

# Data Analysis



Determination of mean foot length  
Woodcut from Jacob Köbel's "Geometrei" published 1575 in Frankfurt

Siegmund Brandt

# Data Analysis

Statistical and Computational Methods  
for Scientists and Engineers

Fourth Edition

Translated by Glen Cowan

Siegmund Brandt  
Department of Physics  
University of Siegen  
Siegen, Germany

Additional material to this book can be downloaded from <http://extras.springer.com>

ISBN 978-3-319-03761-5                      ISBN 978-3-319-03762-2 (eBook)  
DOI 10.1007/978-3-319-03762-2  
Springer Cham Heidelberg New York Dordrecht London

Library of Congress Control Number: 2013957143

© Springer International Publishing Switzerland 2014

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. Exempted from this legal reservation are brief excerpts in connection with reviews or scholarly analysis or material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work. Duplication of this publication or parts thereof is permitted only under the provisions of the Copyright Law of the Publisher's location, in its current version, and permission for use must always be obtained from Springer. Permissions for use may be obtained through RightsLink at the Copyright Clearance Center. Violations are liable to prosecution under the respective Copyright Law.

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.

While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Printed on acid-free paper

Springer is part of Springer Science+Business Media ([www.springer.com](http://www.springer.com))

# Preface to the Fourth English Edition

For the present edition, the book has undergone two major changes: Its appearance was tightened significantly and the programs are now written in the modern programming language Java.

Tightening was possible without giving up essential contents by expedient use of the Internet. Since practically all users can connect to the net, it is no longer necessary to reproduce program listings in the printed text. In this way, the physical size of the book was reduced considerably.

The Java language offers a number of advantages over the older programming languages used in earlier editions. It is object-oriented and hence also more readable. It includes access to libraries of user-friendly auxiliary routines, allowing for instance the easy creation of windows for input, output, or graphics. For most popular computers, Java is either preinstalled or can be downloaded from the Internet free of charge. (See Sect. 1.3 for details.) Since by now Java is often taught at school, many students are already somewhat familiar with the language.

Our Java programs for data analysis and for the production of graphics, including many example programs and solutions to programming problems, can be downloaded from the page [extras.springer.com](http://extras.springer.com).

I am grateful to Dr. Tilo Stroh for numerous stimulating discussions and technical help. The graphics programs are based on previous common work.

Siegen, Germany

Siegmund Brandt



# Contents

<b>Preface to the Fourth English Edition</b>	<b>v</b>
<b>List of Examples</b>	<b>xv</b>
<b>Frequently Used Symbols and Notation</b>	<b>xix</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Typical Problems of Data Analysis . . . . .	1
1.2 On the Structure of this Book . . . . .	2
1.3 About the Computer Programs . . . . .	5
<b>2 Probabilities</b>	<b>7</b>
2.1 Experiments, Events, Sample Space . . . . .	7
2.2 The Concept of Probability . . . . .	8
2.3 Rules of Probability Calculus: Conditional Probability . . . .	10
2.4 Examples . . . . .	11
2.4.1 Probability for $n$ Dots in the Throwing of Two Dice . . . . .	11
2.4.2 Lottery 6 Out of 49 . . . . .	12
2.4.3 Three-Door Game . . . . .	13
<b>3 Random Variables: Distributions</b>	<b>15</b>
3.1 Random Variables . . . . .	15
3.2 Distributions of a Single Random Variable . . . . .	15
3.3 Functions of a Single Random Variable, Expectation Value, Variance, Moments . . . . .	17
3.4 Distribution Function and Probability Density of Two Variables: Conditional Probability . . . . .	25
3.5 Expectation Values, Variance, Covariance, and Correlation . .	27

