

Siegmund Brandt



# Data Analysis

Statistical and Computational Methods  
for Scientists and Engineers

*Fourth Edition*

EXTRAS ONLINE

 Springer

## 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>  | v   |
| <b>List of Examples</b>   | xv  |
| <b>Frequently Used Symbols and Notation</b>   | xix |
| <b>1 Introduction</b>   | 1   |
| 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>  | 7   |
| 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>  | 15  |
| 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<br>Propagation . . . . .                     | 36        |
| <b>4</b> | <b>Computer Generated Random Numbers: The Monte<br/>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<br>Distribution . . . . .                     | 55        |
| 4.8.2    | Generation with the von Neumann Acceptance–Re-<br>jection Technique . . . . .             | 58        |
| 4.9      | Generation of Normally Distributed Random Numbers . . . . .                               | 62        |
| 4.10     | Generation of Random Numbers According<br>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 Samples</b>  | <b>109</b> |
| 6.1 Random Samples. Distribution<br>of a Sample. Estimators . . . . .   | 109        |
| 6.2 Samples from Continuous Populations: Mean<br>and Variance of a Sample . . . . .   | 111        |
| 6.3 Graphical Representation of Samples: Histograms<br>and Scatter Plots . . . . .  | 115        |
| 6.4 Samples from Partitioned Populations . . . . .  | 122        |
| 6.5 Samples Without Replacement from Finite Discrete<br>Populations. Mean Square Deviation. Degrees of<br>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 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<br>Estimators. Sufficient Estimators . . . . .                                 | 157        |
| 7.4 Asymptotic Properties of the Likelihood Function<br>and Maximum-Likelihood Estimators . . . . .                         | 164        |
| 7.5 Simultaneous Estimation of Several Parameters:<br>Confidence Intervals . . . . .  | 167        |
| 7.6 Example Programs . . . . .  | 173        |
| <b>8 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<br>of Freedom . . . . .   | 199        |
| 8.7.2 $\chi^2$ -Test with Reduced Number of Degrees<br>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<br>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<br>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<br>to Constrained Measurements . . . . . | 258        |
| 9.13      | Confidence Region and Asymmetric Errors<br>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<br>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 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 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 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 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 DatanGraphics</b>                                     | <b>431</b> |
| F.1                              | Introductory Remarks . . . . .  | 431        |
| F.2                              | Graphical Workstations: Control Routines . . . . .                          | 431        |
| F.3                              | Coordinate Systems, Transformations<br>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 $A$                       |
| $\mathbf{x}, \mathbf{y}, \xi, \eta, \dots$ | vector variable                   | $Q$          | sum of squares                                     |
| $\mathbf{x}, \mathbf{y}, \dots$            | random variable                   | $s^2, s_x^2$ | sample variance                                    |
| $\mathbf{x}, \mathbf{y}, \dots$            | vector of random variables        | $S$          | estimator  |
| $A, B, C, \dots$                           | matrices                          | $S_c$        | critical region                                    |
| $B$  | bias                              | $t$          | variable of Student's distribution                 |
| $\text{cov}(\mathbf{x}, \mathbf{y})$       | covariance                        | $T$          | Testfunktion                                       |
| $F$  | variance ratio                    | $x_m$        | most probable value (mode)                         |
| $f(x)$                                     | probability density               | $x_{0.5}$    | median   |
| $F(x)$                                     | distribution function             | $x_q$        | quantile   |
| $E(\mathbf{x}) = \hat{x}$                  | mean value, expectation value     | $\bar{x}$    | sample mean  |
| $H$  | hypothesis                        | $\tilde{x}$  | estimator from maximum likelihood or least squares |
| $H_0$                                      | null hypothesis                   |              |  |
| $L, \ell$                                  | 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          | $\sigma(\mathbf{x}) = \Delta(\mathbf{x})$ | standard deviation                          |
| $\phi_0(x), \psi_0(x)$ | probability density and distribution function of the standard normal distribution | $\sigma^2(\mathbf{x})$                    | variance                                    |
|                        |   | $\chi^2$                                  | variable of the $\chi^2$ distribution       |
|                        |   | $\Omega(P)$                               | inverse function of the normal distribution |

# 1. Introduction

## 1.1 Typical Problems of Data Analysis

Every branch of experimental science, after passing through an early stage of qualitative description, concerns itself with quantitative studies of the phenomena of interest, i.e., *measurements*. In addition to designing and carrying out the experiment, an important task is the accurate evaluation and *complete* exploitation of the data obtained. Let us list a few typical problems.

1. A study is made of the weight of laboratory animals under the influence of various drugs. After the application of drug *A* to 25 animals, an average increase of 5 % is observed. Drug *B*, used on 10 animals, yields a 3 % increase. Is drug *A* more effective? The *averages* 5 and 3 % give practically no answer to this question, since the lower value may have been caused by a single animal that lost weight for some unrelated reason. One must therefore study the *distribution* of individual weights and their spread around the average value. Moreover, one has to decide whether the number of test animals used will enable one to differentiate with a certain accuracy between the effects of the two drugs.
2. In experiments on crystal growth it is essential to maintain exactly the ratios of the different components. From a total of 500 crystals, a *sample* of 20 is selected and analyzed. What conclusions can be drawn about the composition of the remaining 480? This problem of sampling comes up, for example, in quality control, reliability tests of automatic measuring devices, and opinion polls.
3. A certain experimental result has been obtained. It must be decided whether it is in contradiction with some predicted theoretical value or with previous experiments. The experiment is used for *hypothesis testing*.

4. A general law is known to describe the dependence of measured variables, but parameters of this law must be obtained from experiment. In radioactive decay, for example, the number  $N$  of atoms that decay per second decreases exponentially with time:  $N(t) = \text{const} \cdot \exp(-\lambda t)$ . One wishes to determine the decay constant  $\lambda$  and its measurement error by making maximal use of a series of measured values  $N_1(t_1)$ ,  $N_2(t_2)$ , .... One is concerned here with the problem of *fitting* a function containing unknown parameters to the data and the determination of the numerical values of the parameters and their errors.

From these examples some of the aspects of data analysis become apparent. We see in particular that the outcome of an experiment is not uniquely determined by the experimental procedure but is also subject to chance: it is a *random variable*. This stochastic tendency is either rooted in the nature of the experiment (test animals are necessarily different, radioactivity is a stochastic phenomenon), or it is a consequence of the inevitable uncertainties of the experimental equipment, i.e., measurement errors. It is often useful to simulate with a computer the variable or stochastic characteristics of the experiment in order to get an idea of the expected uncertainties of the results before carrying out the experiment itself. This *simulation* of random quantities on a computer is called the *Monte Carlo method*, so named in reference to games of chance.

## 1.2 On the Structure of this Book

The basis for using random quantities is the *calculus of probabilities*. The most important concepts and rules for this are collected in Chap. 2. *Random variables* are introduced in Chap. 3. Here one considers distributions of random variables, and parameters are defined to characterize the distributions, such as the expectation value and variance. Special attention is given to the interdependence of several random variables. In addition, *transformations* between different sets of variables are considered; this forms the basis of *error propagation*.

*Generating random numbers on a computer* and the *Monte Carlo method* are the topics of Chap. 4. In addition to methods for generating random numbers, a well-tested program and also examples for generating arbitrarily distributed random numbers are given. Use of the Monte Carlo method for problems of integration and simulation is introduced by means of examples. The method is also used to generate simulated data with measurement errors, with which the data analysis routines of later chapters can be demonstrated.

In Chap. 5 we introduce a number of distributions which are of particular interest in applications. This applies especially to the Gaussian or normal distribution, whose properties are studied in detail.

In practice a distribution must be determined from a finite number of observations, i.e., from a *sample*. Various cases of sampling are considered in Chap. 6. Computer programs are presented for a first rough numerical treatment and graphical display of empirical data. Functions of the sample, i.e., of the individual observations, can be used to estimate the parameters characterizing the distribution. The requirements that a good estimate should satisfy are derived. At this stage the quantity  $\chi^2$  is introduced. This is the sum of the squares of the deviations between observed and expected values and is therefore a suitable indicator of the goodness-of-fit.

