

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 <i>t</i> -Distribution	45
2.4.3	χ^2 -Distribution	46
2.4.4	<i>F</i> -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