3.6	More than Two Variables: Vector and Matrix Notation . . . . .	30
3.7	Transformation of Variables . . . . .	33
3.8	Linear and Orthogonal Transformations: Error Propagation . . . . .	36
<b>4</b>	<b>Computer Generated Random Numbers: The Monte Carlo Method</b>	<b>41</b>
4.1	Random Numbers . . . . .	41
4.2	Representation of Numbers in a Computer . . . . .	42
4.3	Linear Congruential Generators . . . . .	44
4.4	Multiplicative Linear Congruential Generators . . . . .	45
4.5	Quality of an MLCG: Spectral Test . . . . .	47
4.6	Implementation and Portability of an MLCG . . . . .	50
4.7	Combination of Several MLCGs . . . . .	52
4.8	Generation of Arbitrarily Distributed Random Numbers . . . . .	55
4.8.1	Generation by Transformation of the Uniform Distribution . . . . .	55
4.8.2	Generation with the von Neumann Acceptance–Re- jection Technique . . . . .	58
4.9	Generation of Normally Distributed Random Numbers . . . . .	62
4.10	Generation of Random Numbers According to a Multivariate Normal Distribution . . . . .	63
4.11	The Monte Carlo Method for Integration . . . . .	64
4.12	The Monte Carlo Method for Simulation . . . . .	66
4.13	Java Classes and Example Programs . . . . .	67
<b>5</b>	<b>Some Important Distributions and Theorems</b>	<b>69</b>
5.1	The Binomial and Multinomial Distributions . . . . .	69
5.2	Frequency: The Law of Large Numbers . . . . .	72
5.3	The Hypergeometric Distribution . . . . .	74
5.4	The Poisson Distribution . . . . .	78
5.5	The Characteristic Function of a Distribution . . . . .	81
5.6	The Standard Normal Distribution . . . . .	84
5.7	The Normal or Gaussian Distribution . . . . .	86
5.8	Quantitative Properties of the Normal Distribution . . . . .	88
5.9	The Central Limit Theorem . . . . .	90
5.10	The Multivariate Normal Distribution . . . . .	94
5.11	Convolutions of Distributions . . . . .	100
5.11.1	Folding Integrals . . . . .	100
5.11.2	Convolutions with the Normal Distribution . . . . .	103
5.12	Example Programs . . . . .	106

<b>6</b>	<b>Samples</b>	<b>109</b>
6.1	Random Samples. Distribution of a Sample. Estimators . . . . .	109
6.2	Samples from Continuous Populations: Mean and Variance of a Sample . . . . .	111
6.3	Graphical Representation of Samples: Histograms and Scatter Plots . . . . .	115
6.4	Samples from Partitioned Populations . . . . .	122
6.5	Samples Without Replacement from Finite Discrete Populations. Mean Square Deviation. Degrees of Freedom . . . . .	127
6.6	Samples from Gaussian Distributions: $\chi^2$ -Distribution . . . . .	130
6.7	$\chi^2$ and Empirical Variance . . . . .	135
6.8	Sampling by Counting: Small Samples . . . . .	136
6.9	Small Samples with Background . . . . .	142
6.10	Determining a Ratio of Small Numbers of Events . . . . .	144
6.11	Ratio of Small Numbers of Events with Background . . . . .	147
6.12	Java Classes and Example Programs . . . . .	149
<b>7</b>	<b>The Method of Maximum Likelihood</b>	<b>153</b>
7.1	Likelihood Ratio: Likelihood Function . . . . .	153
7.2	The Method of Maximum Likelihood . . . . .	155
7.3	Information Inequality. Minimum Variance Estimators. Sufficient Estimators . . . . .	157
7.4	Asymptotic Properties of the Likelihood Function and Maximum-Likelihood Estimators . . . . .	164
7.5	Simultaneous Estimation of Several Parameters: Confidence Intervals . . . . .	167
7.6	Example Programs . . . . .	173
<b>8</b>	<b>Testing Statistical Hypotheses</b>	<b>175</b>
8.1	Introduction . . . . .	175
8.2	$F$ -Test on Equality of Variances . . . . .	177
8.3	Student's Test: Comparison of Means . . . . .	180
8.4	Concepts of the General Theory of Tests . . . . .	185
8.5	The Neyman–Pearson Lemma and Applications . . . . .	191
8.6	The Likelihood-Ratio Method . . . . .	194
8.7	The $\chi^2$ -Test for Goodness-of-Fit . . . . .	199
8.7.1	$\chi^2$ -Test with Maximal Number of Degrees of Freedom . . . . .	199
8.7.2	$\chi^2$ -Test with Reduced Number of Degrees of Freedom . . . . .	200
8.7.3	$\chi^2$ -Test and Empirical Frequency Distribution . . . . .	200