The *maximum-likelihood method*, discussed in Chap. 7, forms the core of modern statistical analysis. It allows one to construct estimators with optimum properties. The method is discussed for the single and multiparameter cases and illustrated in a number of examples. Chapter 8 is devoted to *hypothesis testing*. It contains the most commonly used  $F$ ,  $t$ , and  $\chi^2$  tests and in addition outlines the general points of test theory.

The *method of least squares*, which is perhaps the most widely used statistical procedure, is the subject of Chap. 9. The special cases of direct, indirect, and constrained measurements, often encountered in applications, are developed in detail before the general case is discussed. Programs and examples are given for all cases. Every least-squares problem, and in general every problem of maximum likelihood, involves determining the minimum of a function of several variables. In Chap. 10 various methods are discussed in detail, by which such a minimization can be carried out. The relative efficiency of the procedures is shown by means of programs and examples.

The analysis of variance (Chap. 11) can be considered as an extension of the  $F$ -test. It is widely used in biological and medical research to study the dependence, or rather to test the independence, of a measured quantity from various experimental conditions expressed by other variables. For several variables rather complex situations can arise. Some simple numerical examples are calculated using a computer program.

Linear and polynomial *regression*, the subject of Chap. 12, is a special case of the least-squares method and has therefore already been treated in Chap. 9. Before the advent of computers, usually only linear least-squares problems were tractable. A special terminology, still used, was developed for this case. It seemed therefore justified to devote a special chapter to this subject. At the same time it extends the treatment of Chap. 9. For example the determination of confidence intervals for a solution and the relation between regression and analysis of variance are studied. A general program for polynomial regression is given and its use is shown in examples.

In the last chapter the elements of time series analysis are introduced. This method is used if data are given as a function of a controlled variable (usually time) and no theoretical prediction for the behavior of the data as a function of the controlled variable is known. It is used to try to reduce the statistical fluctuation of the data without destroying the genuine dependence on the controlled variable. Since the computational work in time series analysis is rather involved, a computer program is also given.

The field of data analysis, which forms the main part of this book, can be called *applied mathematical statistics*. In addition, wide use is made of other branches of mathematics and of specialized computer techniques. This material is contained in the appendices.

In Appendix A, titled “Matrix Calculations”, the most important concepts and methods from *linear algebra* are summarized. Of central importance are procedures for solving systems of linear equations, in particular the singular value decomposition, which provides the best numerical properties.

Necessary concepts and relations of *combinatorics* are compiled in Appendix B. The numerical value of functions of mathematical statistics must often be computed. The necessary formulas and algorithms are contained in Appendix C. Many of these functions are related to the Euler *gamma function* and like it can only be computed with approximation techniques. In Appendix D formulas and methods for gamma and related functions are given. Appendix E describes further methods for numerical differentiation, for the determination of zeros, and for interactive input and output under Java.

The *graphical representation* of measured data and their errors and in many cases also of a fitted function is of special importance in data analysis. In Appendix F a Java class with a comprehensive set of graphical methods is presented. The most important concepts of computer graphics are introduced and all of the necessary explanations for using this class are given.

Appendix G.1 contains *problems* to most chapters. These problems can be solved with paper and pencil. They should help the reader to understand the basic concepts and theorems. In some cases also simple numerical calculations must be carried out. In Appendix G.2 either the solution of problems is sketched or the result is simply given. In Appendix G.3 a number of *programming problems* is presented. For each one an example solution is given.

The set of appendices is concluded with a *collection of formulas* in Appendix H, which should facilitate reference to the most important equations, and with a short collection of statistical tables in Appendix I. Although all of the tabulated values can be computed (and in fact were computed) with the programs of Appendix C, it is easier to look up one or two values from the tables than to use a computer.

## 1.3 About the Computer Programs

For the present edition all programs were newly written in the programming language Java. Since some time Java is taught in many schools so that young readers often are already familiar with that language. Java classes are directly executable on all popular computers – independently of the operating system. The compilation of Java source programs takes place using the Java Development Kit, which for many operating systems, in particular Windows, Linux, and Mac OSX, can be downloaded free of cost from the Internet, <http://www.oracle.com/technetwork/java/index.html>

There are four groups of computer programs discussed in this book. These are

- The data analysis library in the form of the package `datan`,
- The graphics library in the form of the package `datagraphics`,
- A collection of example programs in the package `examples`,
- Solutions to the programming problems in the package `solutions`.

The programs of all groups are available both as compiled classes and (except for `datagraphics`. `DatanGraphics`) also as source files. In addition there is the extensive Java-typical documentation in html format.

Every class and method of the package `datan` deals with a particular, well defined problem, which is extensively described in the text. That also holds for the graphics library, which allows to produce practically any type of line graphics in two dimensions. For many purposes it suffices, however, to use one of 5 classes each yielding a complete graphics.

In order to solve a specific problem the user has to write a short class in Java, which essentially consists of calling classes from the data analysis library, and which in certain cases organizes the input of the user's data and output of the results. The *example programs* are a collection of such classes. The application of each method from the data analysis and graphics libraries is demonstrated in at least one example program. Such example programs are described in a special section near the end of most chapters.

Near the end of the book there is a *List of Computer Programs* in alphabetic order. For each program from the data analysis library and from the graphics library page numbers are given, for an explanation of the program itself, and for one or several example programs demonstrating its use.

The *programming problems* like the example programs are designed to help the reader in using computer methods. Working through these problems should enable readers to formulate their own specific tasks in data analysis

to be solved on a computer. For all programming problems, programs exist which represent a possible solution.

In data analysis, of course, *data* play a special role. The type of data and the format in which they are presented to the computer cannot be defined in a general textbook since it depends very much on the particular problem at hand. In order to have somewhat realistic data for our examples and problems we have decided to produce them in most cases within the program using the Monte Carlo method. It is particularly instructive to simulate data with known properties and a given error distribution and to subsequently analyze these data. In the analysis one must in general make an assumption about the distribution of the errors. If this assumption is not correct, then the results of the analysis are not optimal. Effects that are often decisively important in practice can be “experienced” with exercises combining simulation and analysis.

Here are some short hints concerning the installation of our programs. As material accompanying this book, available from the page [extras.springer.com](http://extras.springer.com), there is a zip file named DatanJ. Download this file, unzip it while keeping the internal tree structure of subdirectories and store it on your computer in a new directory. (It is convenient to also give that directory the name DatanJ.) Further action is described in the file ReadME in that directory.

## 2. Probabilities

### 2.1 Experiments, Events, Sample Space

Since in this book we are concerned with the analysis of data originating from experiments, we will have to state first what we mean by an experiment and its result. Just as in the laboratory, we define an experiment to be a strictly followed procedure, as a consequence of which a quantity or a set of quantities is obtained that constitutes the result. These quantities are continuous (temperature, length, current) or discrete (number of particles, birthday of a person, one of three possible colors). No matter how accurately all conditions of the procedure are maintained, the results of repetitions of an experiment will in general differ. This is caused either by the intrinsic statistical nature of the phenomenon under investigation or by the finite accuracy of the measurement. The possible results will therefore always be spread over a finite region for each quantity. All of these regions for all quantities that make up the result of an experiment constitute the *sample space* of that experiment. Since it is difficult and often impossible to determine exactly the accessible regions for the quantities measured in a particular experiment, the sample space actually used may be larger and may contain the true sample space as a subspace. We shall use this somewhat looser concept of a sample space.

#### Example 2.1: Sample space for continuous variables

In the manufacture of resistors it is important to maintain the values  $R$  (electrical resistance measured in ohms) and  $N$  (maximum heat dissipation measured in watts) at given values. The sample space for  $R$  and  $N$  is a plane spanned by axes labeled  $R$  and  $N$ . Since both quantities are always positive, the first quadrant of this plane is itself a sample space. ■

**Example 2.2:** Sample space for discrete variables

In practice the exact values of  $R$  and  $N$  are unimportant as long as they are contained within a certain interval about the nominal value (e.g.,  $99 \text{ k}\Omega < R < 101 \text{ k}\Omega$ ,  $0.49 \text{ W} < N < 0.60 \text{ W}$ ). If this is the case, we shall say that the resistor has the properties  $R_n$ ,  $N_n$ . If the value falls below (above) the lower (upper) limit, then we shall substitute the index  $n$  by  $-$ ( $+$ ). The possible values of resistance and heat dissipation are therefore  $R_-$ ,  $R_n$ ,  $R_+$ ,  $N_-$ ,  $N_n$ ,  $N_+$ . The sample space now consists of nine points:

$$\begin{aligned} R_- N_-, & \quad R_- N_n, & \quad R_- N_+, \\ R_n N_-, & \quad R_n N_n, & \quad R_n N_+, \\ R_+ N_-, & \quad R_+ N_n, & \quad R_+ N_+. \end{aligned} \quad \blacksquare$$

Often one or more particular subspaces of the sample space are of special interest. In Example 2.2, for instance, the point  $R_n$ ,  $N_n$  represents the case where the resistors meet the production specifications. We can give such subspaces names, e.g.,  $A$ ,  $B$ , ... and say that if the result of an experiment falls into one such subspace, then the *event A* (or  $B$ ,  $C$ , ...) has occurred. If  $A$  has not occurred, we speak of the complementary event  $\bar{A}$  (i.e., not  $A$ ). The whole sample space corresponds to an event that will occur in every experiment, which we call  $E$ . In the rest of this chapter we shall define what we mean by the probability of the occurrence of an event and present rules for computations with probabilities.

