

Yuri A.W. Shardt

Statistics for Chemical and Process Engineers

A Modern Approach

EXTRAS ONLINE

 Springer

Statistics for Chemical and Process Engineers

Yuri A.W. Shardt

Statistics for Chemical and Process Engineers

A Modern Approach



Springer

Yuri A.W. Shardt
Institute of Automation and Complex Systems (AKS)
University of Duisburg-Essen
Duisberg, North Rhine-Westphalia
Germany

ISBN 978-3-319-21508-2 ISBN 978-3-319-21509-9 (eBook)
DOI 10.1007/978-3-319-21509-9

Library of Congress Control Number: 2015950483

Springer Cham Heidelberg New York Dordrecht London
© Springer International Publishing Switzerland 2015

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made.

Printed on acid-free paper

Springer International Publishing AG Switzerland is part of Springer Science+Business Media
(www.springer.com)

Foreword

The need for the development and understanding of large, complex data sets in a wide range of different fields, including economics, chemistry, chemical engineering, and control engineering is very important. In all these fields, the common thread is using these data sets for the development of models to forecast or predict future behaviour. Furthermore, the availability of fast computers has meant that many of the techniques can now be used and tested even on one's own computer. Although there exist a wealth of textbooks available on statistics, they are often lacking in two key respects: application to the chemical and process industry and their emphasis on computationally relevant methods. Many textbooks still contain detailed explanations of how to manually solve a problem. Therefore, the goal of this textbook is to provide a thorough mathematical and statistical background the regression analysis through the use of examples drawn from the chemical and process industries. The majority of the textbook presents the required information using matrices without linking to any particular software. In fact, the goal here is to allow the reader to implement the methods on any appropriate computational device irrespective of their specific availability. Thus, detailed examples, that is, base cases, and solution steps are provided to ease this task. Nevertheless, the textbook contains two chapters devoted to using MATLAB[®] and Excel[®], as these are the most commonly used tools both in industry and in academics. Finally, the textbook contains at the end of each chapter a series of questions divided into three parts: conceptual questions to test the reader's understanding of the material; simple exercise problems that can be solved using pen, paper, and a simple, handheld calculator to provide straightforward examples to test the mechanics and understanding of the material; and computational questions that require modern computational software that challenge and advance the reader's understanding of the material.

This textbook assumes that the reader has completed a basic first-year university course, including univariate calculus and linear algebra. Multivariate calculus, set theory, and numerical methods are useful for understanding some of the concepts,

but knowledge is not required. Basic chemical engineering, including mass and energy balances, may be required to solve some of the examples.

The textbook is written so that the chapters flow from the basic to the most advanced material with minimal assumptions about the background of the reader. Nevertheless, multiple different courses can be organised based on the material presented here depending on the time and focus of the course. Assuming a single semester course of 39 h, the following would be some options:

1. *Introductory Course to Statistics and Data Analysis*: The foundations of statistics and regression are introduced and examined. The main focus would be on Chap. 1: Introduction to Statistics and Data Visualisation, Chap. 2: Theoretical Foundation for Statistical Analysis, and parts of Chap. 3: Regression, including all of linear regression. This course would prepare the student to take the Fundamentals of Engineering Exam in the United States of America, a prerequisite for becoming an engineer there.
2. *Deterministic Modelling and Design of Experiments*: In-depth analysis and interpretation of deterministic models, including design of experiments, is introduced. The main focus would be on Chap. 3: Regression and Chap. 4: Design of Experiments. Parts of Chap. 2: Theoretical Foundation for Statistical Analysis may be included if there is a need to refresh the student's knowledge of background information.
3. *Stochastic Modelling of Dynamic Processes*: In-depth analysis and interpretation of stochastic models, including both time series and prediction error methods, is examined. The main focus would be on Chap. 5: Modelling Stochastic Processes with Time Series Analysis and Chap. 6: Modelling Dynamic Processes. As necessary, information from Chap. 2: Theoretical Foundation for Statistical Analysis and Chap. 3: Regression could be used. The depth in which these concepts would be considered would depend on the orientation of the course: either a theoretical emphasis can be made, by focusing on the theory and proofs, or an application emphasis can be made, by focusing on the practical use of the different results.

As appropriate, material from Chap. 7: Using MATLAB[®] for Statistical Analysis and Chap. 8: Using Excel[®] to do Statistical Analysis could be introduced to show and explain how the students can implement the proposed methods. It should be emphasised that this material should not overwhelm the students nor should it become the main emphasis and hence avoid thoughtful and insightful analysis of the resulting data.

The author would like to thank all those who read and commented on previous versions of this textbook, especially the members of the process control group at the University of Alberta, the students who attended the author's course on process data analysis in the Spring/Summer 2012 semester, and members of the Institute of Automation and Complex Systems (Institute für Automatisierungstechnik und komplexe Systeme) at the University of Duisburg-Essen. The author would specifically wish to thank Profs. Steven X. Ding and Biao Huang for their support,

Oliver Jackson from Springer for his assistance and support, and the Alexander von Humboldt Foundation for the monetary support.

Downloading the data: The data sets, MATLAB[®] files, and Excel[®] templates can be downloaded from <http://extras.springer.com/>. Enter the ISBN of the book, ISBN 978-3-319-21508-2, and you will get the requested information.

