

Yuri A. W. Shardt

# Statistics for Chemical and Process Engineers

A Modern Approach

*Second Edition*

 Springer

# Statistics for Chemical and Process Engineers


Yuri A. W. Shardt

# Statistics for Chemical and Process Engineers

A Modern Approach

Second Edition

 Springer

Yuri A. W. Shardt   
Department of Automation Engineering  
Technical University of Ilmenau  
Ilmenau, Germany

ISBN 978-3-030-83189-9      ISBN 978-3-030-83190-5 (eBook)  
<https://doi.org/10.1007/978-3-030-83190-5>

Microsoft and Excel are trademarks of the Microsoft group of companies. and MATLAB is a registered trademark of The MathWorks, Inc. See <https://www.mathworks.com/trademarks> for a list of additional trademarks

1<sup>st</sup> edition: © Springer International Publishing Switzerland 2015

2<sup>nd</sup> edition: © The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2022, corrected publication 2023

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of 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, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG  
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

# Preface

The need for the development and understanding of large, complex data sets exists in a wide range of different fields, including economics, chemistry, chemical engineering, and control engineering. 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 to 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<sup>®</sup> and Excel<sup>®</sup>, as these are the most commonly used tools both in academics and in industry. 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 course 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<sup>®</sup> for Statistical Analysis and Chap. 8: Using Excel<sup>®</sup> 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, the members of the Institute of Control Engineering and Complex Systems (Institut für Automatisierungstechnik und komplexe Systeme) at the University of Duisburg-Essen, the members of the Department of Automation Engineering (Fachgebiet Automatisierungstechnik) at the Technical University of Ilmenau (Technische Universität Ilmenau), and the students who attended the course "System Identification" at the Technical University of Ilmenau. 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 monetary support. As well, the author would like to thank Ying Deng, Mike Eichhorn, Benedikt Oppeneiger, and M.P. for their help in improving the English version.

Downloading the data: The data sets, MATLAB<sup>®</sup> files, and Excel<sup>®</sup> templates can be downloaded from <https://link.springer.com/book/9783030831899>.

Ilmenau, Germany

Yuri A. W. Shardt

# Symbols and Abbreviations

This section summarises the key symbols and abbreviations found in the book. Please note that despite attempts to limit symbols to single meaning and use, it is occasionally necessary to have multiple meanings assigned to a given symbol. These cases are clarified by pointing out which meaning is used in which chapter. Symbols are always written in italics or using special fonts, while abbreviations are written in uppercase, not in italics, and using normal Latin letters. Symbols and abbreviations are in normal alphabetical order.

## Symbols

$\lfloor \cdot \rfloor$	Round-down function
$A$	$A$ -polynomial
$\mathcal{A}$	Regression matrix (Chaps. 3, 4); state matrix (Chap. 5)
$B$	$B$ -polynomial
$\mathcal{B}$	Input matrix
$\mathfrak{B}(1, q)$	Bernoulli distribution
$\mathfrak{B}(n, q)$	Binomial distribution
$C$	$C$ -polynomial
$\mathbb{C}$	Complex numbers
$\mathcal{C}$	Output matrix
$D$	$D$ -polynomial; seasonal differencing order (both Chap. 5)
$d$	Differencing order
$\mathcal{D}$	Throughput matrix
$D_i$	Cook's Distance
$E$	Expectation operator
$e$	(White, Gaussian) noise
$e_t$	Disturbance signal
$F$	$F$ -polynomial
$f$	Frequency

$\mathcal{F}$	Fisher information matrix
$\mathbb{F}$	Space of all possible events
$f(x)$	Probability density function; probability mass function
$f_{X1}$	Marginal probability density function
$f_{Y X}$	Conditional probability density function
$\mathfrak{F}$	Fourier transform
$\mathfrak{F}(\nu_1, \nu_2), F_{\nu_1, \nu_2}$	$F$ -distribution
$(\omega)$	Power spectrum; spectral density
$G$	General transfer function
$G_a$	Actuator model
$G_c$	Controller model
$G_l$	Disturbance model
$G_p$	Process model
$G_s$	Sensor model
$g_{\hat{\theta}}$	Efficiency
$h$	Impulse response coefficients
$H_0$	Null hypothesis
$H_1$	Alternative hypothesis
$I$	Identity matrix
$\mathcal{J}$	Jacobian matrix
$\mathcal{J}'$	Grand Jacobian matrix
$\mathcal{J}_t$	Kalman smoother gain
$k$	Factor (Chap. 4); discrete time $\in \mathbb{N}$ (Chaps. 5, 6)
$l$	Level
$L(\theta x)$	Likelihood function
$\ell(\theta x)$	Log-likelihood function
$m$	Number of data points for regression, in a time series
$m_i$	Uncentred moment
$\bar{m}_i$	Centred moment
$n$	Number of samples, parameters
$\mathbb{N}$	Set of natural numbers
$\mathcal{N}(\mu, \sigma_2)$	Gaussian (normal) distribution
$n_C$	Number of centre point replicates
$n_R$	Number of replicates
$P$	Probability measure function; order of the seasonal autoregressive polynomial (Chap. 5)
$p$	Order of the autoregressive polynomial (Chap. 5)
$(\lambda)$	Poisson distribution
$P(Y X)$	Conditional probability
$p_l$	Left probability
$p_r$	Right probability
$q$	Order of the moving-average polynomial
$Q$	Order of the seasonal moving-average polynomial
$r$	Residual
$\mathbb{R}$	Set of real numbers