## 2.2 The Concept of Probability

Let us consider the simplest experiment, namely, the tossing of a coin. Like the throwing of dice or certain problems with playing cards it is of no practical interest but is useful for didactic purposes. What is the probability that a “fair” coin shows “heads” when tossed once? Our intuition suggests that this probability is equal to  $1/2$ . It is based on the assumption that all points in sample space (there are only two points: “heads” and “tails”) are equally probable and on the convention that we give the event  $E$  (here: “heads” or “tails”) a probability of unity. This way of determining probabilities can be applied only to symmetric experiments and is therefore of little practical use. (It is, however, of great importance in statistical physics and quantum statistics, where the equal probabilities of all allowed states is an essential postulate of very successful theories.) If no such perfect symmetry exists—which will even be the case with normal “physical” coins—the following procedure seems reason-

able. In a large number  $N$  of experiments the event  $A$  is observed to occur  $n$  times. We define

$$P(A) = \lim_{N \rightarrow \infty} \frac{n}{N} \quad (2.2.1)$$

as the probability of the occurrence of the event  $A$ . This somewhat loose *frequency definition* of probability is sufficient for practical purposes, although it is mathematically unsatisfactory. One of the difficulties with this definition is the need for an infinity of experiments, which are of course impossible to perform and even difficult to imagine. Although we shall in fact use the frequency definition in this book, we will indicate the basic concepts of an axiomatic theory of probability due to KOLMOGOROV [1]. The minimal set of axioms generally used is the following:

- (a) To each event  $A$  there corresponds a non-negative number, its probability,

$$P(A) \geq 0 \quad . \quad (2.2.2)$$

- (b) The event  $E$  has unit probability,

$$P(E) = 1 \quad . \quad (2.2.3)$$

- (c) If  $A$  and  $B$  are *mutually exclusive* events, then the probability of  $A$  or  $B$  (written  $A + B$ ) is

$$P(A + B) = P(A) + P(B) \quad . \quad (2.2.4)$$

From these axioms\* one obtains immediately the following useful results. From (b) and (c):

$$P(\bar{A} + A) = P(A) + P(\bar{A}) = 1 \quad , \quad (2.2.5)$$

and furthermore with (a):

$$0 \leq P(A) \leq 1 \quad . \quad (2.2.6)$$

From (c) one can easily obtain the more general theorem for mutually exclusive events  $A, B, C, \dots$ ,

$$P(A + B + C + \dots) = P(A) + P(B) + P(C) + \dots \quad . \quad (2.2.7)$$

It should be noted that summing the probabilities of events combined with “or” here refers only to mutually exclusive events. If one must deal with events that are not of this type, then they must first be decomposed into mutually exclusive ones. In throwing a die,  $A$  may signify even,  $B$  odd,  $C$  less than 4 dots,  $D$  4 or more dots. Suppose one is interested in the probability for the

---

\*Sometimes the definition (2.3.1) is included as a fourth axiom.

event  $A$  or  $C$ , which are obviously not exclusive. One forms  $A$  and  $C$  (written  $AC$ ) as well as  $AD$ ,  $BC$ , and  $BD$ , which are mutually exclusive, and finds for  $A$  or  $C$  (sometimes written  $A + C$ ) the expression  $AC + AD + BC$ . Note that the axioms do not prescribe a method for assigning the value of a particular probability  $P(A)$ .

Finally it should be pointed out that the word probability is often used in common language in a sense that is different or even opposed to that considered by us. This is subjective probability, where the probability of an event is given by the measure of our belief in its occurrence. An example of this is: “The probability that the party  $A$  will win the next election is  $1/3$ .” As another example consider the case of a certain track in nuclear emulsion which could have been left by a proton or pion. One often says: “The track was caused by a pion with probability  $1/2$ .” But since the event had already taken place and only one of the two kinds of particle could have caused that particular track, the probability in question is either 0 or 1, but we do not know which.

## 2.3 Rules of Probability Calculus: Conditional Probability

Suppose the result of an experiment has the property  $A$ . We now ask for the probability that it also has the property  $B$ , i.e., the probability of  $B$  under the condition  $A$ . We define this *conditional probability* as

$$P(B|A) = \frac{P(AB)}{P(A)} . \quad (2.3.1)$$

It follows that

$$P(AB) = P(A) P(B|A) . \quad (2.3.2)$$

One can also use (2.3.2) directly for the definition, since here the requirement  $P(A) \neq 0$  is not necessary. From Fig. 2.1 it can be seen that this definition is reasonable. Consider the event  $A$  to occur if a point is in the region labeled  $A$ , and correspondingly for the event (and region)  $B$ . For the overlap region both  $A$  and  $B$  occur, i.e., the event  $(AB)$  occurs. Let the area of the different regions be proportional to the probabilities of the corresponding events. Then the probability of  $B$  under the condition  $A$  is the ratio of the area  $AB$  to that of  $A$ . In particular this is equal to unity if  $A$  is contained in  $B$  and zero if the overlapping area vanishes.

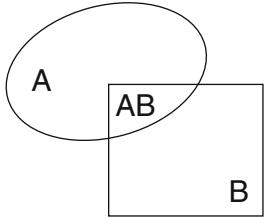
Using conditional probability we can now formulate the *rule of total probability*. Consider an experiment that can lead to one of  $n$  possible mutually exclusive events,

$$E = A_1 + A_2 + \cdots + A_n . \quad (2.3.3)$$

The probability for the occurrence of any event with the property  $B$  is

$$P(B) = \sum_{i=1}^n P(A_i)P(B|A_i) , \quad (2.3.4)$$

as can be seen easily from (2.3.2) and (2.2.7).



**Fig. 2.1:** Illustration of conditional probability.

We can now also define the *independence* of events. Two events  $A$  and  $B$  are said to be independent if the knowledge that  $A$  has occurred does not change the probability for  $B$  and vice versa, i.e., if

$$P(B|A) = P(B) , \quad (2.3.5)$$

or, by use of (2.3.2),

$$P(A|B) = P(A)P(B) . \quad (2.3.6)$$

In general several decompositions of the type (2.3.3),

$$\begin{aligned} E &= A_1 + A_2 + \cdots + A_n , \\ E &= B_1 + B_2 + \cdots + B_m , \\ &\vdots \\ E &= Z_1 + Z_2 + \cdots + Z_\ell , \end{aligned} \quad (2.3.7)$$

are said to be independent, if for all possible combinations  $\alpha, \beta, \dots, \omega$  the condition

$$P(A_\alpha B_\beta \cdots Z_\omega) = P(A_\alpha)P(B_\beta) \cdots P(Z_\omega) \quad (2.3.8)$$

is fulfilled.