Contents

1	Introduction to Statistics and Data Visualisation	1
1.1	Basic Descriptive Statistics	3
1.1.1	Measures of Central Tendency	3
1.1.2	Measures of Dispersion	4
1.1.3	Other Statistical Measures	6
1.2	Data Visualisation	8
1.2.1	Bar Charts and Histograms	9
1.2.2	Pie Charts	10
1.2.3	Line Charts	10
1.2.4	Box-and-Whisker Plots	12
1.2.5	Scatter Plots	13
1.2.6	Probability Plots	13
1.2.7	Tables	18
1.2.8	Sparkplots	19
1.2.9	Other Data Visualisation Methods	19
1.3	Friction Factor Example	21
1.3.1	Explanation of the Data Set	21
1.3.2	Summary Statistics	23
1.3.3	Data Visualisation	24
1.3.4	Some Observations on the Data Set	26
1.4	Further Reading	27
1.5	Chapter Problems	28
1.5.1	Basic Concepts	28
1.5.2	Short Exercises	29
1.5.3	Computational Exercises	29
2	Theoretical Foundation for Statistical Analysis	31
2.1	Statistical Axioms and Definitions	31
2.2	Expectation Operator	37
2.3	Multivariate Statistics	38

2.4	Common Statistical Distributions	43
2.4.1	Normal Distribution	43
2.4.2	Student's t -Distribution	45
2.4.3	χ^2 -Distribution	46
2.4.4	F -Distribution	47
2.4.5	Binomial Distribution	48
2.4.6	Poisson Distribution	50
2.5	Parameter Estimation	50
2.5.1	Considerations for Parameter Estimation	51
2.5.2	Methods of Parameter Estimation	52
2.5.3	Remarks on Estimating the Mean, Variance, and Standard Deviation	57
2.6	Central Limit Theorem	58
2.7	Hypothesis Testing and Confidence Intervals	58
2.7.1	Computing the Critical Value	61
2.7.2	Converting Confidence Intervals	62
2.7.3	Testing the Mean	64
2.7.4	Testing the Variance	67
2.7.5	Testing a Ratio or Proportion	68
2.7.6	Testing Two Samples	69
2.8	Further Reading	79
2.9	Chapter Problems	79
2.9.1	Basic Concepts	79
2.9.2	Short Exercises	80
2.9.3	Computational Exercises	83
	Appendix A2: A Brief Review of Set Theory and Notation	84
3	Regression	87
3.1	Regression Analysis Framework	87
3.2	Regression Models	88
3.2.1	Linear and Nonlinear Regression Functions	90
3.3	Linear Regression	93
3.3.1	Ordinary, Least-Squares Regression	93
3.3.2	Analysis of Variance of the Regression Model	99
3.3.3	Useful Formulae for Ordinary, Least-Squares Regression	102
3.3.4	Computational Example Part I: Determining the Model Parameters	104
3.3.5	Model Validation	107
3.3.6	Computational Example Part II: Model Validation	114
3.3.7	Weighted, Least-Squares Regression	116
3.4	Nonlinear Regression	120
3.4.1	Gauss–Newton Solution for Nonlinear Regression	121
3.4.2	Useful Formulae for Nonlinear Regression	122
3.4.3	Computational Example of Nonlinear Regression	123
3.5	Models and Their Use	126

3.6	Summative Regression Example	126
3.6.1	Data and Problem Statement	127
3.6.2	Solution	127
3.7	Further Reading	131
3.8	Chapter Problems	131
3.8.1	Basic Concepts	131
3.8.2	Short Exercises	132
3.8.3	Computational Exercises	134
Appendix A3: Nonmatrix Solutions to the Linear, Least-Squares Regression Problem		137
A.1 Nonmatrix Solution for the Ordinary, Least-Squares Case		137
A.2 Nonmatrix Solution for the Weighted, Least-Squares Case		139
4	Design of Experiments	141
4.1	Fundamentals of Design of Experiments	141
4.1.1	Sensitivity	142
4.1.2	Confounding and Correlation Between Parameters	142
4.1.3	Blocking	143
4.1.4	Randomisation	145
4.2	Types of Models	145
4.2.1	Model Use	145
4.3	Framework for the Analysis of Experiments	146
4.4	Factorial Design	147
4.4.1	Factorial Design Models	147
4.4.2	Factorial Analysis	150
4.4.3	Selecting Influential Parameters (Effects)	152
4.4.4	Projection	152
4.5	Fractional Factorial Design	157
4.5.1	Notation for Fractional Factorial Experiments	158
4.5.2	Resolution of Fractional Factorial Experiments	158
4.5.3	Confounding in Fractional Factorial Experiments	158
4.5.4	Design Procedure for Fractional Factorial Experiments	166
4.5.5	Analysis of Fractional Factorial Experiments	168
4.5.6	Framework for the Analysis of Factorial Designs	169
4.6	Blocking and Factorial Design	176
4.7	Generalised Factorial Design	178
4.7.1	Obtaining an Orthogonal Basis	179
4.7.2	Orthogonal Bases for Different Levels	180
4.7.3	Sum of Squares in Generalised Factorial Designs	186
4.7.4	Detailed Mixed-Level Example	187