$R^2$	Pearson's coefficient of regression
$r_{critical}$	Critical value
$r_t$	Reference signal
$S$	Sensitivity
$s$	Seasonal order
$\mathbb{S}$	Sample space
$SSE$	Sum of squares due to the error
$SSR$	Sum of squares due to regression
$SSR_i$	Sum of squares due to regression for the $i$ th parameter
$t(v), t_v$	Student's $t$ -distribution
$TSS$	Total sum of squares
$u$	Input
$u_t$	Input signal
$w$	Weight
$X$	Data point; random variable
$x$	Regressor; state (Chaps. 5, 6)
$y$	Observation
$\hat{y}_\infty$	Infinite-horizon predictor / infinite-step-ahead predictor
$y_t$	Output signal
$\hat{y}_{t+\tau t}$	$\tau$ -step-ahead predictor
$z$	Forward shift operator/z-operator
$Z$	Standard normal distribution; Z-score
$\mathbb{Z}$	Set of integers
$z^{-1}$	Backward shift operator/ $z^{-1}$ -operator
$\alpha$	False positive rate; confidence level
$\beta$	False negative rate (Chap. 2); parameter (Chaps. 3, 4)
$\beta_0$	Mean response
$\gamma$	Autocorrelation; skew (Chap. 2)
$\Gamma$	Autocovariance matrix
$\gamma_{ij}$	Experimental coefficients (Chap. 4)
$\gamma_{xy}(\tau)$	Cross-covariance
$\delta$	Bias
$\Delta$	Difference
$\varepsilon$	Error
$\varepsilon_{t+\tau t}$	Prediction error
$\theta$	Regressive coefficients; time delay (Chap. 6)
$\kappa$	Basis function
$\mu$	Mean
$\nu$	Degrees of freedom
$\rho$	Autocovariance
$\rho_{x z}(\tau)$	Partial autocorrelation
$\rho_{x1, x2}$	Correlation
$\rho_{x1z}$	Partial correlation
$\Sigma$	Covariance matrix
$\sigma$	Standard deviation

$\sigma^2$	Variance
$\sigma_{MAD}$	Median absolute deviation
$\tau_s$	Sampling time
$\phi$	Moving-average coefficients
$\chi^2(\nu), \chi^2_\nu$	$\chi^2$ -distribution
$\psi$	Sensitivity function
$\Omega$	Probability space
$\bar{\circ} \text{ (U+0305)}$	Mean value
$\hat{\circ} \text{ (U+0302)}$	Estimated value
$\vec{\circ} \text{ (U+20D7)}$	Vector
$\tilde{\circ} \text{ (U+0303)}$	Normalised value

Abbreviations

AIC	Akaike’s Information Criterion
ANOVA	Analysis of variance
AR	Autoregressive model
ARMA	Autoregressive, moving-average, exogenous model
ARX	Autoregressive exogenous model
BIC	Bayesian or Schwarz Information Criterion
BJ	Box–Jenkins Model
CCD	Central composite design
CDF	Cummulative distribution function
CI	Confidence interval
CTL	Central limit theorem
I	Integrating model
IR	Impulse response model
MA	Moving-average model
ME	Mean error
MSE	Mean-squared error
MVE	Minimum variance estimator
NLARX	Nonlinear autoregressive exogenous model
OE	Output-error model
OLS	Ordinary least squares
PACF	Partial autocorrelation function
pdf	Probability density function
RBS	Random binary signal
SARIMA	Seasonal, autoregressive, integrated, moving-average model
SNR	Signal-to-noise ratio
tf	Transfer function
WLS	Weighted least squares

# Contents

<b>1</b>	<b>Introduction to Statistics and Data Visualisation</b>	<b>1</b>
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	7
1.2	Data Visualisation	8
1.2.1	Bar Charts and Histograms	9
1.2.2	Pie Charts	11
1.2.3	Line Charts	11
1.2.4	Box-and-Whisker Plots	11
1.2.5	Scatter Plots	12
1.2.6	Probability Plots	13
1.2.7	Tables	17
1.2.8	Sparkplots	19
1.2.9	Other Data Visualisation Methods	19
1.3	Friction Factor Example	20
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
<b>2</b>	<b>Theoretical Foundation for Statistical Analysis</b>	<b>31</b>
2.1	Statistical Axioms and Definitions	31
2.2	Expectation Operator	37