## 2.4 Examples

### 2.4.1 Probability for $n$ Dots in the Throwing of Two Dice

If  $n_1$  and  $n_2$  are the number of dots on the individual dice and if  $n = n_1 + n_2$ , then one has  $P(n_i) = 1/6$ ;  $i = 1, 2$ ;  $n_i = 1, 2, \dots, 6$ . Because the two dice are independent of each other one has  $P(n_1, n_2) = P(n_1)P(n_2) = 1/36$ . By

considering in how many different ways the sum  $n = n_i + n_j$  can be formed one obtains

$$\begin{aligned}
 P_2(2) &= P(1, 1) = 1/36 , \\
 P_2(3) &= P(1, 2) + P(2, 1) = 2/36 , \\
 P_2(4) &= P(1, 3) + P(2, 2) + P(3, 1) = 3/36 , \\
 P_2(5) &= P(1, 4) + P(2, 3) + P(3, 2) + P(4, 1) = 4/36 , \\
 P_2(6) &= P(1, 5) + P(2, 4) + P(3, 3) + P(4, 2) \\
 &\quad + P(5, 1) = 5/36 , \\
 P_2(7) &= P(1, 6) + P(2, 5) + P(3, 4) + P(4, 3) \\
 &\quad + P(5, 2) + P(6, 1) = 6/36 , \\
 P_2(8) &= P_2(6) = 5/36 , \\
 P_2(9) &= P_2(5) = 4/36 , \\
 P_2(10) &= P_2(4) = 3/36 , \\
 P_2(11) &= P_2(3) = 2/36 , \\
 P_2(12) &= P_2(2) = 1/36 .
 \end{aligned}$$

Of course, the normalization condition  $\sum_{k=2}^{12} P_2(k) = 1$  is fulfilled.

#### 2.4.2 Lottery 6 Out of 49

A container holds 49 balls numbered 1 through 49. During the drawing 6 balls are taken out of the container consecutively and none are put back in. We compute the probabilities  $P(1), P(2), \dots, P(6)$  that a player, who before the drawing has chosen six of the numbers 1, 2, ..., 49, has predicted exactly 1, 2, ..., or 6 of the drawn numbers.

First we compute  $P(6)$ . The probability to choose as the first number the one which will also be drawn first is obviously  $1/49$ . If that step was successful, then the probability to choose as the second number the one which is also drawn second is  $1/48$ . We conclude that the probability for choosing six numbers correctly in the order in which they are drawn is

$$\frac{1}{49 \cdot 48 \cdot 47 \cdot 46 \cdot 45 \cdot 44} = \frac{43!}{49!} .$$

The order, however, is irrelevant. Since there are  $6!$  possible ways to arrange six numbers in different orders we have

$$P(6) = \frac{6!43!}{49!} = \frac{1}{\binom{49}{6}} = \frac{1}{C_6^{49}} .$$

That is exactly the inverse of the number of combinations  $C_6^{49}$  of 6 elements out of 49 (see Appendix B), since all of these combinations are equally probable but only one of them contains only the drawn numbers.

We may now argue that the container holds two kinds of balls, namely 6 balls in which the player is interested since they carry the numbers which he selected, and 43 balls whose numbers the player did not select. The result of the drawing is a sample from a set of 49 elements of which 6 are of one kind and 43 are of the other. The sample itself contains 6 elements which are drawn without putting elements back into the container. This method of sampling is described by the hypergeometric distribution (see Sect. 5.3). The probability for predicting correctly  $\ell$  out of the 6 drawn numbers is

$$P(\ell) = \frac{\binom{6}{\ell} \binom{43}{6-\ell}}{\binom{49}{6}}, \quad \ell = 0, \dots, 6.$$

### 2.4.3 Three-Door Game

In a TV game show a candidate is given the following problem. Three rooms are closed by three identical doors. One room contains a luxury car, the other two each contain a goat. The candidate is asked to guess behind which of the doors the car is. He chooses a door which we will call  $A$ . The door  $A$ , however, remains closed for the moment. Of course, behind at least one of the other doors there is a goat. The quiz master now opens one door which we will call  $B$  to reveal a goat. He now gives the candidate the chance to either stay with the original choice  $A$  or to choose remaining closed door  $C$ . Can the candidate increase his or her chances by choosing  $C$  instead of  $A$ ?

The answer (astonishing for many) is yes. The probability to find the car behind the door  $A$  obviously is  $P(A) = 1/3$ . Then the probability that the car is behind one of the other doors is  $P(\bar{A}) = 2/3$ . The candidate exhausts this probability fully if he chooses the door  $C$  since through the opening of  $B$  it is shown to be a door without the car, so that  $P(C) = P(\bar{A})$ .

# 3. Random Variables: Distributions

## 3.1 Random Variables

We will now consider not the probability of observing particular events but rather the events themselves and try to find a particularly simple way of classifying them. We can, for instance, associate the event “heads” with the number 0 and the event “tails” with the number 1. Generally we can classify the events of the decomposition (2.3.3) by associating each event  $A_i$  with the real number  $i$ . In this way each event can be characterized by one of the possible values of a *random variable*. Random variables can be *discrete* or *continuous*. We denote them by symbols like  $x, y, \dots$ .

### Example 3.1: Discrete random variable

It may be of interest to study the number of coins still in circulation as a function of their age. It is obviously most convenient to use the year of issue stamped on each coin directly as the (discrete) random variable, e.g.,  $x = \dots, 1949, 1950, 1951, \dots$  ■

### Example 3.2: Continuous random variable

All processes of measurement or production are subject to smaller or larger imperfections or fluctuations that lead to variations in the result, which is therefore described by one or several random variables. Thus the values of electrical resistance and maximum heat dissipation characterizing a resistor in Example 2.1 are continuous random variables. ■

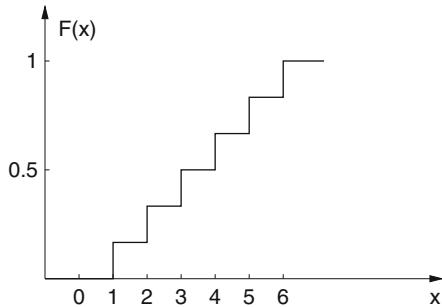
## 3.2 Distributions of a Single Random Variable

From the classification of events we return to probability considerations. We consider the random variable  $x$  and a real number  $x$ , which can assume any value between  $-\infty$  and  $+\infty$ , and study the probability for the event  $x < x$ .

This probability is a function of  $x$  and is called the (*cumulative*) *distribution function* of  $\mathbf{x}$ :

$$F(x) = P(\mathbf{x} < x) . \quad (3.2.1)$$

If  $\mathbf{x}$  can assume only a finite number of discrete values, e.g., the number of dots on the faces of a die, then the distribution function is a step function. It is shown in Fig. 3.1 for the example mentioned above. Obviously distribution functions are always monotonic and non-decreasing.



**Fig. 3.1:** Distribution function for throwing of a symmetric die.

Because of (2.2.3) one has the limiting case

$$\lim_{x \rightarrow \infty} F(x) = \lim_{x \rightarrow \infty} P(\mathbf{x} < x) = P(E) = 1 . \quad (3.2.2)$$

Applying Eqs. (2.2.5)–(3.2.1) we obtain

$$P(\mathbf{x} \geq x) = 1 - F(x) = 1 - P(\mathbf{x} < x) \quad (3.2.3)$$

and therefore

$$\lim_{x \rightarrow -\infty} F(x) = \lim_{x \rightarrow -\infty} P(\mathbf{x} < x) = 1 - \lim_{x \rightarrow -\infty} P(\mathbf{x} \geq x) = 0 . \quad (3.2.4)$$

Of special interest are distribution functions  $F(x)$  that are continuous and differentiable. The first derivative

$$f(x) = \frac{dF(x)}{dx} = F'(x) \quad (3.2.5)$$

is called the *probability density (function)* of  $\mathbf{x}$ . It is a measure of the probability of the event  $(x \leq \mathbf{x} < x + dx)$ . From (3.2.1) and (3.2.5) it immediately follows that

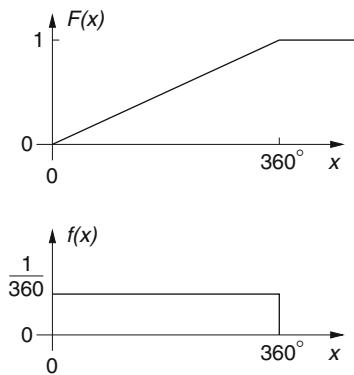
$$P(\mathbf{x} < a) = F(a) = \int_{-\infty}^a f(x) dx , \quad (3.2.6)$$

$$P(a \leq \mathbf{x} < b) = \int_a^b f(x) dx = F(b) - F(a) , \quad (3.2.7)$$

and in particular

$$\int_{-\infty}^{\infty} f(x) dx = 1 \quad . \quad (3.2.8)$$

A trivial example of a continuous distribution is given by the angular position of the hand of a watch read at random intervals. We obtain a constant probability density (Fig. 3.2).



