness of the research, in terms of reproducibility or applicability for other systems, was limited.

Example

EXAMPLE 14.5 ESTIMATION OF THE REACTION COEFFICIENT K AT THE STANDARD TEMPERATURE OF 20°C BASED ON K VALUES OBTAINED AT DIFFERENT TEMPERATURES

Suppose you completed batch experiments at two different liquid temperatures and obtained the respective estimates of the reaction coefficient K. Estimate the temperature coefficient θ *and use it to estimate the K value at the standard temperature of 20°C.*

Data from the two experiments:

- $T_1 = 17°C$ $K_1 = 0.10\ d^{-1}$
- $T_2 = 24°C$ $K_2 = 0.16\ d^{-1}$

Solution:

(a) Determine the temperature coefficient θ

We can take the $\log_{10}$ of Equation 14.20 and raise the result to a power of 10 to obtain

$$\theta = 10^{(\log K_{T_2} - \log K_{T_1})/(T_2 - T_1)} = 10^{(\log 0.16 - \log 0.10)/(24-17)} = 10^{0.0292} = \mathbf{1.069}$$

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

(b) Determine the *K* coefficient at the standard temperature of 20°C

Equation 14.20 can be rearranged to allow us to estimate the value of K at $T = 20°C$:

$$K_{T_{20}} = K_{T_{17}}\theta^{T_{20}-T_{17}} = 0.10 \times 1.069^{20-17} = \mathbf{0.122\ d^{-1}}$$

Note that the estimation of the coefficient θ based on data from experiments done at only two temperatures is not ideal. Ideally, to get a better estimate of the parameter θ, we should complete experiments at a minimum of five different temperatures and use regression analysis to find the best fit value of θ (see Chapter 11). However, if you have data for two different temperatures, it is better for you to report all the data and be transparent in showing how you estimated the K coefficient for the standard temperature of 20°C.

C. 11

14.3.5 Time to reach a certain removal efficiency

Advanced

Based on the K value obtained from batch experiments, under the *assumption of a first-order reaction*, we can compute the reaction time required to achieve a certain removal efficiency. Rearranging Equation 14.9, we can arrive at the following general equation for estimating this required time:

$$t_E = \frac{\ln\left(\frac{1}{1-E}\right)}{K} = \frac{\ln\left(\frac{100}{100-E\%}\right)}{K} \tag{14.22}$$

where

t_E = time required to achieve a certain removal efficiency in a batch experiment following first-order kinetics (d)
E = removal efficiency (as a fraction, and not percentage) = 1 − remaining fraction
$E\%$ = removal efficiency (as percentage) = 100 − remaining percentage
K = first-order reaction coefficient (d^{-1}).

Based on Equation 14.22, Table 14.3 presents the values of $\ln[100/(100 - E\%)] = t_E \cdot K$ for different values of the removal efficiency ($E\%$, expressed as percentage).

Table 14.3 Values of $\ln[100/(100 - E\%)] = t_E \cdot K$ that allow the estimation of the time required to achieve a certain removal efficiency ($E\%$) for a constituent that follows a first-order reaction in a batch reactor, according to Equation 14.22.

$E\%$	50	70	75	80	85	90	95	99	99.9	99.99	99.999	99.9999
$t_E \cdot K$	0.693	1.204	1.386	1.609	1.897	2.303	2.996	4.605	6.908	9.210	11.513	13.816

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Table 14.4 Values of LRV × ln(10) = $t_E \cdot K$ that allow the estimation of the time required to achieve a certain LRV for a constituent that follows a first-order reaction in a batch reactor, according to Equation 14.23.

LRV	1	2	3	4	5	6
$t_E \cdot K$	2.303	4.605	6.908	9.210	11.513	13.816

For instance, in the previous examples given in this section on first-order kinetics, we used $K = 0.10\ d^{-1}$. With this value, and using Equation 14.22 or Table 14.3, we can see that the time required to achieve an efficiency of, say, 50% is $0.693/0.1 = 6.93$ d and to achieve an efficiency of 80% is $1.609/0.1 = 16.09$ d. You can check the consistency of these values in the various examples we used previously to calculate the decay of the constituent over time.

S. 7.2

If we think about the reduction of *E. coli*, where the efficiencies are expressed as log reduction values (LRVs) (see Section 7.2), we can express Equation 14.22 in terms of LRV:

$$t_E = \frac{\text{LRV} \times \ln(10)}{K} = \frac{\text{LRV} \times 2.3026}{K} \tag{14.23}$$

From Equation 14.23, we prepared Table 14.4, which has a similar structure to Table 14.3, but presents removal efficiencies as LRV instead of removal efficiencies (%).

For instance, to achieve 3 log-units reduction, the necessary time is $3 \times 2.303/K = 6.988/K$ (see Table 14.4). This is the same value as the one obtained from Table 14.3 for $E\% = 99.9\%$. This would be already expected, because LRV = 3 is the same as $E\% = 99.9\%$.

We can also see that, under batch conditions, *if we want to double our LRV, we need to double the required time* (for a given value of the reaction coefficient K). If we want to triple our LRV, we need to triple the required time, and so on.

S. 14.3.5

S. 14.4

We should understand that these considerations made in this Section 14.3.5 are **only applicable to batch reactors**, without inflow and outflow. You will also see in Section 14.4 that these considerations are also applicable to reactors that approach plug-flow hydraulics. However, they do *not* apply to flow-through reactors that have any degree of mixing or dispersion in them.

14.3.6 Applicability of reaction coefficients obtained from experiments done with continuous-flow reactors

Advanced

S. 14.3

Section 14.3 was entirely devoted to the derivation of the kinetic coefficient K based on experiments conducted in a **batch reactor**. The next obvious question is: *can we directly apply the K value obtained from a batch reactor experiment to model the performance of a continuous-flow reactor*?

Unfortunately, ***the answer is no***. The K coefficient we obtained in batch reactors will now be called the '**intrinsic coefficient**' because it represents the pure kinetics of the reaction in a reactor that has no inflows and no outflows. Deriving this intrinsic kinetic coefficient is not sufficient to predict the efficiency of a reaction in a flow-through reactor. For us to be able to estimate the removal of a constituent in a continuous-flow reactor, we need to know *not only the* ***kinetics of the constituent removal*** (intrinsic kinetics) *but also the* ***hydraulics of the reactor***.

S. 14.4

Reactor hydraulics is a very important topic that will be described in Sections 14.4 (*idealized* reactors) and 14.5 (*non-idealized* reactors).

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

14.4 IDEALIZED FLOW REGIMENS IN CONTINUOUS-FLOW REACTORS

14.4.1 General concepts

Advanced

In this section, we will describe the two main *idealized* flow regimes used to model the hydraulic pattern in continuous-flow reactors. These two flow regimes will produce different removal efficiencies for the same intrinsic reaction rate coefficient:

- **Complete-mix reactor**, also called completely stirred tank reactor (CSTR), continuous-flow stirred tank reactors (CFSTR) and completely-mixed flow reactor (CMFR)
- **Plug-flow reactor** (PFR)

They are called **idealized** flow regimes because they model a behaviour that is not found in practice (infinite dispersion in the complete-mix model and zero dispersion in the plug-flow model). Some real-life situations may approach one of these idealized flow regimes, and it may be useful for us to assume one of these idealized cases for the purposes of design or modelling, but we should keep in mind that the real-life hydraulics will not be perfectly represented by these idealized cases. We will present here a simple description of these two hydraulic regimes, given their importance in treatment plant and water quality modelling. Further details can be found in the open access references: von Sperling and Chernicharo (2005) and von Sperling (2007).

Table 14.5 presents a short description of these two main hydraulic regimes, and Table 14.6 lists the main operational characteristics of these reactors, including batch reactors, which have been covered in Section 14.3 (Metcalf & Eddy, 2014; Tchobanoglous & Schroeder, 1985). We will *not cover here reactors that have unsaturated conditions* (reactors with a support medium, with some pore spaces occupied by air).

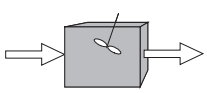
S. 14.3

14.4.2 Idealized plug-flow reactor

Advanced

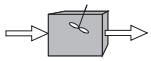
The idealized plug-flow reactor is the one in which each fluid element leaves the tank in the same order of entrance. No single water molecule skips ahead of or falls behind another in the journey through the reactor. The flow occurs as a series of pistons or plugs, moving from upstream to downstream, without mixing

Table 14.5 Characteristics of the two main reactor hydraulic models.

Schematics	Characteristics
Plug-flow	The fluid particles enter the tank continuously at one extremity, pass through the reactor, and are then discharged at the other end, all in the same sequence in which they entered the reactor. The fluid particles move as a piston or a plug, without any longitudinal mixing. The particles maintain their identity and stay in the tank for a period equal to the theoretical hydraulic retention time. This type of flow is *approached* (but not fully matched) in very long reactors with a large length-to-width ratio, in which longitudinal dispersion is minimal. These reactors are also called tubular reactors. Plug-flow reactors are idealized reactors, since the complete absence of longitudinal dispersion is difficult to obtain in practice
Complete-mix 	The particles that enter the tank are immediately dispersed in all the reactor body. The input and output flows are continuous. The fluid particles leave the tank in proportion to their statistical population. Complete-mix can be *approached* in circular or squarish tanks in which the tank's contents are continuously and uniformly distributed. Complete-mix reactors are idealized reactors, since total and identical dispersion is difficult to obtain in practice

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Table 14.6 Operational characteristics of the main idealized reactor types (assuming steady-state conditions).

Reactor type	Schematics	Continuous Flow	Variation of the Concentration with Time (In a Given Position in the Reactor)	Variation of Concentration Along the Reactor (At a Given Time)	Number of Equivalent Complete-mix Reactors	Typical Length/ Width Ratio
Batch reactor		No	Yes	No	–	≈1
Plug-flow		Yes	No	Yes	∞	≫1
Complete-mix		Yes	No	No	1	≈1

S. 14.5.3

Note: For more explanation about the number of equivalent complete-mix reactors, see Section 14.5.3.

between the pistons and without any longitudinal dispersion. Consequently, each water element is exposed to the treatment in the reactor for the exact same amount of time (as in a batch reactor), which is equal to the theoretical hydraulic retention time (HRT) (Arceivala, 1981).

To make this clear, let us explore the analogy between a plug-flow reactor and a batch reactor (Figure 14.7). This is important because you will see in this section that *the kinetic equations for the idealized plug-flow reactor are the same as those for the batch reactor* (which was analysed in detail in the preceding sections of this chapter). Let us hypothesize that a truck is transporting a batch reactor that

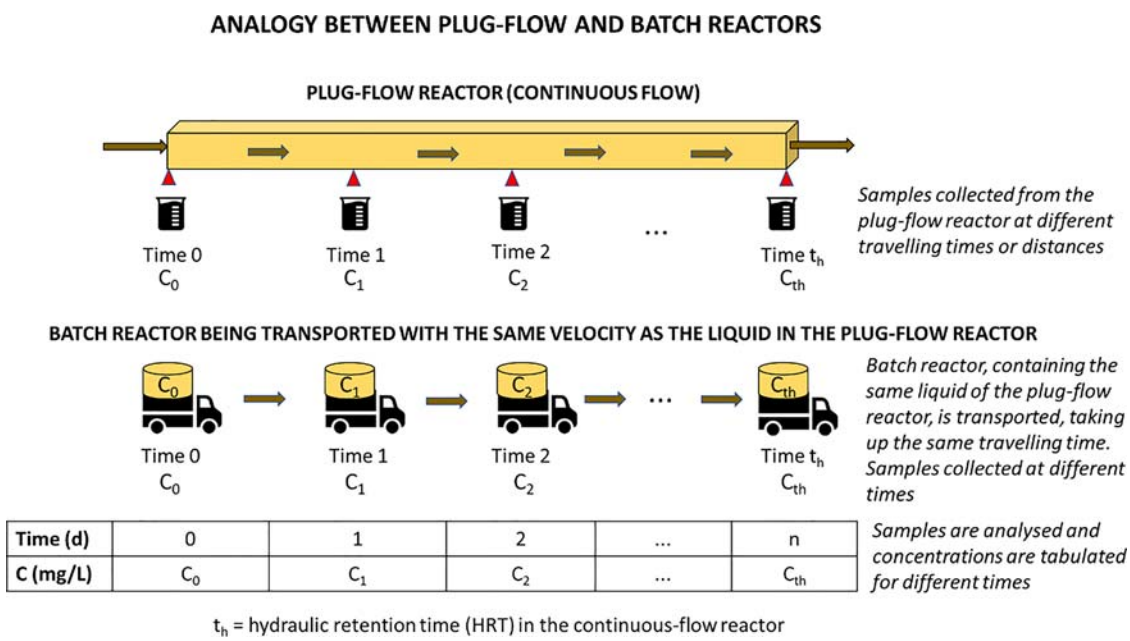


Figure 14.7 Analogy between a plug-flow reactor and a well-mixed batch reactor.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

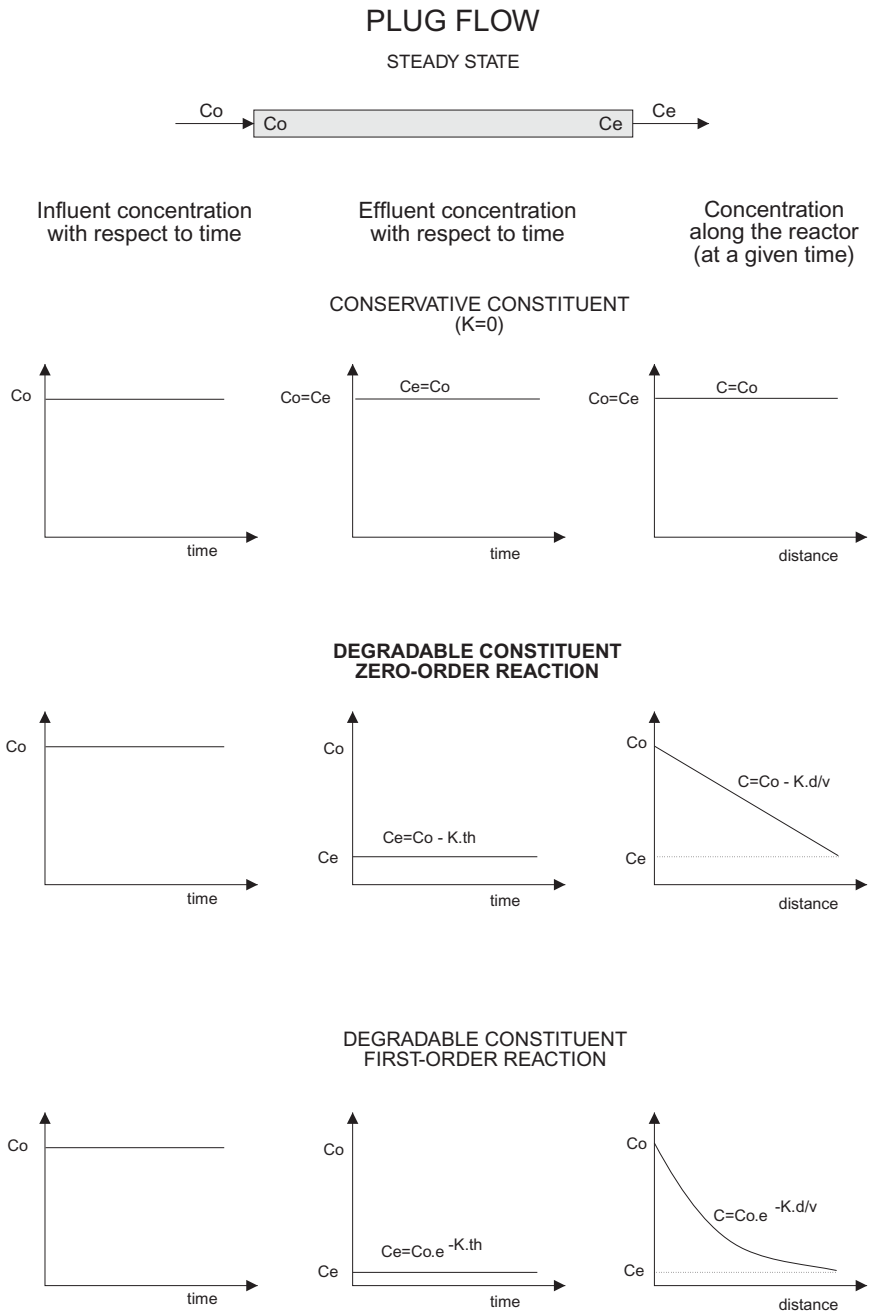


Figure 14.8 Concentration profiles. Ideal plug-flow reactor under steady-state conditions. *Note*: $C =$ concentration at a given time; $C_o =$ influent concentration (also called C_{in} in this book); $C_e =$ effluent concentration (also called C_{out} in this book); $K =$ reaction coefficient; $t_h =$ hydraulic retention time (also called HRT); $d =$ distance (length of the reactor); $v =$ horizontal velocity.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

was filled with the same liquid as the one at the inlet of a plug-flow reactor. The liquid contents are thoroughly mixed in the batch reactor tank. If the truck moves at exactly the same velocity as the liquid in the plug-flow reactor, we can assume that the liquid inside the batch reactor is able to undergo the same reactions as the piston that is moving in the plug-flow reactor. If the conditions are the same, both reactions will be the same. We can understand if you feel uncomfortable with the series of assumptions that we needed to take to make this analogy, but remember that we are talking about idealized reactors and conditions. If you accept this, we can state that, **from the mathematical point of view, a plug-flow reactor behaves like a well-mixed batch reactor**.

Figure 14.8 presents a summary of the concentration profiles with time and position in an ideal plug-flow reactor submitted to a constant influent flow rate and a constant influent concentration (steady-state conditions). If the influent load is varied (dynamic conditions), the derivation of the formula for the plug-flow reactor is more complicated than it is for the complete-mix case. This is because under dynamic conditions, the concentration in the plug-flow reactor varies with respect to time and space in the reactor, while in the complete-mix case, the variation only occurs with respect to time (we have the same concentration at any position within the reactor). That is why complete-mix reactors in series (tanks-in-series) are frequently used to simulate a plug-flow reactor under dynamic (time-varying) conditions, as will be shown in Section 14.5.3.

S. 14.5.3

If the influent (input) flow and concentration are constant, the effluent concentration (output) also remains constant with respect to time. The concentration profile in the tank and, therefore, in the effluent concentration depends on the type and reaction rate of the constituent. Table 14.7 summarizes the main associated equations.

The following generalizations can be made for an idealized plug-flow reactor under steady-state conditions:

- *Conservative substances*: The effluent concentration is equal to the influent concentration.
- *Biodegradable substances with a zero-order reaction*: The removal rate is constant from the inlet to the outlet end of the reactor.
- *Biodegradable substances with a first-order reaction*: Along the reactor, the reaction coefficient (K) is constant, but the concentration decreases gradually while the liquid flows throughout the reactor. At the inlet end of the reactor, the concentration is high, which causes the removal rate to be also high (in first-order reactions, the removal rate is proportional to the concentration). At the outlet end of the reactor, the concentration is reduced and, consequently, the removal rate is lower, that is, more time is required to achieve the same decrease in the concentration.
- *First order or higher reaction orders*: The plug-flow reactor is more efficient than the complete-mix reactor.

Example

EXAMPLE 14.6 ESTIMATION OF THE CONCENTRATION PROFILE AND THE EFFLUENT CONCENTRATION FROM AN IDEALIZED PLUG-FLOW REACTOR

A reactor with extremely elongated dimensions has a volume of 3000 m^3. The influent has the following characteristics: flow = 600 m^3/d; substrate concentration = 200 g/m^3. Calculate the concentration profile along the reactor (assuming an ideal plug-flow reactor under steady-state conditions) for the following situation:

- *Conservative substance ($K = 0$)*
- *Biodegradable substance with first-order removal kinetics ($K = 0.40\ d^{-1}$)*

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Solution:

(a) Hydraulic retention time

The hydraulic retention time (t_h or HRT) is given by

$$t_h = \frac{V}{Q} = \frac{3000\,\text{m}^3}{600\,\text{m}^3/\text{d}} = 5\,\text{d}$$

The travel distance is proportional to the time that has elapsed since the piston or plug first entered the reactor. The piston or plug reaches the end of the reactor when the hydraulic retention time is reached.

(b) Conservative substance

The application of the formula $C = C_0 e^{-K \cdot t}$ for steady-state conditions, with $K = 0$, for various values of t leads to

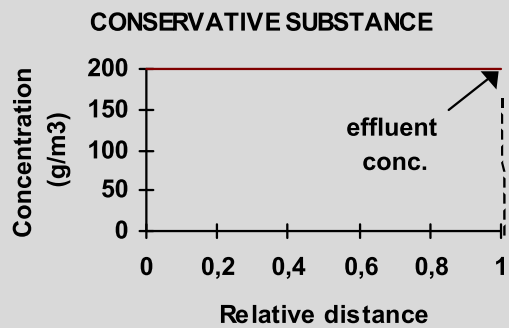
$$C = 200 \cdot e^{-0 \times t}$$

Travel Time (d)	Distance/Total Length	Concentration Along the Tank (g/m^3)
0	0.0	200
1	0.2	200
2	0.4	200
3	0.6	200
4	0.8	200
5	1.0	200

The same values can be obtained through the direct applications of the formula $C = C_o$ (Table 14.7) for conservative substances.

The effluent concentration is the concentration at the end of the hydraulic retention time ($t_h =$ 5 d), that is, 200 g/m^3. The same value can be obtained through the direct application of the formula $C_e = C_0$ (Table 14.7).

The profile of the concentration along the tank is plotted below.



Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

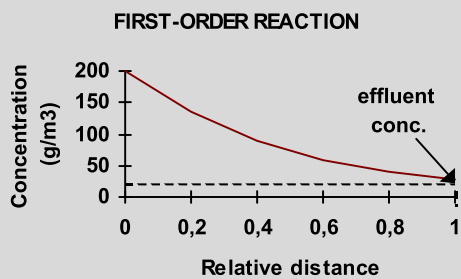
(c) Biodegradable substance (with a first-order reaction)

The application of the formula $C = C_0 e^{-K \cdot t}$ (steady state) for various values of t leads to $C = 200 \cdot e^{-0.40 \times t}$

Travel Time (d)	Distance/Total Length	Concentration Along the Tank (g/m^3)
0	0.0	200
1	0.2	134
2	0.4	90
3	0.6	60
4	0.8	41
5	1.0	27

The effluent concentration is the concentration at the end of the hydraulic retention time ($t_h = 5$ d), that is, 27 g/m^3. The same value can be obtained through the direct application of the formula $C_e = C_0 e^{-K \cdot th}$ (Table 14.7) for first-order reactions under steady-state conditions.

The concentration profile along the tank is plotted below.



Final comment: The estimation of the effluent concentration was made assuming that the reactor behaves like an idealized plug-flow reactor. If K is really the intrinsic kinetic coefficient (obtained from batch experiments), in practice we will observe different experimental results in this continuous-flow reactor, because in real life, *no reactor actually follows the idealized hydraulic behaviour represented by the plug-flow model (though some systems may approach plug-flow)*.

Table 14.7 Ideal plug-flow reactor at steady-state conditions, and equations for the calculation of the concentration along the tank and the effluent concentration.

Reaction	Concentration Along the Reactor (at a given time)	Effluent Concentration
Conservative substance ($r_c = 0$)	$C = C_0$	$C_e = C_0$
Biodegradable substance (zero-order reaction; $r_c = K$)	$C = C_0 - K \cdot d/v$	$C_e = C_0 - K \cdot t_h$
Biodegradable substance (first-order reaction; $r_c = K \cdot C$)	$C = C_0 \cdot e^{-K \cdot d/v}$	$C_e = C_0 \cdot e^{-K \cdot t_h}$

Note: C, concentration at a given point in the reactor (g/m^3 or mg/L); C_0, influent concentration (g/m^3 or mg/L); K, reaction coefficient [(g/m^3)/d for zero-order; d^{-1} for first-order]; d, longitudinal distance along the reactor (m); v, horizontal velocity (m/d); t_h, hydraulic retention time (HRT = volume/flow) (d); r_c, reaction rate of consumption of the constituent [(g/m^3)/d or (mg/L)/d].

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

14.4.3 Idealized complete-mix reactor

Advanced

An idealized complete-mix reactor is one that has continuous-flow and idealized conditions with complete mixing – that is where each water molecule that enters the reactor is instantaneously and totally dispersed. Thus, the reactor contents are considered to be homogeneous, that is, the concentration of any constituent is the same at every single point within the reactor. Likewise, the effluent concentration (C) is the same as the concentration within the reactor.

S. 12.3

The mass balance in the reactor is given by the following equations (see Section 12.3 in Chapter 12):

$$\text{Accumulation} = \text{input} - \text{output} + \text{production} - \text{consumption} \tag{14.24}$$

$$V \cdot \frac{dC}{dt} = Q \cdot C_0 - Q \cdot C + r_p \cdot V - r_c \cdot V \tag{14.25}$$

where

C = concentration at any point in the reactor, equal to the effluent concentration C_{out} (g/m^3 or mg/L)
C_0 = influent concentration C_{in} (g/m^3 or mg/L)
V = reactor volume (m^3)
Q = flow, assuming that inflow = outflow (m^3/d).

Under steady-state conditions, there is no mass accumulation in the reactor, that is, $dC/dt = 0$. In this situation, there is no production of constituents, only the consumption of constituents. Therefore, $r_p = 0$. Dividing the remaining terms by Q, and knowing that HRT $= t_h = V/Q$, the following equations are obtained:

$$0 = Q \cdot C_0 - Q \cdot C - r_c \cdot V \tag{14.26}$$

$$0 = C_0 - C - r_c \cdot t_h \tag{14.27}$$

With the rearrangement of Equation 14.27, concentration profiles along the complete-mix reactor and the effluent concentration under steady-state conditions can be calculated (Figure 14.9).

If the influent (input) flow and concentration are constant, the effluent (output) concentration also remains constant with time. The effluent concentration depends on the reaction rate for the constituent. However, the concentration profile along the reactor depicts a constant concentration, which is in agreement with the assumption that, in a complete-mix reactor, the concentrations are the same at every single point within the tank. Table 14.8 summarises the main equations for an ideal complete-mix reactor. This is unlike the idealized plug-flow reactor, where the concentration is higher at positions closer to the inlet of the reactor and lower at positions closer to the outlet of the reactor.

S. 14.4.4

In comparison with the plug-flow reactor, the effluent concentration is only different for reactions of first order (or higher). For such reaction orders, the complete-mix reactor is less efficient than the plug-flow reactor, for the same hydraulic retention time. This will be further discussed in Section 14.4.4.

The following generalizations can be made for an idealized complete-mix reactor under steady-state conditions:

- *Conservative and biodegradable substances*: The concentration and the removal rate are the same at every point within the reactor. The effluent concentration is thus equal to the concentration within the reactor.
- *Conservative substances*: The effluent concentration is equal to the influent concentration.
- *Biodegradable substances with zero-order reaction*: The effluent concentration is equal to the effluent concentration from a plug-flow reactor with the same retention time (the removal rate is independent of the local constituent concentration).

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

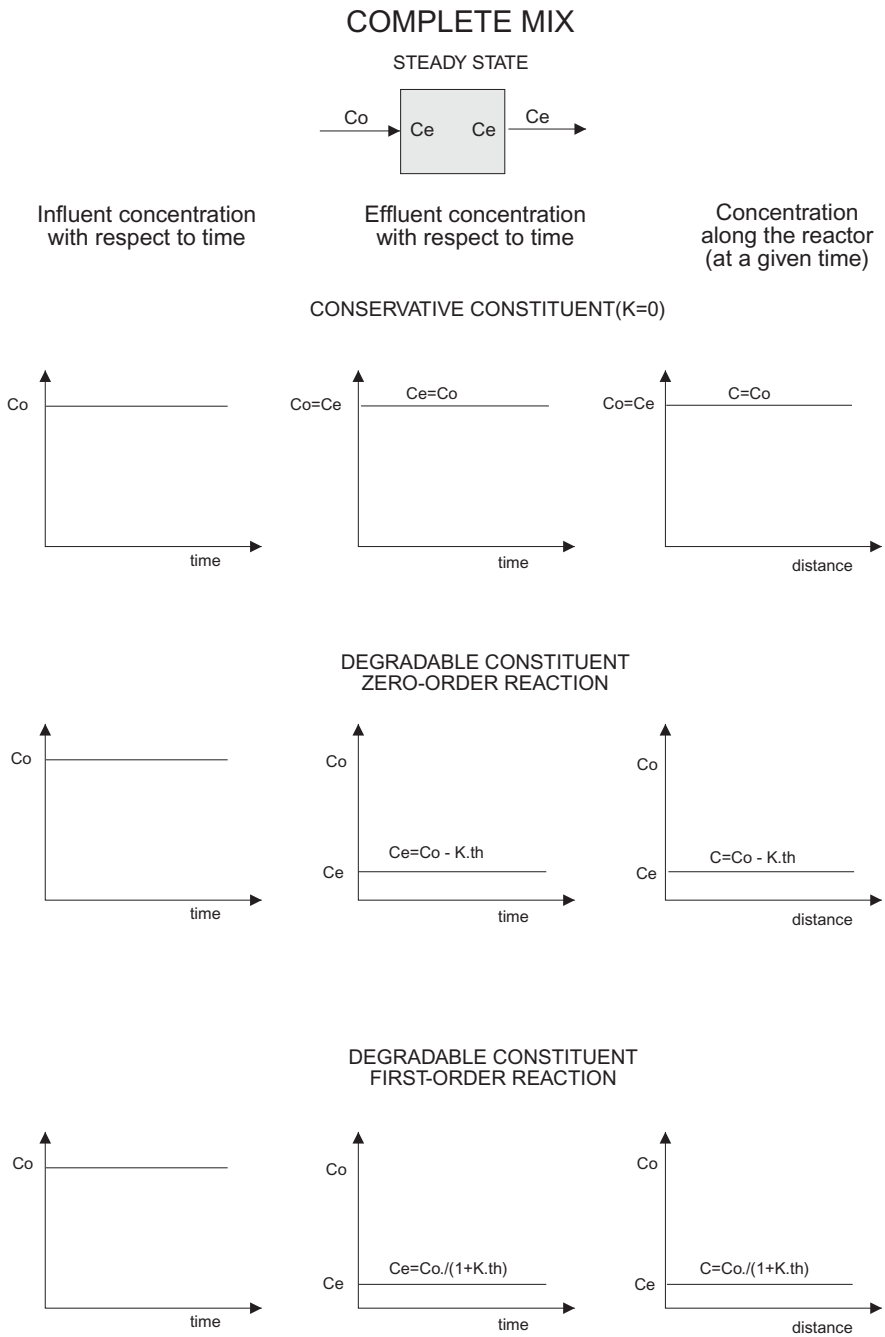


Figure 14.9 Concentration profiles. Ideal complete-mix reactor under steady-state conditions. *Note*: $C =$ concentration at a given time; $C_0 =$ influent concentration (also called C_{in} in this book); $C_e =$ effluent concentration (also called C_{out} in this book); $K =$ reaction coefficient; $t_h =$ hydraulic retention time (also called HRT).

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Table 14.8 Ideal complete-mix reactor at steady-state conditions, and equations for the calculation of the concentration along the tank and the effluent concentration.

Reaction	Concentration Along the Reactor (at a given time)	Effluent Concentration
Conservative substance ($r_c = 0$)	$C = C_0$	$C_e = C_0$
Biodegradable substance (zero-order reaction; $r_c = K$)	$C = C_0 - K \cdot t_h$	$C_e = C_0 - K \cdot t_h$
Biodegradable substance (first-order reaction; $r_c = K \cdot C$)	$C = C_0/(1 + K \cdot t_h)$	$C_e = C_o/(1 + K \cdot t_h)$

Note: C, concentration at a given point in the tank (g/m^3); C_0, influent concentration (g/m^3); K, reaction coefficient [$(g/m^3)/d$ for zero-order; d^{-1} for first-order]; t_h, hydraulic retention time (HRT = volume/flow) (d).

- *Biodegradable substances with first-order or higher-order reactions*: The complete-mix reactor is less efficient than the plug-flow reactor for the same hydraulic retention time. Considering (a) that the removal rate is a function of the local concentration in first- or higher-order reactions and (b) that the concentration at a complete-mix reactor is lower than the average concentration along a plug-flow reactor, then the efficiency of the complete-mix reactor is lower than that of the plug-flow reactor.