8.8	Contingency Tables . . . . .	203
8.9	$2 \times 2$ Table Test . . . . .	204
8.10	Example Programs . . . . .	205
<b>9</b>	<b>The Method of Least Squares</b>	<b>209</b>
9.1	Direct Measurements of Equal or Unequal Accuracy . . . . .	209
9.2	Indirect Measurements: Linear Case . . . . .	214
9.3	Fitting a Straight Line . . . . .	218
9.4	Algorithms for Fitting Linear Functions of the Unknowns . . . . .	222
9.4.1	Fitting a Polynomial . . . . .	222
9.4.2	Fit of an Arbitrary Linear Function . . . . .	224
9.5	Indirect Measurements: Nonlinear Case . . . . .	226
9.6	Algorithms for Fitting Nonlinear Functions . . . . .	228
9.6.1	Iteration with Step-Size Reduction . . . . .	229
9.6.2	Marquardt Iteration . . . . .	234
9.7	Properties of the Least-Squares Solution: $\chi^2$ -Test . . . . .	236
9.8	Confidence Regions and Asymmetric Errors in the Nonlinear Case . . . . .	240
9.9	Constrained Measurements . . . . .	243
9.9.1	The Method of Elements . . . . .	244
9.9.2	The Method of Lagrange Multipliers . . . . .	247
9.10	The General Case of Least-Squares Fitting . . . . .	251
9.11	Algorithm for the General Case of Least Squares . . . . .	255
9.12	Applying the Algorithm for the General Case to Constrained Measurements . . . . .	258
9.13	Confidence Region and Asymmetric Errors in the General Case . . . . .	260
9.14	Java Classes and Example Programs . . . . .	261
<b>10</b>	<b>Function Minimization</b>	<b>267</b>
10.1	Overview: Numerical Accuracy . . . . .	267
10.2	Parabola Through Three Points . . . . .	273
10.3	Function of $n$ Variables on a Line in an $n$ -Dimensional Space . . . . .	275
10.4	Bracketing the Minimum . . . . .	275
10.5	Minimum Search with the Golden Section . . . . .	277
10.6	Minimum Search with Quadratic Interpolation . . . . .	280
10.7	Minimization Along a Direction in $n$ Dimensions . . . . .	280
10.8	Simplex Minimization in $n$ Dimensions . . . . .	281
10.9	Minimization Along the Coordinate Directions . . . . .	284
10.10	Conjugate Directions . . . . .	285
10.11	Minimization Along Chosen Directions . . . . .	287

10.12	Minimization in the Direction of Steepest Descent . . . . .	288
10.13	Minimization Along Conjugate Gradient Directions . . . . .	288
10.14	Minimization with the Quadratic Form . . . . .	292
10.15	Marquardt Minimization . . . . .	292
10.16	On Choosing a Minimization Method . . . . .	295
10.17	Consideration of Errors . . . . .	296
10.18	Examples . . . . .	298
10.19	Java Classes and Example Programs . . . . .	303
<b>11</b>	<b>Analysis of Variance</b>	<b>307</b>
11.1	One-Way Analysis of Variance . . . . .	307
11.2	Two-Way Analysis of Variance . . . . .	311
11.3	Java Class and Example Programs . . . . .	319
<b>12</b>	<b>Linear and Polynomial Regression</b>	<b>321</b>
12.1	Orthogonal Polynomials . . . . .	321
12.2	Regression Curve: Confidence Interval . . . . .	325
12.3	Regression with Unknown Errors . . . . .	326
12.4	Java Class and Example Programs . . . . .	329
<b>13</b>	<b>Time Series Analysis</b>	<b>331</b>
13.1	Time Series: Trend . . . . .	331
13.2	Moving Averages . . . . .	332
13.3	Edge Effects . . . . .	336
13.4	Confidence Intervals . . . . .	336
13.5	Java Class and Example Programs . . . . .	340
	<b>Literature</b>	<b>341</b>
<b>A</b>	<b>Matrix Calculations</b>	<b>347</b>
A.1	Definitions: Simple Operations . . . . .	348
A.2	Vector Space, Subspace, Rank of a Matrix . . . . .	351
A.3	Orthogonal Transformations . . . . .	353
	A.3.1 Givens Transformation . . . . .	354
	A.3.2 Householder Transformation . . . . .	356
	A.3.3 Sign Inversion . . . . .	359
	A.3.4 Permutation Transformation . . . . .	359
A.4	Determinants . . . . .	360
A.5	Matrix Equations: Least Squares . . . . .	362
A.6	Inverse Matrix . . . . .	365
A.7	Gaussian Elimination . . . . .	367
A.8	LR-Decomposition . . . . .	369
A.9	Cholesky Decomposition . . . . .	372