4.8	2^k Factorial Designs with Centre Point Replicates	192
4.8.1	Orthogonal Basis for 2^k Factorial Designs with Centre Point Replicates	193
4.8.2	Factorial Design with Centre Point Example	195
4.9	Response Surface Design	198
4.9.1	Central Composite Design	199
4.9.2	Optimal Design	201
4.9.3	Response Surface Procedure	201
4.10	Further Reading	202
4.11	Chapter Problems	202
4.11.1	Basic Concepts	202
4.11.2	Short Exercises	203
4.11.3	Computational Exercises	205
	Appendix A4: Nonmatrix Approach to the Analysis of 2^k -Factorial Design Experiments	208
5	Modelling Stochastic Processes with Time Series Analysis	211
5.1	Fundamentals of Time Series Analysis	212
5.1.1	Estimating the Autocovariance and Cross-Covariance and Correlation Functions	215
5.1.2	Obtaining a Stationary Time Series	216
5.1.3	Edmonton Weather Data Series Example	216
5.2	Common Time Series Models	219
5.3	Theoretical Examination of Time Series Models	222
5.3.1	Properties of a White Noise Process	223
5.3.2	Properties of a Moving-Average Process	223
5.3.3	Properties of an Autoregressive Process	228
5.3.4	Properties of an Integrating Process	233
5.3.5	Properties of ARMA and ARIMA Processes	235
5.3.6	Properties of the Seasonal Component of a Time Series Model	237
5.3.7	Summary of the Theoretical Properties for Different Time Series Models	239
5.4	Time Series Modelling	240
5.4.1	Estimating the Time Series Model Parameters	241
5.4.2	Maximum-Likelihood Parameter Estimates for ARMA Models	245
5.4.3	Model Validation for Time Series Models	250
5.4.4	Model Prediction and Forecasting Using Time Series Models	253
5.5	Frequency-Domain Analysis of Time Series	259
5.5.1	Fourier Transform	259
5.5.2	Periodogram and Its Use in Frequency-Domain Analysis of Time Series	262

5.6	State-Space Modelling of Time Series	266
5.6.1	State-Space Model for Time Series	266
5.6.2	The Kalman Equation	267
5.6.3	Maximum-Likelihood State-Space Estimates	270
5.7	Comprehensive Example of Time Series Modelling	271
5.7.1	Summary of Available Information	271
5.7.2	Obtaining the Final Univariate Model	272
5.8	Further Reading	273
5.9	Chapter Problems	274
5.9.1	Basic Concepts	275
5.9.2	Short Exercises	276
5.9.3	Computational Exercises	276
	Appendix A5: Data Sets for This Chapter	277
	A5.1: Edmonton Weather Data Series (1882–2002)	277
	A5.2: AR(2) Process Data	281
	A5.3: MA(3) Process Data	282
6	Modelling Dynamic Processes Using System Identification	
	Methods	283
6.1	Control and Process System Identification	284
6.1.1	Predictability of Process Models	287
6.2	Framework for System Identification	291
6.3	Open-Loop Process Identification	292
6.3.1	Parameter Estimation in Process Identification	292
6.3.2	Model Validation in Process Identification	296
6.3.3	Design of Experiments in Process Identification	298
6.3.4	Final Considerations in Open-Loop Process Identification	300
6.4	Closed-Loop Process Identification	303
6.4.1	Indirect Identification of a Closed-Loop Process	305
6.4.2	Direct Identification of a Closed-Loop Process	306
6.4.3	Joint Input-Output Identification of a Closed-Loop Process	308
6.5	Nonlinear Process Identification	309
6.5.1	Transformation of Nonlinear Models: Wiener-Hammerstein Models	310
6.6	Modelling the Water Level in a Tank	310
6.6.1	Design of Experiment	311
6.6.2	Raw Data	313
6.6.3	Linear Model Creation and Validation	314
6.6.4	Nonlinear Model Creation and Validation	318
6.6.5	Final Comments	320
6.7	Further Reading	321

6.8	Chapter Problems	321
6.8.1	Basic Concepts	322
6.8.2	Short Exercises	322
6.8.3	Computational Exercises	324
	Appendix A6: Data Sets for This Chapter	324
	A6.1: Water Level in Tanks 1 and 2 Data	324
7	Using MATLAB® for Statistical Analysis	337
7.1	Basic Statistical Functions	337
7.2	Basic Functions for Creating Graphs	337
7.3	The Statistics and Machine Learning Toolbox	341
7.3.1	Probability Distributions	341
7.3.2	Advanced Statistical Functions	341
7.3.3	Useful Probability Functions	342
7.3.4	Linear Regression Analysis	342
7.3.5	Design of Experiments	342
7.4	The System Identification Toolbox	344
7.5	The Econometrics Toolbox	346
7.6	The Signal Processing Toolbox	346
7.7	MATLAB® Recipes	347
7.7.1	Periodogram	350
7.7.2	Autocorrelation Plot	351
7.7.3	Correlation Plot	352
7.7.4	Cross-Correlation Plot	352
7.8	MATLAB® Examples	354
7.8.1	Linear Regression Example in MATLAB	354
7.8.2	Nonlinear Regression Example in MATLAB	358
7.8.3	System Identification Example in MATLAB	361
7.9	Further Reading	362
8	Using Excel® to Do Statistical Analysis	363
8.1	Ranges and Arrays in Excel	363
8.2	Useful Excel Functions	365
8.2.1	Array Functions in Excel	365
8.2.2	Statistical Functions in Excel	365
8.3	Excel Macros and Security	366
8.3.1	Security in Excel	367
8.4	The Excel Solver Add-In	368
8.4.1	Installing the Solver Add-In	368
8.4.2	Using the Solver Add-In	369
8.5	The Excel Data Analysis Add-In	374
8.6	Excel Templates	376
8.6.1	Normal Probability Plot Template	377
8.6.2	Box-and-Whisker Plot Template	378
8.6.3	Periodogram Template	383