Example

EXAMPLE 14.7 ESTIMATION OF THE CONCENTRATION PROFILE AND THE EFFLUENT CONCENTRATION FROM AN IDEALIZED COMPLETE-MIX REACTOR

A reactor with an approximately square shape and good mixing conditions has the same volume as the reactor in Example 14.6 ($3000\ m^3$). The influent also has the same characteristics of that example (flow = $600\ m^3/d$; influent substrate concentration = $200\ g/m^3$). Calculate the concentration profile along the length (relative distance) of the reactor (assuming an ideal complete-mix reactor under steady-state conditions) for the following situations:

- Conservative substance ($K = 0$)
- Biodegradable substance with first-order removal kinetics ($K = 0.40\ d^{-1}$)

Solution:

(a) Hydraulic retention time

The hydraulic retention time is the same as that which was calculated in Example 14.6, that is, $t_h = 5$ days.

(b) Conservative substance

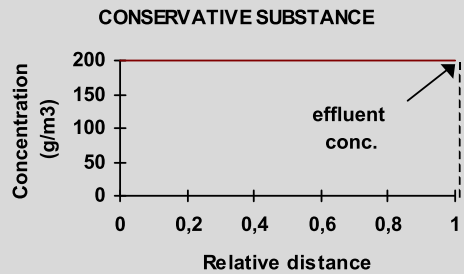
In a complete-mix reactor, the concentration is the same at any point. For a conservative substance, $C = C_0$ (Table 14.8). Hence, for any distance, the concentration is

$$C = 200\ g/m^3$$

The effluent concentration is also equal to $200\ g/m^3$. This value is equal to that which was calculated for the ideal plug-flow reactor.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

The concentration profile along the tank is plotted below.



(c) Biodegradable substance (with a first-order reaction)

At any point in the reactor, the concentration is given by

$$C = \frac{C_0}{1 + K \cdot t_h} = \frac{200}{1 + 0.40 \times 5} = 67\ \text{g/m}^3$$

The effluent concentration is also equal to 67 g/m^3. This value is higher than the value calculated for the plug-flow reactor in Example 14.6 (27 g/m^3), illustrating the fact that a complete-mix reactor is less efficient than a plug-flow reactor, for the same retention time and the same intrinsic (kinetic) reaction rate coefficient.

The concentration profile along the tank is plotted below.

Final comment: The estimation of the effluent concentration was made assuming that the reactor behaves like an idealized complete-mix reactor. If K is really the intrinsic reaction rate coefficient (obtained from batch experiments), in practice we will observe different results, because our existing or designed continuous-flow reactor is not idealized.

S. 14.4.2

S. 14.4.3

S. 14.5.3

If we have **reactors in series**, we can use the equations shown in Sections 14.4.2 and 14.4.3 to predict the effluent concentration from reactor 1, which will be the influent concentration to reactor 2. Then, we use the same equations again and estimate the effluent concentration from reactor 2, which will be the influent concentration to reactor 3, and so on. If all the tanks have the same volume and the removal coefficient K remains the same, we can apply an *overall equation* that takes into account the number of tanks-in-series N. However, we will not present this equation here, because we will preserve it for a different application (see Section 14.5.3).

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

For the *idealized* hydraulic regimens shown here, we can make the following observations **if we split a large reactor into smaller reactors in series** (keeping the same overall volume and the total hydraulic retention time):

- **Idealized plug-flow reactor in series**: *Splitting one plug-flow reactor into smaller idealized plug-flow reactors in series does not alter the removal efficiency* (the liquid piston that travelled and left the first reactor will continue its flow, as a piston, in the second reactor, and so on).
- **Idealized complete-mix reactors in series**: Splitting one complete-mix reactor into smaller complete-mix reactors in series does not alter the removal efficiency if the constituent decays according to zero-order kinetics but increases the removal efficiency if the constituent decays according to first order (or higher) reaction kinetics.
- **Infinite number of complete-mix reactors in series**: If we have an infinite number of reactors in series, we *reproduce the behaviour of a plug-flow reactor* (each infinitesimally small complete-mix reactor behaves like the piston or the plug in the idealized plug-flow reactor).

14.4.4 Deriving kinetic coefficients from existing continuous-flow reactors using idealized hydraulic models (plug-flow and complete-mix)

Advanced

As mentioned before, the theory surrounding reactor hydraulics is explained in detail in our 'open access' companion books (von Sperling & Chernicharo, 2005; von Sperling, 2007), with a stronger emphasis on the design of new systems.

In the present book, our main objective is to apply the fundamental concepts about reactor hydraulics to *assess* existing treatment plants. Therefore, we will now apply this theory to **derive reaction rate coefficients based on monitoring data from an existing continuous-flow reactor** (remember: so far, *we have only seen how to obtain kinetic coefficients based on batch experiments* – see Section 14.3).

S. 14.3

We will only cover **first-order reactions** in this section.

Figure 14.10 shows the strategy you would probably use if you are carrying out a simple experiment. You measure the flow and collect samples from the influent and effluent of a continuous-flow reactor. If you feel you have stable operating conditions, you may take the average of a series of values of the flow (Q), influent concentrations (C_{in} or C_0), and effluent concentrations (C_{out} or C_e). After that, you may assume an idealized hydraulic model to use (plug-flow or complete-mix) and, based on their steady-state equations, you rearrange them to solve for the reaction rate coefficient K'.

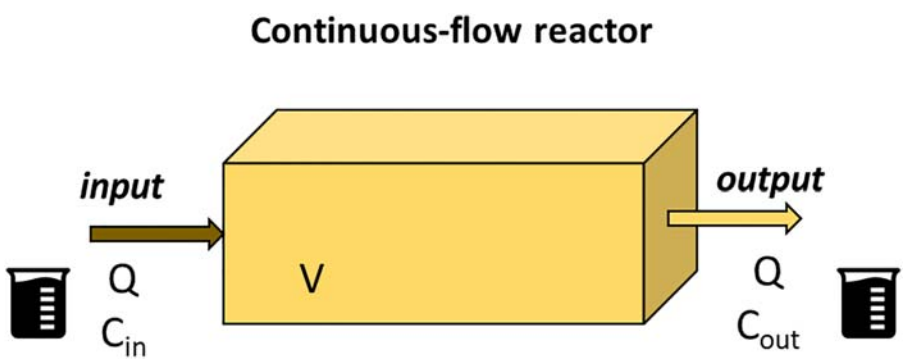


Figure 14.10 Required data for a simple estimation of reaction coefficients from continuous-flow reactors based only on influent and effluent concentrations.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

The related equations are

- **Idealized plug-flow model (first-order reaction)**:

Effluent concentration C_{out}:
$$C_{out} = C_{in} \cdot e^{-K' \cdot t_h} \tag{14.28}$$

Reaction coefficient K':
$$K' = \frac{\ln\left(\frac{C_{in}}{C_{out}}\right)}{t_h} = \frac{-\ln(1-E)}{t_h} \tag{14.29}$$

- **Idealized complete-mix model (first-order reaction)**:

Effluent concentration C_{out}:
$$C_{out} = \frac{C_{in}}{1 + K' \cdot t_h} \tag{14.30}$$

Reaction coefficient K':
$$K' = \frac{\left(\frac{C_{in}}{C_{out}}\right) - 1}{t_h} = \frac{\left(\frac{E}{1-E}\right)}{t_h} \tag{14.31}$$

where

C_{in} = influent concentration, also called C_0 (g/m^3 or mg/L)
C_{out} = effluent concentration, also called C_e (g/m^3 or mg/L)
t_h = hydraulic retention time (d)
K' = reaction coefficient (d^{-1}).

Note that we are now using a **different nomenclature** for the reaction rate coefficient since it is *obtained directly from a continuous-flow reactor*. We are now calling it $\boldsymbol{K'}$ to show that it is different from the 'true intrinsic' kinetic coefficient K which is obtained from batch experiments. Let us explore these differences (von Sperling, 2002).

For a given removal efficiency, the estimation of K' based on the retention time and on the influent and effluent concentrations from an existing reactor leads to the two following divergent situations:

- Adoption of the idealized **complete-mix** model leads to K' values which are **greater** than the intrinsic kinetic coefficient K.
- Adoption of the idealized **plug-flow** model leads to K' values which are **lower** than the intrinsic kinetic coefficient K.

The following example will help to clarify this point. Suppose you monitored an existing reactor and obtained the following average values of a constituent: (a) influent concentration: $C_{in} = 100$ mg/L, (b) effluent concentration: $C_{out} = 10$ mg/L, (c) hydraulic retention time: $t_h = 5$ days. From the influent and effluent concentrations, we can see that the removal efficiency is 90% ($E = 0.90$). If you assume that the reaction is first-order, you can use Equations 14.29 and 14.31 to derive a reaction coefficient (K'_{PF} or K'_{CM}) for the two idealized hydraulic models (plug-flow and complete-mix, respectively):

- **Plug-flow:**

$$K'_{PF} = \frac{\ln\left(\frac{C_{in}}{C_{out}}\right)}{t_h} = \frac{\ln\left(\frac{100}{10}\right)}{5} = \mathbf{0.46\ d^{-1}}$$

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

or

$$K'_{PF} = \frac{-\ln(1-E)}{t_h} = \frac{-\ln(1-0.90)}{5} = \mathbf{0.46\ d^{-1}}$$

- **Complete-mix:**

$$K'_{CM} = \frac{\ln\left(\frac{C_{in}}{C_{out}}\right) - 1}{t_h} = \frac{\ln\left(\frac{100}{10}\right) - 1}{5} = \mathbf{1.80\ d^{-1}}$$

or

$$K'_{CM} = \frac{\left(\frac{E}{1-E}\right)}{t_h} = \frac{\left(\frac{0.90}{1-0.90}\right)}{5} = \mathbf{1.80\ d^{-1}}$$

As it can be seen, for the same reactor, the same influent and effluent concentrations and the same assumed kinetics (first-order), two different K' values are obtained, depending on the hydraulic regime assumed. Which is the correct K' value?

In principle, there should be only one K coefficient, representing the true decay of the constituent, according to its 'intrinsic' kinetics. However, the inadequacy of the idealized hydraulic models for representing the hydraulic pattern in the reactor leads to deviations that occur in practice as we estimate K'. The reason for the differences observed in the example above is that, since idealized complete-mix reactors are the least efficient for first-order removal kinetics, the lower hydraulic efficiency is compensated by a higher K' value. Conversely, since idealized plug-flow reactors are the most efficient reactors, the higher hydraulic efficiency is compensated by a lower K' value.

Depending on the dispersion characteristics in the reactor, these deviations can be very large, inducing considerable errors in our estimate of the true reaction rate coefficient. For instance, if you inadvertently adopt a complete-mix model for a very elongated reactor that would be better represented by a plug-flow equation, you will obtain a K' coefficient that departs considerably from the true intrinsic coefficient K. Conversely, if you adopt a plug-flow model for a well-mixed reactor, you will also obtain K' value that is substantially different from the true intrinsic coefficient K.

These divergences have been the subject of **considerable confusion in the literature**, when expressing K' values. Reported K' values usually show substantial variations, a large part of which can be attributed to inadequate consideration of the hydraulic regime of the reactor. Therefore, if you are reporting reaction coefficients obtained from your experiments, **you need to make everything clear** so that the readers of your work will understand how you obtained the estimated K' value and will have a better idea about the limitations of its application to other systems.

An improvement in your estimates may be achieved if you also collect samples and monitor the concentrations of the constituent at different points **inside the reactor**, instead of simply monitoring at the influent and effluent points. This way, you can make inferences about the behaviour of your constituent with respect to the distance from the inlet to the outlet point, which is a result of both the kinetics and the reactor hydraulics. An example is given in Figure 14.11:

- In the top figure, the reactor is more squarish, with a low length/width ratio. From the samples you collected, you observe that the concentrations were similar, from inlet to outlet, indicating that an approximation to a complete-mix condition would not be far away from reality.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

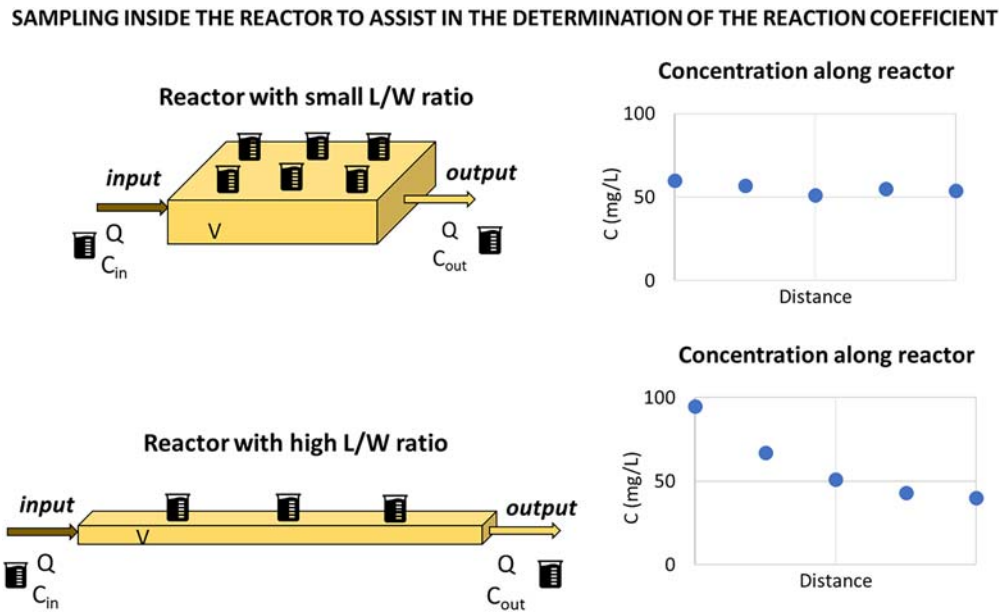


Figure 14.11 Sampling inside the reactor as a means of improving the selection of the hydraulic model and the estimation of the reaction coefficient.

- In the bottom figure, the reactor is elongated, and the samples you collected indicated a decay in the concentration, from inlet to outlet, following the typical exponential curve associated with first-order kinetics. From the two idealized models, the plug-flow would be a much better choice.

From the discussions above, we can see that we would *benefit from having hydraulic models that are not entirely idealized* and that could better reproduce the internal hydrodynamics of our liquid, without resorting to the two extreme assumptions of zero (plug-flow) and infinite (complete-mix) dispersion. This is the subject of Section 14.5, which covers the **plug-flow with dispersion** model and the **apparent tanks-in-series** model.

S. 14.5

14.5 PLUG-FLOW WITH DISPERSION AND APPARENT TANKS-IN-SERIES MODELS

14.5.1 Conversion of the idealized hydraulic models to models that more closely represent reality

Advanced

We saw in Section 14.4 the representation of an existing continuous-flow reactor by the two idealized hydraulic models of plug-flow and complete-mix and the associated limitations of embracing the assumptions of these idealized hydraulic behaviours.

S. 14.4

Now, we will try to get closer to reality and make some adaptations to these two models, leading to two alternative models – namely, the **plug-flow with dispersion** model and the **apparent tanks-in-series** model (see Figure 14.12).

Note the word 'apparent' in the apparent tanks-in-series model: it is really an exercise in abstraction to imagine that a single tank can be represented by a series of tanks; this is why we use the expressions 'apparent' and 'equivalent'. However, this 'imagined' series of tanks is useful, as we will see in

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

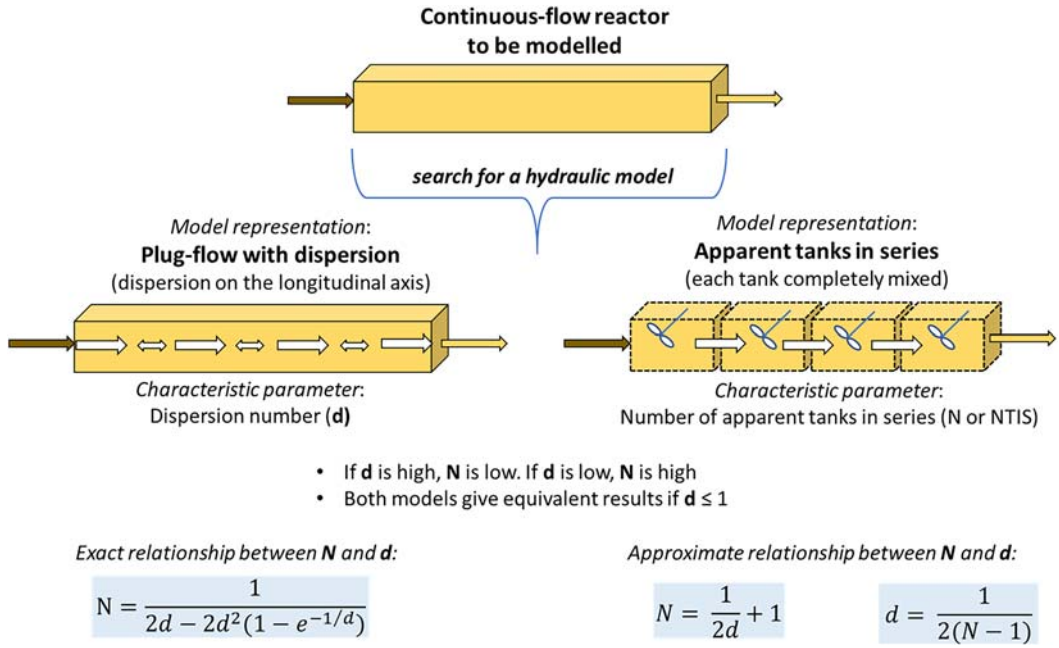


Figure 14.12 Alternative hydraulic representations of a continuous-flow reactor by the plug-flow with dispersion model and apparent tanks-in-series model. *Note*: Dispersion number (d) and number of tanks-in-series (N) are explained in Sections 14.5.2 and 14.5.3. References: Exact relationship between N and d (Levenspiel, 1999); approximate relationship between N and d (Abu-Reesh & Abu-Sharkh, 2003; Elgeti, 1996).

S. 14.5.2

S. 14.5.3

S. 14.5.3

Section 14.5.3. *Both model approaches will yield similar results*, provided the axial dispersion is not too high (provided $d \leq 1$; see below) (Levenspiel, 1999), otherwise we would depart substantially from the underlying assumption of a plug-flow reactor with dispersion.

As seen in Figure 14.12, the main parameters associated with these models are

- Plug-flow with dispersion: *dispersion number* (d)
- Apparent tanks-in-series model: *equivalent number of apparent tanks-in-series (N or NTIS)*

S. 13.2.6

These parameters are obtained by the results of **tracer tests** that must be completed in the reactor being studied. As mentioned in Section 13.2.6, these tests involve adding a conservative *tracer* (*chemical*, *radioactive*, *fluorescent* or *another inert material* or *constituent*) to the inlet and then measuring the distribution of concentrations of that constituent over time at the outlet. This task is laborious, because it involves periodically collecting and analysing samples or measuring effluent concentrations by sensors during a period of approximately three times the theoretical hydraulic retention time. However, it is the best way to obtain the following information from your reactor:

- An estimate of the **dispersion number *d*** for the plug-flow with the dispersion hydraulic model.
- An estimate of the **equivalent NTIS (*N*)** for the apparent tanks-in-series hydraulic model.
- The **true mean hydraulic retention time** (the actual HRT).
- The **volumetric efficiency** (the ratio of the mean HRT and the theoretical HRT; volumetric efficiency is equivalent to the ratio of the 'useful' volume and the total tank volume).

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

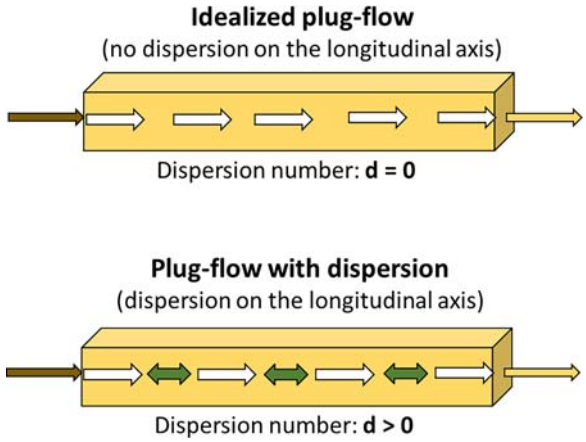


Figure 14.13 Schematic representation of the idealized hydraulic plug-flow model and adaptation to include longitudinal dispersion in the plug-flow with dispersion model.

If you have a value of d, you can convert it into an equivalent value of N. Conversely, if you have a value of N, you can convert it into an equivalent value of d, using the equations shown in Figure 14.12. We will provide more details about these equations in Section 14.5.3.

S. 14.5.3

The description of how to conduct tracer studies is outside the scope of our book. However, this topic is well covered in treatment plant books, including Teefy (1996), Kadlec and Wallace (2009), Metcalf and Eddy (2014) and in chemical reaction engineering textbooks, such as Levenspiel (1999).

14.5.2 Plug-flow with dispersion model

Advanced

The plug-flow with dispersion model is covered in detail in the 'open access' sources von Sperling and Chernicharo (2005) and von Sperling (2007). This hydraulic model is also called the '*dispersed-flow model*', but its characterization as **plug-flow with dispersion** makes it clearer that it is an adaptation of the basic plug-flow model, to take into account the influence of liquid dispersion inside the reactor. In an idealized plug-flow reactor, the 'piston' or 'plug' moves in only one direction, from the inlet to the outlet. However, if there is axial dispersion in the tank, fluid elements may temporarily display other trajectories, including a backward flow toward the inlet (Figure 14.13). As such, one piston or plug (moving forward) may bypass another piston or plug (moving backward). This leads to a residence time distribution (see Section 13.2.1), where some plugs have slightly shorter residence times and other plugs have slightly longer residence times.

S. 13.2.1

S. 14.4

The plug-flow with dispersion model uses the **dispersion number** (d) as its representation of axial or longitudinal dispersion. In the two *idealized* reactors covered in Section 14.4, we have

- *Idealized plug-flow*: **zero dispersion** ($d = 0$)
- *Idealized complete mixing*: **infinite dispersion** ($d = \infty$)

Naturally, reactors found in practice have values of d that are between 0 and ∞. As mentioned before, the value of d can be estimated by tracer tests. Typical values of d or relationships between d and the geometry of the reactor can also be found in the literature (Arceivala, 1981; von Sperling & Chernicharo, 2005) for some reactor types. If you are not undertaking a tracer test, search the literature, but take precautions to consider assuming a d value from a reactor that is similar to the reactor you are studying.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Reactors that have *d* values of *0.2 or less* are closer to *plug-flow*. Conversely, reactors with *d* values of *3.0 or more* can be considered to approach *complete-mixing conditions*. The following factors can affect the extent of dispersion inside a treatment reactor (Arceivala, 1981):

- Scale of the mixing phenomenon
- Geometry of the unit
- Energy introduced per unit volume (mechanical or pneumatic)
- Type and arrangement of the inlets and outlets
- Inflow velocity and its fluctuations
- Density and temperature differences between inflow and reactor contents
- Reynolds number (which is a function of some of the factors listed above).

The analytical solution of the equation for **dispersed flow** (also known as **plug-flow with dispersion**) **for first-order kinetics** was proposed by **Wehner and Wilhem** in 1956. For other reactions that differ from first order, numerical solutions are necessary. The equation for first-order reactions is

$$C_{out} = C_{in} \cdot \frac{4ae^{1/(2d)}}{(1+a)^2 e^{a/(2d)} - (1-a)^2 e^{-a/(2d)}} \qquad (14.32)$$

$$a = \sqrt{1 + 4K' \cdot t_h \cdot d}$$

where

d = dispersion number (–)
t_h = hydraulic retention time (d)
K' = reaction coefficient (d^{-1})
C_{in} = influent concentration (g/m^3)
C_{out} = effluent concentration (g/m^3).

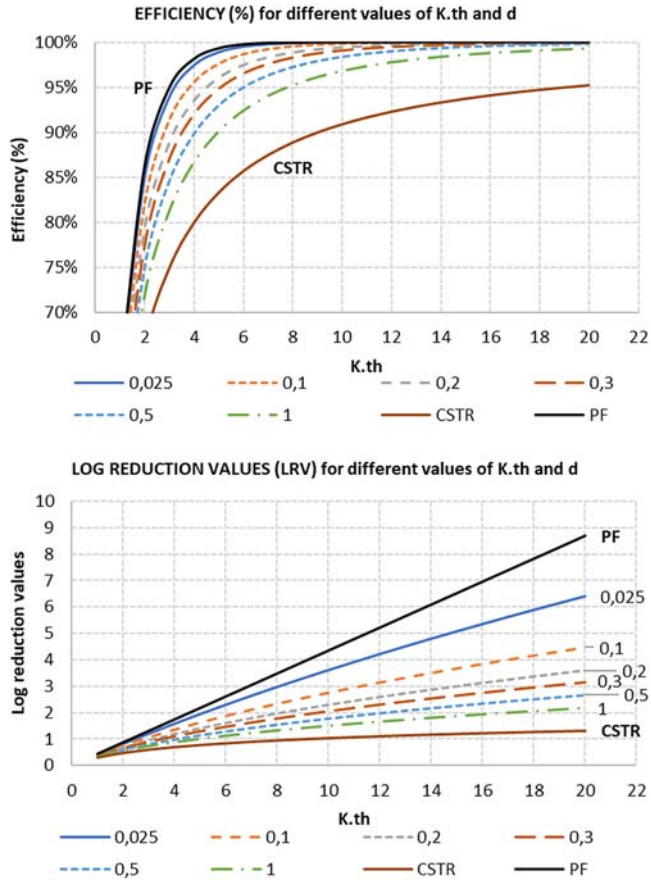
If you want to calculate the removal efficiency *E*, you can algebraically rearrange Equation 14.32 to produce the following equation (the intermediate parameter 'a' will be the same as the one calculated in Equation 14.32):

$$E = 1 - \frac{4ae^{1/(2d)}}{(1+a)^2 e^{a/(2d)} - (1-a)^2 e^{-a/(2d)}} \qquad (14.33)$$

The advantage of these equations is that they allow a continuous solution for different dispersions situated between the idealized limits of plug-flow and complete-mix. When *d* is small, Equation 14.32 gives results very close to the specific equation for the plug-flow idealized case (Equation 14.28). On the other hand, when *d* is very large, Equation 14.32 produces similar values to those obtained from the equation for the complete-mix idealized case (Equation 14.30).

The interpretation of the Wehner–Wilhem equation can be facilitated by the use of graphs, such as the one presented in Figure 14.14. Typically, these graphs have the dimensionless product $K' \cdot t_h$ in the *X*-axis and the removal efficiency in the *Y*-axis. The graph plots a family of curves, each one for a different value of the dispersion number *d*, varying from 0 (plug-flow – PF) to ∞ (complete-mix – CSTR). The purpose of the graph in terms of its application is for you to get a rough visualized idea

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest



Excel

Figure 14.14 Relationship between removal efficiency, the dimensionless pair $K' \cdot t_h$, and the dispersion number d. Top: removal efficiencies as percentages (%). Bottom: log reduction values (LRV).

about the influence of d and $K' \cdot t_h$ on the removal efficiency (or the log reduction value). Some important observations about this are summarized below:

- For an existing reactor with a known volume, if you know the values of K', t_h, and d, then you can estimate the removal **efficiency** using the graph.
- For an existing reactor with a known volume, if you know the values of t_h and d and you have monitoring data that allow you to estimate the removal efficiency for a particular constituent, then you can obtain an estimate of the **reaction coefficient** K'. Just start on the Y-axis at your estimated value of E, travel horizontally until you find the curve with your estimated value of d and then descend vertically to the X-axis, where you will find the value of the product $K' \cdot t_h$. By knowing t_h, you calculate K' by dividing $K' \cdot t_h$ by t_h.

The graphs in Figure 14.14, while useful to help understand the concepts, do not lead to a sufficient precision in your estimates, and so it is better to work with the equations directly. The complex structure of Equation 14.32 can be easily managed in a **spreadsheet**, such as Excel. However, one difficulty remains if you want to rearrange this equation to *estimate the reaction*

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

coefficient K′ from an existing continuous-flow reactor, based on monitoring data for the flow and the influent and effluent concentrations, together with an assumed value for the dispersion number *d*. This equation cannot be rearranged in a way that allows you to solve for *d* directly. However, you can still find a solution for *d* by using the Solver add-in tool in Excel, as illustrated in Example 14.8.

Example

EXAMPLE 14.8 ESTIMATION OF THE K′ COEFFICIENT FOR THE PLUG-FLOW WITH DISPERSION MODEL

S. 14.4.4

You want to estimate the reaction coefficient K' for the plug-flow with dispersion model under the assumption of a first-order reaction, based on monitoring data from a continuous-flow reactor. Based on your monitoring data, you obtained the following information (equal to the values used in Section 14.4.4): (a) influent concentration: $C_{in} = 100$ mg/L; (b) effluent concentration: $C_{out} = 10$ mg/L; (c) hydraulic retention time: $t_h = 5$ days. From the influent and effluent concentrations, you calculated the removal efficiency to be 90% ($E = 0.90$). From the literature, you saw indications that a good estimate for your dispersion number *d* could be 0.40.

Estimate the reaction coefficient K' using an iterative procedure with the Excel Solver tool.

Excel

Note: This example is also available as an Excel spreadsheet.

Solution:

Using a rearrangement of Equation 14.32 and the Excel spreadsheet, after running the Solver tool, we obtain the following value of K' for plug-flow with dispersion:
$\boldsymbol{K' = 0.76\ d^{-1}}$ (plug-flow with dispersion)

S. 14.4.4

Note: In Section 14.4.4, we obtained the following K' values for the two idealized hydraulic models:

$K' = 0.46\ d^{-1}$ (idealized plug-flow)
$K' = 1.80\ d^{-1}$ (idealized complete-mix)

We can see that the estimated value of the K' coefficient for the plug-flow with dispersion model is in between the estimated coefficients derived for the two idealized flow regimes. If the assumption of the dispersion number $d = 0.40$ is close to reality, that is, if the hydraulic model is a close descriptor of the real hydraulic behaviour of the reactor, the resulting K' coefficient estimated using the plug-flow with dispersion model will approach the true 'intrinsic' value of the kinetic coefficient *K*.

14.5.3 Apparent number of tanks-in-series (NTIS) model

Advanced

The tanks-in-series model is also covered in the 'open access' sources von Sperling and Chernicharo (2005) and von Sperling (2007). These other books are primarily focussed on applying this model to design new reactors that are arranged in series. As stated in those books, *based on our knowledge of the reaction coefficient K′ and the adoption of a certain number of reactors in series (N), you calculate the required hydraulic retention time t_h in each reactor (and hence the volume of each reactor) for obtaining a given desired removal efficiency.*

In this section, we will use this same theoretical background, but now with a different application – in coordination with the focus of this book – that is the **derivation of the reaction coefficient K'** from an existing single reactor that you are studying or monitoring.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

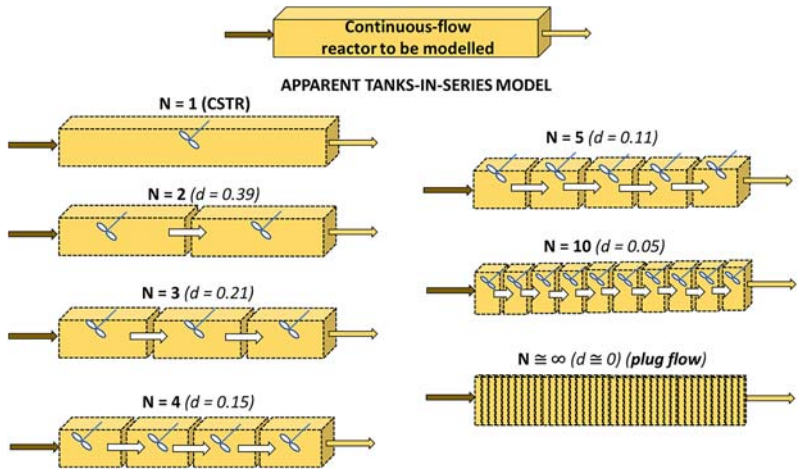


Figure 14.15 Schematic representation of one reactor as N apparent tanks-in-series (NTIS model). The values of N and d provide an indication of the number of tanks-in-series (N) and the corresponding dispersion number (d) for the plug-flow with dispersion model.

You can represent the hydraulics of an existing reactor with a certain number of apparent tanks in series (N or NTIS) and calculate the reaction coefficient K′ based on your observed influent and effluent concentrations (or removal efficiency).

We will first describe some principles of reactors in series, before moving into the application of deriving the coefficient K'. We can say the following about dividing a single reactor into one or more 'apparent' or 'imagined' complete-mix reactors in series:

- If the total volume of the reactor is distributed into an *intermediate number of cells (or tanks)*, the system simulates **dispersed-flow** conditions. When the reactor is subdivided into very few cells, the system is closer to complete-mix. When the reactor is subdivided into a larger number of cells, it is closer to plug-flow conditions.
- If the total volume is distributed in *only one complete-mix reactor*, the system is the same as a conventional idealized **complete-mix** reactor (CSTR reactor).
- Conversely, when the total volume of the reactor is distributed into an *infinite number of complete-mix reactors in series*, the system is equivalent to a single idealized **plug-flow reactor**.