2.3	Multivariate Statistics .....	39
2.4	Common Statistical Distributions .....	43
2.4.1	Normal Distribution .....	44
2.4.2	Student's $t$ -Distribution .....	46
2.4.3	$\chi^2$ -Distribution .....	46
2.4.4	$F$ -Distribution .....	49
2.4.5	Binomial Distribution .....	49
2.4.6	Poisson Distribution .....	51
2.5	Parameter Estimation .....	52
2.5.1	Considerations for Parameter Estimation .....	52
2.5.2	Methods of Parameter Estimation .....	54
2.5.3	Remarks on Estimating the Mean, Variance, and Standard Deviation .....	58
2.6	Central Limit Theorem .....	59
2.7	Hypothesis Testing and Confidence Intervals .....	60
2.7.1	Computing the Critical Value .....	63
2.7.2	Converting Confidence Intervals .....	64
2.7.3	Testing the Mean .....	65
2.7.4	Testing the Variance .....	69
2.7.5	Testing a Ratio or Proportion .....	70
2.7.6	Testing Two Samples .....	71
2.8	Further Reading .....	81
2.9	Chapter Problems .....	82
2.9.1	Basic Concepts .....	82
2.9.2	Short Exercises .....	83
2.9.3	Computational Exercises .....	86
	Appendix A2: A Brief Review of Set Theory and Notation .....	87
<b>3</b>	<b>Regression .....</b>	<b>89</b>
3.1	Regression Analysis Framework .....	89
3.2	Regression Models .....	90
3.2.1	Linear and Nonlinear Regression Functions .....	92
3.3	Linear Regression .....	95
3.3.1	Ordinary, Least-Squares Regression .....	95
3.3.2	Analysis of Variance of the Regression Model .....	101
3.3.3	Useful, Formulae for Ordinary, Least-Squares Regression .....	104
3.3.4	Computational Example Part I: Determining the Model Parameters .....	106
3.3.5	Model Validation .....	110
3.3.6	Computational Example Part II: Model Validation .....	118
3.3.7	Weighted, Least-Squares Regression .....	120

3.4	Nonlinear Regression	124
3.4.1	Gauss–Newton Solution for Nonlinear Regression	125
3.4.2	Useful Formulae for Nonlinear Regression	126
3.4.3	Computational Example of Nonlinear Regression	127
3.5	Models and Their Use	130
3.6	Summative Regression Example	130
3.6.1	Data and Problem Statement	131
3.6.2	Solution	131
3.7	Further Reading	134
3.8	Chapter Problems	135
3.8.1	Basic Concepts	135
3.8.2	Short Exercises	136
3.8.3	Computational Exercises	139
Appendix A3:	Nonmatrix Solutions to the Linear, Least-Squares Regression Problem	142
<b>4</b>	<b>Design of Experiments</b>	147
4.1	Fundamentals of Design of Experiments	147
4.1.1	Sensitivity	147
4.1.2	Confounding and Correlation Between Parameters	148
4.1.3	Blocking	149
4.1.4	Randomization	150
4.2	Types of Models	151
4.2.1	Model Use	151
4.3	Framework for the Analysis of Experiments	152
4.4	Factorial Design	153
4.4.1	Factorial Design Models	153
4.4.2	Factorial Analysis	156
4.4.3	Selecting Influential Parameters (Effects)	158
4.4.4	Projection	159
4.5	Fractional Factorial Design	163
4.5.1	Notation for Fractional Factorial Experiments	164
4.5.2	Resolution of Fractional Factorial Experiments	164
4.5.3	Confounding in Fractional Factorial Experiments	164
4.5.4	Design Procedure for Fractional Factorial Experiments	172
4.5.5	Analysis of Fractional Factorial Experiments	174
4.5.6	Framework for the Analysis of Factorial Designs	175
4.6	Blocking and Factorial Design	182
4.7	Generalized Factorial Design	184
4.7.1	Obtaining an Orthogonal Basis	185
4.7.2	Orthogonal Bases for Different Levels	186
4.7.3	Sum of Squares in Generalized Factorial Designs	193
4.7.4	Detailed Mixed-Level Example	194