8.6.4	Linear Regression Template	385
8.6.5	Nonlinear Regression Template	386
8.6.6	Factorial Design Analysis Template	386
8.7	Excel Examples	388
8.7.1	Linear Regression Example in Excel	389
8.7.2	Nonlinear Regression Example in Excel	391
8.7.3	Factorial Design Examples Using Excel	395
8.8	Further Reading	395
Appendix A: Solution Key		399
Chapter 1		399
Chapter 2		399
Chapter 3		400
Chapter 4		401
Chapter 5		401
Chapter 6		401
References		403
Subject Index		407
Index of Excel and MATLAB Topics		413

List of Figures

Fig. 1.1	(<i>Left</i>) Right-skewed and (<i>right</i>) left-skewed data set	7
Fig. 1.2	(<i>Left</i>) Vertical bar chart and (<i>right</i>) horizontal bar chart	10
Fig. 1.3	Typical histogram	11
Fig. 1.4	Typical pie chart	11
Fig. 1.5	Typical line chart	12
Fig. 1.6	Typical box-and-whisker plots	13
Fig. 1.7	Typical scatter plot	14
Fig. 1.8	Probability plots and the effect of the location parameters (μ and σ^2)	16
Fig. 1.9	Issues with probability plots. (a) Outliers. (b) Tails. (c) Concave behaviour. (d) Rounded to 3 decimal places	17
Fig. 1.10	Nine probability plots of eight samples drawn from a standard normal distribution	18
Fig. 1.11	(<i>Left</i>) Spark bar graph showing the number of times a given fault occurs over the course of many days and (<i>right</i>) sparkline showing the hourly process value for six different variables from a single unit over the course of a day	19
Fig. 1.12	Complex data visualisation example: a cross-correlation plot	20
Fig. 1.13	Complex data visualisation example: combining multiple plot types	21
Fig. 1.14	Scatter plot of the friction factor as a function of Reynolds number for all four runs	25
Fig. 1.15	Box-and-whisker plots for the friction factor experiment for the (<i>left</i>) Reynolds number and (<i>right</i>) friction factor	25
Fig. 2.1	Plot of the probability density function 1 in Example 2.2	35
Fig. 2.2	Probability density function for the normal distribution where $\mu = 0$ and $\sigma = 4$	45
Fig. 2.3	Comparison between the t -distribution with 2 degrees of freedom and the standardised normal distribution	46

Fig. 2.4	Probability density function for the χ^2 -distribution as a function of the degrees of freedom	47
Fig. 2.5	Probability density function for the F -distribution for $\nu_1 = 8$ and $\nu_2 = 10$	48
Fig. 2.6	Probability densities for the two hypotheses	59
Fig. 2.7	Three different distributions and their overlap	60
Fig. 2.8	Confidence intervals and covering a value	61
Fig. 2.9	Difference between (a) left and (b) right probabilities	62
Fig. 3.1	Flow chart for regression analysis	88
Fig. 3.2	Residuals as a function of the (<i>top, left</i>) square root of height, (<i>top, right</i>) mass flow rate, and (<i>bottom</i>) previous residual	115
Fig. 3.3	Normal probability plot of the residuals	115
Fig. 3.4	(<i>Top</i>) Normal probability plots of the residuals and (<i>Bottom</i>) residuals as a function of temperature for (<i>left</i>) linearised and (<i>right</i>) nonlinear models	125
Fig. 3.5	Extrapolation in multivariate analysis	127
Fig. 3.6	Residuals as a function of temperature	129
Fig. 3.7	Normal probability plot of the residuals	129
Fig. 3.8	Residuals as a function of the regressor for the quadratic case	130
Fig. 3.9	Normal probability plot of the residuals for the quadratic case	130
Fig. 3.10	Residuals as a function of current for Question 24	134
Fig. 4.1	Layout of the cages	144
Fig. 4.2	Normal probability plot of parameters (effects) for a 2^4 experiment with significant points highlighted and labelled	153
Fig. 4.3	Normal probability plot of the effects	156
Fig. 4.4	Normal probability plot of the residuals for the reduced model	157
Fig. 4.5	Normal probability plot of the parameters	172
Fig. 4.6	(<i>Top</i>) Normal probability plot of the residuals and (<i>bottom</i>) time series plot of the residuals with the different replicates clearly shown	173
Fig. 4.7	(<i>Top</i>) Normal probability plot of the residuals and (<i>bottom</i>) time series plot of the residuals with the different replicates clearly shown for the model reduced using the F -test	175
Fig. 4.8	Normal probability plot of the parameters for the mixed factorial example	191
Fig. 4.9	Normal probability plot of the residuals	192
Fig. 4.10	Residuals as a function of \hat{y}	192
Fig. 4.11	Time series plot of the residuals	193
Fig. 4.12	Normal probability plot of the residuals for the reduced model	198