If you are familiar with the content of von Sperling and Chernicharo (2005) and von Sperling (2007), you will know that for design purposes, the possibility of having tanks with different volumes in series was considered. However, in our book here, when using the 'apparent' tanks-in-series model, we assume that *all tanks in the series have the same volume and the same hydraulic retention time, with the cumulative volume and cumulative hydraulic retention time equal to that of the reactor being modelled.*

Recall, we are depicting here a **single reactor** which is being imagined or represented as an *apparent series of smaller complete-mix reactors*. Figure 14.15 illustrates some of the possibilities for representing a reactor as N equivalent tanks-in-series ($N = 1, 2, 3, 4, 5, \ldots, 10, \ldots, \infty$). In the figure, we have also listed the corresponding dispersion number (d) for the plug-flow with dispersion model (Section 14.5.2), for you to make a comparison.

S. 14.5.2

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

The **concentration in the final effluent** of a *series of equal-sized reactors* is given by the following equations:

- **Conservative constituent** (non-biodegradable):

$$C_{out} = C_{in} \tag{14.34}$$

- **Constituent being removed according to a zero-order reaction**:

$$C_{out} = C_{in} - K' \cdot t_h \tag{14.35}$$

- **Constituent being removed according to a first-order reaction**:

$$C_{out} = \frac{C_{in}}{\left(1 + K' \cdot \left(\frac{t_h}{N}\right)^N\right)} \tag{14.36}$$

where

N = number of apparent tanks in series (–)
t_h = total hydraulic retention time (d)
K' = reaction coefficient [$(g/m^3)/d$ for the zero-order reaction and d^{-1} for the first-order reaction]
C_{in} = influent concentration (g/m^3)
C_{out} = effluent concentration (g/m^3).

We can see from Equation 14.34 that the final effluent concentration of a **non-biodegradable** constituent is equal to the influent concentration. Also, from Equation 14.35, we can see that the final effluent from a system of N tanks-in-series with a **zero-order reaction** is equal to that from a single complete-mix reactor (with a volume equal to the total volume of all the tanks-in-series) (see Table 14.8). Additionally, it must be noted that this final effluent is also equal to the effluent from a plug-flow reactor (see Table 14.7). This is as expected, considering that in zero-order reactions, the removal rate is independent of the concentration.

We will now devote most of our discussion to the first-order reaction model, which is being analysed in more detail in this chapter. The **removal efficiency** of a constituent that decays according to a **first-order reaction** in a *series of equal-size reactors*, under steady-state conditions, is given by

$$E = 1 - \frac{1}{\left(1 + K' \cdot \left(\frac{t_h}{N}\right)^N\right)} \tag{14.37}$$

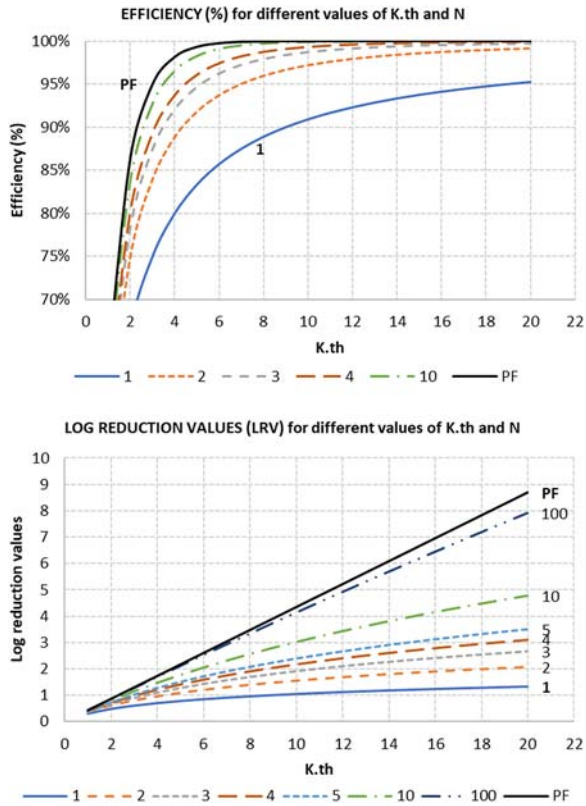
$$\text{LRV} = -\log_{10}(1 - E) \tag{14.38}$$

where

E = removal efficiency = $(C_{in} - C_{out})/C_{in}$
LRV = log reduction value.

Figure 14.16 presents the **removal efficiencies** for first-order kinetics obtained in a reactor that is represented by N apparent equal-sized complete-mix tanks-in-series, as a function of the dimensionless product $K' \cdot t_h$. The great influence of the number of tanks is clearly seen: the higher the number of tanks (N), the greater the efficiency. One tank ($N = 1$) corresponds to the traditional idealized complete-mix

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest



Excel

Figure 14.16 Removal efficiencies for first-order kinetics in a reactor represented by N apparent equal-sized complete-mix tanks-in-series, as a function of the dimensionless product $K' \cdot t_h$. Top: removal efficiencies (E, in %). Bottom: log reduction values (LRV).

reactor (see Equation 14.30). PF stands for plug-flow and represents the idealized plug-flow model, which is the same as a situation of infinite tanks-in-series (see Equation 14.28). The removal efficiencies are also expressed as log reduction values in the bottom graph.

In order to further understand the behaviour of reactors in series, let us analyse Example 14.9, in which a single reactor is represented by different numbers of tanks-in-series ($N = 1$, 2, 5, and 10).

Example

EXAMPLE 14.9 ESTIMATION OF THE CONCENTRATION PROFILE AND THE EFFLUENT CONCENTRATION FROM REACTORS IN SERIES

You want to estimate the effluent concentration and the longitudinal concentration profile of a constituent in a reactor, using the NTIS model for a first-order reaction under steady-state conditions. You have the following information: $C_{in} = 100$ mg/L; total hydraulic retention time: $t_h = 20$ days; removal coefficient $K' = 0.10\ d^{-1}$.

Excel

Note: This example is also available as an Excel spreadsheet.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Solution:

(a) Effluent concentrations

Using Equation 14.28 (*idealized* plug-flow reactor), Equation 14.30 (*idealized* complete-mix reactor), and Equation 14.36 (apparent number of tanks-in-series), you can estimate the effluent concentrations:

- *Idealized plug-flow reactor:*

$$C_{out} = C_{in} \cdot e^{-K' \cdot t_h} = 100 \times e^{-0.10 \times 20} = 13.5\,\text{mg/L}$$

- *Idealized complete-mix reactor:*

$$C_{out} = \frac{C_{in}}{1 + K' \cdot t_h} = \frac{100}{1 + 0.10 \times 20} = 33.3\ \text{mg/L}$$

- *Apparent number of tanks-in-series (for N = 1, 2, 5, and 10):*

$$\boldsymbol{N = 1} \rightarrow C_{out} = \frac{C_{in}}{(1 + K' \cdot (t_h/N))^N} = \frac{100}{(1 + 0.10 \times (20/1))^1} = \mathbf{33.3\,mg/L}$$

$$\boldsymbol{N = 2} \rightarrow C_{out} = \frac{C_{in}}{(1 + K' \cdot (t_h/N))^N} = \frac{100}{(1 + 0.10 \times (20/2))^2} = \mathbf{25.0\,mg/L}$$

$$\boldsymbol{N = 5} \rightarrow C_{out} = \frac{C_{in}}{(1 + K' \cdot (t_h/N))^N} = \frac{100}{(1 + 0.10 \times (20/5))^5} = \mathbf{18.6\,mg/L}$$

$$\boldsymbol{N = 10} \rightarrow C_{out} = \frac{C_{in}}{(1 + K' \cdot (t_h/N))^N} = \frac{100}{(1 + 0.10 \times (20/10))^{10}} = \mathbf{16.2\,mg/L}$$

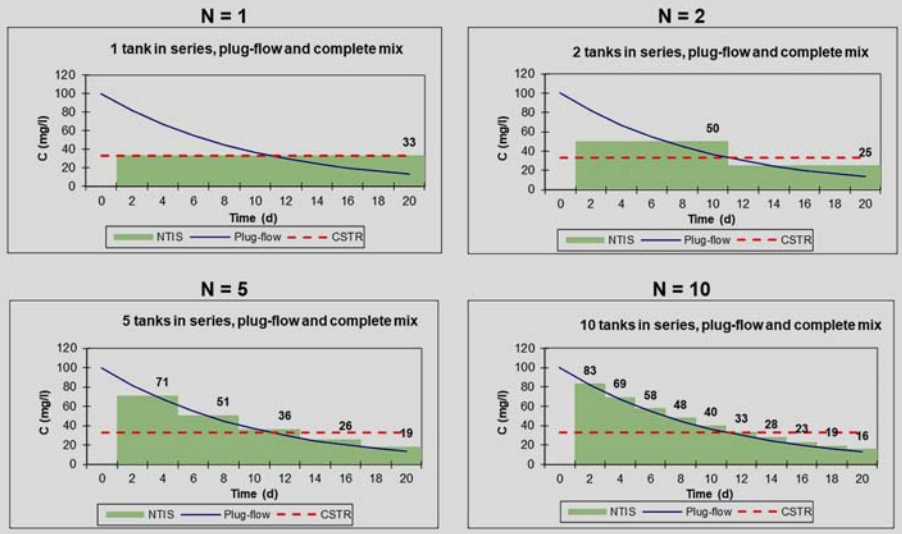
The models using apparent numbers of tanks-in-series of 2, 5, and 10 gave intermediate results between the two idealized reactor models (plug-flow and complete-mix) and are likely to be better descriptors of the reality in the reactor (which can be determined by implementing a tracer test). If you have monitoring data of effluent concentrations, you can infer which value of N leads to the best fit between the estimated and observed values (provided you have confidence in your estimate of the true intrinsic kinetic coefficient K).

(b) Longitudinal concentration profile

You can obtain a better understanding of the behaviour of your reactor if you analyse the change in concentration along the longitudinal axis, from inlet to outlet. The Excel spreadsheet calculates and plots these concentration profiles. In this case, we are assuming that the travelling time is proportional to the travelling distance (constant horizontal velocity).

The following graphs plot the profiles for the two idealized flow regimes (plug-flow and complete-mix) and the apparent number of tanks-in-series ($N = 1$, 2, 5, and 10). The idealized flow regimes are plotted as lines, and the tanks-in-series profiles are plotted as a stepwise decay. The numbers inside the graphs correspond to the values prevailing in each of the reactors in series. The last number is the final effluent, calculated by the NTIS model.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest



You can clearly see that as the number of tanks-in-series increases, the more the longitudinal profile departs from the complete-mix model and approaches the plug-flow model.

If you have monitored the longitudinal profile of the constituent, you will be able to make a much better inference about the appropriateness of the NTIS model and the most adequate number of tanks-in-series compared with the situation in which you have monitored only effluent concentrations.

14.5.4 Deriving kinetic coefficients from existing continuous-flow reactors using the plug-flow with dispersion and the apparent number of tanks-in-series models

Advanced

S. 14.5.2

(a) Determination of K' based on influent and effluent concentrations

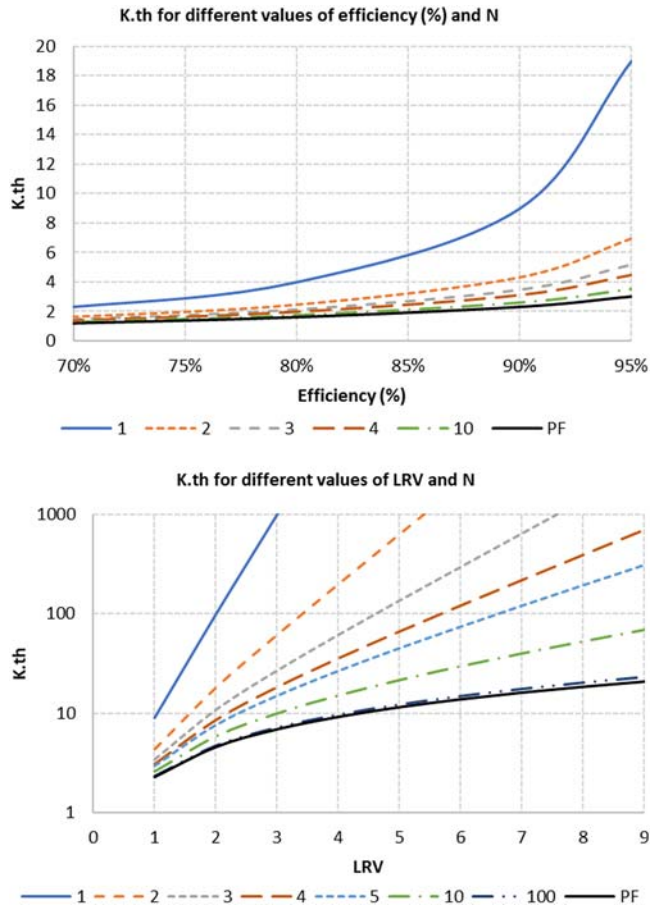
In Section 14.5.2, we saw how to estimate the reaction coefficient K' in the case of the **plug-flow with dispersion model**, if you have *data for the influent (C_{in}) and effluent (C_{out}) concentrations, or the removal efficiencies (E)*. There we observed that there is *no explicit solution* for deriving K' from the general equation due to its complexity. However, we were able to circumvent this limitation by using the Excel Solver tool (see Example 14.8).

Now let us see how to estimate K' when representing our reactor by the **apparent tanks-in-series model** and also *based only on values of C_{in} and C_{out} (or E)*. In this case, we have *explicit solutions*, given by rearrangement of Equations 14.36 and 14.37, that lead to the value of the dimensionless **product $K' \cdot t_h$**:

- Estimation of $K' \cdot t_h$ as a function of N, C_{in}, and C_{out}:

$$K' \cdot t_h = N[(C_{\text{in}/C_{\text{out}}})^{(1/N)} - 1] \tag{14.39}$$

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest



Excel

Figure 14.17 Product $K' \cdot t_h$ for different values of removal efficiency (expressed as percentage or LRV values) and apparent equal-sized complete-mix tanks-in-series (N) for first-order kinetics.

- Estimation of $K' \cdot t_h$ as a function of N and E:

$$K' \cdot t_h = N[(1 - E)^{(-1/N)} - 1] \tag{14.40}$$

- Estimation of $K' \cdot t_h$ as a function of N and LRV:

$$K' \cdot t_h = N[(10^{-\mathrm{LRV}})^{(-1/N)} - 1] \tag{14.41}$$

S. 14.5.2

If you are using the *model with a residual* (refractory, persistent, or non-biodegradable) concentration C^* (see Section 14.3.2), you should use $(C_{in} - C^*)$ and $(C_{out} - C^*)$ in Equation 14.39, and $E = [(C_{in} - C_{out})/(C_{in} - C^*)$ in Equation 14.40.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Using the equations above, you calculate the product $K' \cdot t_h$. Note that this product is shown in several equations and graphs used in this chapter. The usefulness of this calculation can be understood as:

- **If you have t_h, you can calculate K'**. This is the calculation we are interested in this book, allowing the estimation of the reaction coefficient K' based on the total hydraulic retention time in your reactor.
- **If you have K', you can calculate t_h**. This is the calculation we use in designs of new systems (e.g., von Sperling & Chernicharo, 2005; von Sperling, 2007). Based on reported or literature values of the reaction coefficient, you estimate the total hydraulic retention time and hence the required reactor volume ($V = t_h \cdot Q$).

The relationship between the removal efficiency, N, and $K' \cdot t_h$ (Equations 14.40 and 14.41) can be visualized in Figure 14.17. From this figure, you can see why there is so **much confusion in the literature regarding K' values obtained from continuous-flow reactors**: for a given removal efficiency (E or LRV), depending on the number of N you adopt (including $N = 1$ for idealized complete-mix and $N = \infty$ for idealized plug-flow), you end up with completely different values of $K' \cdot t_h$, and hence K' (for a given value of the hydraulic retention time t_h in your reactor). Therefore, remember: *if you calculated K' based on the NTIS model*, you have to **report the number of tanks-in-series (N) you adopted**.

We have mentioned that the representation of a reactor using the *plug-flow with dispersion* model (based on d) is equivalent to the representation of the reactor using the *apparent tanks-in-series model* (based on N). Therefore, it is important to know the **relationship between d and N**. From Levenspiel (1999), we obtain the following equation:

$$N = \frac{1}{2d - 2d^2(1 - e^{-1/d})} \tag{14.42}$$

This equation gives the relationship between N and d, for d up to 1.0 (Levenspiel, 1999). Due to its complexity, there is no explicit solution for obtaining d as a function of N. Table 14.9 presents the resulting values of N as a function of d (directly calculated from Equation 14.42), and Table 14.10 presents the values of d associated with different integer values of N (calculated

Table 14.9 Number of apparent tanks-in-series (N) for different values of the dispersion number d, calculated from Equation 14.42.

d	*N*	*d*	*N*	*d*	*N*
0.01	50.51	0.15	3.92	0.60	1.62
0.02	25.51	0.20	3.12	0.65	1.57
0.03	17.18	0.25	2.65	0.70	1.53
0.04	13.02	0.30	2.35	0.75	1.49
0.05	10.53	0.35	2.13	0.80	1.46
0.06	8.87	0.40	1.98	0.85	1.43
0.07	7.68	0.45	1.86	0.90	1.40
0.08	6.79	0.50	1.76	0.95	1.38
0.09	6.10	0.55	1.69	1.00	1.36
0.10	5.56				

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Table 14.10 Dispersion number d for different values of the number of apparent tanks-in-series (N), calculated using the Solver tool in Equation 14.42.

N	d	N	d	N	d
1		10	0.0528	19	0.0270
2	0.3911	11	0.0477	20	0.0257
3	0.2107	12	0.0436	21	0.0244
4	0.1464	13	0.0401	22	0.0233
5	0.1127	14	0.0371	23	0.0222
6	0.0918	15	0.0345	24	0.0213
7	0.0774	16	0.0323	25	0.0204
8	0.0670	17	0.0303		
9	0.0590	18	0.0286		

using the Solver tool applied to Equation 14.42). Note that Equation 14.37 works even for non-integer values of N. Therefore, even though physical tanks-in-series can only have integer values of N, when we use the apparent tanks-in-series model (Equation 14.37), we can adopt a non-integer value for N (but make this clear in your report).

You can also utilize the **approximate** simple relationships mentioned by Elgeti (1996), shown in Equations 14.43 and 14.44. These equations provide a good fit with the values shown in Tables 14.9 and 14.10, with a difference of less than 15% in the values of N.

$$N = \frac{1}{2d} + 1 \tag{14.43}$$

$$d = \frac{1}{2(N-1)} \tag{14.44}$$

Example

EXAMPLE 14.10 ESTIMATION OF THE K′ COEFFICIENT FOR THE APPARENT TANKS-IN-SERIES MODEL

S. 14.4.4

Suppose that you want to estimate the reaction coefficient K' using the apparent tanks-in-series model under the assumption of a first-order reaction at steady state, using monitoring data from a continuous-flow reactor. Based on your monitoring data, you obtained the following information (equal to those used in Section 14.4.4 and, especially, Example 14.8): (a) influent concentration: $C_{in} = 100$ mg/L; (b) effluent concentration: $C_{out} = 10$ mg/L; (c) hydraulic retention time: $t_h = 5$ days. From the influent and effluent concentrations, you calculated the removal efficiency to be 90% ($E = 0.90$). From the literature, you saw indications that a good estimate for your dispersion number d could be 0.40.

Estimate the reaction coefficient K' by the direct application of Equation 14.40.

Solution:

(a) Estimation of the apparent number of tanks-in-series *N*

You can estimate N based on the value of the dispersion number d. Using Equation 14.42 or Table 14.9, with $d = 0.40$, we obtain $N = 1.98$. In our case, we will round it up to $\mathbf{N = 2}$ (two tanks-in-series), but you can also use the value of $N = 1.98$ directly in the equations.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

(b) Estimation of the reaction coefficient K'

From Equation 14.40, with $E = 0.90$ and $N = 2$, we have

$$K' \cdot t_h = N[(1 - E)^{(-1/N)} - 1] = 2 \times [(1 - 0.90)^{(-1/2)} - 1] = 4.32$$

This is the value of the product $K' \cdot t_h$. Knowing that $t_h = 5$ days, we can estimate K' as

$$K' = \frac{4.32}{5} = \mathbf{0.86\ d^{-1}}$$

The value of K' found in Example 14.8, based on the plug-flow with dispersion model, was 0.76 d^{-1}. For most practical purposes, given all the uncertainty we have in our input data, we could consider that 0.86 and 0.76 are close enough.

We had mentioned before that both approaches are similar and should lead to approximately the same values of K'. However, when d increases, the difference between the estimated K' values using both approaches also increases, and this is why Levenspiel (1999) recommends that we should consider the applicability of either models only up to $d \leq 1.0$. For values of $d > 1.0$, we should use the apparent tanks-in-series model.

(b) Determination of K' based on longitudinal concentration profiles along the reactor length

In Section 14.4.4, we stressed the fact that the derivation of kinetic coefficients using a series of measurements made along the longitudinal axis of the reactor is much better than that based on the measurements of inlet and outlet concentrations alone. This is certainly the case with *non-idealized* hydraulic models (plug-flow with dispersion and apparent tanks-in-series). In this section, we will use an example (Example 14.11) in which samples have been collected at different lengths along the reactor, and then we will fit the *plug-flow with dispersion model*, using Equation 14.32 (see Section 14.5.2).

S. 14.5.2

In the example, we make use of the Solver tool from Excel in order to maximize the goodness-of-fit of the model (as measured by the Coefficient of Determination – see Chapter 15). For this, we allow the following two model parameters to vary in order to try to obtain the best possible fit: dispersion coefficient (d) and reaction coefficient (K').

C. 15

In the associated Excel spreadsheet, we have made a series of comments related to the difficulties associated with the simultaneous estimation of two model parameters, especially if they are potentially correlated: we may end up with similar values of the Coefficients of Determination for different values of d and K'. Therefore, please remember that you must use good judgement when assessing your results, to decide if the model parameters make sense, from a physical and practical point of view. You may be better off using the Solver tool to vary only K' and keep d as a fixed number, based on tracer tests or literature values for reactors with similar geometries and operating conditions.

Our expectation is that, by estimating the reaction coefficient K' based on measurements done along the reactor length, we will adopt a more suitable hydraulic model for the reactor, and our estimated value of K' will be closer to the intrinsic coefficient K (obtained by batch experiments).

Example

EXAMPLE 14.11 ESTIMATION OF THE K′ COEFFICIENT FOR THE DISPERSED-FLOW MODEL BASED ON MEASUREMENTS MADE ALONG THE LONGITUDINAL AXIS OF THE REACTOR

In Example 14.8, you estimated the reaction coefficient K' based on only input and output concentrations. You used the plug-flow with dispersion model under the assumption of a first-order

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

reaction. The data you used were: (a) influent concentration: $C_{in} = 100$ mg/L; (b) effluent concentration: $C_{out} = 10$ mg/L; (c) hydraulic retention time: $t_h = 5$ days. From the literature, you saw indications that your dispersion number d could be adopted as 0.40.

Suppose that you instead collected monitoring data along the reactor length ($L = 10.00$ m), in order to improve your estimate of K'. The data you used in this new estimation are shown as follows:

Sampling Point Along the Reactor Length (m)	Measured Concentration (mg/L)
0.1 (close to inlet end)	96.0
2.0	53.0
4.0	33.0
6.0	18.0
8.0	15.0
10 (outlet)	10.0

Estimate the reaction coefficient K' using an iterative procedure with the Excel Solver tool.

Excel

Note: This example is also available as an Excel spreadsheet.

Solution:

We set up the following computational table, where we estimate the fraction of the reactor length represented by each sampling point (distance from inlet ÷ reactor length). Based on this fraction, we estimate the travelling time to reach each sampling point, from inlet to outlet (t/t_h = fraction of length $\times$ t_h), knowing that the total hydraulic retention time is $t_h = 5.0$ d.

Using Equation 14.32, we calculate the columns of 'a' and 'C_{est}', using $C_{in} = 100$ mg/L and letting the values for K' and d be estimated by the Solver tool.

Fraction of Length	Fraction of HRT	Travelling Time (d)	C_{obs} (mg/L)	Auxiliary Calculation (a)	C_{est} (mg/L)	Error^2
0.01	0.01	0.05	96.0	1.023	96.47	0.2189
0.20	0.20	1	53.0	1.386	53.06	0.0033
0.40	0.40	2	33.0	1.686	31.77	1.5105
0.60	0.60	3	18.0	1.940	20.46	6.0323
0.80	0.80	4	15.0	2.165	13.84	1.3517
1.00	1.00	5	**10.0**	2.368	9.71	0.0865
					(sum error^2) =	9.2031
					CoD =	0.9983

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

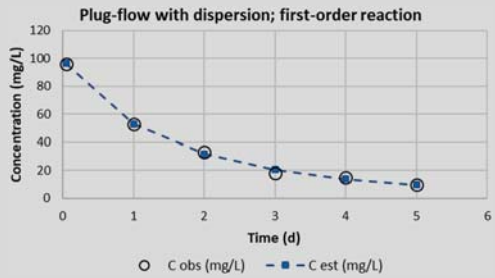
In this example, we simultaneously estimated d and K' using Solver. The best fit (lowest sum of squared errors or highest Coefficient of Determination) was found for the following values:

$\boldsymbol{d = 0.318}$

$\boldsymbol{K' = 0.725\,d^{-1}}$

These values are very close to the ones found in Example 14.8 (d adopted as 0.40; K' found as 0.76 d^{-1}). However, the results could have been different, depending on the internal concentration profile along the longitudinal axis.

The longitudinal profile of observed and estimated concentrations is shown in the figure below.



The fit is very good, which is reflected by the excellent Coefficient of Determination (**CoD = 0.9983**).

In order to allow the utilization of the NTIS model, the value of N that corresponds to $d = 0.318$ is $N = 2.26$ (calculation done using Equation 14.42).

Note that the plug-flow with dispersion model can only be used for $d \leq 1.0$. Also, the correspondence between d and N is confined to this condition.

Now, it is up to you to interpret whether these values of d and N are reasonable, given the geometry and operating conditions of your reactor. What we did was a simple curve-fitting exercise, allowing the model parameters to vary at will. If possible, search for information in the literature, in case you have not completed a tracer test. If you have tracer test data, you should instead use the tracer test results to estimate the value of d (or N) and apply the Solver tool only to find K'. This would allow you to estimate a value of K' that is close to the true intrinsic coefficient K.

14.5.5 Applicability of kinetic coefficients derived under batch and continuous-flow experiments

Advanced

After all the preceding theoretical considerations, we now have the ability to consider the applicability of the two protocols we have covered in this chapter to estimate reaction rate coefficients: (a) experiments conducted in *batch mode* and (b) experiments carried out in *continuous-flow reactors*.

S. 14.3
S. 14.3.6

In Section 14.3, we discussed the estimate of reaction coefficients based on monitoring data of concentrations in completely mixed **batch reactors**. In Section 14.3.6, we emphasized that the reaction coefficients obtained at these circumstances represented the **true intrinsic kinetic coefficient *K***, but they *could not be applied to a continuous-flow reactor if we did not have a suitable hydraulic model of the reactor*. This is particularly the case if we insert the K value directly into the equation of *an idealized reactor*, such as a plug-flow or a complete-mix reactor.

However, if we are confident that we have a **representative hydraulic model** of our reactor (plug-flow with dispersion or equivalent tanks-in-series, particularly when our estimate of d or N is based on data from a tracer test), *we may use the intrinsic kinetic coefficient* in our model to predict the effluent concentration.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Likewise, if we are monitoring the influent and effluent concentrations, we can use the plug-flow with dispersion or equivalent tanks-in-series models to make a better prediction of K', which should be closer to the true intrinsic value K.

S. 13.2

Naturally, we have to pay attention whether *our hydraulic model is really a good descriptor* of the actual hydraulic behaviour in our reactor. In Section 13.2, we covered in detail the concept of hydraulic retention time (HRT or t_h), and in Section 13.2.6, we discussed possible causes for departure from the expected **theoretical hydraulic retention time** ($t_h = V/Q$). Of special importance was the possible occurrence of **dead zones** and **hydraulic short circuiting**, which may substantially affect the **actual retention time** and the **removal efficiency**. If this is the case and if we do not have a good estimate for the hydraulic efficiency of our reactor (e.g., based on a tracer test), then *the K' coefficient determined in our continuous-flow reactor may differ substantially from the intrinsic kinetic coefficient K.*

Of course, if you are able to estimate the reaction coefficient K' based on a good hydraulic model and using the actual HRT, you will get even closer to the true intrinsic kinetic coefficient. But this is **frequently difficult**, unless you devote considerable time to the hydraulic characterization of your reactor, including the undertaking of tracer tests.

S. 14.6

*You and most readers of your report will understand these difficulties, because these difficulties are shared by most of your readers. What you need to do is to be **as clear as possible about how you conducted your experiments and how you implemented your calculations**, such as whether the experiment was conducted at batch or continuous-flow mode, which hydraulic model you used, the value of the dispersion coefficient d or equivalent number of tanks-in-series N, whether you used the theoretical or the actual HRT, and all other assumptions needed to make estimates about removal efficiency (see check-List in Section 14.6). **Remember**: if you went through all the trouble of doing these experiments and calculations, it is because you want your results to be useful to others, who may be able to use them to design new reactors or to assess the performance of other existing systems. If the readers of your report cannot reproduce your work or calculations because the elements were not all clearly described, then most of your effort may have been in vain.*

There are ways to further enhance the hydraulic representation of your reactor, using more advanced models such as **compartmental models** (the segmentation of your tank into reaction zones, slow exchange zones, internal recirculation flows, short circuiting flows) or even more advanced approaches such as **computational fluid dynamics (CFD)**. However, these are outside the scope of this book. Still, you should know that these approaches, especially the latter, are becoming increasingly more popular and common for assessing and modelling the performance of a reactor. In particular, with the increased availability of processing capabilities, the use of CFD (which requires high computing power) is becoming more accessible for research and practice applications.

14.5.6 Utilization of the kinetic coefficient and hydraulic representation for the mathematical modelling of your reactor

Advanced

S. 14.5.5

The kinetic coefficients and hydraulic representation of our reactor are useful for allowing the estimation of effluent concentrations, not only for our reactor, but also, hopefully, for new reactors that are being designed. For new reactors, the designer must understand clearly the applicability and limitations of the model parameters you derived, as discussed in Section 14.5.5, so that the model can be adapted and used for design purposes.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

S. 12.1

Basically, the estimation of effluent concentrations can be done for these two modelling conditions (see Section 12.1):

- **Steady-state conditions**. More frequently used for design purposes.
- **Dynamic-state conditions**. More frequently used for operational control.

(a) Application for steady-state conditions

This is the application that we covered in greater detail in this chapter. Depending on the reactor model adopted, the estimation of effluent concentrations under **steady-state conditions** can be done using equations previously presented in this chapter. Of course, you need to decide which reaction order you will use. For the most frequent case of **first-order reactions**, for which we have dedicated the majority of examples in this chapter, the equations you can use are summarized below:

Model	First order reaction	Equation
Batch reactor model	$C_{out} = C_{in} \cdot e^{-K \cdot t}$	Equation 14.9
Idealized plug-flow model	$C_{out} = C_{in} \cdot e^{-K' \cdot t_h}$	Equation 14.28
Idealized complete-mix model	$C_{out} = \dfrac{C_{in}}{1 + K' \cdot t_h}$	Equation 14.30
Plug-flow with dispersion model	$C_{out} = C_{in} \cdot \dfrac{4ae^{1/(2d)}}{(1+a)^2 e^{a/(2d)} - (1-a)^2 e^{-a/(2d)}}$ $a = \sqrt{1 + 4K' \cdot t_h \cdot d}$	Equation 14.32
Apparent tanks-in-series model	$C_{out} = \dfrac{C_{in}}{(1 + K' \cdot (t_h/N))^N}$	Equation 14.36

(b) Application for dynamic-state conditions

Application for *dynamic conditions is always more complex*, because we need to model not only variations along the position in our reactor (or as a function of the travelling time from inlet to outlet) but also variations over the operating or observation time (time now, in the next hour, in the next day, in the next week, and so on).

Our approach here will be essentially simple so that you have a grasp on the underlying principles of dynamic modelling. If you understand these basic elements, and if you like the subject, you can move into more complex mathematical models, such as those mentioned in Section 14.1.