**Fig. 3.2:** Distribution function and probability density for the angular position of a watch hand.

### 3.3 Functions of a Single Random Variable, Expectation Value, Variance, Moments

In addition to the distribution of a random variable  $x$ , we are often interested in the distribution of a function of  $x$ . Such a function of a random variable is also a random variable:

$$y = H(x) \quad . \quad (3.3.1)$$

The variable  $y$  then possesses a distribution function and probability density in the same way as  $x$ .

In the two simple examples of the last section we were able to give the distribution function immediately because of the symmetric nature of the problems. Usually this is not possible. Instead, we have to obtain it from experiment. Often we are limited to determining a few characteristic parameters instead of the complete distribution.

The *mean* or *expectation value* of a random variable is the sum of all possible values  $x_i$  of  $x$  multiplied by their corresponding probabilities

$$E(x) = \hat{x} = \sum_{i=1}^n x_i P(x = x_i) \quad . \quad (3.3.2)$$

Note that  $\hat{x}$  is not a random variable but rather has a fixed value. Correspondingly the expectation value of a function (3.3.1) is defined to be

$$E\{H(\mathbf{x})\} = \sum_{i=1}^n H(x_i) P(\mathbf{x} = x_i) . \quad (3.3.3)$$

In the case of a continuous random variable (with a differentiable distribution function), we define by analogy

$$E(\mathbf{x}) = \hat{x} = \int_{-\infty}^{\infty} x f(x) dx \quad (3.3.4)$$

and

$$E\{H(\mathbf{x})\} = \int_{-\infty}^{\infty} H(x) f(x) dx . \quad (3.3.5)$$

If we choose in particular

$$H(\mathbf{x}) = (\mathbf{x} - c)^\ell , \quad (3.3.6)$$

we obtain the expectation values

$$\alpha_\ell = E\{(\mathbf{x} - c)^\ell\} , \quad (3.3.7)$$

which are called the  *$\ell$ -th moments* of the variable about the point  $c$ . Of special interest are the *moments about the mean*,

$$\mu_\ell = E\{(\mathbf{x} - \hat{x})^\ell\} . \quad (3.3.8)$$

The lowest moments are obviously

$$\mu_0 = 1 , \quad \mu_1 = 0 . \quad (3.3.9)$$

The quantity

$$\mu_2 = \sigma^2(\mathbf{x}) = \text{var}(\mathbf{x}) = E\{(\mathbf{x} - \hat{x})^2\} \quad (3.3.10)$$

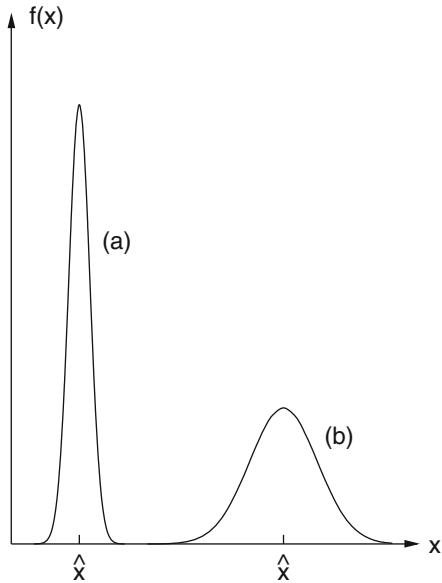
is the lowest moment containing information about the average deviation of the variable  $\mathbf{x}$  from its mean. It is called the *variance* of  $\mathbf{x}$ .

We will now try to visualize the practical meaning of the expectation value and variance of a random variable  $\mathbf{x}$ . Let us consider the measurement of some quantity, for example, the length  $x_0$  of a small crystal using a microscope. Because of the influence of different factors, such as the imperfections of the different components of the microscope and observational errors, repetitions of the measurement will yield slightly different results for  $\mathbf{x}$ . The individual measurements will, however, tend to group themselves in the neighborhood of the true value of the length to be measured, i.e., it will

be more probable to find a value of  $x$  near to  $x_0$  than far from it, providing no systematic biases exist. The probability density of  $x$  will therefore have a bell-shaped form as sketched in Fig. 3.3, although it need not be symmetric. It seems reasonable – especially in the case of a symmetric probability density – to interpret the expectation value (3.3.4) as the best estimate of the true value. It is interesting to note that (3.3.4) has the mathematical form of a center of gravity, i.e.,  $\hat{x}$  can be visualized as the  $x$ -coordinate of the center of gravity of the surface under the curve describing the probability density.

The variance (3.3.10),

$$\sigma^2(x) = \int_{-\infty}^{\infty} (x - \hat{x})^2 f(x) dx , \quad (3.3.11)$$



**Fig.3.3:** Distribution with small variance (a) and large variance (b).

which has the form of a moment of inertia, is a measure of the width or dispersion of the probability density about the mean. If it is small, the individual measurements lie close to  $\hat{x}$  (Fig. 3.3a); if it is large, they will in general be further from the mean (Fig. 3.3b). The positive square root of the variance

$$\sigma = \sqrt{\sigma^2(x)} \quad (3.3.12)$$

is called the *standard deviation* (or sometimes the *dispersion*) of  $x$ . Like the variance itself it is a measure of the average deviation of the measurements  $x$  from the expectation value.

Since the standard deviation has the same dimension as  $x$  (in our example both have the dimension of length), it is identified with the error of the measurement,

$$\sigma(\mathbf{x}) = \Delta x \quad .$$

This definition of measurement error is discussed in more detail in Sects. 5.6 – 5.10. It should be noted that the definitions (3.3.4) and (3.3.10) do not provide completely a way of calculating the mean or the measurement error, since the probability density describing a measurement is in general unknown.

The third moment about the mean is sometimes called *skewness*. We prefer to define the dimensionless quantity

$$\gamma = \mu_3 / \sigma^3 \quad (3.3.13)$$

to be the skewness of  $\mathbf{x}$ . It is positive (negative) if the distribution is skew to the right (left) of the mean. For symmetric distributions the skewness vanishes. It contains information about a possible difference between positive and negative deviation from the mean.

We will now obtain a few important rules about means and variances. In the case where

$$H(\mathbf{x}) = c\mathbf{x} \quad , \quad c = \text{const} \quad , \quad (3.3.14)$$

it follows immediately that

$$\begin{aligned} E(c\mathbf{x}) &= cE(\mathbf{x}) \quad , \\ \sigma^2(c\mathbf{x}) &= c^2\sigma^2(\mathbf{x}) \quad , \end{aligned} \quad (3.3.15)$$

and therefore

$$\sigma^2(\mathbf{x}) = E\{(\mathbf{x} - \hat{\mathbf{x}})^2\} = E\{\mathbf{x}^2 - 2\mathbf{x}\hat{\mathbf{x}} + \hat{\mathbf{x}}^2\} = E(\mathbf{x}^2) - \hat{\mathbf{x}}^2 \quad . \quad (3.3.16)$$

We now consider the function

$$\mathbf{u} = \frac{\mathbf{x} - \hat{\mathbf{x}}}{\sigma(\mathbf{x})} \quad . \quad (3.3.17)$$

It has the expectation value

$$E(\mathbf{u}) = \frac{1}{\sigma(\mathbf{x})} E(\mathbf{x} - \hat{\mathbf{x}}) = \frac{1}{\sigma(\mathbf{x})} (\hat{\mathbf{x}} - \hat{\mathbf{x}}) = 0 \quad (3.3.18)$$

and variance

$$\sigma^2(\mathbf{u}) = \frac{1}{\sigma^2(\mathbf{x})} E\{(\mathbf{x} - \hat{\mathbf{x}})^2\} = \frac{\sigma^2(\mathbf{x})}{\sigma^2(\mathbf{x})} = 1 \quad . \quad (3.3.19)$$

The function  $\mathbf{u}$  – which is also a random variable – has particularly simple properties, which makes its use in more involved calculations preferable. We will call such a variable (having zero mean and unit variance) a *reduced variable*. It is also called a standardized, normalized, or dimensionless variable.

Although a distribution is mathematically most easily described by its expectation value, variance, and higher moments (in fact any distribution can be completely specified by these quantities, cf. Sect. 5.5), it is often convenient to introduce further definitions so as to better visualize the form of a distribution.