4.8	2 <sup>k</sup> -Factorial Designs with Centre Point Replicates . . . . .	200
4.8.1	Orthogonal Basis for 2 <sup>k</sup> -Factorial Designs with Centre Point Replicates . . . . .	200
4.8.2	Factorial Design with Centre Point Example . . . . .	202
4.9	Response Surface Design . . . . .	205
4.9.1	Central Composite Design . . . . .	207
4.9.2	Optimal Design . . . . .	208
4.9.3	Response Surface Procedure . . . . .	209
4.10	Further Reading . . . . .	209
4.11	Chapter Problems . . . . .	210
4.11.1	Basic Concepts . . . . .	210
4.11.2	Short Exercises . . . . .	211
4.11.3	Computational Exercises . . . . .	213
Appendix A4:	Nonmatrix Approach to the Analysis of 2 <sup>k</sup> -Factorial Design Experiments . . . . .	216
<b>5</b>	<b>Modelling Stochastic Processes with Time-Series Analysis . . . . .</b>	<b>219</b>
5.1	Fundamentals of Time-Series Analysis . . . . .	220
5.1.1	Estimating the Auto- and Cross-Covariance and Correlation Functions . . . . .	223
5.1.2	Obtaining a Stationary Time Series . . . . .	224
5.1.3	Edmonton Weather Data Series Example . . . . .	224
5.2	Common Time-Series Models . . . . .	227
5.3	Theoretical Examination of Time-Series Models . . . . .	231
5.3.1	Properties of a White-Noise Process . . . . .	232
5.3.2	Properties of a Moving-Average Process . . . . .	232
5.3.3	Properties of an Autoregressive Process . . . . .	237
5.3.4	Properties of an Integrating Process . . . . .	242
5.3.5	Properties of ARMA and ARIMA Processes . . . . .	244
5.3.6	Properties of the Seasonal Component of a Time-Series Model . . . . .	246
5.3.7	Summary of the Theoretical Properties for Different Time-Series Models . . . . .	249
5.4	Time-Series Modelling . . . . .	249
5.4.1	Estimating the Time-Series Model Parameters . . . . .	250
5.4.2	Maximum Likelihood Parameter Estimates for ARMA Models . . . . .	255
5.4.3	Model Validation for Time-Series Models . . . . .	260
5.4.4	Model Prediction and Forecasting Using Time-Series Models . . . . .	263
5.5	Frequency-Domain Analysis of Time Series . . . . .	269
5.5.1	Fourier Transform . . . . .	269
5.5.2	The Periodogram and Its Use in the Frequency-Domain Analysis of Time Series . . . . .	272

5.6	State-Space Modelling of Time Series .....	276
5.6.1	State-Space Model for Time Series .....	277
5.6.2	The Kalman Equation .....	277
5.6.3	Maximum Likelihood State-Space Estimates .....	280
5.7	Comprehensive Example of Time-Series Modelling .....	281
5.7.1	Summary of Available Information .....	281
5.7.2	Obtaining the Final Univariate Model .....	282
5.8	Further Reading .....	284
5.9	Chapter Problems .....	285
5.9.1	Basic Concepts .....	285
5.9.2	Short Exercises .....	286
5.9.3	Computational Exercises .....	287
	Appendix A5: Data Sets for This Chapter .....	288
<b>6</b>	<b>Modelling Dynamic Processes Using System Identification</b>	
	<b>Methods</b> .....	301
6.1	Control and Process System Identification .....	302
6.1.1	Predictability of Process Models .....	305
6.2	Framework for System Identification .....	309
6.3	Open-Loop Process Identification .....	309
6.3.1	Parameter Estimation in Process Identification .....	311
6.3.2	Model Validation in Process Identification .....	314
6.3.3	Design of Experiments in Process Identification .....	316
6.3.4	Final Considerations in Open-Loop Process Identification .....	318
6.4	Closed-Loop Process Identification .....	321
6.4.1	Indirect Identification of a Closed-Loop Process .....	323
6.4.2	Direct Identification of a Closed-Loop Process .....	324
6.4.3	Joint Input–Output Identification of a Closed-Loop Process .....	326
6.5	Nonlinear Process Identification .....	327
6.5.1	Transformation of Nonlinear Models: Wiener–Hammerstein Models .....	328
6.6	Modelling the Water Level in a Tank .....	328
6.6.1	Design of Experiment .....	329
6.6.2	Raw Data .....	331
6.6.3	Linear Model Creation and Validation .....	331
6.6.4	Nonlinear Model Creation and Validation .....	336
6.6.5	Final Comments .....	337
6.7	Further Reading .....	338
6.8	Chapter Problems .....	339
6.8.1	Basic Concepts .....	339
6.8.2	Short Exercises .....	340
6.8.3	Computational Exercises .....	341
	Appendix A6: Data Sets for This Chapter .....	342