S. 12.3

In Section 12.3 (Chapter 12), we presented basic concepts of mass balance in a reactor. For a complete-mix reactor, we showed the general mass balance equation:

$$V \cdot \frac{dC}{dt} = Q_{in} \cdot C_{in} - Q_{out} \cdot C_{out} + r_p \cdot V - r_c \cdot V \tag{14.45}$$

where

C = concentration of the constituent inside the reactor at a time t (g/m^3)
Q_{in} = input flow (m^3/d)
Q_{out} = output flow (m^3/d)

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

C_{in} = input concentration of the constituent (g/m^3)
C_{out} = output concentration of the constituent (g/m^3)
V = volume of the reactor or volume element of any reactor (m^3)
t = time (d)
r_p = reaction rate of production of the constituent ((g/m^3)/d)
r_c = reaction rate of consumption of the constituent consumed ((g/m^3)/d).

If we divide each term of the equation by the volume V, we obtain this alternative representation:

$$\frac{dC}{dt} = \frac{Q_{in} \cdot C_{in}}{V} - \frac{Q_{out} \cdot C_{out}}{V} + r_p - r_c \tag{14.46}$$

Knowing that the hydraulic retention time (HRT or t_h) is equal to V/Q, we can say that, mathematically speaking, $Q/V = 1/t_h$ and restate the equation in the following form:

$$\frac{dC}{dt} = \frac{C_{in}}{t_h} - \frac{C_{out}}{t_h} + r_p - r_c \tag{14.47}$$

Equations 14.46 and 14.47 are equivalent. Equation 14.46 is probably conceptually easier to understand, since we make flows and concentrations explicit. Equation 14.47 is simpler to present. You can choose the one you prefer in your simulations.

In the example given in this section, we will analyse only the *removal* of a constituent, for the sake of simplicity. However, the general equation should be used if the constituent you are studying can be subject to removal and production reactions. With this simplification, we obtain the following equation:

$$\frac{dC}{dt} = \frac{C_{in} - C_{out}}{t_h} - r_c \tag{14.48}$$

For a *first-order reaction*, we substitute the reaction rate r_c by $K' \cdot C$, and remembering that in a complete-mix reactor the concentration inside the tank (C) is equal to the effluent concentration (C_{out}), we end up with

$$\boxed{\frac{dC}{dt} = \frac{C_{in} - C_{out}}{t_h} - K' \cdot C_{out}} \tag{14.49}$$

The utilization of this equation will be greatly simplified if we represent the hydraulics of our reactor using the **apparent number of tanks-in-series** (NTIS) model. This is because each reactor in the series is assumed to be completely mixed, and so we do not need to model the concentrations at different positions in each reactor: the effluent concentration from one reactor will be the prevailing concentration inside this reactor. Our differential equation (Equation 14.49) will have only the variation with respect to time.

We can apply Equation 14.49 to estimate the effluent concentration from the first 'apparent' reactor, which will be the influent concentration to the second 'apparent' reactor. The effluent from the second reactor will be the influent to the third reactor, and so on.

S. 14.2.3

The integration of the equation can be done using a *simple numerical procedure*, such as *Euler's method* (see Section 14.2.3 for a description of its working principles, applicability, and limitations).

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

According to Euler's method, the stepwise calculation of the concentration over time can be done as

$$C \text{ at time } t+1 = C \text{ at time } t + (\text{input} - \text{output} - \text{removal}) \tag{14.50}$$

$$C_{t+1} = C_t + \left[\frac{C_{in} - C_{out}}{t_h} - K' \cdot C_{out}\right]\Delta t \tag{14.51}$$

where

Δt = time step for the numerical integration.

This procedure will become clearer if you follow Example 14.12, and especially if you use its associated Excel spreadsheet. This example can be modified to represent time in terms of other units besides days, such as hours or minutes (provided we are consistent in the units of the other variables and model parameters). In the example, to reduce the errors associated with Euler's numerical integration, we divided each day into 200 time-intervals, that is, our time step Δt was $1/200 = 0.005$ days. The example is based on the representation of the reactor as five apparent tanks-in-series ($N = 5$). You can use the spreadsheet for any value of N between 1 and 5, provided you insert the correct number of N in the appropriate cell, give the values of the initial concentrations in the reactors you are considering, and plot only the correct number of reactors.

Example

EXAMPLE 14.12 UNDERTAKE THE DYNAMIC SIMULATION OF THE CONCENTRATION OF A CONSTITUENT IN A REACTOR USING THE NTIS MODEL

You have a reactor with a total volume of 10,000 m^3. Based on hydraulic studies, you came to the conclusion that this reactor can be represented as five complete-mix apparent tanks-in-series ($N = 5$). From your kinetic studies, you obtained an estimate of the reaction coefficient $K' = 0.25\ d^{-1}$ (first-order reaction). The reactor received a fixed concentration of 100 g/m^3 at the influent point for a period of 20 days. The influent flow rate was constant during the first 10 days, at 500 m^3/d, but then starting at day 11, it doubled to 1000 m^3/d and then remained at this value until day 20.

Estimate the effluent concentrations throughout these 20 days.

Excel

Note: This example is also available as an Excel spreadsheet.

Solution:

(a) Initial conditions

To run any dynamic model, we need to specify the initial conditions, that is, the concentrations of the variables to be modelled at the beginning of the simulation period (time $t = 0$). When we are studying a reactor for which we have measurements of the estimated variable and want to use a model to fit the experimental data, we can use the values measured at the time that corresponds to $t = 0$. When we are doing a pure simulation (as is the case now), frequently we specify the initial conditions that coincide with the steady-state values.

In our case, since we are simulating five apparent tanks-in-series, we need the initial concentrations of the constituent at all five apparent tanks. We will use the steady-state values, which can be calculated using Equation 14.36:

$$C_{out} = \frac{C_{in}}{(1 + K' \cdot (t_h/N))^N}$$

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

For each apparent tank in the series, our calculations will use the following information:

- Reaction coefficient: $K' = 0.25\ \text{d}^{-1}$
- Volume of one apparent tank (V_1) = total volume/number of apparent tanks = V/N = 10,000 $\text{m}^3/5 = 2000\ \text{m}^3$
- Hydraulic retention time at each apparent tank: $t_h = V_1/Q = (2000\ \text{m}^3)/(500\ \text{m}^3/\text{d}) = 4.0$ d (this is the value to be included in Equation 14.36, without needing to divide it further by N)

For apparent tank 1, with $N = 1$ (using Equation 14.36), the steady-state concentration will be

$$C_1 = \frac{100}{(1 + 0.25 \times 4.0)^1} = 50\ \text{g/m}^3$$

For apparent tank 2, with $N = 2$

$$C_2 = \frac{100}{(1 + 0.25 \times 4.0)^2} = 25\ \text{g/m}^3$$

Doing similar calculations, we obtain $C_3 = 12.5\ \text{g/m}^3$, $C_4 = 6.25\ \text{g/m}^3$, and $C_5 = 3.13\ \text{g/m}^3$.

(b) Structure of the numerical integration procedure

The concentration at each time step will be calculated using Euler's numerical integration (Equation 14.51). It is better that you follow the procedure in the accompanying Excel spreadsheet, because there are 200 calculations performed for each day, for each apparent tank.

In this spreadsheet, we divide each day into 200 time steps. Therefore, each step of integration is $\Delta t = (1\ \text{day})/200 = 0.005$ d.

For you to understand the procedure of the numerical integration, we will show you the concentration at apparent tank 1, at one time step after $t = 0$ (initial conditions). From Equation 14.51, we have

$$C_{t+1} = C_t + \left[\frac{C_{\text{in}} - C_{\text{out}}}{t_h} - K' \cdot C_{\text{out}}\right]\Delta t = 50 + \left[\frac{100 - 50}{4.0} - 0.25 \times 50\right] \times 0.005$$

$$= 50.00 + (12.50 - 12.50) \times 0.005 = 50.00 + 0 \times 0.005$$

$$= 50.00\ \text{g/m}^3$$

At time $t + 1$, we obtained exactly the same concentration as the preceding one, at time t, that is, 50.00 g/m^3. This may seem frustrating, but you should remember that in the first 10 days of our simulation, we had steady-state conditions (all input variables were the same), and therefore the output concentrations were not supposed to change.

However, in this example, at day 11, our flow doubled, from 500 to 1000 m^3/d, and remained at this level until day 20. From day 11, the hydraulic retention time in each tank was reduced to $(2000\ \text{m}^3)/(1000\ \text{m}^3/\text{d}) = 2.0$ d. Therefore, we expect that the concentrations will change.

In the first time step of day 11, that is, at time $t + \Delta t = 11 + 0.005$ d, the concentration in **apparent tank 1** changed to

$$C_{t+1} = C_t + \left[\frac{C_{\text{in}} - C_{\text{out}}}{t_h} - K' \cdot C_{\text{out}}\right]\Delta t = 50 + \left[\frac{100 - 50}{2.0} - 0.25 \times 50\right] \times 0.005$$

$$= 50.00 + (25.00 - 12.50) \times 0.005 = 50.00 + 12.50 \times 0.005 = 50.06\ \text{g/m}^3$$

This is the value you will find in the spreadsheet, for **apparent tank 1**, day 11, step 1. For the next time steps, we just follow the same procedure, making progressive calculations based on the preceding time step.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

At exactly this same time (day 11, step 1), in **apparent tank 2**, the concentration will be changed, as a result of the instantaneous change in apparent tank 1. Let us remember that the influent to apparent tank 2 is the effluent from apparent tank 1. Therefore, we can make the same calculation, having as influent concentration at time $t + \Delta t = 11 + 0.005$ d the value of 50.06 g/m^3:

$$C_{t+1} = C_t + \left[\frac{C_{in} - C_{out}}{t_h} - K' \cdot C_{out}\right]\Delta t = 25.00 + \left[\frac{50.06 - 25.00}{2.0} - 0.25 \times 25.00\right] \times 0.005$$

$$= 25.00 + (12.53 - 6.25) \times 0.005 = 25.00 + 6.28 \times 0.005 = 25.03 \text{ g/m}^3$$

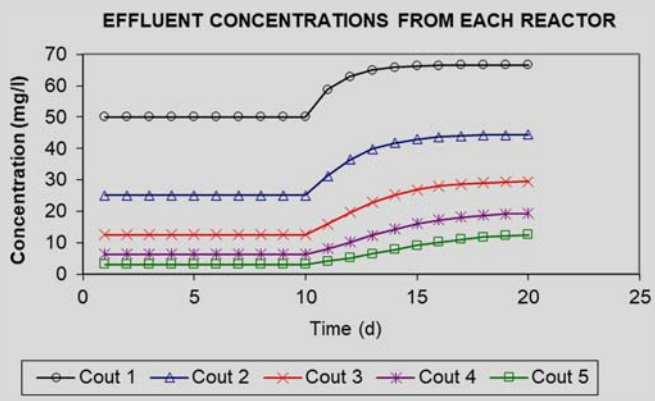
This is the value you will find in the spreadsheet, for **apparent tank 2**, day 11, step 1. For the next time steps, we just follow the same procedure.

(c) Simulation of the concentrations at each apparent tank along the 20 days

Now that you understood how the calculations were done, we can show the results of the concentration values along the 20 days. The exercise we are doing is what is called a '*step increase*': inflow increased as a step at day 11 and remained elevated at this new level.

The figure below shows the time series of concentrations in each of the five apparent tanks in the series. As expected, the concentrations decrease, from apparent tank 1 to apparent tank 5.

Also, you can see that, from days 1 to 10, all values were fixed at steady state. On day 11, the concentrations increased, and this increase continued in the movement towards the new steady state. If you do the steady-state calculations (as we did in item 'a', but with $t_h = 2.0$ d in each apparent tank), you will see that the new steady-state values are: $C_1 = 100/(1 + 0.25 \times 2.00) = 66.67$ g/m^3; $C_2 = 66.67/(1 + 0.25 \times 2.00) = 44.45$ g/m^3; $C_3 = 29.63$ g/m^3, $C_4 = 19.75$ g/m^3, and $C_5 = 13.17$ g/m^3. The conditions between the two steady states are called *transient conditions.*



(d) Use of the model for different conditions

Now that your model is established, you can change the input data (input variables, model coefficients, number of apparent tanks, reactor volume, etc.), see what results you obtain and interpret them. We give you the following suggestions regarding modifications in the influent flows and concentrations:

- *Conduct a simulation where you keep all Q_{in} and C_{in} values constant throughout the 20 days*. Observe values at steady state along the whole period.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

- *Change Q_{in} and/or C_{in} for only one day and return to the previous values after this day (simulation of influent peak conditions)*. Observe the peak in concentration and the subsequent return to steady-state conditions.
- *Change Q_{in} and/or C_{in} for one day and keep the next days with these new values (step-change)*. Observe the transient to the new steady-state conditions. Compare the influence on the effluent concentrations when you double the influent flow (keeping the same concentration) and when you double the influent concentration (keeping the same flow): even though the input load will be the same, the impact will be different.
- *Change Q_{in} and C_{in} to have different values on all days, simulating actual dynamic conditions*. This is a situation that approaches real-life conditions.

If you understood the working principles of a simple dynamic model, as shown here, and you enjoyed its potential to give you a closer representation of what happens inside actual reactors, you may want to explore this subject in more depth. We suggest you to consult the vast literature available on mathematical modelling. If practiced judiciously, keeping in mind the limitations associated with the difficulties in getting a good representation of real life, the utilization of mathematical models can open new roads to your research and reveal important elements of the system you are studying.

14.6 CHECK-LIST FOR YOUR REPORT

- ✓ Make sure you described clearly all the assumptions you used in the derivation of the kinetic coefficients and the description of the hydraulic model for the reactor.
- ✓ Specify whether the experiment was conducted in batch or in continuous-flow mode.
- ✓ Specify the reaction order (0, 1, 2, or other) you assumed and how you obtained it (experiments, literature, etc.).
- ✓ If you obtained the kinetic coefficient *K* from a batch experiment, describe all operating conditions (reactor volume, duration of experiment, frequency of sampling or measurement, etc.) and whether you calculated the coefficient by linearization of the observed values or using non-transformed data.
- ✓ Report the liquid temperature used in your experiments and for which temperature are you reporting the coefficient value (standard temperature of 20°C or a different liquid temperature). If you converted your *K* value to the standard temperature of 20°C, specify which temperature coefficient θ you used and how you obtained the estimate for that parameter θ.
- ✓ If you obtained the kinetic coefficient based on measurements in continuous-flow reactors, mention whether you made the calculations based on influent and effluent concentrations only or based on internal measurements taken along the reactor length. Specify whether you are using average values (flows and concentrations) from a longer monitoring campaign or single values collected on a single day.
- ✓ For continuous-flow reactors, specify which hydraulic model you used (idealized plug-flow, idealized complete-mix, plug-flow with dispersion, or apparent number of tanks-in-series). For the non-idealized models, describe the basis for your estimate of the values of the dispersion number (*d*) or the number of tanks-in-series (*N*) (tracer studies, literature, etc.).
- ✓ Provide the dimensions of your reactor: length, width, and depth.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

- ✓ If your reactor has a support medium, give the characteristics of the medium (specific diameter and porosity). Do not use the methods here for unsaturated reactors, in which the pore spaces are filled with air (unless you made adaptations to the model, which are well described in your report).
- ✓ Make it clear whether you used the theoretical HRT (V/Q) or the actual mean HRT (based on tracer tests) for your calculations. If your reactor has a support medium, make sure that your calculation of HRT is based on the volume occupied by liquid (volume $\times$ porosity).
- ✓ If you completed a dynamic simulation, specify the numerical integration method you used.
- ✓ Describe all other assumptions made and the detailed methodological steps of your study.

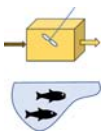
Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Chapter 15

Model application, calibration, and verification

This chapter starts with general introductory concepts on water quality and treatment plant modelling. After that we cover in a simplified way model calibration, including procedures for assessing goodness-of-fit. Finally, model verification is discussed, presenting the properties required for model residuals and the verification of compliance with these properties.



The contents in this chapter are applicable to both treatment plant monitoring and water quality monitoring.

CHAPTER CONTENTS

15.1 Concepts Involved in Water Quality and Treatment Plant Modelling **596**
15.2 Model Calibration **602**
15.3 Model Verification (Analysis of Residuals) **618**
15.4 Check-List for Your Report **630**

© 2020 The Authors. This is an Open Access book chapter distributed under the terms of the Creative Commons Attribution Licence (CC BY-NC-ND 4.0), which permits copying and redistribution for non-commercial purposes with no derivatives, provided the original work is properly cited (https://creativecommons.org/licenses/by-nc-nd/4.0/). This does not affect the rights licensed or assigned from any third party in this book. The chapter is from the book Assessment of Treatment Plant Performance and Water Quality Data: A Guide for Students, Researchers and Practitioners, Marcos von Sperling, Matthew E. Verbyla and Sílvia M. A. C. Oliveira (Authors).
doi: 10.2166/9781780409320_0595

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

15.1 CONCEPTS INVOLVED IN WATER QUALITY AND TREATMENT PLANT MODELLING

15.1.1 A simple concept of mathematical models

Advanced

To enhance your level of understanding of the system you are studying, you may decide to search for relationships between variables, and you may start to use mathematical models to explain the behaviour of your system. This can involve one or more of the following approaches:

C. 11
C. 12
C. 13
C. 14

- Use models based on **regression analysis** (see Chapter 11).
- Use **models from the literature**, probably using concepts of mass balances (Chapter 12), loading rates (Chapter 13), and reaction kinetics and reactor hydraulics (Chapter 14).
- **Develop your own model**, using the concepts listed above.

We will *not discuss model structures*, because they are inherently linked to the process you want to represent, which is out of the scope of this book. For more about model structures, you should consult the relevant literature on treatment plant or water quality modelling, as mentioned elsewhere in this book.

However, regardless of the modelling approach used, there are certain concepts that are universal – the application of these concepts will be covered in this chapter. We will present **introductory concepts of modelling** (this Section 15.1) and then we will go into a more practical and applied introduction to **model calibration** (Section 15.2) and **model verification** (Section 15.3).

S. 15.2
S. 15.3

Let us now start with some basic concepts related to modelling. In the literature, there are several definitions of 'mathematical models'. We do not present here a formal definition, but we simply present the following important concepts related to a **model** (Lee, 1973):

- *A model is a representation of reality.*
- *A model is a simplified and generalized translation of what appears to be the most important characteristics of a real-world situation.*
- *A model is an abstraction of the reality used to achieve conceptual clarity – to reduce the variety and complexity of the real world to a level that can be understood and represented.*

Environmental systems are extremely complex, and **environmental models** aim to *represent this complex reality* ***as it is observed or measured***. However, a model cannot possibly represent all the complexity of multiple interactions that exist in environmental systems (several of which are arguably not measurable or quantifiable).

Mathematical models are composed of (a) a *theoretical structure*, represented by mathematical equations, (b) numerical values of the *parameters* (coefficients) of the equations, and (c) input and output data, often comprising field or laboratory *observations/measurements*, and relating external factors with the system response. Figure 15.1 presents a simplified schematic representation of a model.

Often, readers can be intimidated by the word 'mathematical' in the expression 'mathematical model', imagining unsurpassed complexities. Difficulties (not unsurpassed) could be associated with several

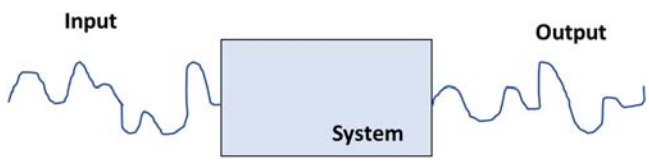


Figure 15.1 Schematic representation of a model.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

complex models with a high degree of mathematical sophistication, but one must remember that **models can also be simple**! The equation of a straight line, $\boldsymbol{Y = a + bX}$, is a mathematical model and, incidentally, an extremely useful model. In this linear equation, $\boldsymbol{Y}$ and $\boldsymbol{X}$ are the **variables** of the model, and **a** and **b** are the **parameters** (coefficients) of the model. You have noted that in this book, whenever possible, simplified approaches are presented, but of course, a basic understanding of mathematics is also essential.

Advanced

15.1.2 A procedure for modelling

The technical literature is full of resources to support your modelling endeavours. You should consult the literature for resources that present concepts such as 'good modelling practice' or 'best modelling practice'. For instance, for modelling the activated sludge process, an International Water Association (IWA) task group published a technical report on good modelling practice (IWA Task Group on Good Modelling Practice, 2012).

Figure 15.2 presents a simplified flowchart of the sequence used to develop a mathematical model. In the literature, there are different variants of the flow of activities, but this figure depicts the main steps. It should be noted that the modelling process is often iterative by nature, with returns to previous stages for

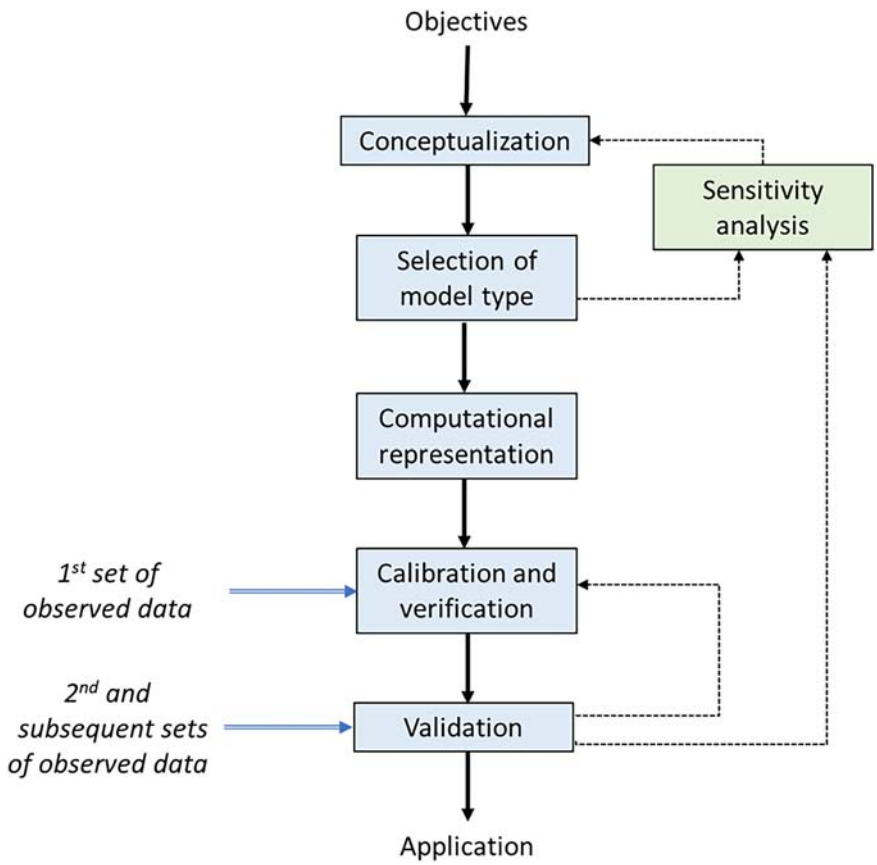


Figure 15.2 Flowchart of the stages of developing a mathematical model (adapted from Beck, 1983).

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

readjustment of the model structure and/or the values of the coefficients. There are other variants of this flowchart, such as that presented by Jakeman *et al.* (2018).

The most important items associated with the flowchart shown in Figure 15.2 are described below (von Sperling, 2014).

Objectives. Initially, you need to state very clearly the objectives of your mathematical modelling exercise. These objectives define the structure of the model to be used and establish the necessary efforts for field and laboratory works.

Conceptualization. The first step in the modelling procedure is to structure the concept of the system, such as the physical representation of the water body or the treatment reactor and the selection of the system boundaries and the variables to be measured and modelled.

Selection of the model type. There are different types of mathematical models linked to distinct objectives and involving varying degrees of complexity. Model-type selection is a crucial step in the modelling exercise.

Computational representation. After having selected the type of model, you should structure it in terms of mathematical equations, defining it with analytical or numerical solutions. The present book presents some computational representations of chemical or biochemical conversion processes based on the literature. We have used both analytical solutions and numerical integrations (see Chapter 14).

C. 14

Calibration and verification. The purpose of calibrating a model is to obtain a good fit between the observed (measured) data and estimated values (calculated by the model), by means of *varying the values of the model parameters (coefficients)*. In this step, it is necessary to evaluate 'how good is the fit of the model'. In the literature, there are distinct interpretations of the concept of model **verification**. We adopt here the concept presented by Beck (1983), according to which the verification is the determination that the 'correct' model was obtained from a single set of experimental data. For this purpose, verification covers the **analysis of residuals** (difference between the observed values and the estimated values), which must comply with certain properties. Other authors (e.g., Thomann, 1982) understand verification as the test of the calibrated model with additional data, preferably under different conditions (e.g., different flows, loads). This concept, however, is more similar to the concept of validation, which is discussed below. The conditions under which your model has been calibrated and verified must be specified so that the user can be aware of the model's applicability range. The calibration and verification steps are discussed in Sections 15.2 and 15.3.

S. 15.2

Validation. The validation step corresponds to the *evaluation of the model fit under conditions that are different from those used for the calibration*. For this purpose, *one or more independent experimental data sets are used*, and these data must be distinct from the data used for calibration. If the model does not give a good fit to the new data, you should reanalyse its structure and/or try a new calibration. If the model gives a good fit to the new data, you could consider the model to be validated. However, *a model can never be unconditionally validated*, in the sense that its results will never represent reality under all scenarios. There may always be other conditions (not accounted for) for which the model may not perform well. In other words, *a model can always be invalidated but can never be completely and unconditionally validated*. Even with these constraints, naturally you will gain much more confidence in the model if it shows good performance with new data during the validation step. However, because the validation stage involves the conception and implementation of new experiments, it is unfortunately not frequently practiced.

S. 15.3

Sensitivity analysis. In different stages of the model development, you could evaluate the model structure and its parameter set by means of a sensitivity analysis. Based on this analysis, you can infer the magnitude of the model's response to the given input parameters (variables or coefficients). A model may be very sensitive to some inputs (meaning the response value changes considerably with small changes in the input value) and it may be less sensitive to others (changes in the input value only cause

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

very slight changes in the response value). Consequently, a sensitivity analysis allows you to judge if there is a need to obtain more accurate values of the input data.

Application. After having gone through the previous steps, and if you detect that the model has an appropriate structure and parameter values, you can apply the model for your particular purposes and conditions.

We recognize the fact that for some modelling applications, there is no time and/or resources to obtain experimental data. For this, an existing model structure may be used, and typical values of the model coefficients may be adopted based on the literature. If you are using a model for **pure simulation** (no comparison with experimental data), there are no steps of calibration, verification, and validation. Naturally, you should have special care in selecting the most appropriate values for the parameters, which is why you should have thorough knowledge of the system and of the model structure. The interpretation of the output data from the model should reflect these additional uncertainties.

15.1.3 Definition of the model objectives

Advanced

The modelling approach you use will depend very much on the objectives you establish for your model. Some possibilities are

- **Research**
  - Improving the understanding of the system
  - Deriving model coefficients
  - Testing different operational conditions
  - Replacing lab- or pilot-scale studies
- **Management/planning**
  - Long-term planning (prediction of future conditions)
  - Assessment of compliance with quality standards
  - Waste load allocation (planning of required removal efficiencies and consent discharges)
- **Real-time control**
  - Evaluation of transient phenomena (peak input loads, rainfall events, and change in operating conditions)
  - Evaluation of seasonal variations
  - Integrated control of water bodies and treatment plant

15.1.4 Model conceptualization

Advanced

The stage of model conception is dedicated to the definition of the physical representation of the system and the formulation of the intervening variables and parameters. Figure 15.3 presents a typical definition of a system and its variables.

Input variables (measured). These variables are also known as input disturbances and forcing functions. In water quality and treatment plant modelling, they are typically the influent flow rates and concentrations. These variables can be measured, and the measured values are used in the simulations.

Input disturbances (not measured). These are disturbances to the system, because they are not measured or they are unknown. A variable also may not be measured; for example, if there is no technology available for its measurement or if it is not included in the monitoring programme because of limitations of time or resources.

State variables. Process state variables characterize the essential properties and the behaviour of the process or the system, such as time and space functions. Typically, in the case of water quality and treatment plant models, they are the concentrations of the modelled constituents.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

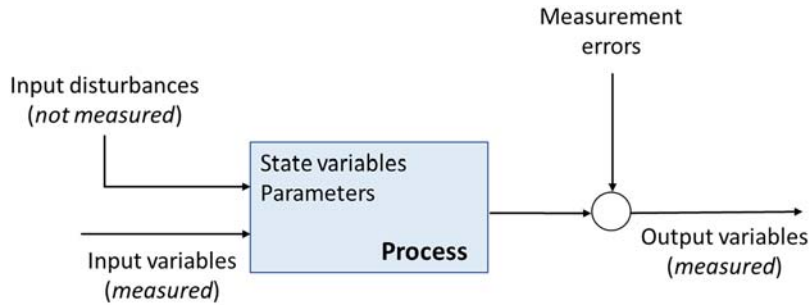


Figure 15.3 Schematic representation of a system and variables for a model (adapted from Beck, 1983).

Output variables (measured). The measured output variables are mostly the state variables (e.g., ammonia) or aggregate variables (e.g., total nitrogen, representing the sum of the nitrogen fractions – organic, ammoniacal, nitrite, and nitrate).

Measurement errors. Measurement errors can be random or systematic and can be derived from limitations in field measurements, sample collection procedures, laboratory analysis, and measuring instruments (see Chapters 3 and 4). These errors, to a greater or lesser extent, are inherent in the measurement of the output variables and prevent them from being an absolutely accurate representation of the state variables.

C. 3
C. 4

Parameters (coefficients). These are the parameters or coefficients of the equations that represent the physical, chemical, and biochemical reactions of the models. Examples are the kinetic coefficients covered in Chapter 14. Often, they are called model *constants*, but we should recognize that most kinetic coefficients vary over space or time, that is, their values are not necessarily constant.

C. 14

15.1.5 Selection of the model type

Advanced

The classification of the different **types of models** reflects their basic structure and key objectives. There are different ways of classifying models. A few of the more common ways are described below.

S. 15.1.3

- **Model for research versus model for management versus model for control:** This classification, mainly based on Beck (1983), was discussed in Section 15.1.3.
- **Distributed model versus lumped model: Distributed models**, or distributed parameter models, are those in which variations in the quantities of constituents are continuous functions of time and space. An example is the advection-diffusion model for transporting substances along a reactor. A distributed model representing the variations in the three dimensions (x, y, z) is an advanced way of representing the behaviour of the constituents in a reactor or water body. **Lumped models** aggregate parts of the system description into a finite volume, in which the quality of the water is assumed to be uniform and independent of the position. The representation of complete-mix reactors (see Chapter 14) is an example of this type of model. When models describe variations over time (dynamic conditions), distributed models are often represented by *partial differential equations* (function of time and space), whereas lumped models are represented by *ordinary differential equation*s (function of time only) being, therefore, easier to solve mathematically.

C. 14

- **Mechanistic model versus black-box model: Mechanistic models** (also called conceptual models, physically based models, or internally descriptive models) are models that incorporate a description of the internal mechanisms (variables and parameters) of the process behaviour. In contrast, **black-box**

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Figure 15.4 Schematic comparison between mechanistic and black-box models.

models (also called empirical models, statistical models) do not seek explicit references to what occurs within the process and focus only on what is measurable: the input and output variables of the model (see Figure 15.4). Often, black-box models are based on fittings made by *regression analyses* between the output variable and the input variables. Both approaches are useful within their context. Mechanistic models, although more difficult to represent, are more useful to allow a better understanding of the behaviour of the system you are studying. Black-box models are simple to structure (see regression analysis in Chapter 11), but their applicability is usually confined within the boundaries of the system you investigated. As a matter of fact, these two categories of models characterize the two extremes of the spectrum, and most models are located within these two boundaries (characterizing what one could call '**grey-box models**').

C. 11

- **Steady-state model versus dynamic-state model:** These two approaches have already been described in Section 12.1. Environmental and loading conditions in a water body or treatment plant reactor naturally vary over time, and to represent these variations we need **dynamic models**. However, the representation of systems that vary with space and time is more complex, and for this reason, simplifications are often introduced, in the sense of assuming that all variables of the model are constant over time. These models are called **steady-state models**. Figure 15.5 compares the two model types. The models in the steady state are typically more used for planning and design purposes, whereas the models in the dynamic state are more applied for process control.

S. 12.1

- **Deterministic model versus stochastic model:** *Determinism* is the principle that the phenomena are linked to each other by rigid rules of causality and universal laws that exclude chance and indeterminacy so that if we are capable of knowing the present state, we could also predict the future and reconstitute the past. This general definition emphasizes the assumption, in the case of mathematical modelling, that one has a perfect knowledge of the behaviour of the system. Of course, this rigour cannot be observed in practice, and the models continue to be simplified representations of reality. **Stochastic models** (probabilistic, with a random component) incorporate uncertainty in the measurements, parameters, and variables. A stochastic model is reduced to a **deterministic model** if the stochastic disturbances of input and the random errors of measurement are assumed to be equal to zero and that the parameters are known exactly (instead of being estimates formulated in terms of statistical distributions).