The *mode*  $x_m$  (or *most probable value*) of a distribution is defined as that value of the random variable that corresponds to the highest probability:

$$P(x = x_m) = \max . \quad (3.3.20)$$

If the distribution has a differentiable probability density, the mode, which corresponds to its maximum, is easily determined by the conditions

$$\frac{d}{dx} f(x) = 0 , \quad \frac{d^2}{dx^2} f(x) < 0 . \quad (3.3.21)$$

In many cases only one maximum exists; the distribution is said to be *unimodal*. The *median*  $x_{0.5}$  of a distribution is defined as that value of the random variable for which the distribution function equals 1/2:

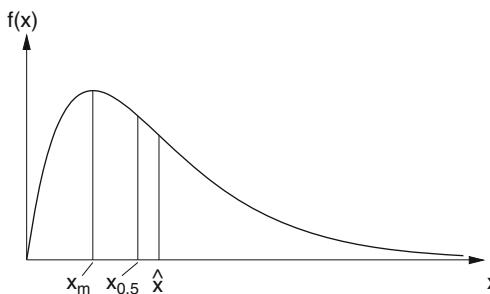
$$F(x_{0.5}) = P(x < x_{0.5}) = 0.5 . \quad (3.3.22)$$

In the case of a continuous probability density Eq. (3.3.22) takes the form

$$\int_{-\infty}^{x_{0.5}} f(x) dx = 0.5 , \quad (3.3.23)$$

i.e., the median divides the total range of the random variable into two regions each containing equal probability.

It is clear from these definitions that in the case of a unimodal distribution with continuous probability density that is symmetric about its maximum, the values of mean, mode, and median coincide. This is not, however, the case for asymmetric distributions (Fig. 3.4).



**Fig. 3.4:** Most probable value (mode)  $x_m$ , mean  $\hat{x}$ , and median  $x_{0.5}$  of an asymmetric distribution.

The definition (3.3.22) can easily be generalized. The quantities  $x_{0.25}$  and  $x_{0.75}$  defined by

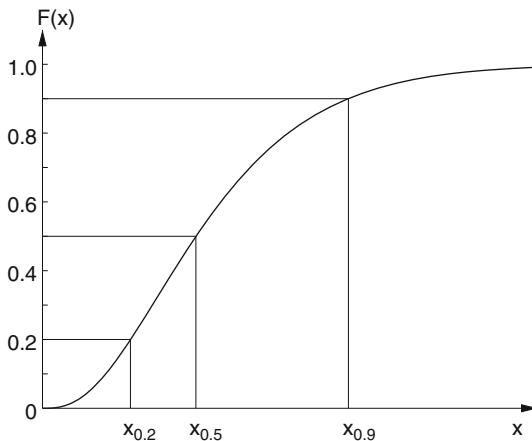
$$F(x_{0.25}) = 0.25 , \quad F(x_{0.75}) = 0.75 \quad (3.3.24)$$

are called lower and upper *quartiles*. Similarly we can define *deciles*  $x_{0.1}, x_{0.2}, \dots, x_{0.9}$ , or in general *quantiles*  $x_q$ , by

$$F(x_q) = \int_{-\infty}^{x_q} f(x) dx = q \quad (3.3.25)$$

with  $0 \leq q \leq 1$ .

The definition of quantiles is most easily visualized from Fig. 3.5. In a plot of the distribution function  $F(x)$ , the quantile  $x_q$  can be read off as the abscissa corresponding to the value  $q$  on the ordinate. The quantile  $x_q(q)$ , regarded as a function of the probability  $q$ , is simply the inverse of the distribution function.



**Fig. 3.5:** Median and quantiles of a continuous distribution.

### Example 3.3: Uniform distribution

We will now discuss the simplest case of a distribution function of a continuous variable. Suppose that in the interval  $a \leq x < b$  the probability density of  $x$  is constant, and it is zero outside of this interval:

$$\begin{aligned} f(x) &= c & , & a \leq x < b & , \\ f(x) &= 0 & , & x < a & , & x \geq b & . \end{aligned} \quad (3.3.26)$$

Because of (3.2.8) one has

$$\int_{-\infty}^{\infty} f(x) dx = c \int_a^b dx = c(b-a) = 1$$

or

$$\begin{aligned} f(x) &= \frac{1}{b-a} & , & a \leq x < b & , \\ f(x) &= 0 & , & x < a & , & x \geq b & . \end{aligned} \quad (3.3.27)$$

The distribution function is

$$\begin{aligned} F(x) &= \int_a^x \frac{dx}{b-a} = \frac{x-a}{b-a} \quad , \quad a \leq x < b \quad , \\ F(x) &= 0 \quad , \quad x < a \quad , \\ F(x) &= 1 \quad , \quad x \geq b \quad . \end{aligned} \tag{3.3.28}$$

By symmetry arguments the expectation value of  $x$  must be the arithmetic mean of the boundaries  $a$  and  $b$ . In fact, (3.3.4) immediately gives

$$E(x) = \bar{x} = \frac{1}{b-a} \int_a^b x dx = \frac{1}{2} \frac{1}{(b-a)} (b^2 - a^2) = \frac{b+a}{2} \quad . \tag{3.3.29}$$

Correspondingly, one obtains from (3.3.10)

$$\sigma^2(x) = \frac{1}{12} (b-a)^2 \quad . \tag{3.3.30}$$

The uniform distribution is not of great practical interest. It is, however, particularly easy to handle, being the simplest distribution of a continuous variable. It is often advantageous to transform a distribution function by means of a transformation of variables into a uniform distribution or the reverse, to express the given distribution in terms of a uniform distribution. This method is used particularly in the “Monte Carlo method” discussed in Chap. 4. ■

#### Example 3.4: Cauchy distribution

In the  $(x, y)$  plane a gun is mounted at the point  $(x, y) = (0, -1)$  such that its barrel lies in the  $(x, y)$  plane and can rotate around an axis parallel to the  $z$  axis (Fig. 3.6).

The gun is fired such that the angle  $\theta$  between the barrel and the  $y$  axis is chosen at random from uniform distribution in the range  $-\pi/2 \leq \theta < \pi/2$ , i.e., the probability density of  $\theta$  is

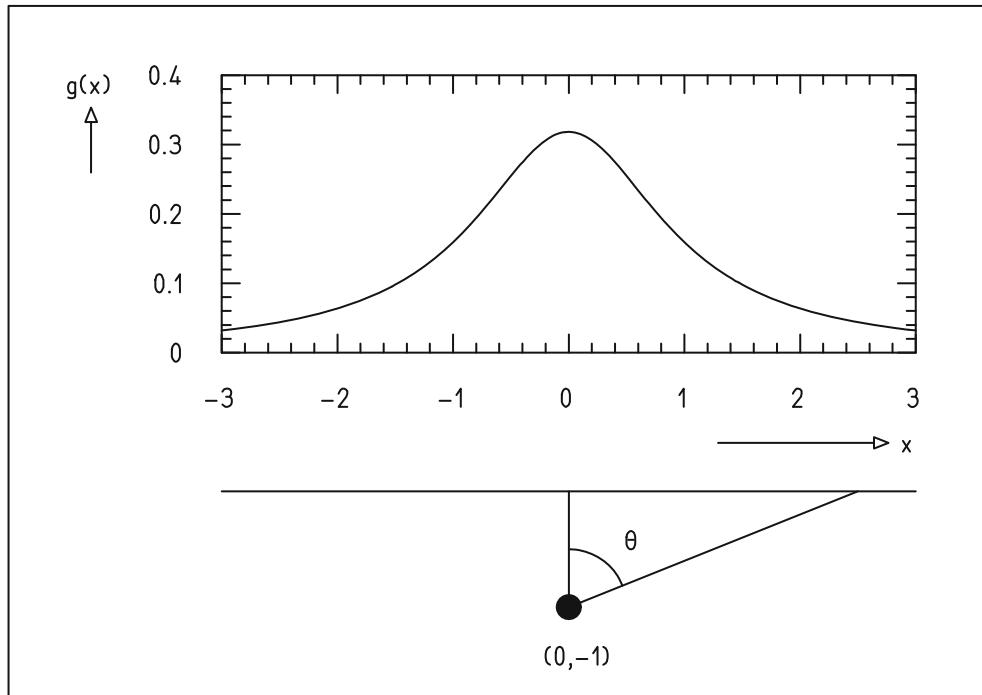
$$f(\theta) = \frac{1}{\pi} \quad .$$

Since

$$\theta = \arctan x \quad , \quad \frac{d\theta}{dx} = \frac{1}{1+x^2} \quad ,$$

we find by the transformation (cf. Sect. 3.7)  $\theta \rightarrow x$  of the variable for the probability density in  $x$

$$g(x) = \left| \frac{d\theta}{dx} \right| f(\theta) = \frac{1}{\pi} \frac{1}{1+x^2} \quad . \tag{3.3.31}$$



**Fig. 3.6:** Model for producing a Cauchy distribution (below) and probability density of the Cauchy distribution (above).