Fig. 4.13	Residuals for the reduced model as a function of \hat{y}	198
Fig. 4.14	Residuals for the reduced model as a function of x_1	199
Fig. 4.15	Residuals for the reduced model as a function of x_2	199
Fig. 5.1	Time series plot of the mean summer temperature in Edmonton	217
Fig. 5.2	Autocorrelation plot for the mean summer temperature in Edmonton. The thick dashed lines show the 95% confidence intervals for the given data set	218
Fig. 5.3	Partial autocorrelation plot for the mean summer temperature in Edmonton. The thick dashed lines show the 95% confidence intervals for the given data set	218
Fig. 5.4	Cross-correlation between the mean summer temperature (y) and the mean spring temperature (x) in Edmonton. The thick dashed lines show the 95% confidence intervals for the given data set	219
Fig. 5.5	(<i>Left</i>) Time series plot of the given moving-average process and (<i>right</i>) autocorrelation plot for the same process	227
Fig. 5.6	(<i>Left</i>) Time series plot of the given autoregressive process and (<i>right</i>) autocorrelation plot for the same process	232
Fig. 5.7	Partial autocorrelation plot for (<i>left</i>) AR(1) and (<i>right</i>) MA(2) processes	233
Fig. 5.8	(<i>Top</i>) Time series plot, (<i>middle</i>) autocorrelation plot, and (<i>bottom</i>) partial autocorrelation plot for (<i>left</i>) integrating and (<i>right</i>) AR(1) with $\alpha = -0.98$ processes	234
Fig. 5.9	Time series plot of the ARMA process	236
Fig. 5.10	(<i>Left</i>) Autocorrelation plot and (<i>right</i>) partial autocorrelation plot for the ARMA process	237
Fig. 5.11	(<i>Left</i>) Autocorrelation plot and (<i>right</i>) partial autocorrelation plot for the seasonal autoregressive process	238
Fig. 5.12	(<i>Left</i>) Autocorrelation plot and (<i>right</i>) partial autocorrelation plot for the seasonal moving-average process	239
Fig. 5.13	(<i>Left</i>) Autocorrelation plot and (<i>right</i>) partial autocorrelation plot for the seasonal integrating process	239
Fig. 5.14	(<i>Left</i>) Normal probability plot and (<i>right</i>) autocorrelation plot for the residuals	252
Fig. 5.15	Measured and one-step-ahead forecast temperatures as a function of years since 1882	253
Fig. 5.16	Periodograms for three simple cases: (<i>left</i>) single cosine, (<i>middle</i>) single sine, and (<i>right</i>) both cosine and sine together	263
Fig. 5.17	Process with a seasonal component of 3 samples: (<i>left</i>) integrator, (<i>middle</i>) autoregressive, and (<i>right</i>) white noise	264
Fig. 5.18	A seasonal moving-average process with a seasonal component of 3 and (<i>left</i>) $\beta_1 = -0.95$, (<i>middle</i>) $\beta_1 = -0.5$, and (<i>right</i>) $\beta_1 = 0.5$	264

Fig. 5.19	Periodograms for (<i>left</i>) spring, (<i>middle</i>) summer, and (<i>right</i>) winter of the Edmonton temperature series	265
Fig. 5.20	Periodogram for the differenced summer temperature series	265
Fig. 5.21	(<i>Left</i>) Residual analysis for the final temperature model: autocorrelation plot of the residuals and (<i>right</i>) normal probability plot of the residuals	272
Fig. 5.22	Predicted and measured mean summer temperature using the final model	273
Fig. 5.23	(<i>Top</i>) Periodogram, (<i>bottom, left</i>) autocorrelation plot, and (<i>bottom, right</i>) partial autocorrelation plot for an unknown process	277
Fig. 6.1	Block diagram of the control system	284
Fig. 6.2	Generic open-loop process	285
Fig. 6.3	System identification framework	292
Fig. 6.4	Estimating parameters using a step test	301
Fig. 6.5	Estimating the time delay using (<i>left</i>) the cross-correlation plot and (<i>right</i>) the impulse response method	302
Fig. 6.6	(<i>Left</i>) Ideal behaviour for the response for the step-up and step-down check and (<i>right</i>) ideal behaviour for the response for the proportional test	303
Fig. 6.7	Block diagram for a closed-loop process	304
Fig. 6.8	Schematic of the four-tank system	311
Fig. 6.9	Level in Tank 1: (<i>left</i>) Step change in u_1 and (<i>right</i>) step change in u_2	312
Fig. 6.10	The signals and heights as a function of time	314
Fig. 6.11	Impulse responses for Tank 1 level (<i>left</i>) for u_1 and (<i>right</i>) for u_2	315
Fig. 6.12	(<i>Top</i>) Autocorrelation plot for the residuals and (<i>bottom</i>) cross-correlation plots between the inputs (<i>left</i>) u_1 and (<i>right</i>) u_2 and the residuals for the initial linear model	316
Fig. 6.13	Predicted and experimental tank levels for the initial linear model	317
Fig. 6.14	(<i>Top</i>) Autocorrelation plot for the residuals and (<i>bottom</i>) cross-correlation plots between the inputs (<i>left</i>) u_1 and (<i>right</i>) u_2 and the residuals for the final linear model	317
Fig. 6.15	Predicted and experimental tank levels for the final linear model	318
Fig. 6.16	(<i>Top</i>) Autocorrelation plot for the residuals and (<i>bottom</i>) cross-correlation plots between the inputs (<i>left</i>) u_1 and (<i>right</i>) u_2 and the residuals for the nonlinear model	319
Fig. 6.17	Predicted and experimental tank level for the nonlinear model	320
Fig. 6.18	Estimating time delay: (<i>left</i>) cross-correlation plot and (<i>right</i>) impulse response coefficients	323