Figure 15.5 Schematic comparison between steady-state models and dynamic models.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

15.1.6 Properties required for mathematical models

Advanced

The following **properties are required for mathematical models** (Kauark Leite & Nascimento, 1993):

Rational coherence. There must be coherence of the transcription that is made between the objects or phenomena perceived and its translation into the chosen theoretical representation.

Fitting to experimental data. The model is expected to be able to generate results that approach the experimental (measured or observed) data. However, the mere fitting to a set of data is not sufficient for the model to be considered as fully adequate (see Section 15.1.2 for a discussion about the difficulties of validating a model).

S. 15.1.2

Unicity and identifiability. It is possible to construct different models representing the same phenomenon, but the justification of adopting a given model should not be based solely on the fitting to the experimental data. We should seek to obtain a unique model for a given level of representation (unicity of system representation) and fit the model parameters from the experimental data, obtaining a single set of parameters (parameter identifiability). We should pay attention to the fact that environmental models can be quite complex and with limited identifiability, with distinct sets of coefficient values leading to a similar fit to the experimental data.

Parsimony. The concept of parsimony is that if a simple model is sufficient, no other more complex model is needed. Parsimony or minimality concerns the economy of the means used (minimum number of variables and parameters) or, still, the principle of the reduction of the arbitrary. The principle of minimality does not simply mean that we will eliminate complexity but rather that we should be prudent and reduce redundancy. Of course, the principle is desirable, but it must be practiced with discretion, in order not to hinder the advancement of scientific knowledge.

Falsifiability. Falsifiability concerns the possibility of invalidating hypotheses that have been assumed for the construction of the model. A hypothesis is falsifiable if the logic allows the existence of a statement or a series of statements that, if shown to be true, would contradict the hypothesis or indicate that it is false. Every hypothesis or a whole system of hypotheses must satisfy this fundamental condition. A hypothesis must be falsifiable, that is, it must be possible for us to invalidate it. Of course, in the case of mathematical modelling of environmental systems, the available experimental means may be insufficient to prove the inadequacy of a hypothesis.

Predictive power. This relates to the rational coherence of the model and the observations used in its construction. Predictive power is related to the extent of its validity domain. A model is more justified the broader its field of applicability is shown to be, as detected *a posteriori*. In this sense, black-box models, even if they occasionally led to a better fitting to the experimental data, have less predictive power than mechanistic models.

15.2 MODEL CALIBRATION

15.2.1 General aspects of model calibration

Basic

An overview of the various stages involved in the development of a model has been presented in Section 15.1. There, we saw that the development of a model involves several sequential steps. Furthermore, we saw that an iterative process is used until we have what we consider to be the most appropriate model. This section deals with one of these stages, related to model calibration.

S. 15.1

The **calibration** stage, or the estimation of the parameters (coefficients) that are part of the equations, is an essential step in the development of any mathematical model. The techniques used for the calibration of *well-defined systems* (i.e., those for which a valid structure is known for the model and for which the model and parameters can be determined accurately) cannot be applied to *ill-defined models* (Whitehead and

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

O'Connell, 1984). In this sense, we should take into account the fact that *most environmental systems are typically poorly defined*. Therefore, we should keep in mind that some calibration techniques used for well-defined, non-environmental systems may not always be applicable to our environmental models.

There are important **limitations of environmental models**, which can make calibration challenging: (a) non-linearity of the equations, (b) difficulty in representing the systems in a real scale, (c) difficulty in quantifying biochemical reactions, (d) high number of parameters and state variables in various current models, and (e) identifiability problems in several model equations.

To use the calibration methods presented in this section, we assume that there are **observed (measured) data** of the state variables (e.g., dissolved oxygen (DO), biochemical oxygen demand (BOD), nitrogen (N), phosphorus (P), coliforms, etc.), which allow comparison with the **data estimated** by the model. For a visual example of what calibration does for a model, see Figure 15.6, which presents an example of data and a model for a biological reactor (in this case, a river). For this case, imagine that you have a model for the concentration of DO in a river. But then imagine that you also have obtained measured values of DO at distances of 5, 27.5, and 47.5 km from a particular point (the data from these different distances are represented as small circles). The panel on the left of Figure 15.6 shows the DO profile with non-calibrated values of the model coefficients (e.g., perhaps something you have taken from literature or parameters from another study). It can be seen that, in this case, the model fitting is very poor, since the simulated values differ greatly from your measured values. The panel on the right of Figure 15.6 shows the DO profile after calibration of the model with your data (adequate modification of the model coefficients). It is possible to observe, visually, the improved fit of the model to your experimental data.

The most common way to estimate model parameters is to **minimize something known as an objective function**. This function typically represents the *sum of the squares of the residuals* (SSRs) (where the ***residual*** *is the difference between the observed value and the estimated value*). This is the procedure we used in Chapter 14 to derive reaction coefficients (zero- or first-order reactions) based on regression analysis or iterative methods.

C. 14

The correct use of this procedure of minimizing the objective function requires compliance with several criteria, specifically regarding **properties of the residuals**. If these properties are not satisfied, then the model may not be appropriate. Additionally, there is no guarantee that the optimization algorithm will find the overall minimum. Instead, the optimization may find one or more local minimum values. Unfortunately, the simultaneous optimization and satisfaction of criteria for residuals in the case of environmental systems is more an exception than a rule. The **lack of identifiability**, which is one of the most important limitations for modelling environmental systems, may occur when there is a *high correlation between the parameters of the model*. Thus, different calibrations based on the minimization

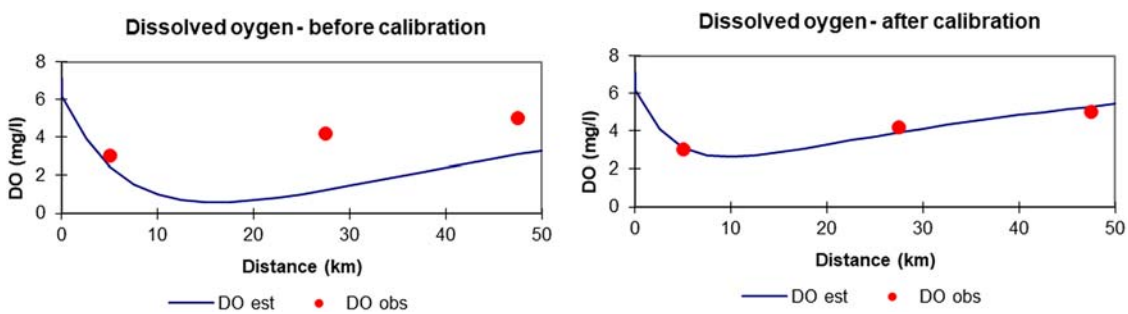


Figure 15.6 Examples of the profile of dissolved oxygen in a river, before and after model calibration.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

of the error function can lead to totally different values of the parameters, due to the fact that they are correlated.

S. 15.2.2

Section 15.2.2 covers the traditional calibration methods (minimization of the error function), while Section 15.3 is dedicated to model verification (analysis of the residuals).

S. 15.3

15.2.2 Calibration by minimization of the residuals

Basic

As mentioned above, a **residual** is the *difference between the observed (measured) value and the value estimated (calculated) by the model*. We prefer to call it a *residual*, instead of an *error*, because a residual is based on a comparison with a 'measured' value of our variable, and an error would be based on a comparison with a 'true' value of the variable. In principle, when modelling environmental systems, we cannot guarantee that the measurements represent reality (we cannot guarantee that we have 'true' values – the best we can do is to rely on 'measured' values). However, if you read some of the literature on modelling, you may also find the use of the word 'error' for what we are calling 'residuals' here in this book. With that considered, you do not need to worry so much about the semantics, as long as you understand the concepts. As a matter of fact, to be consistent with most of the literature, sometimes we will use the word '**error**' in the description of some goodness-of-fit indicators. The concept of residual was also discussed in Chapter 11, in the context of regression analysis.

A model will present a *perfect fit if all the estimated values are exactly the same as all of the measured values*. In this case, the *residuals will be equal to zero*, and of course the sum of the residuals will also be equal to zero.

This situation of a perfect fit for a model seldom occurs, especially in the case of environmental models. In practice, it is usual that there are residuals from our estimate, both positive and negative. For example, at one measuring point in a reactor, you might obtain an observed BOD concentration of 8.0 mg/L and an estimated (modelled) BOD concentration of 6.0 mg/L. In this case, the residual is equal to $8.0 - 6.0 = 2.0$ mg/L. On the other hand, at another measuring point, you might get an observed BOD concentration of 5.0 mg/L and an estimated (modelled) BOD concentration of 7.0 mg/L. In this case, the residual is equal to $5.0 - 7.0 = -2.0$ mg/L. When adding the values of the two residuals, you obtain $2.0 + (-2.0) = 0.0$ mg/L, giving the false impression that, as the sum of the residuals is zero, model fitting is perfect. For this reason, we work with the square of the residuals to make the values always positive and lead to a total sum of the residuals that is always greater than zero (and that would be equal to zero only in the case of a perfect fit, which is almost never the case). In the present example, you would get the following **sum of the squares of the residuals, SSR** (also known as **sum of the squares for error – SSE**): $(2.0)^2 + (-2.0)^2 = 4.0 + 4.0 = 8.0$. In summary, the sum of the squares of the residuals is defined as follows:

$$\text{SSR} = \sum (Y_{\text{obs}} - Y_{\text{est}})^2 \tag{15.1}$$

where

SSR = sum of the squares of the residuals (also called sum of squares for error SSE)
Y_{obs} = observed (measured) value
Y_{est} = estimated (calculated) value.

C. 11

C. 15

In Chapter 11, which deals with regression analysis, we also referred to Y_{obs} as Y and Y_{est} as $\hat{Y}$. In that chapter, we analyzed the relationship between any two (or more) variables X (independent) and one Y (dependent). In this chapter, most of our interest is to fit estimated (calculated) data (Y_{est}) to observed

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

(measured) data (Y_{obs}), based on general model equations (based on regression analysis or not), and this is why we use the nomenclature of Y_{obs} and Y_{est}, to make things even clearer.

You can do the calibration **manually**, in an *informal* or *subjective* manner, varying the values of the parameters until the sum of the squares of the residuals decreases to the point that you consider the fit to be acceptable. This manual approach is often employed because of its simplicity and can lead to satisfactory effects, especially if the model structure is simple and if there are very few coefficients to be estimated. However, there is no guarantee that the best set of parameters will be obtained, and often there will be influence of the identifiability problems discussed above.

You can also do the calibration in a *formal* or *objective* way using an **automated** process, by means of some optimization method that systematically searches possible values of the coefficients and by means of an algorithm, which converges on the set of values that leads to the smallest sum of the squares of the residuals. There are several minimization algorithms that can be used. In our book, we have used the Excel Solver tool to carry out this procedure (see examples in Chapter 14). Consult the Excel manual on the different algorithms that are used by the Solver tool.

C. 14

A challenge we face when doing simultaneous calibration of several parameters is the fact that the **parameters may be correlated among themselves**, and different combinations of parameter values may lead to similar values of the objective function we are using (e.g., in this case, the minimization of the sum of the squares of the residuals). In some cases, you may *take out one parameter from the automated procedure* and use a value from the *literature* so that there are fewer parameters to be simultaneously estimated. Otherwise, you may do the simultaneous estimations in a *stepwise manner, in blocks of parameters*, instead of having all of them varying together at the same time.

The procedure used for this automatic minimization relies on a mathematical algorithm, without any guarantee that the final values of the parameters will have a physical meaning. Therefore, **you should be in control of this procedure** and **establish constraints** (minimum and maximum allowable values, for instance, based on the literature) as necessary for the parameter values so that you do not end up with values that you know are not acceptable or do not have any physical meaning in the real world. The Solver tool in Excel allows you to establish constraints for the values that are to be varied.

You should also remember that the optimization methods *do not guarantee that a **global minimum** has been obtained* (e.g., in this case, the smallest possible value of the sum of the squares of the residuals). As these methods work with convergence processes, it is possible that the algorithm stopped at a **local** (and not global) minimum value. One way to check if the convergence procedure got stuck at a local minimum is to perform the optimization several times, each time using different initial (seed) values for the coefficients. If the convergence always ends on the same parameter values, we have a better indication (though not absolute certainty) that we are reaching the *global* minimum.

You should also never forget to be aware of the **number of parameters you are trying to estimate and the number of observed data** points (your sample size). If your sample size is small (you have few data points), and the number of parameters in your model is approaching the number of data points in your data set, then you may mathematically obtain a good match between observed and estimated values. However, this will not necessarily indicate that you made a good and reliable calibration. You should have either more data points or fewer model parameters.

15.2.3 Evaluation of the goodness-of-fit of the model

Drawing inference about the overall quality of the adjustment between the estimated data and the observed data is always difficult, and there is no single criterion that fully answers the question: '*How good is my model?*' We analyze the following different approaches that can be used to infer the quality of the model fit (Thomann, 1982).

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Basic

(a) Graphical visualization

S. 15.2.1

The interpretation of graphs in which the **observed values and the estimated values are plotted in a sequence** over distance or time will be of great value to you. In Figure 15.6 (Section 15.2.1), regardless of any statistical analysis, you clearly see that the graph on the right leads to a much better fitting than the left chart. This type of chart, with distance or time in the X axis, may also reveal stretches of your reactor or periods of time in which the simulation was unsuccessful and there are possible anomalies in the results.

A typical structure of a graph with observed and estimated values plotted along a sequence (time or position) is shown in Figure 15.7. Take a look at this figure. After a brief visual interpretation, you may infer that, although the model does not provide a perfect fit (as is mostly the case with environmental systems), the model *does* pick up the main upward and downward trends of the observed data. It is now up to you to **interpret the goodness of this fit with respect to your model objectives**: do you expect your model to simply give a good representation of average observed values, or overall trends in the data, or do you need the model to be able to pick up extreme values, such as peaks and dips in the data? These are different objectives, and your assessment of how successful the model fit is will depend on your expectations for what the model should be able to do.

Basic

(b) Coefficient of Determination (CoD)

One of the most useful statistical indicators for model goodness-of-fit is given by the **Coefficient of Determination (CoD)**. This coefficient reflects the relationship between the sum of the squares of the residuals and the total variance of the observed data and is mathematically defined as follows:

$$\mathrm{CoD} = 1 - \frac{\sum_{i=1}^{n} (Y_{\mathrm{obs}\,i} - Y_{\mathrm{est}\,i})^2}{\sum_{i=1}^{n} (Y_{\mathrm{obs}\,i} - Y_{\mathrm{obs\ mean}})^2} \tag{15.2}$$

where

C. 11

$Y_{\mathrm{obs}\ i}$ = observed value at time or position i in the data sequence
$Y_{\mathrm{est}\ i}$ = estimated value at time or position i in the data sequence (in Chapter 11, which deals with regression analysis, we also call Y_{est} as $\hat{Y}$)
$Y_{\mathrm{obs\ mean}}$ = mean of observed values
n = number of data.

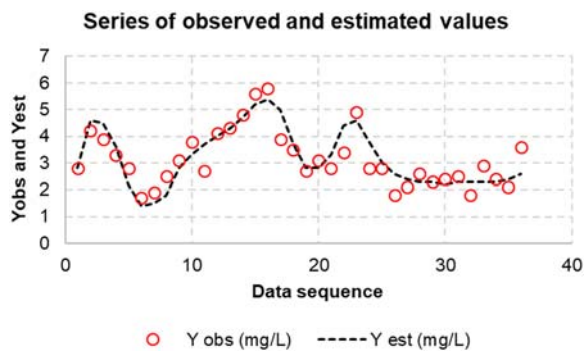


Figure 15.7 Structure of a graph plotting observed and estimated values along a sequence of time of simulation or position in a reactor.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Interpretation of the Coefficient of Determination CoD

- The **numerator** of the equation is the *sum of the square of residuals (SSR, also called SSE)* and the **denominator** is *associated with the variance of the observed values*.
- CoD values may vary between $-\infty$ **and** $+1$.
- For positive CoD values ($0 \leq \text{CoD} \leq 1$), the value represents the **fraction of the total variance of the observed data that is explained by the model**.
- **CoD equal to 1** indicates a **perfect fit** between the observed and estimated data.
- **CoD equal to zero** indicates that the model fit is equivalent to that of a *horizontal line passing through the average value* of the observed data.
- **Negative CoD** indicates that the model *fit is worse than that of a model consisting of a horizontal line* passing through the average of the observed data.
- **CoD** is *influenced by the variability of the observed data* (the variance of data, which is associated with the denominator of the equation).
- In models based on **regression analysis** (see Chapter 11), the **CoD values are equal to R^2** (squared correlation coefficient r) and are **always** $\geq$ **0**. In the case of models that are **not based on regression analysis**, the **CoD values may be negative** (in contrast, R^2 values are never negative).

C. 11

C. 11

Please read again the last observation above. You have studied regression analysis in Chapter 11 and saw the concept of the correlation coefficient (r) and its value raised to the power 2 (r^2 or R^2), which we called Coefficient of Determination there. The concept of CoD here is similar, with the main distinction that it can be negative, if your model is not based on regression analysis and performs worse than a regression-based model. In summary, we have

- *Model based on regression analysis*: **Coefficient of correlation** (***r***) varies from -1 **to** $+1$ (the closer the value is to -1 or $+1$, the stronger is the linear relationship between the two variables X and Y).
- *Model based on regression analysis*: **Coefficient of Determination** ($\boldsymbol{r^2}$ or $\boldsymbol{R^2}$) varies from **zero to** $+1$ (the closer the value is to $+1$, the better is the fit of the regression-based model estimates $\hat{Y}$ to the values of the dependent variable Y).
- *General model, not based on regression analysis*: **Coefficient of Determination** (**CoD**) varies from **–infinite to** $+1$ (the closer the value is to $+1$, the better is the fit of the non-regression-based model estimates Y_{est} to the values of the observed data Y_{obs}).

Example 15.1 shows how to calculate the CoD based on observed and estimated (modelled) values. To keep the example simple, we use a sequence of only five data points. We show the structure of the calculation and also a short-cut version, using the Excel functions **SUMXMY2** (numerator of Equation 15.2) and **DEVSQ** (denominator of Equation 15.2).

Example

EXAMPLE 15.1 EXAMPLE OF THE CALCULATION OF THE COEFFICIENT OF DETERMINATION (CoD)

Based on experimental data collection and modelling studies, you obtained the observed values (Y_{obs}) and estimated values (Y_{est}) listed in the following table. Calculate the Coefficient of Determination (CoD).

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Data Sequence	Y_{obs} (mg/L)	Y_{est} (mg/L)
1	3.00	3.40
2	5.00	4.60
3	3.50	3.80
4	2.50	2.10
5	4.00	4.50

Excel

Note: This example is also available as an Excel spreadsheet.

Solution:

(a) Calculation of the sum of the squares of the residuals

We now calculate the numerator of Equation 15.2.

Data Sequence	Y_{obs}	Y_{est}	Residual$^2 = (Y_{obs} - Y_{est})^2$
1	3.00	3.40	$(3.00 - 3.40)^2 = 0.16$
2	5.00	4.60	$(5.00 - 4.60)^2 = 0.16$
3	3.50	3.80	$(3.50 - 3.80)^2 = 0.09$
4	2.50	2.10	$(2.50 - 2.10)^2 = 0.16$
5	4.00	4.50	$(4.00 - 4.50)^2 = 0.25$
Sum			0.82

(b) Calculation of the sum of the squares of the deviations of the observed data from their mean value

We now calculate the denominator of Equation 15.2. To do this, we need to determine the mean of the observed data, which is

$$Y_{obs\ mean} = \frac{3.00 + 5.00 + 3.50 + 2.50 + 4.00}{5} = 3.60\ mg/L$$

Data Sequence	Y_{obs}	$(Y_{obs} - Y_{obs\ mean})^2$
1	3.00	$(3.00 - 3.60)^2 = 0.36$
2	5.00	$(5.00 - 3.60)^2 = 1.96$
3	3.50	$(3.50 - 3.60)^2 = 0.01$
4	2.50	$(2.50 - 3.60)^2 = 1.21$
5	4.00	$(4.00 - 3.60)^2 = 0.16$
Sum		3.70

(c) Calculation of the Coefficient of Determination

From Equation 15.2, we obtain

$$CoD = 1 - \frac{\Sigma(Y_{obs} - Y_{est})^2}{\Sigma(Y_{obs} - Y_{obs\ mean})^2} = 1 - \frac{0.82}{3.70} = 0.7784$$

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

The interpretation is that 77.84% of the variance of the observed data is explained by the model.

(d) Calculation of the Coefficient of Determination using Excel functions

We can use Excel functions to calculate CoD directly, without the need of the intermediate calculations.

Numerator of Equation 15.2: SUMXMY2 (cells with Y_{obs}; cells with Y_{est}) = 0.82.

Denominator of Equation 15.2: DEVSQ (cells with Y_{obs}) = 3.70.

$$\text{CoD} = 1 - \frac{\Sigma(Y_{obs} - Y_{est})^2}{\Sigma(Y_{obs} - Y_{obs\ mean})^2} = 1 - \frac{0.82}{3.70} = 0.7784$$

CoD is probably the **most complete indicator** of a model's goodness-of-fit. However, *judgments based purely on CoD statistics can sometimes be misleading*, because the **CoD values are greatly influenced by the variability or stability of the observed data**. When the observed data present little variability (or, more specifically, small variance), either along the reactor location or over time, the denominator in Equation 15.2 is small, and thus it is more difficult to obtain high CoD values.

This can become even more complex if these relatively stable data are influenced by *measurement noises*, which of course are not expected to be reproduced by the model. If, however, the *observed data present increasing or decreasing trends, then it is easier to achieve higher CoD values*, as long as the model follows the trends reasonably well (von Sperling, 1990). For this reason, *in your experimental design, you may plan to purposefully* ***introduce disturbances*** *in the system you are modelling to obtain experimental data with greater variability and to facilitate model calibration*.

These situations are illustrated in Table 15.1, which presents distinct simulations with different interpretations of the CoD. The results represent hypothetical observations and simulations of a constituent with respect to time or with respect to distance along a reactor. We use only five data points to make the analysis easier to understand.

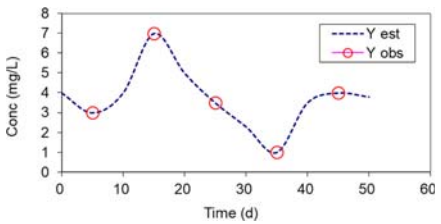
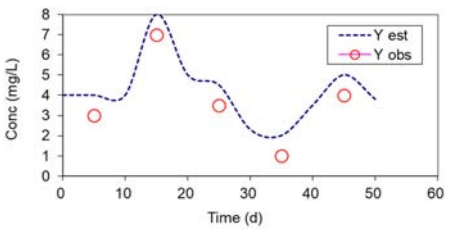
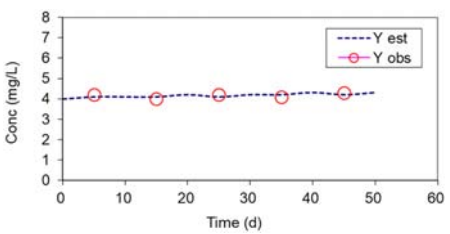
- **Case 1**. The simulation leads to a *perfect fit* between the estimated and observed data and the CoD is equal to 1.0000 as a result.
- **Case 2**. The simulation leads to a *large and systematic residual* (simulated values are always 1.0 mg/L higher than the observed values). However, the observed data series has high variability, which causes the CoD to be high, because the simulated series has been able to follow the main trends of the observed series, even with a high fixed residual.
- **Case 3**. The *residual is small* (0.1 mg/L). However, since the observed data series is relatively stable, it is more difficult to obtain a high CoD value. In this case, the CoD is close to zero, and is much lower than the CoD was for case 2, in which the residual was greater, but the observed series was less stable.
- **Case 4**. The observed data series is the same as in case 3, that is, with low variability. The *residual is intermediate* (0.5 mg/L) *and systematic* (estimated values are always lower than those observed). These conditions cause the CoD to be very low (in this case, negative, and much lower than in case 2, although the residuals are smaller now).
- **Case 5**. Here, the model is *simply the equation of a line that passes through the mean of the observed data* (3.7 mg/L), that is, $Y_{est} = Y_{obs\ mean}$. In this situation, the CoD is exactly equal to zero.

Despite the above limitations, the **CoD statistic is still probably the best criterion for judging the adherence of a model to a data set**. The CoD is directly related to the sum of the squares of the errors

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

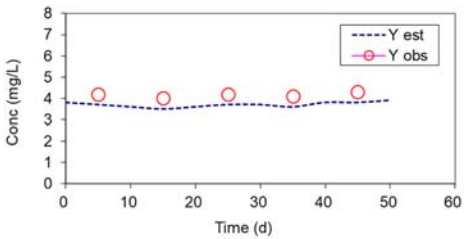
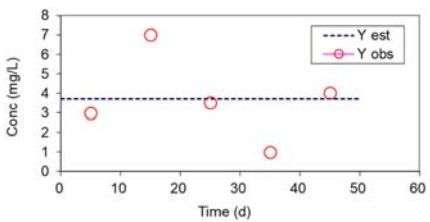
Table 15.1 Interpretation of the Coefficient of Determination (CoD) in different simulations.

Case	Time (d), Estimated Values (Y_{est}), and Observed Values (Y_{obs})				Observed and Estimated Values as a Function of Time	CoD
	Time (d)	Y_{est} (mg/L)	Y_{obs} (mg/L)	$(Y_{obs} - Y_{est})^2$		
1	0.0	4.00				1.0000
	5.0	3.00	3.00	0.000		
	10.0	4.00				
	15.0	7.00	7.00	0.000		
	20.0	5.00				
	25.0	3.50	3.50	0.000		
	30.0	2.00				
	35.0	1.00	1.00	0.000		
	40.0	3.50				
	45.0	4.00	4.00	0.000		
	50.0	3.80				
2	0.0	4.00				0.7340
	5.0	4.00	3.00	1.000		
	10.0	4.00				
	15.0	8.00	7.00	1.000		
	20.0	5.00				
	25.0	4.50	3.50	1.000		
	30.0	2.30				
	35.0	2.00	1.00	1.000		
	40.0	3.50				
	45.0	5.00	4.00	1.000		
	50.0	3.80				
3	0.0	4.00				0.0385
	5.0	4.10	4.20	0.010		
	10.0	4.10				
	15.0	4.10	4.00	0.010		
	20.0	4.20				
	25.0	4.10	4.20	0.010		
	30.0	4.20				
	35.0	4.20	4.10	0.010		
	40.0	4.30				
	45.0	4.20	4.30	0.010		
	50.0	4.30				

(*Continued*)

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Table 15.1 Interpretation of the Coefficient of Determination (CoD) in different simulations (*Continued*).

Case	Time (d), Estimated Values (Y_{est}), and Observed Values (Y_{obs})				Observed and Estimated Values as a Function of Time	CoD
	Time (d)	Y_{est} (mg/L)	Y_{obs} (mg/L)	$(Y_{obs} - Y_{est})^2$		
4	0.0	3.80				−23.04
	5.0	3.70	4.20	0.250		
	10.0	3.60				
	15.0	3.50	4.00	0.250		
	20.0	3.60				
	25.0	3.70	4.20	0.250		
	30.0	3.70				
	35.0	3.60	4.10	0.250		
	40.0	3.80				
	45.0	3.80	4.30	0.250		
	50.0	3.90				
5	0.0	3.70				0.0000
	5.0	3.70	3.00	0.490		
	10.0	3.70				
	15.0	3.70	7.00	10.890		
	20.0	3.70				
	25.0	3.70	3.50	0.040		
	30.0	3.70				
	35.0	3.70	1.00	7.290		
	40.0	3.70				
	45.0	3.70	4.00	0.090		
	50.0	3.70				

(numerator of Equation 15.2), and therefore, *a parameter estimation process that aims to maximize the CoD is equivalent to one that minimizes the sum of the squares of the errors*. If the sum of the squares of the errors is an *absolute* measure of adherence, the CoD is a *relative* measure and can be used to compare the results of other simulations (as long as the limitations described above are taken into account).

Advanced

(c) Root-mean-square residual (RMSR or RMSE)

Root mean square residual (RMSR) or **root-mean-square error** (**RMSE**) is another indicator of a model's goodness-of-fit. It is based on the computation of the square root of the sum of the squared residuals (Equation 15.1) divided by the number of data points (the observed data that has been used to fit the model), as shown in Equation 15.3.

$$\mathrm{RMSR} = \sqrt{\frac{\sum_{i=1}^{n} (Y_{\mathrm{obs}\,i} - Y_{\mathrm{est}\,i})^2}{n}} \tag{15.3}$$

where

$Y_{\mathrm{obs}\ i}$ = observed value at time or position i in the data sequence
$Y_{\mathrm{est}\ i}$ = estimated value ($\hat{Y}$) at time or position i in the data sequence
n = number of data (number of pairs Y_{obs} and Y_{est}).

RMSR (or RMSE) displays good statistical behaviour and provides a direct measurement of the model residual. If RMSR is divided by the mean of the observed variable (**RMSR/$Y_{\mathrm{obs\ mean}}$**), it gives an indication of the *relative magnitude of the error* (Thomann, 1982). For a given variable, CoD and RMSR are directly related (both have the same numerator in the equation), but CoD has the additional advantage of allowing relative comparisons between different variables.

Advanced

(d) Relative residual (or relative error)

The computation of the relative residual (or relative error) is shown in Equation 15.4. Because this statistic is relative, it can be used to compare models. However, it does not behave well for low values of $Y_{\mathrm{obs\ mean}}$ and does not take into account the variability in the data (Thomann, 1982).

$$\mathrm{RR}(\%) = 100 \cdot \frac{|(Y_{\mathrm{obs\ mean}} - Y_{\mathrm{est\ mean}})|}{Y_{\mathrm{obs\ mean}}} \tag{15.4}$$

where

RR = relative residual (also called relative error)
$Y_{\mathrm{obs\ mean}}$ = mean of the observed values
$Y_{\mathrm{est\ mean}}$ = mean of the estimated values.

The numerator uses the absolute value of $Y_{\mathrm{obs\ mean}} - Y_{\mathrm{est\ mean}}$.

Advanced

(e) Relation between estimated and observed values

Another frequently used method to visually interpret the goodness-of-fit of a model is by graphing a **scatter-plot of the observed data versus the estimated data** (Y_{obs} versus Y_{est}). Figure 15.8 presents an example of such a graph. Not only are the points plotted here but you will notice that we also include a line with a slope of 1:1 (an angle of **45**°). This line indicates a **perfect theoretical fitting** – *points plotted on top of this line are those where the estimated value is exactly equal to the observed value*. Points plotted above the line indicate that the model *underestimated* the measured value, and points plotted below the line indicate that the model

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

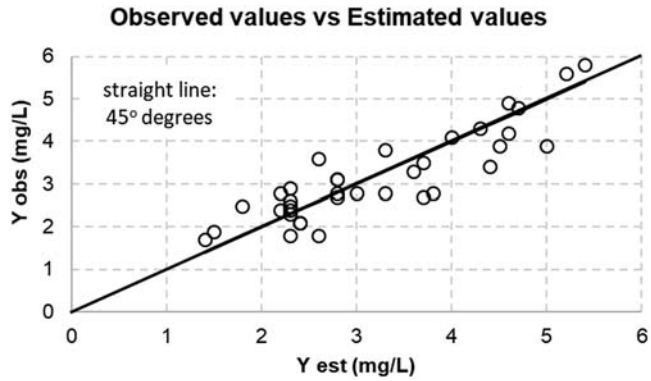


Figure 15.8 Example of a scatter-plot showing Y_{obs} versus Y_{est} (using the same data of Figure 15.7) and the 45° line (slope of 1:1).

overestimated the measured value. This graph is useful because you can use it to identify regions where your model may be overestimating or underestimating the observed data.

However, frequently researchers go beyond this simple interpretation and try to evaluate the goodness-of-fit of the model by completing a **linear regression analysis between the estimated data and the observed data**. The assumption usually considered is that if the R^2 coefficient of the regression analysis is close to 1, the fit is good. However, you should be **very careful** to not arrive at any **risky conclusions**, as illustrated in the cases exemplified in Table 15.2.