<b>7</b>	<b>Using MATLAB® for Statistical Analysis</b>	<b>357</b>
7.1	Basic Statistical Functions	357
7.2	Basic Functions for Creating Graphs	357
7.3	The Statistics and Machine Learning Toolbox	358
7.3.1	Probability Distributions	358
7.3.2	Advanced Statistical Functions	358
7.3.3	Advanced Probability Functions	358
7.3.4	Linear Regression Analysis	361
7.3.5	Design of Experiments	361
7.4	The System Identification Toolbox	361
7.5	The Econometrics Toolbox	366
7.6	The Signal Processing Toolbox	366
7.7	MATLAB® Recipes	367
7.7.1	Periodogram	367
7.7.2	Autocorrelation Plot	372
7.7.3	Correlation Plot	373
7.7.4	Cross-Correlation Plot	374
7.8	MATLAB® Examples	376
7.8.1	Linear Regression Example in MATLAB®	376
7.8.2	Nonlinear Regression Example in MATLAB®	380
7.8.3	System Identification Example in MATLAB®	383
7.9	Further Reading	385
<b>8</b>	<b>Using Excel® to Do Statistical Analysis</b>	<b>387</b>
8.1	Ranges and Arrays in Excel®	387
8.2	Useful Excel® Functions	389
8.2.1	Array Functions in Excel®	389
8.2.2	Statistical Functions in Excel®	389
8.3	Excel® Macros and Security	390
8.3.1	Security in Excel®	391
8.4	The Excel® Solver Add-In	391
8.4.1	Installing the Solver Add-In	391
8.4.2	Using the Solver Add-In	393
8.5	The Excel® Data Analysis Add-In	395
8.6	Excel® Templates	397
8.6.1	Normal Probability Plot Template	397
8.6.2	Box-and-Whisker Plot Template	400
8.6.3	Periodogram Template	402
8.6.4	Linear Regression Template	404
8.6.5	Nonlinear Regression Template	405
8.6.6	Factorial Design Analysis Template	406
8.7	Excel® Examples	407
8.7.1	Linear Regression Example in Excel®	407

8.7.2	Nonlinear Regression Example in Excel®	409
8.7.3	Factorial Design Examples Using Excel®	413
8.8	Further Reading	415
<b>Correction to: Statistics for Chemical and Process Engineers</b>		<b>C1</b>
<b>Appendix A: Solution Key</b>		<b>417</b>
<b>References</b>		<b>421</b>
<b>Index</b>		<b>425</b>
<b>MATLAB and EXCEL Functions</b>		<b>431</b>

# List of Figures

Fig. 1.1	(Left) Right-skewed and (right) Left-skewed data set	6
Fig. 1.2	(Left) Vertical bar chart and (right) Horizontal bar chart	10
Fig. 1.3	Typical histogram	10
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 ( $\mu$ and $\sigma^2$ )	16
Fig. 1.9	Issues with probability plots	17
Fig. 1.10	Nine probability plots of eight samples drawn from a standard normal distribution	18
Fig. 1.11	(Left) Spark bar graph showing the number of times a given fault occurs over the course of many days and (right) 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 (left) Reynolds number and (right) Friction factor	25
Fig. 2.1	Plot of the probability density function 1 in Example 2.2	36
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 two degrees of freedom and the standardised normal distribution	47

Fig. 2.4	Probability density function for the $\chi^2$ -distribution as a function of the degrees of freedom	48
Fig. 2.5	Probability density function for the $F$ -distribution for $\nu_1 =$ 8 and $\nu_2 = 10$	50
Fig. 2.6	Probability densities for the two hypotheses	61
Fig. 2.7	Three different distributions and their overlap	62
Fig. 2.8	Confidence intervals and covering a value ( $ME = r_{crit}\sigma_\theta$ )	63
Fig. 2.9	Difference between ( <b>left</b> ) left and ( <b>right</b> ) right probabilities	64
Fig. 3.1	Flowchart for regression analysis	90
Fig. 3.2	Residuals as a function of the (top, left) Amount of compound A, (top, right) Yield strength, and (bottom) Previous residual	118
Fig. 3.3	Normal probability plot of the residuals	119
Fig. 3.4	(top) Normal probability plots of the residuals and (bottom) Residuals as a function of temperature for (left) Linearised and (right) Nonlinear models	129
Fig. 3.5	Extrapolation in multivariate analysis	131
Fig. 3.6	Residuals as a function of temperature	132
Fig. 3.7	Normal probability plot of the residuals	133
Fig. 3.8	Normal probability plot of the residuals for the quadratic case	133
Fig. 3.9	Residuals as a function of the regressor for the quadratic case	134
Fig. 3.10	Residuals as a function of current (for Question 24)	137
Fig. 4.1	Layout of the cages	150
Fig. 4.2	Normal probability plot of parameters (effects) for a $2^4$ experiment with significant points highlighted and labelled	158
Fig. 4.3	Normal probability plot of the effects	162
Fig. 4.4	Normal probability plot of the residuals for the reduced model	163
Fig. 4.5	Normal probability plot of the parameters	178
Fig. 4.6	(Top) Normal probability plot of the residuals and (bottom) Time-series plot of the residuals with the different replicates clearly shown	179
Fig. 4.7	(Top) Normal probability plot of the residuals and (bottom) Time-series plot of the residuals with the different replicates clearly shown for the model reduced using the $F$ -test	181
Fig. 4.8	Normal probability plot of the parameters for the mixed factorial example	198
Fig. 4.9	Normal probability plot of the residuals	198
Fig. 4.10	Residuals as a function of $\hat{y}$	199
Fig. 4.11	Time-series plot of the residuals	199
Fig. 4.12	Normal probability plot of the residuals for the reduced model	205