Fig. 6.19	Model validation for the open-loop case: (<i>left</i>) cross-correlation between the input and the residuals and (<i>right</i>) autocorrelation of the residuals	323
Fig. 6.20	Model validation for the closed-loop case: (<i>left</i>) cross-correlation between the input and the residuals and (<i>right</i>) autocorrelation of the residuals	324
Fig. 7.1	Linear regression example: MATLAB plots of the (<i>top, left</i>) normal probability plot of the residuals, (<i>top, centre</i>) residuals as a function of y , (<i>top, right</i>) residuals as a function of the first regressor, x_1 , (<i>bottom, left</i>) residuals as a function of x_2 , (<i>bottom, centre</i>) residuals as a function of \hat{y} , and (<i>bottom, right</i>) a time series plot of the residuals	357
Fig. 7.2	Linear regression example: MATLAB plots of the (<i>top, left</i>) normal probability plot of the residuals, (<i>top, right</i>) residuals as a function of Π , (<i>bottom, left</i>) residuals as a function of \hat{y} , and (<i>bottom, right</i>) a time series plot of the residuals	360
Fig. 8.1	Naming a range (Excel 2007)	364
Fig. 8.2	Warning when dealing with a file with a macro in Excel 2003	367
Fig. 8.3	Security warning when macros are present (Excel 2010 or newer)	368
Fig. 8.4	Security warning when macros are present (Excel 2007). The inset shows the window that appears after clicking options	368
Fig. 8.5	Navigating to the Solver installation menu (Excel 2013)	370
Fig. 8.6	Installing Solver	371
Fig. 8.7	Location of the Solver and Data Analysis add-ins (Excel 2013)	371
Fig. 8.8	Main Solver window (Excel 2010 or newer)	371
Fig. 8.9	Add constraint window	372
Fig. 8.10	(<i>Left</i>) Solver found a solution and (<i>right</i>) Solver failed to find a solution (one possible result)	372
Fig. 8.11	Solver option window (Excel 2010 or newer)	373
Fig. 8.12	Solver window (Excel 2007 or older)	374
Fig. 8.13	Solver options (Excel 2007 or older)	374
Fig. 8.14	Data Analysis window (Excel 2010 or newer)	375
Fig. 8.15	Fourier analysis window (Excel 2010 or newer)	375
Fig. 8.16	(<i>Left</i>) Inserting a row and (<i>right</i>) column (Excel 2013)	376
Fig. 8.17	Normal probability plot data (the formulae given are those placed in the first row, and they would then be dragged down into each of the remaining rows)	377
Fig. 8.18	Resulting normal probability plot	378
Fig. 8.19	Box-and-whisker plot in Excel	379
Fig. 8.20	Creating the initial graph for a box-and-whisker plot (Excel 2013). The arrows provide the sequence of events to follow	380

Fig. 8.21	Adding error bars (Excel 2013). The arrows provide the sequence of events to follow	381
Fig. 8.22	Changing the fill and border options (Excel 2013). The arrows provide the sequence of events to follow	382
Fig. 8.23	Periodogram template layout (Excel 2013). The inset shows how to initialise the Fourier analysis function	384
Fig. 8.24	Sample full and half periodograms	385
Fig. 8.25	Linear regression template	385
Fig. 8.26	Nonlinear regression template. The inset shows how to set up the Solver (Excel 2013)	387
Fig. 8.27	Analysis of factorial experiments template	388
Fig. 8.28	Linear regression example: Data Analysis results	390
Fig. 8.29	(<i>Left</i>) Linear regression example: normal probability and (<i>right</i>) time series plots. The circled point is a potential outlier	391
Fig. 8.30	Linear regression example: Data Analysis results after removing the outlier	391
Fig. 8.31	Linear regression example: (<i>left</i>) normal probability and (<i>right</i>) time series plots after removing outliers	392
Fig. 8.32	Nonlinear regression example: Excel spreadsheet results	393
Fig. 8.33	Nonlinear regression example: (<i>left</i>) normal probability plot and (<i>right</i>) time series plot of the residuals	394
Fig. 8.34	Factorial design: full factorial example	396
Fig. 8.35	Factorial design: mixed-level example	397
Fig. 8.36	Factorial design: combined factorial and centre point example	398

List of Tables

Table 1.1	Summary of the main properties of the measures of central tendency	4
Table 1.2	Summary of the main properties of the measures of dispersion	5
Table 1.3	Typical table formatting	19
Table 1.4	Data from friction factor experiments	22
Table 1.5	Summary statistics for the friction factor data set	23
Table 1.6	Computing quartiles with different software packages	27
Table 1.7	Reactor fault types by shift (for Question 23)	30
Table 1.8	Steam control data with two different methods (for Question 24)	30
Table 2.1	Useful properties of the normal distribution	44
Table 2.2	Useful properties of the Student's t -distribution	46
Table 2.3	Useful properties of the χ^2 -distribution	47
Table 2.4	Useful properties of the F -distribution	48
Table 2.5	Useful properties of the binomial distribution	49
Table 2.6	Useful properties of the Poisson distribution	50
Table 2.7	Different software and the probability values they return	63
Table 2.8	Summary of the required critical values, bounds, and confidence intervals for testing hypotheses about the mean	64
Table 2.9	Summary of the required critical values, bounds, and confidence intervals for testing hypotheses about the variance	67
Table 2.10	Summary of the required critical values, bounds, and confidence intervals for testing hypotheses about a ratio	69
Table 2.11	Summary of the required critical values and bounds for testing hypotheses about a difference when the true variances are known	70