Table 15.2 showcases some examples of situations where the model fit is inadequate, but where there is still a high R^2 value for the linear regression between the estimated and observed values. To make the interpretation simple, we use an example data set composed of only three points. In the graphs on the left are plotted the observed and estimated concentrations in the data sequence (along the distance of the reactor or sampling time). In the graphs on the right, we present the typical linear regression graphs frequently used in this analysis (linear regression between the estimated and the observed values). In these last charts, the 1:1 (45°) line (shown as a dashed line) would indicate a perfect fit. If all the points are located exactly on top of this line, then $Y_{obs} = Y_{est}$ for all cases. *In the linear regression analysis, where* $Y_{obs} = a + b \cdot Y_{est}$, *the coefficient a (intercept) should be equal to zero and the coefficient b (slope) should be equal to 1, in case we have a perfect fit, i.e.,* $Y_{obs} = 1.0\ Y_{est}$. However, considerable confusion lies on this simple analysis, and **frequently inadequate interpretations** are made, as illustrated below.

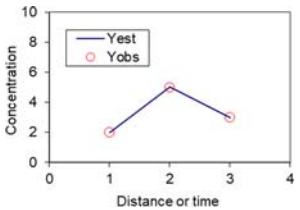
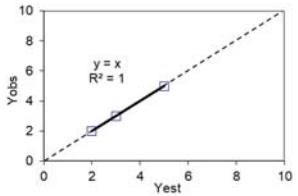
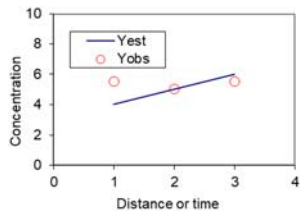
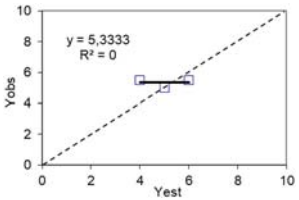
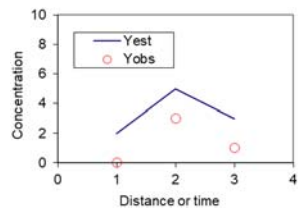
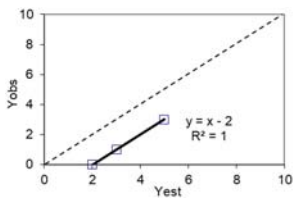
Only case 1 depicts a suitable fit (in this case, perfect) of the model. Case 2 shows a totally inappropriate fit, and this aspect is well portrayed by the R^2 regression coefficient of $Y_{obs} \times Y_{est}$, which, in this situation, is equal to zero. The other cases also showcase inadequate fittings (a or b values different from 0 and 1, respectively). However, despite of the poor fittings, the R^2 value is equal to 1 in all situations, which may potentially lead some people to incorrectly think that the model provides a perfect fit. Note that, in the charts on the right, the points are not on top of the 1:1 slope (45°) line.

Note that we are not suggesting that you should avoid this approach of plotting $Y_{obs} \times Y_{est}$. Rather, our point here is to emphasize the fact that **you should not rely solely on the interpretation of the R^2 value obtained from a linear regression between Y_{obs} and Y_{est}**. If you produce a simple scatter-plot between Y_{obs} and Y_{est} (as the one shown in Figure 15.8), it may be useful for you to identify how your model represents the experimental data. All plots

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

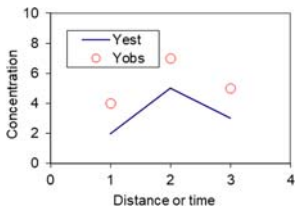
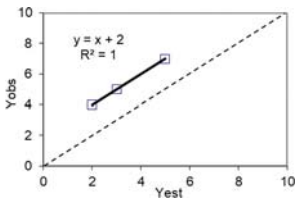
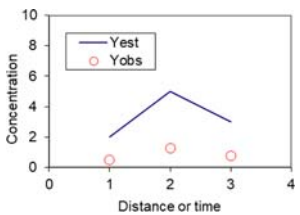
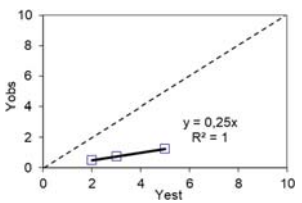
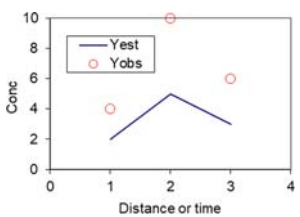
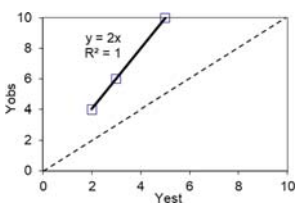
Table 15.2 Examples of situations with inadequate model fittings but with misleadingly perfect fitting ($R^2 = 1$) in the linear regression analysis between estimated and observed values ($Y_{obs} = a + b \cdot Y_{est}$).

Case	Distance or Time and Estimated (Y_{est}) and Observed (Y_{obs}) Values	Observed and Estimated Values as a Function of Distance or Time	Linear Regression Between Y_{est} and Y_{obs}	Comment
1	d or t Yest Yobs 1 2 2 2 5 5 3 3 3			Perfect fitting $a = 0$ $b = 1$ $R^2 = 1$
2	d or t Yest Yobs 1 4 5.5 2 5 5.0 3 6 5.5			Totally inadequate fitting. $a \neq 0$ $b \neq 1$ $R^2 = 0$
3	d or t Yest Yobs 1 2 0 2 5 3 3 3 1			Inadequate fitting. Estimated values are systematically greater than the observed values (constant residual). In this example, all Y_{est} values are 2 mg/L higher than the corresponding Y_{obs}. $a < 0$ $b = 1$ $R^2 = 1$

(*Continued*)

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Table 15.2 Examples of situations with inadequate model fittings but with misleadingly perfect fitting ($R^2 = 1$) in the linear regression analysis between estimated and observed values ($Y_{obs} = a + b \cdot Y_{est}$) (*Continued*)

Case	Distance or Time and Estimated (Y_{est}) and Observed (Y_{obs}) Values	Observed and Estimated Values as a Function of Distance or Time	Linear Regression Between Y_{est} and Y_{obs}	Comment
4	d or t Yest Yobs 1 2 4 2 5 7 3 3 5			Inadequate fitting. Estimated values are systematically lower than the observed values (constant residual). In this example, all Y_{est} values are 2 mg/L lower than the corresponding Y_{obs}. $a > 0$ $b = 1$ $R^2 = 1$
5	d or t Yest Yobs 1 2 0,5 2 5 1.25 3 3 0.75			Inadequate fitting. Estimated values are systematically greater than the observed values (constant ratio Y_{obs}/Y_{est}). In this example, all Y_{est} values are four times larger than the corresponding Y_{obs}. $a = 0$ $b < 1$ $R^2 = 1$
6	d or t Yest Yobs 1 2 4 2 5 10 3 3 6			Inadequate fitting. Estimated values are systematically lower than the observed values (constant ratio Y_{obs}/Y_{est}). In this example, all Y_{est} values are half of the corresponding Y_{obs}. $a = 0$ $b > 1$ $R^2 = 1$

in Table 15.2 (except for case 1 with the straight line of best fit) are useful and they would help you visually reveal how your model is representing the measured data and the regions with overestimations or underestimations.

15.2.4 Sensitivity analysis

Advanced

As commented in this chapter, for studies involving mathematical modelling, it is important to understand the effect of the magnitude of your input data (*variables or coefficients*). Collecting data for certain parameters sometimes requires a large investment of time and resources, and even more so if high precision is required. Thus, before completing the field and laboratory sampling efforts, analyses, and measurements (which involves expending time and resources), it is often worthwhile to assess whether or not it is important to have very precise values of the input data, or if you can get by with collecting these measurements with less precision without affecting too much the model results. To do this, you must complete what is known as a **sensitivity analysis**.

If the sensitivity analysis shows that a *small variation in the input data* causes a *large variation in the output results*, then your conclusion is that the model is ***sensitive*** to this specific input parameter, and efforts should be concentrated on obtaining its *reliable measurement with precision* (based on field work, lab work, or with an extensive literature review). On the other hand, if large variations in the input data cause only small variations in the output results, then there are no justifications to complete such costly work to determine this input data with high precision, and it might not be worth it to collect these data at all.

The following are different techniques that can be used to carry out a sensitivity analysis:

- Informal sensitivity analysis
- Parameter perturbation (one-at-a-time method)
- First-order sensitivity analysis
- Monte Carlo simulation

(a) Informal sensitivity analysis

A simple way to perform a sensitivity analysis for a given input parameter is to **run the model with different values of this input data** (*specified by you*) and **interpret whether the model results (output data) change much in response to this variation**. The specification of the different values of the input data, as well as the interpretation about the sensitivity of the model, will depend on your knowledge, experience, and good sense.

Figure 15.9 presents an example of two scenarios of the simulation of dissolved oxygen (DO) concentrations along a river stretch. Both graphs portray the simulated values of DO over a distance of 50 km, using the same set of parameter values (in each graph there are three simulations, each one with different values of one of the model input parameters, the reaeration coefficient). The left-hand graph indicates a model that has high sensitivity to this particular parameter. The right-hand graph, using the same set of parameter values, indicates a model that has low sensitivity to this parameter. The reason is that, in each scenario (left and right), the conditions were different: the left one simulated the discharge of untreated wastewater in the river, while the right one represented the condition of discharge of treated wastewater. In the latter case, the coefficient of reaeration had little influence on the DO concentration in the river. Because the model equations are non-linear (not shown here), you probably would not be able to anticipate these conclusions if you had not run the model once-at-a-time with the different parameter values and the different scenarios.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

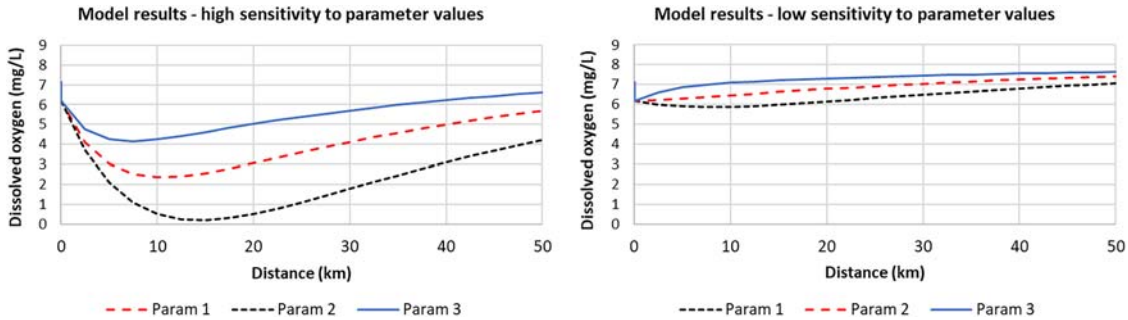


Figure 15.9 Two different scenarios of a sensitivity analysis. Each scenario is composed of three simulations, each one with a different value of a model coefficient.

(b) Parameter perturbation (one-at-a-time method)

Instead of varying a parameter depending on your judgement, as shown in 'a' above, you can complete a more formal assessment by systematically introducing specific variations in its value. For instance, you may vary a parameter value by 1% or by 10%, while keeping all other terms constant, and then see the resulting influence on the model output. The corresponding variations of the state variables reflect the sensitivity of the solution to the varied parameter. This is called **parameter perturbation**.

(c) First-order sensitivity analysis

To complete a **first-order sensitivity analysis**, you must take the derivative of the model function with respect to the parameter to estimate the model's sensitivity to it. This procedure is more complex, and further details can be found in Chapra (1997).

(d) Monte Carlo simulation

Monte Carlo simulation can be used for both **sensitivity analysis** and **uncertainty analysis**. Let us first comment a little on the concept of uncertainty analysis (von Sperling, 1990, 2014). Usually, we tend to focus on the model structure and the fit with the model output. Frequently, we do not question the reliability associated with the **input data**, whether they are variables or parameters (coefficients) of the system. It is not uncommon to see decisions taken, involving very high costs, supported by the results of a model, for which there is *insufficient reliability*, starting with the input data used to generate the model outputs. The following components are responsible for introducing uncertainty in the input data of a model:

- Errors in the estimation of an input data (based on literature values, surveys, personal experience, etc.)
- Sampling errors
- Errors in measurement, calibration, or laboratory analysis
- Errors in transcription or transfer of results from lab analysis or field measurements
- Errors in the estimation of future input data (in the case of a model that simulates future conditions)

Thus, we can observe that even traditionally unquestioned data used to run a model (such as measurements and lab results) are subject to a component of **uncertainty**. However, this variability in the input data can be incorporated into the interpretation of the results of the model, through the so-called uncertainty analysis.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

One of the techniques used to complete an uncertainty analysis is Monte Carlo simulation. This technique, in addition to allowing the completion of an uncertainty analysis, also allows for the completion of a sensitivity analysis and the expression of the model results in probabilistic terms (not simply as single deterministic values or point estimates). Therefore, someone using the model can make a managerial decision based on an indication of the probability of success or failure.

The essence of the Monte Carlo simulation is to run the model a large number of times (e.g., 1000 or 10,000), instead of carrying out the model simulation only once. In each run, we use a different set of values of the input data we are analysing. Each value is randomly generated, according to a selected distribution (uniform, normal, and log-normal), within a predefined range (minimum and maximum values) or criteria (e.g., mean and standard deviation). The more complex the model is, and the greater the number of input data, the larger is the required number of model runs or Monte Carlo simulations. As a result of the Monte Carlo method, we will obtain thousands of different independent model outputs, each associated with a different combination of model inputs, and this information can be used to perform statistical analyses that lead to the following results and conclusions:

- **Expression of results in probabilistic terms**. For instance, we can conclude that, based on the model results, we have a probability of, say, 70% of not complying with regulatory standards.
- **Determination of the sensitivity of the model results to the input data**. Based on regression analysis or hypothesis tests, we can use the thousands of model results to infer whether our model's output is sensitive to a particular input parameter.

There is plenty of literature on the technique of Monte Carlo simulation, and we will not cover it further here. Therefore, if you find that it can be useful for your model studies, due to its power and inherent simplicity, you should go in more depth and consult the relevant literature.

15.3 MODEL VERIFICATION (ANALYSIS OF RESIDUALS)

15.3.1 Required properties for the residuals

Advanced

As stated previously, the residuals correspond to the observed values of Y minus the estimated values of Y ($Y_{obs} - Y_{est}$). The analysis of the residuals can provide important information about the performance of our model, and about the possible need for data transformations or changes in the model structure, which may be necessary to improve the model's explanatory capacity. The formal analysis of our model's performance should also include the assessment of residuals, which is also known as the **model verification** step (Beck, 1983).

Completing a **residual analysis** is not practiced in many treatment plant and water quality modelling studies. *In most cases, authors only report goodness-of-fit graphs* of the observed and estimated (modelled) data, *and maybe some goodness-of-fit indicators*, such as the Coefficient of Determination (CoD). However, we would like to stress the importance of including a **residual analysis in your modelling study**, so that your model can be more robust. In your report, you may decide whether the description of the residual analysis is incorporated into the body of the text or if it is instead placed in an appendix or as supplementary material.

The analysis of residuals is greatly facilitated by a **visual evaluation of graphs**, in which the residuals are plotted in a sequence. Figure 15.10 illustrates different possible situations of residual plots and demonstrates how they should be interpreted in terms of the model performance (Nascimento *et al.*, 1996). In the figure, the X axis represents the data sequence, either represented by time (in case we have a time series of data) or distance (in case we are monitoring different positions in a reactor), and the Y axis represents the residuals.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

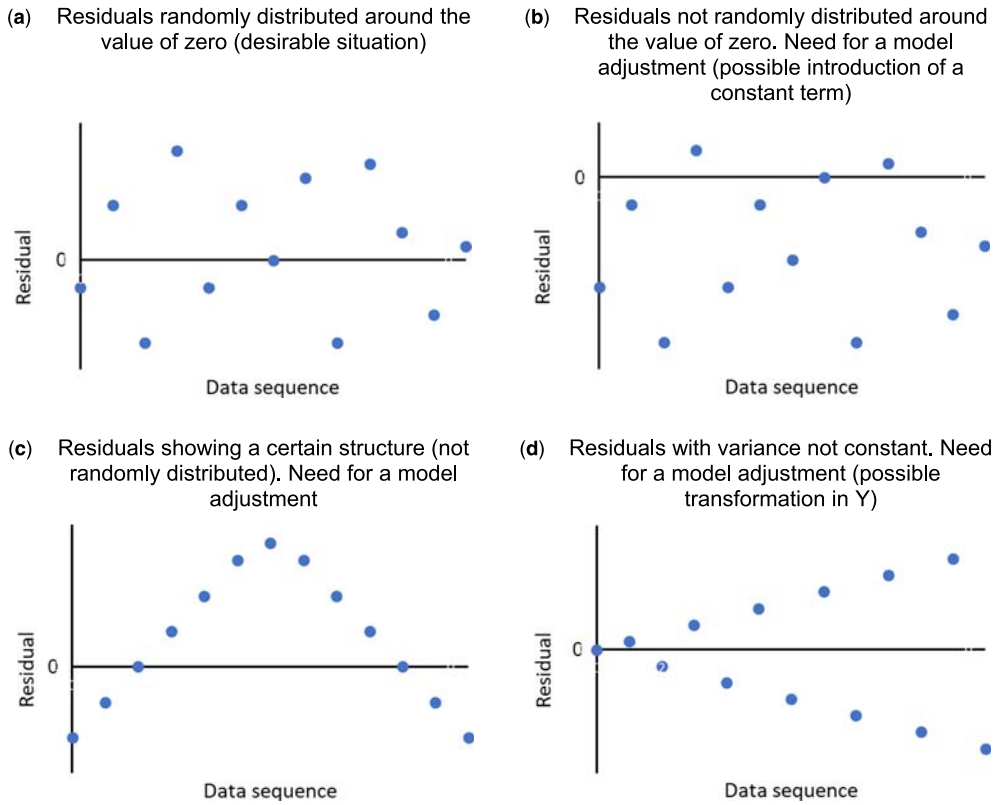


Figure 15.10 Scatter-plot of residuals along a data sequence of time (simulation of a time series) or distance (simulation along a reactor's length).

Besides the visual interpretation of the residual plots, the following **assumptions related to the residuals** must be satisfied in the model verification stage (Beck, 1983; von Sperling, 1990):

- The residuals should be randomly distributed around the mean, and the probability **distribution** of the residuals should be **normal**.
- The **mean** of the probability distribution of the residuals should be **zero**.
- The **variance** of the distribution of the residuals should be **constant** (e.g., with respect to time, distance, or sequence of samples).
- The **residuals** should be **independent** from each other, showing **no autocorrelation** (the residual at a given time should not be correlated with residuals in previous or subsequent time periods).
- The series of residuals should **not be correlated with other series of residuals** associated with other modelled variables.
- The series of residuals should **not be correlated with the series of the input variables** to the model.

In this book, we will show you the basics of assessing compliance with these criteria. However, you should consult statistical textbooks if you want to expand your knowledge and learn more advanced concepts

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

related to residual analysis. In most books, these methods are frequently included in the description of regression analysis. We also covered residual analysis in Section 11.5.4, Chapter 11, that deals with regression analysis (go there for further discussions on this topic). However, **we can apply the same principles here for any model, whether or not it is based on regression analysis**. The last two items in the list shown above will not be addressed here, because we would need more specific knowledge about the model being used and its variables and input data, which is not the case for this chapter, since we are not covering any particular model.

S. 11.5.4

15.3.2 Assessing the normality of the distribution of residuals

Advanced

You can assess if your model's residuals follow a normal distribution using the same approach that is presented in Section 8.2.8. Additionally, Example 15.2 presents an application of these concepts for the analysis of residuals. In summary, we have the following tools that can be used to check for the normality of residuals:

S. 8.2.8

- Graphical analysis
  - Normal probability plots
  - Q–Q plots (quantile–quantile plots)
- Interpretation of the skewness coefficient
- Statistical tests for normality or goodness-of-fit tests

S. 8.2.8

In Section 8.2.8, we presented a method for using graphical analysis to assess adherence to the normal distribution, using **normal probability plots** and **Q–Q plots**, based on fitting the points to a straight line. The graphs have similar concepts but are presented in different ways:

- The Q–Q plot shows the Z quantiles on both axes (theoretical quantiles on the X axis and quantiles of the measured data on the Y axis).
- The normal probability plot shows the theoretical quantiles on the X axis and the values of the residuals on the Y axis. Some people prefer to invert the positions of the X and Y axes for this plot.

The **skewness coefficient** will assist you to analyse the symmetry of the data (the skewness coefficient for data that follow a normal distribution should be equal to zero). For a **right-skewed** distribution, the skewness value is **positive**, and for a **left-skewed** distribution, the value is **negative**. The Excel function for skewness is SKEW. Consult Section 8.2.6 for a more detailed description of the concept of skewness.

S. 8.2.6

S. 8.2.8

Testing for normality can also be done using statistical hypothesis tests, such as the **Shapiro–Wilk** test. A short description was given in Section 8.2.8, but because the implementation of this test is somewhat complex, we did not go into much detail (as this type of statistical test is outside the scope of this book), but rather we suggested that you use a statistical software to implement it. The main information you need to look for is the resulting ***p*-value** from the test. The p-value should be interpreted in comparison with a specified **significance level (α)**. Usually a significance level of $\alpha = 0.05$ (5%) is used, implying a confidence level of 0.95 (95%). The interpretation of the p-value from a Shapiro–Wilk test (or any equivalent test) is as follows:

- If the ***p*-value is less than the significance level (α)** (e.g., p-value < 0.05), then the distribution of your residuals is *significantly different from the normal distribution*.
- If the ***p*-value is greater than or equal to the significance level (α)** (e.g., p-value ≥ 0.05), then the distribution of your residuals is ***not*** *significantly different from the normal distribution*.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

15.3.3 Testing whether the residual mean is significantly different from zero

Advanced

We have mentioned that we expect that the **mean of our model residuals should be equal to zero**. To determine this with confidence, we need to use a hypothesis test (in this case, a one-sample two-tailed test). Chapter 10 describes this test in detail, including the parametric Z and t tests and also the non-parametric tests.

C. 10

Here, we will use the ***t*-test** to demonstrate. One of the requirements for the t-test is that the underlying population from which our sample was obtained is *approximately normal*. Since normality of the data is one of the required properties of model residuals (see section above about the Shapiro–Wilk test), we can consider that this requirement is fulfilled. However, if you still prefer to use non-parametric tests, you can use those suggested in Chapter 10.

C. 10

We use a **two-tailed test** because we have rejection regions on both sides of the mean value, since our alternative hypothesis is that the mean is different from zero, and we do not have any strong reason to believe that it should be lower or that it should be higher (consult Chapter 10 for more about this concept of rejection regions). We will use the traditional significance level of 5% ($\alpha = 0.05$), that is, a confidence level of 95%, and establish our null and alternative hypotheses as follows:

C. 10

Test if residual mean is significantly different from zero

- Null hypothesis $\boldsymbol{H_0}$**: mean = 0**
- Alternative hypothesis $\boldsymbol{H_a}$**: mean ≠ 0**

Significance level for the test: $\boldsymbol{\alpha = 0.05}$

Interpretation:

- If ***p*-value** $<$ **0.05**: Reject null hypothesis that mean = zero (equivalent to supporting the alternative hypothesis, and stating that mean ≠ zero).
- If ***p*-value** $\geq$ **0.05**: Do not reject null hypothesis that mean = zero (note that we cannot say that we accept the null hypothesis that mean = zero).

See Example 15.2 to see how to apply the t-test for a residual's analysis.

15.3.4 Checking whether the variance is constant (homoscedasticity of variance)

Advanced

Here, for the sake of simplicity, we will not complete specific tests for assessing the homoscedasticity of the variance of our model's residuals. Our interpretation will be **visual**, based on plots of the residuals. Take the plot in Figure 15.10(d). We can clearly see that the variance here is not constant along the data sequence. In fact, it starts small and then increases with respect to the X axis. For a **model based on regression analysis**, this would be an indication that we should **make a transformation in our observed data** to be used in the regression analysis. Transforming the data by taking the log value or the square root are two examples that may help to stabilize the variance here. Other common transformations include inverting the data (1/data) and taking the squared value of the data.

15.3.5 Evaluating the existence of autocorrelation in the residuals

Advanced

To visually assess the independence of the errors, the **residuals should be plotted in the order in which the data were collected**. It may be that the residual is related to the previous residual, in case our observed variable has been obtained as a time series or a series of values along the position in a reactor. If this

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

type of relationship exists between consecutive residuals, the graph of residual versus data sequence usually displays a cyclic pattern, known as autocorrelation. When plotting the residuals in a sequential order, if there is a positive autocorrelation, there will be a sequence of residuals with the same sign, and thus it is possible to detect an apparent pattern. Figure 15.11 shows an example of a residual plot that indicates autocorrelation, because we can detect a cyclic pattern in the series of residuals, with a sequence of positive values followed by a sequence of negative values.

A **formal assessment** of independence and autocorrelation involves more advanced concepts and procedures that go beyond the scope of our book. We would like our residuals to follow a random pattern, in which there are no autocorrelations. This may involve *removing trends* in the residual series by processes of *non-seasonal decomposition*, aiming to make the new series stationary. One such process is called *first-order differencing*, which is where we subtract the series of residuals by the same series with a lag of one-period (one interval in our data sequence). Environmental data are also subject to seasonality (daily cycles of hourly variations or annual cycles of monthly variations). Seasonality also influences the analysis of autocorrelation, which may require that we complete some procedures of *seasonal decomposition* to remove the cyclic pattern. If we remove trend and seasonality, we can do more advanced analysis based on the so-called *autocorrelation function (ACF)*. Statistical software that has a time-series component is capable of completing this type of analysis.

S. 11.4

In our book, we present the calculation of ACF and the associated plot (autocorrelogram) for any time series and, as a consequence, also for our residuals. Go to Section 11.4, where autocorrelation is discussed, and insert your sequence of residual values in the associated spreadsheet and see how to interpret it. See Example 11.6, where this calculation has been performed using the residuals from Example 15.2.

Given the relatively complex nature of the aspects listed above, we will describe here a simple approach for assessing the autocorrelation of a series, using the **Durbin–Watson** (DW) procedure. This statistic measures the *correlation between each residual and the residual for the preceding time period* throughout the data sequence. The statistic is calculated as follows:

$$\mathrm{DW} = \frac{\sum_{i=2}^{n} (e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2} \tag{15.5}$$

where

DW = Durbin–Watson statistic
e_i = residual at position i in the sequence
e_{i-1} = residual at position $i - 1$ in the sequence.

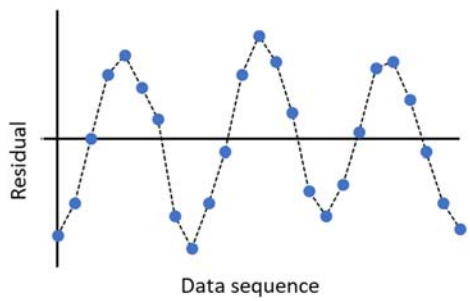


Figure 15.11 Data series of residuals showing a cyclic pattern and indications of autocorrelation.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

We should note that this evaluation is based on a **first-order autocorrelation analysis**, that is, it is based on the *relation between one residual value and its preceding value in the sequence*. We do not cover here autocorrelations associated with seasonality, in which we would need to study lags of several data intervals (for instance, hourly values that are correlated with values obtained 24 h before, because of daily cyclical patterns). The Durbin-Watson test is usually presented in textbooks when analyzing residuals arising from models based on **regression analysis**.

The application of the Durbin-Watson statistic is demonstrated in Example 15.2. The numerator represents the **sum of the squared difference between two successive residuals** and can be calculated using the Excel function **SUMXMY2** (of the residual sequence with one lag, covering the sequence from 2 to n; residual sequence without lag, covering the sequence from 1 to $n-1$). The denominator is the **sum of the squares of the residuals** and can be calculated using the Excel function **SUMSQ** (of the residual sequence, from 1 to n). Using Equation 15.5, we can obtain values ranging from 0 up to 4, which can be interpreted as follows:

- If residuals are **positively correlated**, DW approaches **0**.
- If residuals are **negatively correlated** (which happens less frequently), DW approaches **4**.
- If there is **little autocorrelation**, DW approaches **2**.
- **In most cases**, when DW is between 1.5 and 2.5, there are usually no indications of autocorrelation. However, we can perform a more careful assessment using critical values, as demonstrated below.

Even though these concepts may seem simple, in practice it may be difficult to *interpret some of the intermediate DW values* and draw conclusions about whether they indicate a strong or a weak autocorrelation. To help with this the Durbin-Watson statistics are supported by a **look-up table**, which presents two reference values: $\boldsymbol{d_L}$ **(lower critical value)** and $\boldsymbol{d_U}$ **(upper critical value)**. The tabulated values of d_L and d_U vary depending on the number of data points (i.e., the sample size n), the number of independent variables included in the model (p), and the significance level (usually adopted as $\alpha = 0.05$). For a simple linear regression model (e.g., $y = a + b \cdot x$), $p = 1$.

We will not present the full table here, but just a summary of it, with ranges of d_L and d_U values that are sufficient for the purposes of our interpretation (see Table 15.3). For instance, if we have 36 data

Table 15.3 Values of d_L and d_U for the interpretation of first-order autocorrelation based on Durbin-Watson statistics, for different numbers of independent variables (p) and sample sizes (n). Significance level $\alpha = 0.05$.

Number of Independent Variables (p)	Range of Sample Size	d_L	d_U
$p = 1$	$15 \leq n < 30$	1.08–1.34	1.36–1.49
	$30 \leq n < 50$	1.35–1.53	1.49–1.60
	$50 \leq n < 100$	1.53–1.65	1.60–1.69
$p = 2$	$15 \leq n < 30$	0.95–1.27	1.54–1.56
	$30 \leq n < 50$	1.28–1.45	1.57–1.63
	$50 \leq n < 100$	1.46–1.63	1.63–1.72
$p = 3$	$15 \leq n < 30$	0.82–1.20	1.75–1.65
	$30 \leq n < 50$	1.21–1.41	1.65–1.67
	$50 \leq n < 100$	1.42–1.61	1.67–1.74

Notes: Within a range of sample size, the lower the value of n, the lower are the values of d_L and d_U. For instance, in the first row, for $p = 1$ and $n = 15$, $d_L = 1.08$, and for $n = 29$, $d_L = 1.34$.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

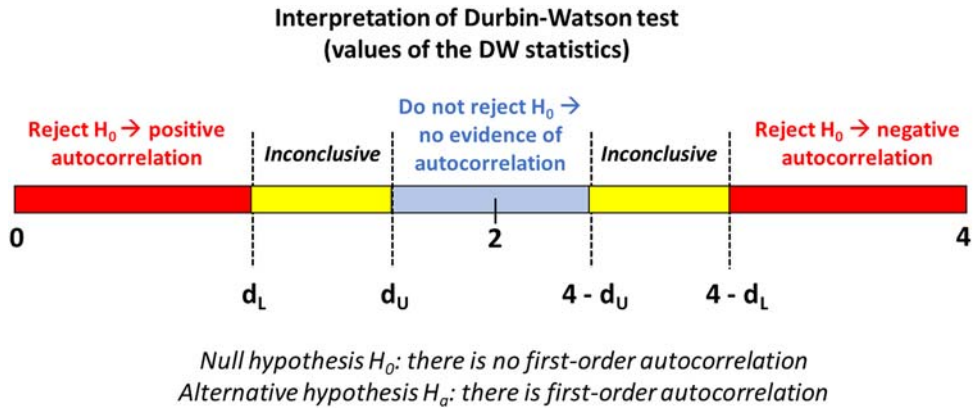


Figure 15.12 Interpretation of the Durbin-Watson test based on the relative position of DW, compared with d_L and d_U. *Source:* adapted from Brooks (2014).

points ($n = 36$) and our model has only one independent variable ($p = 1$), from the table we see that d_L is between 1.35 and 1.53, say, $d_L = 1.40$. Similarly, d_U will be between 1.49 and 1.60, say, $d_U = 1.53$.

This summary-table was constructed based on a complete table presented at Levine *et al.* (1988).

With the values of DW, d_L, and d_U, we can interpret the likelihood that our residuals series is autocorrelated. Figure 15.12 shows a simple schematic of this interpretation. The null and alternative hypotheses (H_0 and H_a) that we use are shown in the figure. For instance, using the example values shown in the preceding paragraph ($d_L = 1.40$ and $d_U = 1.53$), we can have the following possibilities (at 5% significance level): (a) if DW < 1.40: there is positive autocorrelation; (b) if DW is between 1.40 and 1.53: the test is inconclusive; (c) if DW is between 1.53 and 2.47 ($4 - 1.53 = 2.47$): there is no evidence of autocorrelation; (d) if DW is between 2.48 and 2.60 ($4 - 1.40 = 2.60$): the test is inconclusive; (e) if DW > 2.60: there is negative autocorrelation.