A.10	Pseudo-inverse Matrix . . . . .	375
A.11	Eigenvalues and Eigenvectors . . . . .	376
A.12	Singular Value Decomposition . . . . .	379
A.13	Singular Value Analysis . . . . .	380
A.14	Algorithm for Singular Value Decomposition . . . . .	385
A.14.1	Strategy . . . . .	385
A.14.2	Bidiagonalization . . . . .	386
A.14.3	Diagonalization . . . . .	388
A.14.4	Ordering of the Singular Values and Permutation . . . . .	392
A.14.5	Singular Value Analysis . . . . .	392
A.15	Least Squares with Weights . . . . .	392
A.16	Least Squares with Change of Scale . . . . .	393
A.17	Modification of Least Squares According to Marquardt . . . . .	394
A.18	Least Squares with Constraints . . . . .	396
A.19	Java Classes and Example Programs . . . . .	399
<b>B</b>	<b>Combinatorics</b>	<b>405</b>
<b>C</b>	<b>Formulas and Methods for the Computation of Statistical Functions</b>	<b>409</b>
C.1	Binomial Distribution . . . . .	409
C.2	Hypergeometric Distribution . . . . .	409
C.3	Poisson Distribution . . . . .	410
C.4	Normal Distribution . . . . .	410
C.5	$\chi^2$ -Distribution . . . . .	412
C.6	$F$ -Distribution . . . . .	413
C.7	$t$ -Distribution . . . . .	413
C.8	Java Class and Example Program . . . . .	414
<b>D</b>	<b>The Gamma Function and Related Functions: Methods and Programs for Their Computation</b>	<b>415</b>
D.1	The Euler Gamma Function . . . . .	415
D.2	Factorial and Binomial Coefficients . . . . .	418
D.3	Beta Function . . . . .	418
D.4	Computing Continued Fractions . . . . .	418
D.5	Incomplete Gamma Function . . . . .	420

D.6	Incomplete Beta Function . . . . .	420
D.7	Java Class and Example Program . . . . .	422
<b>E</b>	<b>Utility Programs</b>	<b>425</b>
E.1	Numerical Differentiation . . . . .	425
E.2	Numerical Determination of Zeros . . . . .	427
E.3	Interactive Input and Output Under Java . . . . .	427
E.4	Java Classes . . . . .	428
<b>F</b>	<b>The Graphics Class <code>DatanGraphics</code></b>	<b>431</b>
F.1	Introductory Remarks . . . . .	431
F.2	Graphical Workstations: Control Routines . . . . .	431
F.3	Coordinate Systems, Transformations and Transformation Methods . . . . .	432
F.3.1	Coordinate Systems . . . . .	432
F.3.2	Linear Transformations: Window – Viewport . . . . .	433
F.4	Transformation Methods . . . . .	435
F.5	Drawing Methods . . . . .	436
F.6	Utility Methods . . . . .	439
F.7	Text Within the Plot . . . . .	441
F.8	Java Classes and Example Programs . . . . .	441
<b>G</b>	<b>Problems, Hints and Solutions, and Programming Problems</b>	<b>447</b>
G.1	Problems . . . . .	447
G.2	Hints and Solutions . . . . .	456
G.3	Programming Problems . . . . .	470
<b>H</b>	<b>Collection of Formulas</b>	<b>487</b>
<b>I</b>	<b>Statistical Tables</b>	<b>503</b>
	<b>List of Computer Programs</b>	<b>515</b>
	<b>Index</b>	<b>517</b>