A distribution with this probability density (in our example of the position of hits on the  $x$  axis) is called the *Cauchy distribution*.

The expectation value of  $x$  is (taking the principal value for the integral)

$$\hat{x} = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{x \, dx}{1+x^2} = 0 \quad .$$

The expression for the variance,

$$\begin{aligned} \int_{-\infty}^{\infty} x^2 g(x) \, dx &= \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{x^2 \, dx}{1+x^2} = \frac{1}{\pi} (x - \arctan x) \Big|_{x=-\infty}^{x=\infty} \\ &= \frac{2}{\pi} \lim_{x \rightarrow \infty} (x - \arctan x) \quad , \end{aligned}$$

yields an infinite result. One says that the variance of the Cauchy distribution does not exist.

One can, however, construct another measure for the width of the distribution, the *full width at half maximum ‘FWHM’* (cf. Example 6.3). The function  $g(x)$  has its maximum at  $x = \hat{x} = 0$  and reaches half its maximum value at the points  $x_a = -1$  and  $x_\ell = 1$ . Therefore,

$$\Gamma = 2$$

is the full width at half maximum of the Cauchy distribution. ■

**Example 3.5:** Lorentz (Breit–Wigner) distribution

With  $\hat{x} = a = 0$  and  $\Gamma = 2$  we can write the probability density (3.3.31) of the Cauchy distribution in the form

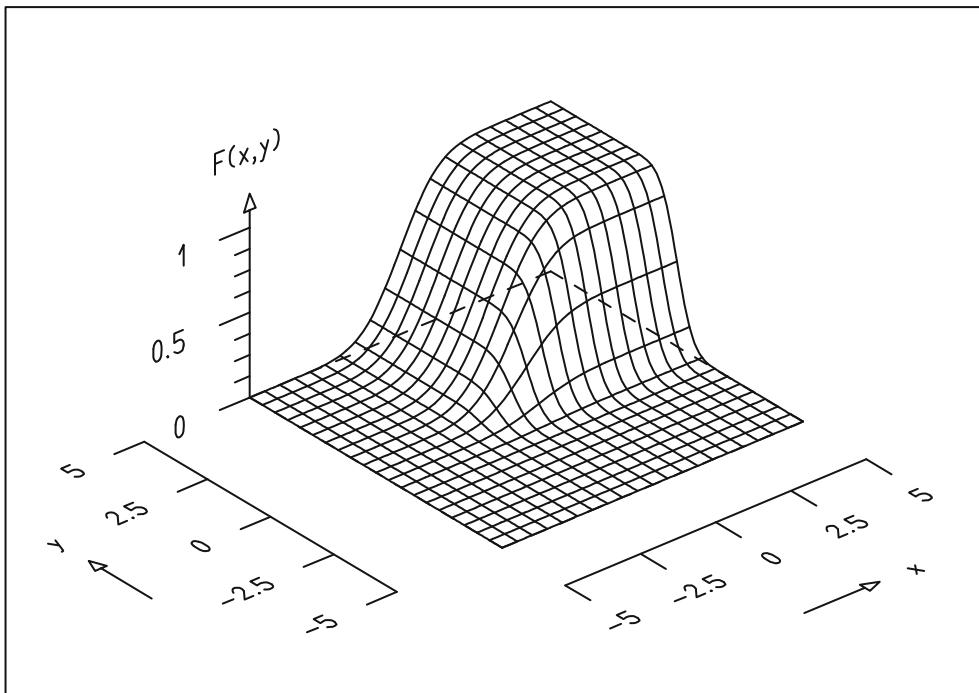
$$g(x) = \frac{2}{\pi \Gamma} \frac{\Gamma^2}{4(x - a)^2 + \Gamma^2} . \quad (3.3.32)$$

This function is a normalized probability density for all values of  $a$  and full width at half maximum  $\Gamma > 0$ . It is called the probability density of the *Lorentz* or also *Breit–Wigner distribution* and plays an important role in the physics of resonance phenomena. ■

### 3.4 Distribution Function and Probability Density of Two Variables: Conditional Probability

We now consider two random variables  $x$  and  $y$  and ask for the probability that both  $x < x$  and  $y < y$ . As in the case of a single variable we expect there to exist of a *distribution function* (see Fig. 3.7)

$$F(x, y) = P(x < x, y < y) . \quad (3.4.1)$$



**Fig.3.7:** Distribution function of two variables.

We will not enter here into axiomatic details and into the conditions for the existence of  $F$ , since these are always fulfilled in cases of practical interest. If  $F$  is a differentiable function of  $x$  and  $y$ , then the *joint probability density* of  $\mathbf{x}$  and  $\mathbf{y}$  is

$$f(x, y) = \frac{\partial}{\partial x} \frac{\partial}{\partial y} F(x, y) \quad . \quad (3.4.2)$$

One then has

$$P(a \leq x < b, c \leq y < d) = \int_a^b \left[ \int_c^d f(x, y) dy \right] dx \quad . \quad (3.4.3)$$

Often we are faced with the following experimental problem. One determines approximately with many measurements the joint distribution function  $F(x, y)$ . One wishes to find the probability for  $\mathbf{x}$  without consideration of  $\mathbf{y}$ . (For example, the probability density for the appearance of a certain infectious disease might be given as a function of date and geographic location. For some investigations the dependence on the time of year might be of no interest.)

We integrate Eq. (3.4.3) over the whole range of  $y$  and obtain

$$P(a \leq x < b, -\infty < y < \infty) = \int_a^b \left[ \int_{-\infty}^{\infty} f(x, y) dy \right] dx = \int_a^b g(x) dx \quad ,$$

where

$$g(x) = \int_{-\infty}^{\infty} f(x, y) dy \quad (3.4.4)$$

is the probability density for  $\mathbf{x}$ . It is called the *marginal probability density* of  $\mathbf{x}$ . The corresponding distribution for  $\mathbf{y}$  is

$$h(y) = \int_{-\infty}^{\infty} f(x, y) dx \quad . \quad (3.4.5)$$

In analogy to the independence of events [Eq. (2.3.6)] we can now define the *independence of random variables*. The variables  $\mathbf{x}$  and  $\mathbf{y}$  are said to be independent if

$$f(x, y) = g(x)h(y) \quad . \quad (3.4.6)$$

Using the marginal distributions we can also define conditional probability for  $\mathbf{y}$  under the condition that  $\mathbf{x}$  is known,

$$P(y \leq y < y + dy | x \leq x \leq x + dx) \quad . \quad (3.4.7)$$

We define the *conditional probability density* as

$$f(y|x) = \frac{f(x, y)}{g(x)} \quad , \quad (3.4.8)$$

so that the probability of Eq. (3.4.7) is given by

$$f(y|x) dy \quad .$$

The rule of total probability can now also be expressed for distributions:

$$h(y) = \int_{-\infty}^{\infty} f(x, y) dx = \int_{-\infty}^{\infty} f(y|x) g(x) dx . \quad (3.4.9)$$

In the case of independent variables as defined by Eq. (3.4.6) one obtains directly from Eq. (3.4.8)

$$f(y|x) = \frac{f(x, y)}{g(x)} = \frac{g(x)h(y)}{g(x)} = h(y) . \quad (3.4.10)$$

This was expected since, in the case of independent variables, any constraint on one variable cannot contribute information about the probability distribution of the other.