Example

S. 15.2

S. 15.3

EXAMPLE 15.2 ANALYSIS OF THE MODEL RESIDUALS

Carry out a residual analysis based on the observed and estimated values listed in the following table. Follow the procedures described in Sections 15.2 and 15.3.

Data Sequence	Y_{obs} (mg/L)	Y_{est} (mg/L)	Data Sequence	Y_{obs} (mg/L)	Y_{est} (mg/L)
1	2.8	2.8	19	2.7	2.8
2	4.2	4.6	20	3.1	2.8
3	3.9	4.5	21	2.8	3.3
4	3.3	3.6	22	3.4	4.4
5	2.8	2.2	23	4.9	4.6
6	1.7	1.4	24	2.8	3.8

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Data Sequence	Y_{obs} (mg/L)	Y_{est} (mg/L)	Data Sequence	Y_{obs} (mg/L)	Y_{est} (mg/L)
7	1.9	1.5	25	2.8	3.0
8	2.5	1.8	26	1.8	2.6
9	3.1	2.8	27	2.1	2.4
10	3.8	3.3	28	2.6	2.3
11	2.7	3.7	29	2.3	2.3
12	4.1	4.0	30	2.4	2.2
13	4.3	4.3	31	2.5	2.3
14	4.8	4.7	32	1.8	2.3
15	5.6	5.2	33	2.9	2.3
16	5.8	5.4	34	2.4	2.3
17	3.9	5.0	35	2.1	2.4
18	3.5	3.7	36	3.6	2.6

Excel

Note: This example is also available as an Excel spreadsheet.

Solution:

(a) Calculation of the residuals and the squares of the residuals

The full calculation of the residuals and of the squares of the residuals is given in the Excel spreadsheet. We show here only the structure of the calculation.

Data Sequence	Y_{obs} (mg/L)	Y_{est} (mg/L)	Residual ($Y_{obs} - Y_{est}$)	Residual Squared
1	2.8	2.8	0.0	0.00
2	4.2	4.6	−0.4	0.16
3	3.9	4.5	−0.6	0.36
4	3.3	3.6	−0.3	0.09
5	2.8	2.2	0.6	0.36
…	…	…	…	…
35	2.1	2.4	−0.3	0.09
36	3.6	2.6	1.0	1.00
			Sum	**9.73**

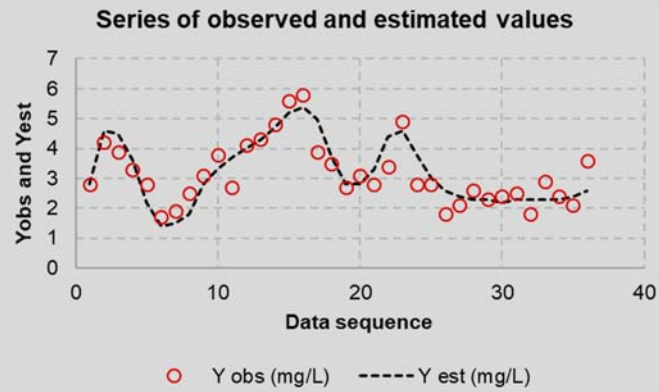
The basic statistics of the residuals are

Number of data (n)	36
Mean	−0.042
Standard deviation	0.526

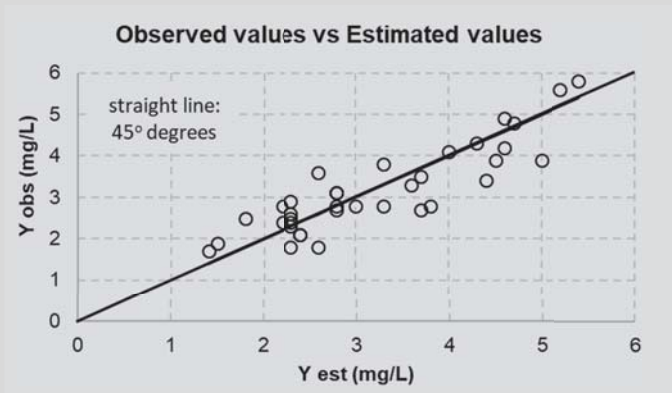
(b) Goodness-of-fit of the model simulations

The plot of the observed and estimated values is shown below. Visually, we can infer that the estimated values follow the main trends of the observed data.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest



The scatter-plot of the observed versus estimated values is shown in the graph below, along with a line with a slope of 1:1 (45° angle).



The main goodness-of-fit statistics are presented below. We will not show the full calculations, but you can find them in the associated Excel spreadsheet.

- Sum of the squares of the residuals (SSR) – Equation 15.1 (calculation shown in the table in item 'a' of this example):

$$SSR = \Sigma(Y_{obs} - Y_{est})^2 = \mathbf{9.73}$$

- Coefficient of Determination (CoD) – Equation 15.2:

$$CoD = 1 - \frac{\Sigma(Y_{obs} - Y_{est})^2}{\Sigma(Y_{obs} - Y_{obs\ mean})^2} = 1 - \frac{9.73}{37.69} = \mathbf{0.7418}$$

- Root-mean-square residual (RMSR) – Equation 15.3:

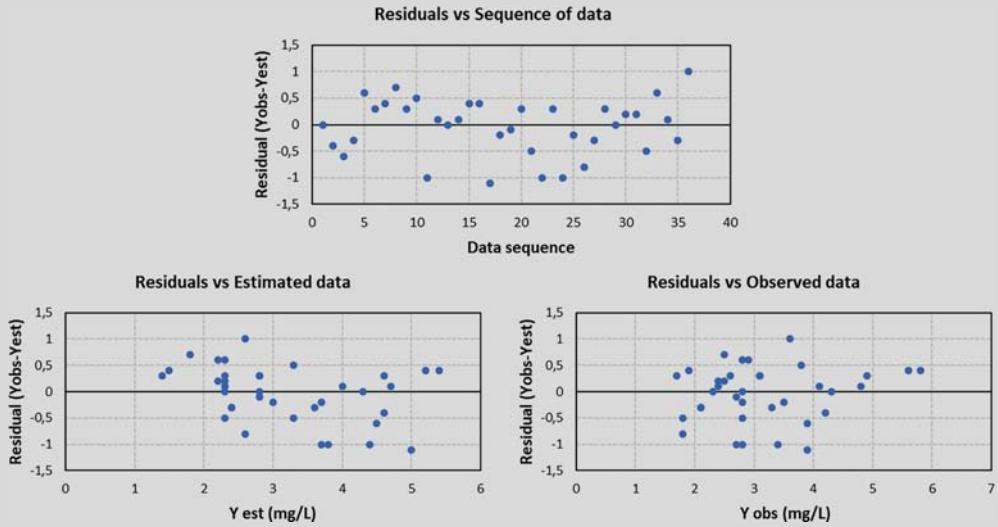
$$RMSR = \sqrt{\frac{\sum (Y_{obs} - Y_{est})^2}{n}} = \sqrt{\frac{9.73}{36}} = \mathbf{0.5199}$$

(c) Residual plots

A series of plots representing the residuals as a data sequence and as a function of the observed and estimated values is shown below. No abnormal behaviour can be detected.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Therefore, we can state that this requirement has been satisfied.

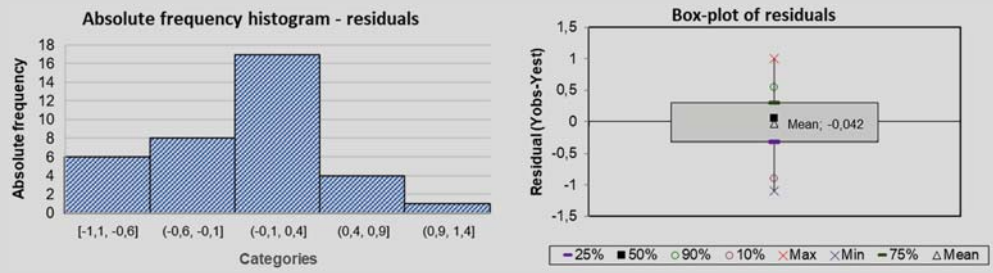


(d) Assessment of the adherence of the distribution of the residuals to the normal distribution

S. 15.3.2

The sequence of plots and calculations follows the description presented in Section 15.3.2.

The frequency histogram and the box-plot assist us in the visual interpretation of the adherence of the residual distribution to the normal distribution. The histogram does not show strong deviations from the typical bell-shaped curve from the theoretical normal distribution, and the box-plot does not reveal any substantial departure from normality.

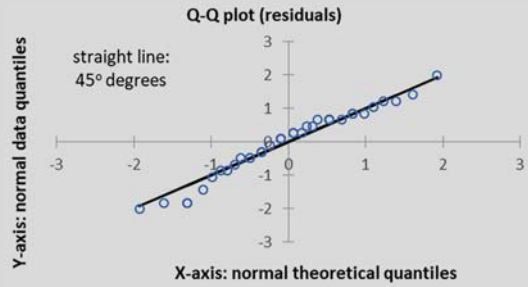


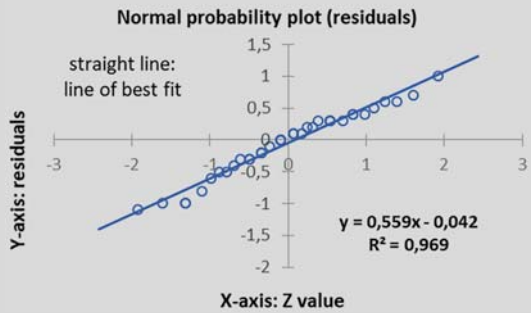
The **skewness coefficient** of the residuals, using the Excel function SKEW, is **−0.420**. The skewness coefficient of a theoretical normal distribution is zero.

The two main plots for assessing adherence to the normal distribution (Q–Q plot and normal probability plot) are shown below. No substantial deviations from the expected behaviour of a theoretical normal distribution can be seen.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Therefore, we can state that this requirement has been satisfied.





If we undertake a statistical test for assessing adherence to the normal distribution, we can state the conclusions in a more formal way. We carried out the **Shapiro–Wilk** test using a statistical software (calculations not shown here neither in the Excel spreadsheet) and obtained the *p*-value of **0.2229**. Since this *p*-value is $\geq$0.05, we can conclude that the distribution of the residuals is *not significantly different from a normal distribution.*
Therefore, we can state that this requirement has been satisfied.

(e) Evaluate whether the mean of the residuals is significantly different from zero

To test the property that the mean of the residuals should be equal to zero, we apply the *t-test* (see Excel spreadsheet). The hypotheses we establish are

- Null hypotheses $\boldsymbol{H_0}$**: mean** $=$ **0**
- Alternative hypothesis $\boldsymbol{H_a}$**: mean** $\neq$ **0**

We obtain the following result:
***p*-value: 0.637**.
Since this value is greater than 0.05, we can say that, at the 5% significance level, *we cannot reject the hypothesis that the mean of the residuals is equal to zero*.
Therefore, we can state that this requirement has been satisfied.

(f) Assess whether the variance of the data is constant

We will use here a simple visual interpretation of the plot of the data sequence of the residuals (first plot on item 'c' of this example). Apparently, the residuals are approximately evenly distributed around the zero value, without any trends of increasing variance. In this example, we have not specified our model structure, and therefore, we do not have an independent variable to use in additional plots of the residuals. However, if we use the estimated and observed values in the

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

plots (second and third plots in item 'c'), we see that the points are distributed mainly as clouds around the zero value, without any marked narrowing or widening, suggesting that the variance (based on the square of the residuals) appears to be constant.
Therefore, we can state that this requirement has been satisfied.

(g) Assess whether the residuals are autocorrelated

S. 15.3.4

The assessment of autocorrelation of the residuals is done using the Durbin-Watson test (see Section 15.3.4), which evaluates first-order autocorrelation (whether the residuals in a data sequence are correlated with the residuals immediately before in the data sequence). The Durbin-Watson statistic is given by Equation 15.5:

$$DW = \frac{\sum_{i=1}^{n} (e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}$$

To calculate the numerator and denominator of the equation, we will use Excel functions (see Excel spreadsheet for this example).

- Numerator: sum of the squared difference between two successive residuals: Excel function **SUMXMY2** (of the residual sequence with one lag, covering the sequence from 2 to n; of the residual sequence without lag, covering the sequence from 1 to $n-1) = 18.10$.

  For you to understand how this calculation was done, let us take the residual values shown in section 'a' of this example:

Data Sequence (i)	Residual at i	Residual at i-1	$(\text{Residual}_i - \text{Residual}_{i-1})^2$
1	0.0	-	
2	-0.4	0.0	$[-0.4-(0.0)]^2 = 0.16$
3	-0.6	-0.4	$[-0.6-(-0.4)]^2 = 0.04$
4	-0.3	-0.6	$[-0.3-(-0.6)]^2 = 0.09$
5	0.6	-0.3	$[0.6-(-0.3)]^2 = 0.81$
...	...	...	…
35	-0.3	0.1	$[-0.3-(0.1)]^2 = 0.16$
36	1.0	-0.3	$[1.0-(-0.3)]^2 = 1.69$
Sum			**18.10**

- Denominator of the DW statistic: sum of the squares of the residuals (SSR), which was calculated in the table in item 'a' of this example as **9.73**. Its calculation can also be done using the Excel function **SUMSQ** (of the residual sequence, from 1 to n).

With these values, we can calculate the DW statistic:

$$DW = \frac{\sum_{i=1}^{n} (e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2} = \frac{18.10}{9.73} = 1.86$$

S. 15.3.4

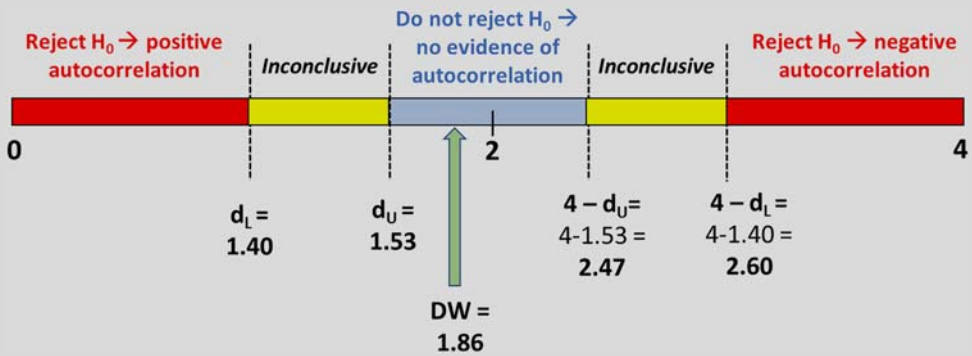
In most cases, when the *DW statistic is in the range of 1.5 and 2.5, there are usually no indications of autocorrelation*. However, we can perform a more careful assessment, using the critical values d_L and d_U, as presented in Section 15.3.4. In our example here, we have not described the structure of the model we are using. Therefore, we do not know the number of independent variables in the model. However, let us assume that we are doing a simple linear regression (only one independent variable). If this were the case, then we can obtain the values of d_L and d_U from Table 15.3 as follows:

$p = 1$ (only one independent variable)
$n = 36$ (use the range between 30 and 50)

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

$d_L = 1.35–1.53 \rightarrow$ let us assume 1.40
$d_U = 1.49–1.60 \rightarrow$ let us assume 1.53

Inserting DW, d_L, and d_U in Figure 15.12, we obtain the figure shown below. From it, we can see that DW is situated in a region which indicates that there is no evidence of autocorrelation. *Therefore, we can state that this requirement has been satisfied.*



S. 11.4

If you wish, you can plot the autocorrelogram of the residuals, as described in Section 11.4.

(h) Overall assessment

From the overall analysis of residuals that we undertook in this example, we can conclude that the residuals from our model are complying with the required conditions, which is a positive result regarding the verification of the model.

15.4 CHECK-LIST FOR YOUR REPORT

- ✓ If you are using a model, make sure you describe it properly in your publication. If you are applying a widely known model that is already extensively described in the literature, it is possible that you do not need to present its full structure, but rather you can refer to the publications. However, if you are applying a less-known model, or if you developed your own model, you will need to describe it fully, including all of the equations and all of the data used to calibrate it, so that other people may be able to implement and use it themselves.
- ✓ In any case, you need to clearly present all input data used (input variables and model parameters) and show how you obtained them. A convenient form of presenting the data is to summarize all of the values in a table. See Chapter 4 for more details about storing and publishing your data in appropriate formats and outlets.
- ✓ Regarding the parameter values, make it clear whether you used literature values or completed your own model calibration. If you adopted the latter strategy, indicate the procedure used for calibrating the model. Report any limitations associated with this procedure.
- ✓ Make sure you present the most important graphs of observed and estimated values for the variable you are studying. The key graphs may be inserted in the body of the text, while other less important charts may go into an Appendix or Supplementary Material.

C. 4

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

- ✓ Present suitable indicators of goodness-of-fit, such as the Coefficient of Determination (CoD), and interpret them in the report. Do not rely only on visual interpretation of model fitting. Reduce subjectivity, because the readers may have a different opinion from you when they look at your plots.
- ✓ If possible, try to include statements associated with your model verification (residual analysis). You may not have space to present all of the analyses and graphs in your report, but you may state whether your residuals complied with the required properties, and you can present the residual analysis in a summarized way in an Appendix or in Supplementary Material.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

References

ABNT (1987). NBR9897. Planejamento de amostragem de efluentes líquidos e corpos receptores. (Planning of liquid effluent and receiving bodies sampling). Associação Brasileira de Normas Técnicas (in Portuguese).

Abu-Reesh I. M. and Abu-Sharkh B. F. (2003). Comparison of axial dispersion and tanks-in-series models for simulating the performance of enzyme reactors. *Industrial & Engineering Chemistry Research*, **42**, 5495–5505.

ACTION STAT (2019). Statistical software. Manual. www.portalaction.com.br (accessed 23 February 2019) (in Portuguese).

APHA (2017). Standard Methods for the Examination of Water and Wastewater, 23rd edn, American Public Health Association, Washington, DC.

Arceivala S. J. (1981). Wastewater Treatment and Disposal. Marcel Dekker, New York.

Armbruster D. A. and Pry T. (2008). Limit of blank, limit of detection and limit of quantitation. *The Clinical Biochemist Reviews*, **29**(Suppl. 1), S49.

Austin B. J., Scott J. T., Daniels M. and Haggard B. E. (2016). Water Quality Reporting Limits, Method Detection Limits, and Censored Values: What Does It All Mean? Arkansas Water Resources Center, FS-2016-01, Fayetteville, AR, 8 pp.

Barnett V. (2004). Environmental Statistics: Methods and Applications. John Wiley & Sons, Inc., New York. ISBN: 978-0-471-48971-9.

Beck M. B. (1983). A procedure for modeling. In: Mathematical Modeling of Water Quality: Streams, Lakes and Reservoirs, G. T. Orlob (ed.), John Wiley & Sons, New York, pp. 11–41.

Benefield L. D. and Randall C. W. (1980). Biological Process Design for Wastewater Treatment. Prentice-Hall, EUA, Upper Saddle River, NJ, 526 p.

Berthouex P. M. and Hunter W. G. (1975). Treatment plant monitoring programs: a preliminary analysis. *Journal of Water Pollution Control Federation*, **47**(8), 2143–2156.

Berthouex P. M. and Hunter W. G. (1981). Simple statistics for interpreting environmental data. *Journal of Water Pollution Control Federation*, **53**(2), 167–175.

Berthouex P. M. and Hunter W. G. (1983). How to construct reference distributions to evaluate treatment plant effluent quality. *Journal of Water Pollution Control Federation*, **55**(12), 1417–1424.

Bertolo (2019). Frequency distributions. www.bertolo.pro.br/FinEst/Estatistica/Planilhas/distribs.htm (accessed 23 February 2019) (in Portuguese).

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Box G. E. P., Jenkins G. M., Reinsel G. C. and Ljung G. M. (2015). Time Series Analysis: Forecasting and Control, 5th edn, Wiley, New York, 712 p. ISBN 1118675029.

Brooks C. (2014). Introductory Econometrics for Finance, 3rd edn, Cambridge University Press, Cambridge, 740 p.

Burr I. W. (1976). Statistical Quality Control Methods, Marcel Dekker, Inc., New York, Vol. **16**, 522 p.

Cantor A., Kiparsky M., Kennedy R., Hubbard S., Bales R., Pecharroman C. L., Guivetchi K., McCready C. and Darling G. (2018). Data for water decision making: informing the implementation of California's open and transparent water data act through research and engagement. *Center for Law, Energy & The Environment Publications*, **56**. https://scholarship.law.berkeley.edu/cleepubs/56

Chapra S. C. (1997). Surface Water Quality Modeling. WCB/McGraw-Hill, New York, 844 p.

Charles K. J., Ashbolt N. J., Roser D. J., McGuinness R. and Deere D. A. (2005). Effluent quality from 200 on-site sewage systems: design values for guidelines. *Water Science and Technology*, **51**(10), 163–169.

Cheng S. W. and Xie H. (2000). Control charts for lognormal data. *Tamkang Journal of Science and Engineering*, **3**(3), 131–137.

Chernicharo C. A. L. and Bressani T. (2019). Anaerobic Reactors for Sewage Treatment: Design, Construction and Operation. IWA Publishing, London, 399 p.

Chow V. T., Maidment D. R. and Mays L. W. (1988). Applied Hydrology, McGraw-Hill, New York, 572 p.

Cohen J. (1988). Statistical Power Analysis for the Behavioral Sciences, 2nd edn, Lawrence Erlbaum, Hillsdale, NJ.

Crites R. and Tchobanoglous G. (2006). Small and Decentralized Wastewater Management Systems. McGraw-Hill, Boston, MA.

Dean R. B. and Forsythe S. L. (1976a). Estimating the reliability of advanced waste treatment. Part 1. *Water & Sewage Works*, **123**(6), 87–89.

Dean R. B. and Forsythe S. L. (1976b). Estimating the reliability of advanced waste treatment. Part 2. *Water & Sewage Works*, **123**(7), 57–60.

Dotro G., Langergraber G., Molle P., Nivala J., Puigagut J., Stein O. and Von Sperling M. (2017). Treatment Wetlands, Biological Wastewater Treatment Series. IWA Publishing, London, Vol. **7**, 154 p.

Elgeti K. (1996). A new equation for correlating a pipe flow reactor with a cascade of mixed reactors. *Chemical Engineering Science*, **51**, 5077–5080.

Farrugia P., Petrisor B. A., Farrokhyar F. and Bhandari M. (2010). Research questions, hypotheses and objectives. *Canadian Journal of Surgery*, **53**(4), 278.

Ferrell E. B. (1958). Control charts for lognormal universes. *Industrial Quality Control*, **15**, 4–6.

Gilbert R. O. (1987). Statistical Methods for Environmental Pollution Monitoring. John Wiley & Sons, Inc., New York, 320 p.

Halsey L. G., Curran-Everett D., Vowler S. L. and Drummond G. B. (2015). The fickle p-value generates irreproducible results. *Nature Methods*, **12**(3), 179–185.

Hammer M. J. and Hammer M. J., Jr (2012). Water and Wastewater Technology, 7th edn, Pearson, London.

Henze M., Van Loosdrecht M. C. M., Ekama G. A. and Brdjanovic D. (2008). Biological Wastewater Treatment. Principles, Modelling and Design. IWA Publishing, London, 511 p.

Hines W. W., Montgomery D. C., Goldsman D. M. and Borror C. M. (2003). Probability and Statistics in Engineering, 4th edn, John Wiley and Sons, New York, 672 p.

IWA Task Group on Good Modelling Practice (2012). Guidelines for using activated sludge models. Scientific and Technical Report No. 22, IWA Publishing, London, 312 p.

IWA Task Group on Mathematical Modelling for Design and Operation of Biological Wastewater Treatment (2000). Activated Sludge Models ASM1, ASM2, ASM2d and ASM3. IWA Publishing, London, 130 p.

Jakeman A. J., EL Sawah S., Cuddy S., Robson B., Mcintyre N. and Cook F. (2018). QWMN Good Modelling Practice Principles. The State of Queensland (Department of Environment and Science), Queensland, https://www.des.qld.gov.au/science/documents/qwmn-good-modelling-practice-principles.pdf.

Joffe A. D. and Sichel H. S. A. (1968). Chart for sequentially testing observed arithmetic means from lognormal populations against a given standard. *Technometrics*, **10**(3), 605–612.

Kadlec R. H. and Wallace S. D. (2009). Treatment Wetlands, 2nd edn, CRC Press, Boca Raton, FL.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Kauark Leite L. A. and Nascimento N. O. (1993). Développement, utilisation et incertitudes des modèles conceptuels en hydrologie. In: Modélisation du Comportement des Polluants dans les Hydrosystemes. Ministère de l'Environnement, Paris, Vol. **1**, pp. 191–219.

Lee C. (1973). Models in planning. In: An Introduction to the Use of Quantitative Models in Planning. Pergamon Press, Oxford.

Levenspiel O. (1999). Chemical Reaction Engineering, 3rd edn, John Wiley & Sons, Inc., New York.

Levine D. M., Berenson M. L. and Stephan D. (1998). Statistics for Managers Using Microsoft Excel. Prentice Hall, Upper Saddle River, NJ

Limpert E., Stahel W. A. and Abbt M. (2001). Log-normal distributions across the sciences: keys and clues. *BioScience*, **51**(5), 341–352.

Manser N. D., Wald I., Ergas S. J., Izurieta R. and Mihelcic J. R. (2015). Assessing the fate of *Ascaris suum* ova during mesophilic anaerobic digestion. *Environmental Science and Technology*, **49**, 3128–3135.

Melo L. D. V. (2019). Avaliação estatística de desempenho de estações de tratamento de água do Brasil, em função da tecnologia, do porte e do tipo de manancial (Statistical Evaluation of the Performance of Water Treatment Plants in Brazil, Depending on the Technology, Size and Source). PhD thesis, Federal University of Minas Gerais, Brazil (in Portuguese).

Melo L. D. V., Oliveira M. D., Libanio M. and Oliveira S. C. (2015). Applicability of statistical tools for evaluation of water treatment plants. *Desalination and Water Treatment*, **55**(30), 14024–2015.

Mendenhall W. and Sincich T. (1988). Statistics for the Engineering and Computer Sciences. Dellen Publishing Company, San Francisco, CA, 1036 p.

Mendenhall W. and Sincich T. (2012). A Second Course in Statistics: Regression Analysis, 7th edn, Prentice Hall, Upper Saddle River, NJ, 816 p. ISBN-10: 0321691695. ISBN-13: 978-0321691699.

Metcalf & Eddy (2003). Wastewater Engineering: Treatment and Reuse. McGraw-Hill, New York, 1819 p.

Metcalf & Eddy (2014). Wastewater Engineering: Treatment and Resource Recovery, 5th edn, Metcalf & Eddy/AECOM, New York, 2018 p.

Meijer S. C. F. and Brdjanovic D. (2012). A Practical Guide to Activated Sludge Modeling. UNESCO-IHE Lecture Notes, UNESCO-IHE, Delft, 277 p.

Mihelcic J. R. and Zimmerman J. B. (2014). Environmental Engineering: Fundamentals, Sustainability, Design, 2nd edn, Wiley, New York.

Modarres R., Gastwirth J. L. and Ewens W. (2005). A cautionary note on the use of non-parametric tests in the analysis of environmental data. *Environmetrics*, **16**(4), 319–419.

Montgomery D. G. (2009). Introduction to Statistical Quality Control, 6th edn, Wiley, New York, 734 p.

Morrison J. (1958). The lognormal distribution in quality control. *Applied Statistics*, **7**(3), 160–172.

Naguettini M. and Pinto E. J. A. (2007). Hidrologia estatística. CPRM – Serviço Geológico do Brasil, Belo Horizonte, 561 p (in Portuguese).

Nascimento N. O., Naghettini M., Héller L. and Von Sperling M. (1996). Investigação científica em engenharia sanitária e ambiental. Parte 3: Análise estatística de dados e de modelos, *Engenharia Sanitária e Ambiental (ABES)*, **1**(4), 152–168 (in Portuguese).

Niku S., Schroeder E. D. and Samaniego F. J. (1979). Performance of activated sludge process and reliability-based design. *Journal Water Pollution Control Association*, **51**(12), 2841–2857.

Niku S., Schroeder E. D., Tchobanoglous G. and Samaniego F. J. (1981). Performance of Activated Sludge Process: Reliability, Stability and Variability. Environmental Protection Agency, EPA Grant No R805097-01, Washington, D.C., pp. 1–124.

Niku S., Schroeder E. D. and Haugh R. S. (1982). Reliability and stability of trickling filter processes. *Journal Water Pollution Control Association*, **54**(2), 129–134.

Oliveira S. M. A. C. (2017). Apostila. Tratamento estatístico de dados ambientais (Lecture notes: statistical treatment of environmental data). Federal University of Minas Gerais (in Portuguese).

Oliveira S. M. A. C. and Gomes L. L. (2011). Consequências da utilização de métodos de substituição de valores censurados nos resultados das análises de dados de monitoramento ambiental. Congresso

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Brasileiro de Engenharia Sanitária e Ambiental, 26–29 September 2011, Porto Alegre, Brazil, Vol. **26** (in Portuguese).

Oliveira S. M. A. C. and Von Sperling M. (2008). Reliability analysis of wastewater treatment plants. *Water Research*, **42**, 1182–1194.

Oliveira S. M. A. C. and Von Sperling M. (2009). Gráficos de controle da qualidade de efluentes de estações de tratamento de esgotos. Congresso Brasileiro de Engenharia Sanitária e Ambiental, 20–24 September 2009, Recife, Brazil, Vol. **25** (in Portuguese).

Oliveira S. M. A. C. and Von Sperling M. (2011). Performance evaluation of different wastewater treatment technologies operating in a developing country. *Journal of Water, Sanitation and Hygiene for Development*, **1**(1), 37–56.

Oliveira S. C., Souki I. and Von Sperling M. (2012). Lognormal behaviour of untreated and treated wastewater constituents. *Water Science and Technology*, **65**(4), 596–603. doi: 10.2166/wst.2012.899.

Ott W. R. (1995). Environmental Statistics and Data Analysis. CRC Press LLC, Boca Raton, FL, 313 pp.

Ott R. L. and Longnecker M. (2010). An Introduction to Statistical Methods and Data Analysis, 6th edn, Brooks/Cole, Cengage Learning, Belmont, CA, 1273 p. ISBN-10: 0495017582 | ISBN-13: 978-0495017585.

Pecson B. M., Barrios J. A., Jimenez B. E. and Nelson K. L. (2007). The effects of temperature, pH, and ammonia concentration on the inactivation of Ascaris eggs in sewage sludge. *Water Research*, **41**, 2893–2902.

Potvin C. and Roff D. A. (1993). Distribution-free and robust statistical methods: viable alternative to parametric statistics. *Ecology*, **74**(6), 1617–1628.

Rose J. B. and Jiménez-Cisneros B. (eds) (2019). The Global Water Pathogens Project. Michigan State University, UNESCO, E. Lansing, MI. http://www.waterpathogens.org/

Sawyer C. N. and Mc Carty P. L. (1978). Chemistry for Environmental Engineering, 3rd edn, Mc Graw-Hill, Inc, New York, 532 p.

Schiermeier Q. (2018). For the record: making project data freely available is vital for open science. *Nature*, **555**, 403–405.

Shaban S. A. (1988). Chapter 10. Applications in industry. In: Lognormal Distributions: Theory and Applications, E. L. Crow and K. Shimizu (eds), Marcel Dekker, Inc., New York, Vol. **88**, pp. 279–281. ISBN 0-8247-7803-0.

Shore H. (1998). A new approach to analysing non-normal quality data with application to process capability analysis. *International Journal of Production Research*, **36**(7), 1917–1933.

Shore H. (2000). Three approaches to analyze quality data originating in non-normal populations. *Quality Engineering*, **13**(2), 277–291.

SKYMARK (2019). Normal probability plot: does your data follow the standard bell curve? http://www.skymark.com/resources/tools/normal_test_plot.asp (accessed 25 April 2019).

Sokal R. R. and Rohlf F. J. (1995). Biometry, 3rd edn, Freeman and Company, New York, NY, 887 p. ISBN: 0716724111.

Sokal R. R. and Rohlf F. J. (2012). Biometry, 4th edn, WH Freeman and Company, New York, NY.

Statistics How To (2019). Studentized range distribution. https://www.statisticshowto.datasciencecentral.com/studentized-range-distribution/#qtable (accessed 29 July 2019).

Sullivan G. M. and Feinn R. (2012). Using effect size—or why the p-value is not enough. *Journal of Graduate Medical Education*, (September), **4**(3),279–282.

Tchobanoglous G. and Schroeder E. D. (1985). Water Quality: Characteristics, Modeling, Modification. Addison-Wesley, Reading, MA.