# List of Examples

2.1	Sample space for continuous variables . . . . .	7
2.2	Sample space for discrete variables . . . . .	8
3.1	Discrete random variable . . . . .	15
3.2	Continuous random variable . . . . .	15
3.3	Uniform distribution . . . . .	22
3.4	Cauchy distribution . . . . .	23
3.5	Lorentz (Breit–Wigner) distribution . . . . .	25
3.6	Error propagation and covariance . . . . .	38
4.1	Exponentially distributed random numbers . . . . .	57
4.2	Generation of random numbers following a Breit–Wigner distribution . . . . .	57
4.3	Generation of random numbers with a triangular distribution . . .	58
4.4	Semicircle distribution with the simple acceptance–rejection method . . . . .	59
4.5	Semicircle distribution with the general acceptance–rejection method . . . . .	61
4.6	Computation of $\pi$ . . . . .	65
4.7	Simulation of measurement errors of points on a line . . . . .	66
4.8	Generation of decay times for a mixture of two different radioactive substances . . . . .	66
5.1	Statistical error . . . . .	74
5.2	Application of the hypergeometric distribution for determination of zoological populations . . . . .	77
5.3	Poisson distribution and independence of radioactive decays . . .	80
5.4	Poisson distribution and the independence of scientific discoveries . . . . .	81
5.5	Addition of two Poisson distributed variables with use of the characteristic function . . . . .	84

5.6	Normal distribution as the limiting case of the binomial distribution . . . . .	92
5.7	Error model of Laplace . . . . .	92
5.8	Convolution of uniform distributions . . . . .	102
5.9	Convolution of uniform and normal distributions . . . . .	104
5.10	Convolution of two normal distributions. “Quadratic addition of errors” . . . . .	104
5.11	Convolution of exponential and normal distributions . . . . .	105
6.1	Computation of the sample mean and variance from data . . . . .	114
6.2	Histograms of the same sample with various choices of bin width . . . . .	117
6.3	Full width at half maximum (FWHM) . . . . .	119
6.4	Investigation of characteristic quantities of samples from a Gaussian distribution with the Monte Carlo method . . . . .	119
6.5	Two-dimensional scatter plot: Dividend versus price for industrial stocks . . . . .	120
6.6	Optimal choice of the sample size for subpopulations . . . . .	125
6.7	Determination of a lower limit for the lifetime of the proton from the observation of no decays . . . . .	142
7.1	Likelihood ratio . . . . .	154
7.2	Repeated measurements of differing accuracy . . . . .	156
7.3	Estimation of the parameter $N$ of the hypergeometric distribution . . . . .	157
7.4	Estimator for the parameter of the Poisson distribution . . . . .	162
7.5	Estimator for the parameter of the binomial distribution . . . . .	163
7.6	Law of error combination (“Quadratic averaging of individual errors”) . . . . .	163
7.7	Determination of the mean lifetime from a small number of decays . . . . .	166
7.8	Estimation of the mean and variance of a normal distribution . . . . .	171
7.9	Estimators for the parameters of a two-dimensional normal distribution . . . . .	172
8.1	$F$ -test of the hypothesis of equal variance of two series of measurements . . . . .	180
8.2	Student’s test of the hypothesis of equal means of two series of measurements . . . . .	185
8.3	Test of the hypothesis that a normal distribution with given variance $\sigma^2$ has the mean $\lambda = \lambda_0$ . . . . .	189
8.4	Most powerful test for the problem of Example 8.3 . . . . .	193
8.5	Power function for the test from Example 8.3 . . . . .	195
8.6	Test of the hypothesis that a normal distribution of unknown variance has the mean value $\lambda = \lambda_0$ . . . . .	197