Fig. 4.13	Residuals for the reduced model as a function of $\hat{y}$ .....	205
Fig. 4.14	Residuals for the reduced model as a function of $x_1$ .....	206
Fig. 4.15	Residuals for the reduced model as a function of $x_2$ .....	206
Fig. 5.1	Time-series plot of the mean summer temperature in Edmonton .....	225
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 .....	226
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 .....	227
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 .....	228
Fig. 5.5	(left) Time-series plot of the given moving-average process and (right) Autocorrelation plot for the same process .....	236
Fig. 5.6	(left) Time-series plot of the given autoregressive process and (right) Autocorrelation plot for the same process .....	241
Fig. 5.7	Partial autocorrelation plot for (left) AR(1) and (right) MA(2) processes .....	242
Fig. 5.8	(top) Time-series plot, (middle) Autocorrelation plot, and (bottom) Partial autocorrelation plot for (left) Integrating and (right) AR(1) with $\alpha = -0.98$ processes .....	243
Fig. 5.9	Time-series plot of the ARMA process .....	246
Fig. 5.10	(left) Autocorrelation plot and (right) Partial autocorrelation plot for the ARMA process .....	246
Fig. 5.11	(left) Autocorrelation plot and (right) Partial autocorrelation plot for the seasonal autoregressive process ....	248
Fig. 5.12	(left) Autocorrelation plot and (right) Partial autocorrelation plot for the seasonal moving-average process .....	248
Fig. 5.13	(left) Autocorrelation plot and (right) Partial autocorrelation plot for the seasonal integrating process .....	248
Fig. 5.14	(left) Normal probability plot and (right) autocorrelation plot for the residuals .....	262
Fig. 5.15	Measured and one-step-ahead forecast temperatures as a function of years since 1882 .....	263
Fig. 5.16	Periodograms for three simple cases: (left) Single cosine, (middle) Single sine, and (right) Both cosine and sine together .....	274
Fig. 5.17	Process with a seasonal component of three samples: (left) Integrator, (middle) Autoregressive, and (right) White noise ....	274

Fig. 5.18	A seasonal moving-average process with a seasonal component of 3 and (left) $\beta_1 = -0.95$ , (middle) $\beta_1 = -0.5$ , and (right) $\beta_1 = 0.5$ .....	275
Fig. 5.19	Periodograms for (left) Spring, (middle) Summer, and (right) Winter of the Edmonton temperature series .....	275
Fig. 5.20	Periodogram for the differenced summer temperature series ....	276
Fig. 5.21	Residual analysis for the final temperature model: (left) Autocorrelation plot of the residuals and (right) Normal probability plot of the residuals .....	283
Fig. 5.22	Predicted and measured mean summer temperature using the final model .....	283
Fig. 5.23	(top) Periodogram, (bottom, left) Autocorrelation plot, and (bottom, right) partial autocorrelation plot for an unknown process .....	288
Fig. 6.1	Block diagram of the control system .....	302
Fig. 6.2	Generic open-loop process .....	303
Fig. 6.3	System identification framework .....	310
Fig. 6.4	Estimating parameters using a step test .....	319
Fig. 6.5	Estimating the time delay using (left) The cross-correlation plot and (right) The impulse response method .....	320
Fig. 6.6	(left) Ideal behaviour for the response for the step-up and step-down check and (right) Ideal behaviour for the response for the proportional test .....	321
Fig. 6.7	Block diagram for a closed-loop process .....	322
Fig. 6.8	Schematic of the four-tank system .....	329
Fig. 6.9	Level in Tank 1: (left) Step change in $u_1$ and (right) Step change in $u_2$ .....	330
Fig. 6.10	The signals and heights as a function of time .....	332
Fig. 6.11	Impulse responses for Tank 1 level (left) For $u_1$ and (right) For $u_2$ .....	333
Fig. 6.12	(top) Autocorrelation plot for the residuals and (bottom) Cross-correlation plots between the inputs (left) $u_1$ and (right) $u_2$ and the residuals for the initial linear model .....	334
Fig. 6.13	Predicted and experimental tank levels for the initial linear model .....	334
Fig. 6.14	(top) Autocorrelation plot for the residuals and (bottom) Cross-correlation plots between the inputs (left) $u_1$ and (right) $u_2$ and the residuals for the final linear model .....	335
Fig. 6.15	Predicted and experimental tank levels for the final linear model .....	335
Fig. 6.16	(top) Autocorrelation plot for the residuals and (bottom) Cross-correlation plots between the inputs (left) $u_1$ and (right) $u_2$ and the residuals for the nonlinear model .....	337
Fig. 6.17	Predicted and experimental tank level for the nonlinear model .....	338