Tchobanoglous G., Stensel H., Tsuchihashi R., Burton F., Abu-Orf M., Bowden G. and Pfrang W. (2014). Wastewater Engineering: Treatment and Resource Recovery, 5th edn, Metcalf and Eddy & AECOM, McGraw-Hill, Boston, MA.

Teefy S. (1996). Tracer Studies in Water Treatment Facilities: A Protocol and Case Studies. American Water Works Association, Denver, CO, 152 p. ISBN 0898678579.

Thomann R. V. (1982). Verification of water quality models. *Journal of Environmental Engineering Division, ASCE*, **108**(EE5), 923–940.

UNITED STATES CODE (1974). Safe Drinking Water Act. 42 U.S.C. §300(f)(1)(C)(i).

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

US EPA (2005). Quality Assurance Project Plan for Monitoring of Surface Water at the Eagle Valley Reservation. Eagle Valley Environmental Program, Eagle Valley Band of Indians, Eagle Valley Reservation, Shadowland, CA. https://www.epa.gov/sites/production/files/2015-06/documents/module3_0.pdf (accessed 5 May 2019).

US EPA (2017). Operating Procedure: Field Sampling Quality Control. No. SESDPROC-011-R5. Athens, GA. https://www.epa.gov/sites/production/files/2017-07/documents/field_sampling_quality_control011_af.r5.pdf (accessed 5 May 2019).

US EPA (2018). Overview of Total Maximum Daily Loads (TMDLs). Impaired Waters and TMDLs. https://www.epa.gov/tmdl/overview-total-maximum-daily-loads-tmdls (accessed 5 May 2019).

Van Haandel A. C. and Van der Lubbe J. (2012). Handbook of Biological Wastewater Treatment: Design and Optimisation of Activated Sludge Systems. IWA Publishing, London, 770 p.

Van Loosdrecht M. C. M., Nielsen P. H., Lopez-Vazquez C. M. and Brdjanovic D. (2016). Experimental Methods in Wastewater Treatment. IWA Publishing, London, 360 p.

Von Sperling M. (1990). Optimal Management of the Oxidation Ditch Process. PhD thesis, Imperial College, University of London, 371 p.

Von Sperling M. (1999). A critical analysis of classical design equations for waste stabilization ponds and other waste treatment systems. *Water Environment Research*, **71**(6), 1240–1243.

Von Sperling M. (2002). Relationship between first-order decay coefficients in ponds, according to plug flow, CSTR and dispersed flow regimens. *Water Science and Technology*, **45**(1), 17–24.

Von Sperling M. (2005). Modelling of coliform removal in 186 facultative and maturation ponds around the world. *Water Research*, **39**, 5261–5273.

Von Sperling M. (2007). Basic Principles of Wastewater Treatment. Biological Wastewater Treatment Series. IWA Publishing, London, Vol. **2**, 200 p.

Von Sperling M. (2014). *Princípios do tratamento biológico de águas residuárias*. In: Estudos e modelagem da qualidade da água de rios. In: Editora UFMG, 2nd edn, Vol. **7**, Belo Horizonte, 592 p. ISBN 9788542300802 (in Portuguese).

Von Sperling M. and Chernicharo C. A. L. (2005). Biological Wastewater Treatment in Warm Climate Regions, **Two volumes**, IWA Publishing, London, 1496 p.

Von Sperling M., Heller L. and Nascimento N. O. (1996). Investigação científica em engenharia sanitária e ambiental. Parte 2: a análise preliminar dos dados. *Engenharia Sanitária e Ambiental (ABES)*, Ano 1, **1**(3), 115–124 (in Portuguese).

Von Sperling M., Verbyla M. E. and Mihelcic J. R. (2018). Understanding pathogen reduction in sanitation systems: units of measurement, expressing changes in concentrations, and kinetics. In: Global Water Pathogen Project, J. B. Rose and B. Jiménez-Cisneros (eds). http://www.waterpathogens.org (C. Haas, J. R. Mihelcic and M. E. Verbyla (eds). Part 4. Management of Risk from Excreta and Wastewater) http://www.waterpathogens.org/book/understanding-pathogen-reduction-sanitation-systems-units-measurement-expressing-changes Michigan State University, UNESCO, E. Lansing, MI. https://doi.org/10.14321/waterpathogens.54.

Whitehead P. G. and O'Connel P. E. (ed.) (1984). Water quality modeling, forecasting and control. Proceedings of an International Workshop. Institute of Hydrology, Wallingford. Report No. 88. 123 p.

Wickham H. (2014). Tidy data. Journal of Statistical Software, **59**(10), 1–23.

Wilkinson M. D. et al. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, **3**, 160018. doi: 10.1038/sdata.2016.18.

Zar J. H. (1999). Biostatistical Analysis, 4th edn. Prentice Hall, Inc., Upper Saddle River, NJ. ISBN 013081542-x.

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Index

A

accuracy, 10, 40, 48, 67, 69, 81–82, 88, 106, 168, 184, 202, 404, 443, 448, 471, 497, 539, 548–549
amplitude, 11, 142, 295–296, 298, 302–304, 306
analysis of variance (ANOVA), 15, 51, 63, 317, 324, 324, 371, 373–374, 377–382, 385–386, 390, 395, 404, 442, 446–450, 457–458, 462, 471
arithmetic mean, 97–98, 119, 129, 131, 133–138, 140, 142–143, 146, 148, 172, 189, 200, 203–205, 224–225, 229–230, 233–235, 239, 271, 273, 276–277, 279, 301, 304, 306, 308, 310, 374–377
asymmetry, 13, 172, 207, 210, 217, 220, 225, 239, 305, 467
autocorrelation, 16, 20, 293, 397–398, 429, 436–440, 453, 466, 468, 476, 619, 621–624, 629–630

B

batch, 18–19, 83, 203, 293, 512–513, 531–533, 535–538, 541, 543–544, 547–549, 553–557, 559, 561, 565–567, 583, 585–587, 592
beta distribution, 214, 235–236, 238–239
box plot, 11, 14, 100, 149, 151–152, 172–174, 225, 239, 244–245, 454, 467, 506
box-and-whisker plot, 11, 100, 149

C

calculated values, 10, 69–72, 74, 91, 94, 101, 168, 189, 200, 218, 227, 260, 267, 269, 469, 540
categorical data, 152, 310
censored data, 10, 12, 95–96, 98, 117–120, 122, 142, 150, 181, 188–190, 193, 206
central tendency, 10, 12, 51, 95, 97–98, 101, 107, 112–113, 117–118, 122, 128–136, 138, 143, 146, 150, 172, 181, 200–206, 223, 229, 233–235, 302, 353, 378
coefficient of determination, 20, 398, 441, 446, 450, 454, 457, 460, 466, 469–471, 473, 475–476, 547–549, 583, 585, 606–611, 618, 626, 631
coefficient of variation, 11, 48–49, 88, 95, 98, 104, 115, 144–146, 212–214, 229–230, 233–234, 270, 273, 277, 279
column chart, 11, 151, 176–177, 300, 429, 436
complete-mix, 18–19, 490, 510, 531–533, 556–557, 559, 562–569, 571–572, 574–581, 585, 587–589, 592, 600
completely-mixed, 490, 497, 510, 532, 556, 585, 588
compliance, 3–4, 6, 9, 14, 40–41, 44–45, 47, 50, 53, 55, 63, 65, 85, 87, 100, 149, 167, 170, 195, 197, 199, 223, 241–245, 247–250, 255, 257, 259–260, 263, 268–270, 272–274, 276–279, 281, 310, 315, 318, 327, 338, 438, 459, 595, 599, 603, 619

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

confidence interval, 16, 34, 51, 83–87, 91–93, 97, 106, 119, 208, 211, 217, 220, 222, 226, 336, 340, 344, 355–356, 395, 408–409, 415, 417, 440, 442, 449–455, 460, 463–464, 476–477
conformity, 14, 170, 200, 241–248, 250, 257–258, 260, 269, 271, 273, 282, 310–312, 315
continuous flow, 82, 293, 508, 511, 557
control chart, 14, 85, 223, 241, 243, 281–289, 291–303, 305–308, 310, 312–313, 315
control chart for individual measurements, 14, 306–307, 310
control chart for means, 14, 282–283, 286–287, 291, 293–294, 296–300, 302–303, 305, 310
control chart for proportion of failures, 14
correlation, 4, 16, 20, 26, 37, 52, 59, 63, 96, 119, 175–176, 208, 211, 219–220, 222, 264, 293, 397–399, 402–410, 412–440, 445–446, 449–450, 453, 455–456, 460, 466, 468–470, 472, 476, 603, 607, 619, 621–624, 629–630
correlation coefficient, 16, 397–398, 402, 404–410, 414–430, 432–434, 436–437, 440, 445–446, 449–450, 466, 469, 476, 607
cross-correlation, 16, 59, 397–398, 429–430, 436, 440, 476

D

database, 10, 72–75, 79–80, 96, 98, 110, 118, 123
dead zone, 18, 514–515, 519, 522, 586
detection limit, 10, 40, 48, 69, 87–89, 117–122, 136, 150, 186, 188–190, 193, 222
dose, 24
Dunn test, 15, 52, 317, 324, 371, 390–391, 394
dynamic state, 17, 479–481, 483, 491, 495, 497, 601

E

equalization, 8, 28–30, 288, 487

F

fitting a distribution, 217–219, 226–228
flow, 3, 8, 10, 18–19, 21–38, 40, 44, 46, 53–55, 57–62, 74–77, 81–83, 95, 98–101, 103, 105, 107, 117, 123, 128–129, 139–142, 152, 154, 162–164, 170, 177–179, 182, 188, 197, 203, 208, 239, 263, 270, 282, 293, 338, 435–436, 480–489, 491–496, 500, 502, 504–522, 524–527, 529, 531–533, 555–559, 561–562, 564–572, 574–579, 581–583, 585–592, 597–599
flow rate, 8, 21–28, 30–38, 40, 46, 55, 57, 60–62, 74–76, 81, 152, 197, 263, 484, 500, 508, 559, 589, 599
food-to-microorganism ratio, 18, 523–524, 528
frequency analysis, 14, 30, 63, 241, 243, 263, 265, 267, 269, 281, 315
frequency distribution, 11–13, 98, 100, 129–130, 135, 151–152, 165–170, 181, 200, 204–209, 217–218, 221–222, 226–227, 239, 263, 272, 315, 352, 365
frequency polygon, 11, 151–153, 168–169, 173, 208–209, 217–219, 227–228, 236, 238

G

geometric mean, 10, 13, 32, 38, 95, 97–98, 104, 106, 129–131, 134–138, 146–148, 150, 200–206, 223–233, 235, 239, 301–302, 304, 306, 308, 310
geometric standard deviation, 11, 13, 146–148, 223–233, 239, 301–304, 306, 308, 310
goodness-of-fit, 13, 19, 124–125, 218–220, 239, 267, 305, 398, 441, 459, 466, 475–476, 547, 583, 595, 604–606, 609, 612–613, 618, 620, 625–626, 631
graphs, 2, 7–8, 11, 14, 28, 33–34, 74, 95–96, 98–100, 134, 151–154, 156–160, 165, 167, 169–170, 172–173, 175–180, 204, 208, 212, 218, 220, 224–225, 233, 239, 244–246, 260–261, 263, 267–269, 280, 305, 309, 315, 395, 414, 435, 450, 455, 466, 472, 486–487, 497, 506, 529, 535, 542, 547, 572–573, 578, 581, 606, 613, 616, 618, 620, 630–631

H

histogram, 11, 98, 100, 130, 151–152, 165–169, 172–173, 204–205, 208–209, 217–218, 227, 281, 341, 627
hydraulic loading rate, 18, 44, 55, 499–502, 517–521, 528
hydraulic retention time (HRT), 8, 18, 18, 28, 36, 36–37, 37–38, 58–59, 482, 491, 493, 499, 501–502, 507, 507–509, 509–510, 510–511, 511–512, 512–513, 513–514, 514–517, 517–518, 518, 524–527, 527–529, 529, 556–558, 558, 560, 560–562, 562–563, 563–564, 566–567, 570, 570, 572, 574–577, 581–582, 584, 586, 586, 588, 590
hypothesis test, 3–4, 15, 65, 85, 100, 119, 174, 208, 211, 219–220, 222, 243, 247–250, 253, 255, 258, 315, 317–318, 320, 322–331, 334, 336–337, 340, 348, 361, 363, 371, 394–395, 405–408, 410–412, 414, 417, 421, 424, 426–427, 445, 449, 476, 618, 620–621

I

idealized hydraulic model, 19, 566–569, 574, 583

K

Kruskal-Wallis test, 15, 63, 317, 371, 377, 386, 388, 390, 391, 394, 395

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

L

lag phase, 19, 550, 552
limit of detection, 88, 152
linear regression, 16, 52, 91, 93, 397–401, 440–442, 444–445, 447, 450, 452, 455, 457, 459–461, 468–473, 475–476, 542, 613–615, 623, 629
loading, 4, 7–8, 16–18, 21–24, 28, 44, 55–56, 63, 71, 74–75, 80–81, 96, 99, 102–103, 105, 107–111, 114, 139, 150, 188, 210, 397–398, 425, 468, 480–481, 484, 499–508, 517–523, 525, 527–529, 532, 596, 601
log-normal distribution, 13–14, 30–32, 97, 130–131, 135–136, 147–148, 167–168, 204–205, 207–210, 214, 222–235, 239, 241, 243, 258, 263–265, 267–269, 271–281, 291, 300–310, 315, 321, 323
logarithm, 12, 32, 134, 136–138, 160–161, 180, 183–184, 186, 226–227, 233, 273, 472–473, 475, 534, 545

M

mass balance, 4, 17, 23, 28–29, 38, 53, 74, 135, 139, 142, 188, 192, 398, 401, 479–480, 487–494, 496–498, 526, 532–533, 562, 587, 596
mass loading rate, 17–18, 75, 99, 103, 105, 139, 425, 499–502, 506, 520–523, 528, 532
mathematical model, 4, 19, 117, 124, 192, 436, 482, 490–491, 533, 586–587, 592, 596–598, 601–602, 616
mean, 3–4, 7, 10, 12–15, 18, 20, 28, 31–38, 43, 47–48, 50–51, 54, 58–59, 64–66, 70, 72–73, 81–86, 88, 90, 93, 95–98, 100, 103–116, 118–119, 121–122, 125, 129–138, 140, 142–150, 152, 156–157, 170, 172–173, 179–180, 182–183, 189–191, 196–198, 200–206, 208–209, 211–218, 222–239, 244, 246–252, 255–257, 265–289, 291–308, 310–311, 313, 317–318, 320–321, 323–327, 329–331, 336–358, 361, 363–366, 369, 371–384, 390–395, 398, 400, 408, 410, 413, 415–416, 429–430, 436, 443, 445–447, 450–454, 459–462, 464–465, 467, 473–474, 482–483, 490–491, 495, 499, 501, 504, 506–507, 509–510, 512–514, 516–517, 519–520, 524–526, 529, 536–537, 550–552, 559, 566, 569–570, 593, 598, 600, 602, 605–606, 608–609, 612, 614–615, 618–619, 621, 625–626, 628
median, 10, 15, 51, 95, 97–98, 104–111, 113–116, 118–119, 121, 129–131, 133–136, 138, 148, 150, 171–173, 200–206, 211, 217, 223–225, 229, 232–235, 250–256, 275–276, 279, 318, 320–321, 324, 348, 358–363, 366–367, 369–371, 385, 388–395, 398, 506
metadata, 10, 50, 69, 73, 80–81, 94
missing data, 10–11, 75, 95, 98, 116–117, 126, 156, 159, 201, 203
mode, 4, 7, 10, 14, 16, 18–20, 52, 79, 92, 117–118, 124–125, 129–131, 165, 192, 208, 210–211, 217–218, 223, 229, 231, 233–235, 263, 292, 397–398, 400–402, 404, 406, 436–438, 440–444, 446, 449–450, 452–455, 459–461, 466, 470–477, 480–482, 489–491, 497, 512–514, 520, 531–533, 535, 537–539, 541–542, 550–552, 555–556, 561, 566–571, 574–583, 585–589, 591–593, 595–607, 609, 612–621, 623–625, 628–631
model calibration, 19, 475, 595–596, 602–603, 609, 630
moment matching, 207, 214, 235–236
monitoring data, 2–4, 7, 10–11, 13–15, 95, 99–100, 103, 109–110, 112, 114, 116–117, 123, 129, 134, 151–153, 172, 191, 196, 201, 205, 207–208, 211, 217, 220, 222, 235, 241, 243, 260–263, 265, 268, 271–272, 275–276, 292, 294, 306, 315, 317, 338, 342, 373, 398, 400, 425, 430–431, 566, 573–574, 578, 582, 584–585
multiple linear regression, 16, 397–398, 400, 442, 470–473

N

non-conformity, 14, 200, 241, 243–246, 248, 250, 257–258, 260, 269, 271, 282, 310–312, 315
non-idealized hydraulic model, 19, 583
non-linear regression, 16, 397–398, 400–401, 473
non-parametric, 14–15, 51–52, 209, 219, 222, 241, 250–256, 315, 317–318, 320–324, 342, 348, 358–359, 362–363, 366, 368–369, 371, 377, 385, 388, 390–391, 394–395, 397, 419, 421, 427, 476, 621
normal distribution, 13–14, 30–32, 52, 83–85, 88–89, 97, 121–122, 130–131, 135–136, 144, 147–148, 167–168, 204–205, 207–220, 222–236, 238–239, 241, 243, 250–251, 258–259, 263–269, 271–281, 285–287, 291, 293, 296–297, 299–310, 315, 321–323, 330, 333–336, 339, 341–342, 349, 353, 359–361, 374, 377, 410, 419, 445, 454, 468, 620, 627–628
norms, 6, 411

O

one-sample hypothesis test, 247, 249, 255, 258, 315, 317–318, 323, 340
one-tailed, 14, 249–250, 253, 256, 329, 332–335, 338, 342, 346, 348, 350–352, 354, 357–358, 361, 364–365, 368–369
outliers, 10, 95–96, 98, 123–128, 135, 142, 150, 172, 201, 293, 403–404, 454–455, 476

P

parametric, 14–15, 51–52, 208–209, 219–220, 222, 241, 250–256, 258, 315, 317–318, 320–324, 338, 342, 348–349, 352, 358–359, 361–364, 366, 368–369, 371,

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

377, 380, 382, 385, 388, 390–391, 394–395, 397, 419, 421, 427, 476, 621
Pearson, 16, 372, 397, 402, 404, 412, 419, 421, 423–427, 429, 431–432, 436–437, 450, 476
Pearson correlation coefficient, 419, 421, 423–424, 426, 436–437, 476
percentile graph, 11, 14, 98, 100, 151, 169–171, 173, 244–246, 260, 263
percentiles, 11, 14, 51, 83, 95, 98, 100, 104, 119, 125, 128, 148–150, 152–153, 169–173, 212, 230–231, 245–246, 273, 279
pie chart, 11, 99, 151–153, 176, 178–180
plots, 11, 14, 30, 32, 74, 100, 122, 141, 149, 151, 158, 168, 172, 174–176, 210, 212–214, 219–222, 224–225, 228, 238–239, 244–245, 267, 285, 353, 378, 395, 402–403, 406, 453–454, 473, 476, 542–543, 572, 578, 613, 618–621, 626–629, 631
plug-flow, 18–19, 58–59, 531–533, 555–559, 561–562, 564–572, 574–579, 581, 583, 585–587, 592
plug-flow with dispersion, 18–19, 531, 533, 569–572, 574–575, 579, 581, 583, 585–587, 592
Poisson distribution, 14, 208
polynomial regression, 16
power, 9, 39, 54, 63–67, 75, 96, 129, 136, 143, 147, 220, 222, 224, 231, 273, 279, 303, 321, 323, 327–329, 336, 345–348, 356–357, 361, 363, 381, 385, 395, 398, 450, 472–473, 475, 481, 553, 586, 602, 607, 618
precision, 10, 26, 40, 48–49, 64, 67, 69–70, 81–83, 86–88, 90, 93, 218, 336, 345, 355, 453, 573, 616
prediction interval, 33–35, 82, 85, 87–88, 90–93, 106, 234, 282–283, 285–286, 450–452, 464–465, 477
process, 2–4, 6, 8–19, 23–24, 28, 36, 41, 43, 45–46, 48–50, 53, 59, 63, 66, 70, 74, 80, 86–90, 96–97, 100, 107, 109–110, 113–114, 123, 152, 154, 166, 176–179, 187–188, 192–198, 208, 244, 254, 271–272, 281–285, 287–289, 291, 293–296, 299, 301–302, 305, 310, 313, 315, 318, 341, 344, 347, 350, 368, 398, 401, 426, 435–438, 450, 455, 484, 488, 497, 503–505, 510, 512, 517, 519, 521, 523–527, 532–533, 586, 596–602, 605, 612, 622
process control, 14, 281–282, 293, 301, 601
proportions, 14, 51–52, 63, 241, 257–260, 310–311, 313, 324

Q

qualitative data, 11, 50, 151–152, 176
quality assurance, 9, 39–41, 62, 67, 80, 94
quality control, 9, 39–41, 45, 47–49, 67, 80, 94, 223, 243, 281, 286, 293–295, 301, 315

R

R^2 value, 92–93, 441, 450, 454, 471, 473, 476, 547, 607, 613
raw data, 10, 69–72, 74, 79, 81, 94, 101, 154
reaction, 4, 17–19, 49, 57, 74, 398, 401, 487–490, 494, 496, 506, 512, 514, 527, 531–539, 541–556, 558–559, 561–569, 571–577, 579, 581–590, 592, 596, 600, 603
reaction order, 18–19, 531, 533–535, 541–544, 547, 549, 559, 562, 587, 592
reduction, 12, 26, 44–45, 47, 55, 181–182, 184–188, 191–192, 194, 205–206, 281, 347, 508, 516–517, 519, 522, 537, 555, 573, 576–577, 602
refractory fraction, 19, 549–550
reliability analysis, 14, 63, 241, 243, 270–272, 275–276, 279, 315
removal, 10–14, 16, 19, 36–37, 45, 63, 74–77, 87, 95–96, 98–100, 103–105, 107–108, 110–111, 113–114, 118, 129, 131, 149–152, 154, 156, 169–170, 174, 180–201, 204–206, 208, 213–214, 217, 235–237, 239, 243–246, 271–272, 279, 289, 347–348, 397–398, 469, 484, 493, 505, 512, 514, 516–517, 519, 521, 524–525, 527, 531–533, 535, 537–538, 549, 554–556, 559, 562, 564–568, 572–577, 579–582, 586, 588–589, 599
removal coefficient (K), 19, 19, 96, 135, 150, 294–298, 301–302, 311–312, 373–377, 379–383, 385–386, 389–391, 429, 436, 471, 532, 534–537, 537–555, 558, 560–561, 563–565, 565–568, 572–577, 577–592
removal efficiency, 12, 19, 37, 74, 108, 113, 149, 152, 170, 181–183, 185–190, 192–201, 204–206, 214, 236–237, 469, 493, 505, 516–517, 524–525, 537, 554, 566–567, 572–576, 580–582, 586
replicates, 9–10, 48–49, 70, 72, 82, 88, 172
residuals, 16, 19–20, 52, 436, 438–440, 442–443, 446–447, 452–453, 455–459, 466–468, 472, 475–476, 550, 595, 598, 603–609, 612, 618–631
residuals analysis, 16, 19, 438, 440, 452, 456, 459, 466, 468, 472, 476

S

sample, 4, 9, 14–15, 22, 39–45, 47–67, 70–72, 74, 77, 80–88, 90–93, 96–98, 101, 104–105, 116–117, 119, 122–123, 129, 131, 134, 136, 139, 142, 144, 148–150, 152, 154, 156–157, 165–168, 170–172, 182, 186, 189, 200–204, 206, 208, 210–211, 215–216, 223, 236, 241, 243–245, 247, 249–260, 263–264, 281, 283–289, 291–297, 300–303, 306, 310–313, 315, 317–318, 320–321, 323–331, 336–342, 344–350, 352–378, 380–395, 404–410, 412–420, 422, 429, 431, 433, 435, 442–445, 450, 453–456, 460–461, 464–466, 471, 476, 514, 534, 566, 568–570, 583, 600, 605, 619, 621, 623

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

sample collection, 9, 39, 43, 47, 49–50, 54, 60–62, 70, 74, 83, 104–105, 123, 600
sample size, 9, 39, 42, 62–67, 83, 85, 87, 96, 98, 166, 251, 253, 256–260, 283, 285, 293, 295–297, 300, 302–303, 306, 313, 315, 321, 327, 329, 331, 337, 340–342, 344–348, 356–361, 377, 381, 386, 390, 393–395, 406–407, 410, 412, 414–415, 420, 422, 476, 605, 623
sampling, 9, 39, 42–44, 48–50, 53–55, 57–60, 62, 83, 85, 96, 107, 114, 116, 126, 142, 152, 174, 182, 204, 225, 268, 285, 293–294, 320, 326, 329, 341, 350, 363, 431–432, 435, 437, 569, 584, 592, 613, 616–617
sampling (spatial aspects of sampling), 9, 9, 39, 42–44, 48–50, 53–55, 57–60, 62, 83, 85, 96, 107, 114, 116, 126, 142, 152, 174, 182, 204, 225, 268, 285, 293–294, 320, 326, 329, 341, 350, 363, 431–432, 435, 437, 569, 584, 592, 613, 616–617
scatter plot, 11, 99–100, 151–153, 158, 175–176, 398, 400, 402–404, 412, 432, 435, 440–442, 445–446, 456–457, 459, 464–466, 470, 472–473, 476
sensitivity, 20, 598–599, 616–618
sensitivity analysis, 20, 598–599, 616–618
short circuiting, 18, 586
significant digits, 7, 69, 90–92, 202, 218
significant figures, 10, 69, 75, 90–91, 93–94, 106, 256
simple linear regression, 16, 397–398, 400, 440, 445, 447, 455, 459–461, 468, 470–471, 623, 629
skewness, 13, 204, 217, 220, 222–223, 230, 235, 454, 620, 627
sludge age, 18, 499–500, 525–527
Spearman, 16, 397, 419–423, 427–429, 431–432, 436, 476
Spearman correlation coefficient, 16, 419–420, 428–429
standard deviation, 11, 13, 33, 35, 48, 51, 64–67, 70, 72, 82–85, 87–91, 95, 97–98, 104, 106–109, 114–115, 118–119, 121–122, 143–148, 150, 189, 200, 211–218, 223–239, 251–252, 255, 257, 265–275, 277–279, 284–289, 291, 293–297, 299, 301–304, 306, 308, 310, 331, 337–342, 344–347, 349, 352, 354–356, 361, 364, 366, 407, 430, 453, 618, 625
standard normal variable, 13, 215, 231, 265, 286, 303, 333, 407, 411
standards, 4, 14, 44–45, 47–48, 57, 74, 85, 92, 100, 119, 135, 149, 152, 170, 195, 200, 206, 208, 223, 241–244, 247, 249–250, 260, 271–274, 276–277, 279, 310, 312, 318, 338, 500, 505, 599, 618
statistical power, 39, 63–66, 222, 321, 328–329, 336, 345–346, 348, 357, 395
statistics, 2–4, 6–19, 28, 45, 50, 69–70, 72, 74, 83, 88, 94–111, 113–119, 121–122, 125, 133–134, 146, 150–154, 165, 172, 175–176, 181–182, 188, 202, 204, 208, 211, 217, 226, 235, 250, 255, 258, 277, 298, 304, 312, 318, 320, 323, 327, 330–331, 334–336, 343, 345, 350, 353, 358, 361–363, 366, 369, 385–386, 392, 398, 440, 442, 454–455, 457–458, 466, 471–472, 506, 609, 623, 625–626
steady state, 17, 344–345, 479–483, 491, 525, 561, 582, 591, 601
summary tables, 2–4, 10, 95–96, 101–103, 106–107, 113–114, 150, 225, 239, 394, 498, 529
surface loading rate, 518–520, 521–523
symmetry, 13, 107, 172–173, 204, 207, 209–210, 217, 220, 222–223, 225, 235, 239, 305, 315, 467, 620

T

t-test, 13–15, 45, 51, 63–66, 219–220, 239, 249–256, 267, 317, 324, 329, 331, 334, 336, 338, 340–356, 358–359, 361, 363–366, 371, 373, 380, 395, 405–407, 410–411, 414, 421, 423–424, 430, 445, 448, 463, 467, 472, 620–621, 628
tanks-in-series, 18–19, 514, 531, 533, 559, 565, 569–570, 574–583, 585–589, 592
targets, 4, 14, 100, 241–243, 271, 310
temperature, 19, 42, 44, 46, 50, 54–56, 58, 62–63, 74, 103, 197–198, 270, 288, 425–429, 480, 506, 551–554, 572, 592
theoretical HRT, 18, 36, 508–511, 513–515, 570, 593
time series, 11, 14, 59, 98–100, 117, 121, 124–126, 128, 151–152, 154–159, 162–163, 198–200, 244, 270, 286, 429, 432, 435–437, 439, 453, 472, 476, 497, 506, 542, 550, 591, 618–619, 621–622
Tukey test, 15, 324, 371, 374, 380–383, 394
two-sample hypothesis test, 320
two-tailed, 15, 249, 252–254, 256, 259, 329–335, 338–344, 346, 348, 350–353, 355, 357–358, 361–362, 364–369, 371, 405, 407, 411, 414, 417–418, 420–421, 423, 448, 459, 463, 621

U

uncertainty, 10, 47, 69, 82–83, 86–87, 93, 97, 106, 217, 225–226, 282, 485, 583, 601, 617–618

V

variability, 4, 10, 48–49, 54, 57, 69–70, 82–83, 86–88, 90, 96, 106–107, 117–118, 131, 142, 144, 146, 154, 175, 225, 247, 258, 271, 276, 281–282, 288–289, 292–293, 300, 310, 315, 318, 326, 344, 353, 356, 377, 446–447, 466, 607, 609, 612, 617
variance, 11, 15, 20, 51–52, 82, 97–98, 106, 143–146, 235, 317, 324, 346, 349–352, 354–355, 371, 373–374, 376–378, 380, 382, 385, 404, 407, 442, 445–446, 448,

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

450, 452–453, 455, 467, 476, 606–607, 609, 619, 621, 628–629
volumetric loading rate, 74

W
water balance, 8, 17, 38, 479, 481, 483–485, 487–488, 490–494, 507–508
weighted average, 10, 95, 98, 129, 138–140, 142, 350
Wilcoxon, 14–15, 51, 63, 250–256, 317, 324, 348, 358, 363, 366–369
Wilcoxon signed-rank test, 14–15, 51, 63, 250, 252, 254–256, 317, 324, 363, 366–368
Wilcoxon-Mann-Whitney U-test, 15, 317, 358

Z
Z test, 14, 317, 324, 338

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest

Assessment of Treatment Plant Performance and Water Quality Data

A GUIDE FOR STUDENTS, RESEARCHERS AND PRACTITIONERS

Marcos von Sperling, Matthew E. Verbyla and Sílvia M. A. C. Oliveira

This book presents the basic principles for evaluating water quality and treatment plant performance in a clear, innovative and didactic way, using a combined approach that involves the interpretation of monitoring data associated with (i) the basic processes that take place in water bodies and in water and wastewater treatment plants and (ii) data management and statistical calculations to allow a deep interpretation of the data.

This book is problem-oriented and works from practice to theory, covering most of the information you will need, such as (a) obtaining flow data and working with the concept of loading, (b) organizing sampling programmes and measurements, (c) connecting laboratory analysis to data management, (e) using numerical and graphical methods for describing monitoring data (descriptive statistics), (f) understanding and reporting removal efficiencies, (g) recognizing symmetry and asymmetry in monitoring data (normal and log-normal distributions), (h) evaluating compliance with targets and regulatory standards for effluents and water bodies, (i) making comparisons with the monitoring data (tests of hypothesis), (j) understanding the relationship between monitoring variables (correlation and regression analysis), (k) making water and mass balances, (l) understanding the different loading rates applied to treatment units, (m) learning the principles of reaction kinetics and reactor hydraulics and (n) performing calibration and verification of models.

The major concepts are illustrated by 92 fully worked-out examples, which are supported by 75 freely-downloadable Excel spreadsheets. Each chapter concludes with a checklist for your report. If you are a student, researcher or practitioner planning to use or already using treatment plant and water quality monitoring data, then this book is for you!

75 freely-downloadable Excel spreadsheets are available for download through the IWA Publishing website (https://doi.org/10.2166/9781780409320).

iwapublishing.com
@IWAPublishing

ISBN: 9781780409313 (paperback)
ISBN: 9781780409320 (eBook)

ISBN 978-1-78040-931-3
90000
9 781780 409313

Downloaded from http://iwaponline.com/ebooks/book-pdf/643390/wio9781780409320.pdf
by guest