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

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® and Excel®, 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® 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, the members of the Institute of Control Engineering and Complex Systems (Institute 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® files, and Excel® templates can be downloaded from https://link.springer.com/book/9783030831899.

Ilmenau, Germany Yuri A. W. Shardt

The original version of this book has been revised. There are some scientific errors in this publication from Chapter 1 to 7. The correction to this book is available at https://doi.org/10.1007/978-3-030-83190-5_9

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. This 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
$\mathscr{B}$	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

$\mathscr{F}$	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
(ω)	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 ϵ $\mathbb{N}$ (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
$\mathfrak{N}(\mu, \sigma_2)$	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)
(λ)	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 ith parameter
$t(\nu)$, t_ν	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
$\hat{y}_{t+\tau\|t}$	τ-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
α	False positive rate; confidence level
β	False negative rate (Chap. 2); parameter (Chaps. 3, 4)
β_0	Mean response
γ	Autocorrelation; skew (Chap. 2)
Γ	Autocovariance matrix
γ_{ij}	Experimental coefficients (Chap. 4)
$\gamma_{XY}(\tau)$	Cross-covariance
δ	Bias
Δ	Difference
ε	Error
$\varepsilon_{t+\tau\|t}$	Prediction error
θ	Regressive coefficients; time delay (Chap. 6)
κ	Basis function
μ	Mean
ν	Degrees of freedom
ρ	Autocovariance
$\rho_{X\|Z}(\tau)$	Partial autocorrelation
$\rho_{X1,\,X2}$	Correlation
$\rho_{XY\|Z}$	Partial correlation
Σ	Covariance matrix
σ	Standard deviation

σ^2	Variance
σ_{MAD}	Median absolute deviation
τ_s	Sampling time
ϕ	Moving-average coefficients
$\chi^2(\nu)$, $\chi^2{}_\nu$	χ^2-distribution
ψ	Sensitivity function
Ω	Probability space
◌̅ (U+0305)	Mean value
◌̂(U+0302)	Estimated value
◌⃗ (U+20D7)	Vector
◌̃(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

List of Figures

List of Tables

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 spurned 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 Neolatinate 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

Y. A. W. Shardt, *Statistics for Chemical and Process Engineers*,
https://doi.org/10.1007/978-3-030-83190-5_1

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 χ^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