Fig. 6.18	Estimating time delay: (left) Cross-correlation plot and (right) Impulse response coefficients	340
Fig. 6.19	Model validation for the open-loop case: (left) Cross-correlation between the input and the residuals and (right) Autocorrelation of the residuals	341
Fig. 6.20	Model validation for the closed-loop case: (left) Cross-correlation between the input and the residuals and (right) Autocorrelation of the residuals	341
Fig. 7.1	Linear regression example: MATLAB® plots of the (top) Normal probability plot of the residuals, Residuals as a function of $y$ , and Residuals as a function of the first regressor, $x_1$ ; and (bottom) Residuals as a function of $x_2$ , Residuals as a function of $\hat{y}$ , and A time-series plot of the residuals	379
Fig. 7.2	Nonlinear regression example: MATLAB® plots of the (top) Normal probability plot of the residuals and Residuals as a function of $\Pi$ ; and (bottom) Residuals as a function of $\hat{y}$ and A time-series plot of the residuals	383
Fig. 8.1	Naming a range (Excel® 2019)	388
Fig. 8.2	Security warning when macros are present (Excel® 2019)	391
Fig. 8.3	Navigating to the Solver installation menu (Excel® 2019)	392
Fig. 8.4	Installing Solver	392
Fig. 8.5	Location of the Solver and Data Analysis add-ins (Excel® 2019)	393
Fig. 8.6	Main Solver window (Excel® 2016 or newer)	393
Fig. 8.7	Add Constraint window	394
Fig. 8.8	(left) Solver found a solution and (right) Solver failed to find a solution (one possible result)	395
Fig. 8.9	Solver Option Window (Excel® 2016 or newer)	396
Fig. 8.10	Data Analysis window (Excel® 2016 or newer)	396
Fig. 8.11	Fourier Analysis window (Excel® 2016 or newer)	397
Fig. 8.12	Inserting a (left) row and (right) column (Excel® 2019)	398
Fig. 8.13	Normal probability plot data (The formulae given are those placed in the first row, they would then be dragged down into each of the remaining rows.)	399
Fig. 8.14	Resulting normal probability plot	399
Fig. 8.15	Box-and-whisker plot in Excel®	400
Fig. 8.16	Creating the initial graph for a box-and-whisker plot (Excel® 2019). The arrows provide the sequence of events to follow	401
Fig. 8.17	Adding error bars (Excel® 2019). The arrows provide the sequence of events to follow	401
Fig. 8.18	Changing the fill and border options (Excel® 2019). The arrows provide the sequence of events to follow	402

Fig. 8.19	Periodogram template layout (Excel® 2019). The inset shows how to initialise the Fourier analysis function	403
Fig. 8.20	Sample full and half periodograms	403
Fig. 8.21	Linear regression template	404
Fig. 8.22	Nonlinear regression template. The inset shows how to set up the Solver (Excel® 2019)	405
Fig. 8.23	Analysis of factorial experiments template	406
Fig. 8.24	Linear regression example: Data analysis results	408
Fig. 8.25	Linear regression example: (left) Normal probability and (right) Time-series plots. The circled point is a potential outlier	409
Fig. 8.26	Linear regression example: Data analysis results after removing the outlier	409
Fig. 8.27	Linear regression example: (left) Normal probability and (right) Time-series plots after removing outliers	410
Fig. 8.28	Nonlinear regression example: Excel® spreadsheet results	412
Fig. 8.29	Nonlinear regression example: (left) Normal probability plot and (right) Time-series plot of the residuals	412
Fig. 8.30	Factorial design: Full factorial example	413
Fig. 8.31	Factorial design: Mixed-level example	414
Fig. 8.32	Factorial design: Combined factorial and centre point example	414

# List of Tables

Table 1.1	Summary of the main properties of the measures of central tendency .....	3
Table 1.2	Summary of the main properties of the measures of dispersion .....	5
Table 1.3	Typical table formatting .....	18
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 .....	26
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 .....	45
Table 2.2	Useful properties of the Student's $t$ -distribution .....	47
Table 2.3	Useful properties of the $\chi^2$ -distribution .....	48
Table 2.4	Useful properties of the $F$ -distribution .....	49
Table 2.5	Useful properties of the binomial distribution .....	50
Table 2.6	Useful properties of the Poisson distribution .....	52
Table 2.7	Different software and the probability values they return .....	64
Table 2.8	Summary of the required critical values, bounds, and confidence intervals for testing hypotheses about the mean .....	66
Table 2.9	Summary of the required critical values, bounds, and confidence intervals for testing hypotheses about the variance .....	69
Table 2.10	Summary of the required critical values, bounds, and confidence intervals for testing hypotheses about a ratio .....	70
Table 2.11	Summary of the required critical values and bounds for testing hypotheses about a difference when the true variances are known .....	72