Table 2.12	Summary of the required critical values and bounds for testing hypotheses about a difference when the true variances are unknown, but assumed equal	70
Table 2.13	Summary of the required critical values, bounds, and confidence intervals for testing hypotheses about a paired mean value	71
Table 2.14	Summary of the required critical values and bounds for testing hypotheses about the two variances	77
Table 2.15	Summary of the required critical values and bounds for testing hypotheses about two proportions	78
Table 3.1	Height and flow rate data	104
Table 3.2	Sample, normal probability plots	109
Table 3.3	Sample scatter plots	109
Table 3.4	Sample, predicted as a function of true value plots	112
Table 3.5	Calculating Cook's distance	116
Table 3.6	Replicated data for determining the weights	119
Table 3.7	Weights for the example	120
Table 3.8	Reaction rate data	124
Table 3.9	Peak power and temperature	128
Table 3.10	Current and voltage for an unknown resistor (for Question 24)	134
Table 3.11	Freezing point of different ethylene glycol – water mixtures (for Question 27)	135
Table 3.12	Gas chromatography calibration data (for Question 28)	136
Table 3.13	Time constant (τ) as a function of the tank height (h) (for Question 29)	136
Table 3.14	Partial pressures of toluene at different temperatures (for Question 31)	137
Table 4.1	Factorial design data for a plant distillation column	155
Table 4.2	Design for the fractional factorial experiment	167
Table 4.3	Preparing beef stew ration data	171
Table 4.4	Reduced model statistics for beef stew ration example	173
Table 4.5	Model parameters and statistical scores for the beef stew ration model reduced using the F -test	174
Table 4.6	Design for a blocked, full factorial experiment	177
Table 4.7	Optimising the performance of a bottling process	187
Table 4.8	F -test values—values in bold are significant at the 95% level	190
Table 4.9	Improving chemical plant yield data set	195
Table 4.10	F -test values—values in bold are significant at the 95% level	197
Table 4.11	Design for the fractional factorial experiment (for Question 22)	204
Table 4.12	Dry soup variability data (for Question 29)	206

Table 4.13	Tool life data (for Question 30)	207
Table 4.14	Crystal optimisation data (for Question 31)	208
Table 5.1	Summary of the theoretical properties of different time series models	240
Table 5.2	Autocovariance and partial autocorrelation data (for Question 24)	276
Table 5.3	Edmonton Weather Data Series (1882–2002)	278
Table 5.4	Sample data for the AR(2) process	281
Table 5.5	Sample data for the MA(3) process	282
Table 6.1	Steady-state parameter values for the system	311
Table 6.2	Summary of the values used to obtain the time constants, where τ_p is the time constant, h is the height, θ the time delay, and t is the time. The subscript ss_1 refers to the initial steady-state values and ss_2 the final steady-state height. Subscripts b and c refer to specified time instants	312
Table 6.3	Water tank data set	325
Table 7.1	Basic statistics functions	338
Table 7.2	Basic plotting functions (functions followed by an asterisk (*) require the Statistics and Machine Learning Toolbox)	338
Table 7.3	Useful formatting options	340
Table 7.4	Probability distribution functions	341
Table 7.5	Advanced statistical functions	342
Table 7.6	Useful probability functions	342
Table 7.7	Linear regression functions	343
Table 7.8	Design of experiment functions	344
Table 7.9	System Identification Toolbox: Functions for creating the data object	346
Table 7.10	System Identification Toolbox: Functions for creating a model	347
Table 7.11	System Identification Toolbox: Functions for validating a model	348
Table 7.12	System Identification Toolbox: Functions for designing a system identification experiment	348
Table 7.13	Econometrics Toolbox: Functions for creating the data object	348
Table 7.14	Econometrics Toolbox: Functions for creating various correlation plots	349
Table 7.15	Econometrics Toolbox: Functions for estimating model parameters	349
Table 7.16	Econometrics Toolbox: Functions for validating the model	349
Table 7.17	Signal Processing Toolbox: Functions for analysing signals ...	349
Table 7.18	Fitting the virial equation (MATLAB example)	355
Table 7.19	Equilibrium cell volume data (MATLAB example)	358

Table 8.1	Excel array functions	365
Table 8.2	Excel statistical functions	366
Table 8.3	Fitting the virial equation (Excel example)	389
Table 8.4	Equilibrium cell volume data (Excel example)	392
Table A.1	Answers for question 27 in Chap. 2	400

Chapter 1

Introduction to Statistics and Data Visualisation

*Εἰκὸς γὰρ γίνεσθαι πολλὰ καὶ παρὰ τὸ εἰκός.
It is likely that unlikely things should happen.*

Aristotle, Poetics, 1456a, 24

Although it is a common perception that statistics seeks to quantify and categorise uncertainty and unlikely events, it is actually a much broader and more general field. In fact, statistics is the science of collecting, analysing, interpreting, and displaying data in an objective manner. Built on a strong foundation in probability, the application of statistics has expanded to consider such topics as curve fitting, game theory, and forecasting. Its results are applied in many different fields, including biology, market research, polling, economics, cryptography, chemistry, and process engineering.