8.7	$\chi^2$ -test for the fit of a Poisson distribution to an empirical frequency distribution . . . . .	202
9.1	Weighted mean of measurements of different accuracy . . . . .	212
9.2	Fitting of various polynomials . . . . .	223
9.3	Fitting a proportional relation . . . . .	224
9.4	Fitting a Gaussian curve . . . . .	231
9.5	Fit of an exponential function . . . . .	232
9.6	Fitting a sum of exponential functions . . . . .	233
9.7	Fitting a sum of two Gaussian functions and a polynomial . . . . .	235
9.8	The influence of large measurement errors on the confidence region of the parameters for fitting an exponential function . . . . .	241
9.9	Constraint between the angles of a triangle . . . . .	245
9.10	Application of the method of Lagrange multipliers to Example 9.9 . . . . .	249
9.11	Fitting a line to points with measurement errors in both the abscissa and ordinate . . . . .	257
9.12	Fixing parameters . . . . .	257
9.13	$\chi^2$ -test of the description of measured points with errors in abscissa and ordinate by a given line . . . . .	259
9.14	Asymmetric errors and confidence region for fitting a straight line to measured points with errors in the abscissa and ordinate . . . . .	260
10.1	Determining the parameters of a distribution from the elements of a sample with the method of maximum likelihood . . . . .	298
10.2	Determination of the parameters of a distribution from the histogram of a sample by maximizing the likelihood . . . . .	299
10.3	Determination of the parameters of a distribution from the histogram of a sample by minimization of a sum of squares . . . . .	302
11.1	One-way analysis of variance of the influence of various drugs . . . . .	310
11.2	Two-way analysis of variance in cancer research . . . . .	318
12.1	Treatment of Example 9.2 with Orthogonal Polynomials . . . . .	325
12.2	Confidence limits for linear regression . . . . .	327
13.1	Moving average with linear trend . . . . .	335
13.2	Time series analysis of the same set of measurements using different averaging intervals and polynomials of different orders . . . . .	338
A.1	Inversion of a $3 \times 3$ matrix . . . . .	369
A.2	Almost vanishing singular values . . . . .	381
A.3	Point of intersection of two almost parallel lines . . . . .	381
A.4	Numerical superiority of the singular value decomposition compared to the solution of normal equations . . . . .	384
A.5	Least squares with constraints . . . . .	398



# Frequently Used Symbols and Notation

$x, y, \xi, \eta, \dots$	(ordinary) variable	$P(A)$	probability of the event $A$
$\mathbf{x}, \mathbf{y}, \boldsymbol{\xi}, \boldsymbol{\eta}, \dots$	vector variable		
$\mathbf{x}, \mathbf{y}, \dots$	random variable	$Q$	sum of squares
$\mathbf{x}, \mathbf{y}, \dots$	vector of random variables	$\mathbf{s}^2, \mathbf{s}_x^2$	sample variance
$A, B, C, \dots$	matrices	$\mathbf{S}$	estimator
$B$	bias	$S_c$	critical region
$\text{cov}(\mathbf{x}, \mathbf{y})$	covariance	$t$	variable of Student's distribution
$F$	variance ratio	$T$	Testfunktion
$f(x)$	probability density	$x_m$	most probable value (mode)
$F(x)$	distribution function	$x_{0,5}$	median
$E(\mathbf{x}) = \hat{x}$	mean value, expectation value	$x_q$	quantile
$H$	hypothesis	$\bar{\mathbf{x}}$	sample mean
$H_0$	null hypothesis	$\tilde{\mathbf{x}}$	estimator from maximum likelihood or least squares
$L, \ell$	likelihood functions	$\alpha$	level of significance
$L(S_c, \lambda)$	operating characteristic function	$1 - \alpha$	level of confidence
$M(S_c, \lambda)$	power (of a test)	$\lambda$	parameter of a distribution
$M$	minimum function, target function	$\varphi(t)$	characteristic function

$\phi(x), \psi(x)$  probability density and distribution function of the normal distribution

$\phi_0(x), \psi_0(x)$  probability density and distribution function of the standard normal distribution

$\sigma(\mathbf{x}) = \Delta(\mathbf{x})$  standard deviation

$\sigma^2(\mathbf{x})$  variance

$\chi^2$  variable of the  $\chi^2$  distribution

$\Omega(P)$  inverse function of the normal distribution