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 . . . . .	73
Table 2.13	Summary of the required critical values, bounds, and confidence intervals for testing hypotheses about a paired mean value . . . . .	74
Table 2.14	Summary of the required critical values and bounds for testing hypotheses about the two variances . . . . .	79
Table 2.15	Summary of the required critical values and bounds for testing hypotheses about two proportions . . . . .	80
Table 3.1	Yield strength data . . . . .	106
Table 3.2	Sample, normal probability plots . . . . .	112
Table 3.3	Sample scatter plots . . . . .	113
Table 3.4	Sample, predicted as a function of true value plots . . . . .	115
Table 3.5	Calculating Cook's distance . . . . .	119
Table 3.6	Replicated data for determining the weights . . . . .	123
Table 3.7	Weights for the example . . . . .	123
Table 3.8	Reaction rate data . . . . .	128
Table 3.9	Peak power and temperature . . . . .	132
Table 3.10	Current and voltage for an unknown resistor (for Question 24) . . . . .	137
Table 3.11	Height and flow rate data . . . . .	138
Table 3.12	Freezing point of different ethylene glycol—water mixtures (for Question 28) . . . . .	139
Table 3.13	Gas chromatography calibration data (for Question 29) . . . . .	140
Table 3.14	Time constant ( $\tau$ ) as a function of the water level ( $h$ ) (for Question 30) . . . . .	140
Table 3.15	Partial pressures of toluene at different temperatures (for Question 32) . . . . .	141
Table 4.1	Factorial design data for a plant distillation column . . . . .	160
Table 4.2	Design for the fractional factorial experiment . . . . .	173
Table 4.3	Preparing beef stew ration data . . . . .	177
Table 4.4	Reduced model statistics for beef stew ration example . . . . .	178
Table 4.5	Model parameters and statistical scores for the beef stew ration model reduced using the $F$ -test . . . . .	180
Table 4.6	Design for a blocked, full factorial experiment. All experiments with (—) in the final column would be run on one day and those with a (+) in the final column would be run on another day . . . . .	183
Table 4.7	Optimizing the performance of a bottling process . . . . .	194
Table 4.8	$F$ -test values—values in bold are significant at the 95% level . . . . .	197
Table 4.9	Improving chemical plant yield data set . . . . .	202
Table 4.10	$F$ -test values—values in bold are significant at the 95% level . . . . .	204

Table 4.11	Design for the fractional factorial experiment (for Question 22) .....	212
Table 4.12	Dry soup variability data (for Question 29) .....	214
Table 4.13	Tool life data (for Question 30) .....	215
Table 4.14	Crystal optimisation data (for Question 31) .....	215
Table 5.1	Summary of the theoretical properties of different time-series models .....	249
Table 5.2	Autocovariance and partial autocorrelation data (for Question 24) .....	287
Table 5.3	Edmonton Weather Data Series (1882–2002) .....	290
Table 5.4	Sample data for the AR(2) process .....	293
Table 5.5	Sample data for the MA(3) process .....	297
Table 6.1	Steady-state parameter values for the system .....	329
Table 6.2	Summary of the values used to obtain the time constants, where $\tau_p$ is the time constant, $h$ is the height, $\theta$ is 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 .....	330
Table 6.3	Water tank data set .....	342
Table 7.1	Basic statistics functions .....	358
Table 7.2	Basic plotting functions (functions followed by an asterisk (*) require the Statistics and Machine Learning Toolbox) .....	359
Table 7.3	Useful formatting options .....	362
Table 7.4	Probability distribution functions .....	363
Table 7.5	Advanced statistical functions .....	363
Table 7.6	Advanced probability functions .....	363
Table 7.7	Linear regression functions .....	364
Table 7.8	Design of experiment functions .....	365
Table 7.9	System Identification Toolbox: Functions for creating the data object .....	367
Table 7.10	System Identification Toolbox: Functions for creating a model .....	368
Table 7.11	System Identification Toolbox: Functions for validating a model .....	369
Table 7.12	System Identification Toolbox: Functions for designing a system identification experiment .....	369
Table 7.13	Econometrics Toolbox: Functions for creating the data object .....	370
Table 7.14	Econometrics Toolbox: Functions for creating various correlation plots .....	370
Table 7.15	Econometrics Toolbox: Functions for estimating model parameters .....	370
Table 7.16	Econometrics Toolbox: Functions for validating the model ...	371

Table 7.17	Signal Processing Toolbox: Functions for analysing signals .....	371
Table 7.18	Fitting the virial equation (MATLAB <sup>®</sup> example) .....	377
Table 7.19	Equilibrium cell volume data (MATLAB <sup>®</sup> example) .....	380
Table 8.1	Excel <sup>®</sup> array functions .....	389
Table 8.2	Excel <sup>®</sup> statistical functions .....	390
Table 8.3	Fitting the virial equation (Excel <sup>®</sup> example) .....	408
Table 8.4	Equilibrium cell volume data (Excel <sup>®</sup> example) .....	411
Table A.1	Answers for Question 27 in Chap. 2 .....	418

# 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 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 1300, 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, among others, works by 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  $\chi^2$ -distribution, principal component analysis, design of experiments, analysis of variance, and the 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, the 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