Basic statistical methods have been traced back to the earliest times in such forms as the collection of data regarding a farmer's livestock; the amount, quality, and type of grain in the city granaries; or the phases of the moon by early astronomers. With these simple data sets, graphs could be created, summary values could be computed, and patterns could be detected and used. Greek philosophers, such as Aristotle (384–322 B.C), pontificated on the meaning of probability and its different realisations. Meanwhile, ancient astronomers, such as Ptolemy (c. A.D. 90–168) and Al-Biruni (973–1048), were developing methods to deal with the randomness and inherent errors in their astronomical measurements. By the start of the late Middle Ages around 1,300, rudimentary probability was being developed and applied to break codes. With the start of the seventeenth century and spurred by a general interest in games of chance, the foundations of statistics probability were developed by Abraham de Moivre (1667–1754), Blaise Pascal (1623–1662), and Jacob Bernoulli (1655–1705). These scientists sought to resolve and determine optimal strategies for such games of chance. The nascent nation states also took a strong interest in the collection and interpretation of economic and demographic information. In fact, the word *statistics*, first used by the German philosopher Gottfried Achenwall (1719–1772) in 1749, is derived from the Neolatin term *statisticum collegium*, meaning *council of the state*, referring to the fact that even then the primary use of the collected information was to provide insight (*council*) about the nation state (Varberg 1963). In the early nineteenth century, work by

amongst others Johann Carl Friedrich Gauss (1777–1855), Pierre-Simon Laplace (1749–1827), and Thomas Bayes (1701–1761) led to the development of new theoretical and practical ideas. Theoretically, the grounding of statistics in probability theory, especially the development of the Gaussian distribution, allowed for many practical applications, including curve fitting and linear regression. Subsequent work, by such researchers as Andrei Kolmogorov (1903–1987) and Andrei Markov (1856–1922), solidified the theoretical underpinning and developed new ways of understanding randomness and methods for quantifying its behaviour. From these foundations, Karl Pearson (1857–1936) and Ronald Fisher (1890–1962) developed hypothesis testing, the χ^2 -distribution, principal component analysis, design of experiments, analysis of variance, and method of maximum likelihood, which continue to be used today. Subsequently, these ideas were used by George Box (1919–2013), Gwilym Jenkins (1932–1982), and Lenart Ljung (1946–) to develop stochastic modelling and advanced probabilistic models with applications in economics, biology, and process control. With the advent of computers, many of the previously developed methods can now be realised efficiently and quickly to analyse enormous amounts of data. Furthermore, the increasing availability of computers has led to the use of new methods, such as Monte Carlo simulations and bootstrapping.

Even though statistics still remains solidly applied to the study of economics and demographics, it has broadened its scope to cover almost every human endeavour. Some of the earliest modern applications were to the design and analysis of agricultural experiments to show which fertilisers and watering methods were better despite uncontrollable environmental differences, for example, amount of sunlight received or local soil conditions. Later these methods were extended to analyse various genetic experiments. Currently, with the use of powerful computers, it is possible to process and unearth unexpected statistical relationships in a data set given many thousands of variables. For example, advertisers can now accurately predict changes in consumer behaviour based on their purchases over a period of time.

Another area where statistics is used greatly is the chemical process industry, which seeks to understand and interpret large amounts of industrial data obtained from a given (often, chemical) process in order to achieve a safer, more environmentally friendly, and more profitable plant. The process industry uses a wide range of statistics, ranging from simple descriptive methods through to linear regression and on to complex topics such as system identification and data mining. In order to appreciate the more advanced methods, there is a need to thoroughly understand the fundamentals of statistics. Therefore, this chapter will start the exploration with some fundamental results in statistical analysis of data sets coupled with a thorough analysis of the different methods for visualising or displaying data. Subsequent chapters will provide a more theoretical approach and cover more complex methods that will always come back to use the methods presented here. Finally, as a side note, it should be noted that the focus of this book is on presenting methods that can be used with modern computers. For these reasons, heavy emphasis will be made on matrices and generalised approaches to solving the problems. However, except for

the last two chapters dedicated to MATLAB[®] and Excel[®], little to no emphasis will be placed on any specific software as a computational tool; instead the theoretical and implementation aspects will be examined in depth.

1.1 Basic Descriptive Statistics

The most basic step in statistical analysis of a data set is to describe it descriptively, that is, to compute properties associated with the data set and to display the data set in an informative manner. A data set consists of a finite number of *samples* or data points. In this book, a data set will be denoted using either set notation, that is, $\{x_1, x_2, \dots, x_n\}$ or vector notation, that is, as $\vec{x} = \langle x_1, x_2, \dots, x_n \rangle$. Set notation is useful for describing and listing the elements of a data set, while vector notation is useful for mathematical manipulation. The size of the data set is equal to n . The most common descriptive statistics include measures of *central tendency* and *dispersion*.

1.1.1 Measures of Central Tendency

Measures of central tendency provide some information about the most common value in the data set. The basic measures of central tendency include the *mean*, *mode*, and *median*. Since the most common such measure is the *mean*, which is often colloquially called the average, all of these measures are often referred to as *averages*. A summary of the basic properties of these measures is provided in Table 1.1.

The *mean* is a measure of the central value of the set of numbers. It is often denoted as an overbar ($\bar{}$) over a variable, for example, the mean of \vec{x} would be written as \bar{x} . The most common **mean** is simply the sum of all the values divided by the total number of data points, n , that is,

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} \quad (1.1)$$

Alternatively, a weighted mean can be computed, where for each value a weight w is assigned, that is,

$$\bar{x} = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i} \quad (1.2)$$