### 3.5 Expectation Values, Variance, Covariance, and Correlation

In analogy to Eq. (3.3.5) we define the expectation value of a function  $H(\mathbf{x}, \mathbf{y})$  to be

$$E\{H(\mathbf{x}, \mathbf{y})\} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} H(x, y) f(x, y) dx dy . \quad (3.5.1)$$

Similarly, the variance of  $H(\mathbf{x}, \mathbf{y})$  is defined to be

$$\sigma^2\{H(\mathbf{x}, \mathbf{y})\} = E\{[H(\mathbf{x}, \mathbf{y}) - E(H(\mathbf{x}, \mathbf{y}))]^2\} . \quad (3.5.2)$$

For the simple case  $H(\mathbf{x}, \mathbf{y}) = a\mathbf{x} + b\mathbf{y}$ , Eq. (3.5.1) clearly gives

$$E(a\mathbf{x} + b\mathbf{y}) = aE(\mathbf{x}) + bE(\mathbf{y}) . \quad (3.5.3)$$

We now choose

$$H(\mathbf{x}, \mathbf{y}) = x^\ell y^m \quad (\ell, m \text{ non-negative integers}) . \quad (3.5.4)$$

The expectation values of such functions are the  $\ell m$ th *moments* of  $\mathbf{x}, \mathbf{y}$  about the origin,

$$\lambda_{\ell m} = E(x^\ell y^m) . \quad (3.5.5)$$

If we choose more generally

$$H(\mathbf{x}, \mathbf{y}) = (\mathbf{x} - a)^\ell (\mathbf{y} - b)^m , \quad (3.5.6)$$

the expectation values

$$\alpha_{\ell m} = E\{(\mathbf{x} - a)^\ell (\mathbf{y} - b)^m\} \quad (3.5.7)$$

are the  $\ell m$ -th moments about the point  $a, b$ . Of special interest are the moments about the point  $\lambda_{10}, \lambda_{01}$ ,

$$\mu_{\ell m} = E\{(\mathbf{x} - \lambda_{10})^\ell (\mathbf{y} - \lambda_{01})^m\} . \quad (3.5.8)$$

As in the case of a single variable, the lower moments have a special significance, in particular,

$$\begin{aligned}\mu_{00} &= \lambda_{00} = 1 , \\ \mu_{10} &= \mu_{01} = 0 ;\end{aligned}$$

$$\begin{aligned}\lambda_{10} &= E(\mathbf{x}) = \widehat{\mathbf{x}} , \\ \lambda_{01} &= E(\mathbf{y}) = \widehat{\mathbf{y}} ;\end{aligned} \quad (3.5.9)$$

$$\begin{aligned}\mu_{11} &= E\{(\mathbf{x} - \widehat{\mathbf{x}})(\mathbf{y} - \widehat{\mathbf{y}})\} = \text{cov}(\mathbf{x}, \mathbf{y}) , \\ \mu_{20} &= E\{(\mathbf{x} - \widehat{\mathbf{x}})^2\} = \sigma^2(\mathbf{x}) , \\ \mu_{02} &= E\{(\mathbf{y} - \widehat{\mathbf{y}})^2\} = \sigma^2(\mathbf{y}) .\end{aligned}$$

We can now express the variance of  $a\mathbf{x} + b\mathbf{y}$  in terms of these quantities:

$$\begin{aligned}\sigma^2(a\mathbf{x} + b\mathbf{y}) &= E\{[(a\mathbf{x} + b\mathbf{y}) - E(a\mathbf{x} + b\mathbf{y})]^2\} \\ &= E\{[a(\mathbf{x} - \widehat{\mathbf{x}}) + b(\mathbf{y} - \widehat{\mathbf{y}})]^2\} \\ &= E\{a^2(\mathbf{x} - \widehat{\mathbf{x}})^2 + b^2(\mathbf{y} - \widehat{\mathbf{y}})^2 + 2ab(\mathbf{x} - \widehat{\mathbf{x}})(\mathbf{y} - \widehat{\mathbf{y}})\} , \\ \sigma^2(a\mathbf{x} + b\mathbf{y}) &= a^2\sigma^2(\mathbf{x}) + b^2\sigma^2(\mathbf{y}) + 2ab \text{cov}(\mathbf{x}, \mathbf{y}) .\end{aligned} \quad (3.5.10)$$

In deriving (3.5.10) we have made use of (3.3.14). As another example we consider

$$H(\mathbf{x}, \mathbf{y}) = \mathbf{x}\mathbf{y} . \quad (3.5.11)$$

In this case we have to assume the independence of  $\mathbf{x}$  and  $\mathbf{y}$  in the sense of (3.4.6) in order to obtain the expectation value. Then according to (3.5.1) one has

$$\begin{aligned}E(\mathbf{x}\mathbf{y}) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x y g(x)h(y) dx dy \\ &= \left( \int_{-\infty}^{\infty} x g(x) dx \right) \left( \int_{-\infty}^{\infty} y h(y) dy \right)\end{aligned} \quad (3.5.12)$$

or

$$E(\mathbf{x}\mathbf{y}) = E(\mathbf{x})E(\mathbf{y}) . \quad (3.5.13)$$

While the quantities  $E(\mathbf{x}), E(\mathbf{y}), \sigma^2(\mathbf{x}), \sigma^2(\mathbf{y})$  are very similar to those obtained in the case of a single variable, we still have to explain the meaning

of  $\text{cov}(x, y)$ . The concept of *covariance* is of considerable importance for the understanding of many of our subsequent problems. From its definition we see that  $\text{cov}(x, y)$  is positive if values  $x > \hat{x}$  appear preferentially together with values  $y > \hat{y}$ . On the other hand,  $\text{cov}(x, y)$  is negative if in general  $x > \hat{x}$  implies  $y < \hat{y}$ . If, finally, the knowledge of the value of  $x$  does not give us additional information about the probable position of  $y$ , the covariance vanishes. These cases are illustrated in Fig. 3.8.

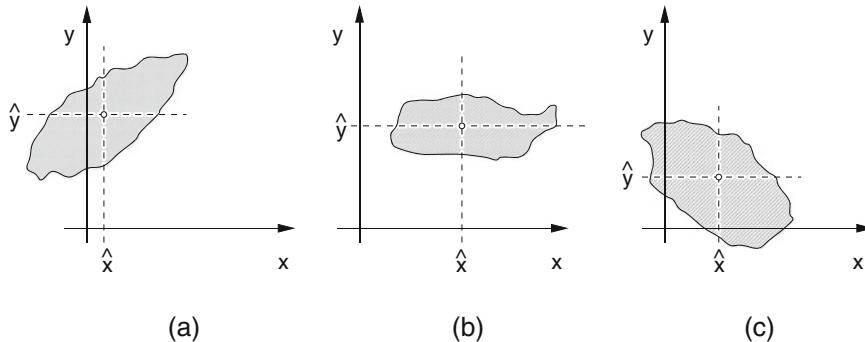
It is often convenient to use the *correlation coefficient*

$$\rho(x, y) = \frac{\text{cov}(x, y)}{\sigma(x)\sigma(y)} \quad (3.5.14)$$

rather than the covariance.

Both the covariance and the correlation coefficient offer a (necessarily crude) measure of the mutual dependence of  $x$  and  $y$ . To investigate this further we now consider two reduced variables  $u$  and  $v$  in the sense of Eq. (3.3.17) and determine the variance of their sum by using (3.5.9),

$$\sigma^2(u+v) = \sigma^2(u) + \sigma^2(v) + 2\rho(u, v)\sigma(u)\sigma(v) \quad . \quad (3.5.15)$$



**Fig. 3.8:** Illustration of the covariance between the variables  $x$  and  $y$ . (a)  $\text{cov}(x, y) > 0$ ; (b)  $\text{cov}(x, y) \approx 0$ ; (c)  $\text{cov}(x, y) < 0$ .

From Eq. (3.3.19) we know that  $\sigma^2(u) = \sigma^2(v) = 1$ . Therefore we have

$$\sigma^2(u+v) = 2(1 + \rho(u, v)) \quad (3.5.16)$$

and correspondingly

$$\sigma^2(u-v) = 2(1 - \rho(u, v)) \quad . \quad (3.5.17)$$

Since the variance always fulfills

$$\sigma^2 \geq 0 \quad , \quad (3.5.18)$$

it follows that

$$-1 \leq \rho(u, v) \leq 1 . \quad (3.5.19)$$

If one now returns to the original variables  $x, y$ , then it is easy to show that

$$\rho(u, v) = \rho(x, y) . \quad (3.5.20)$$

Thus we have finally shown that

$$-1 \leq \rho(x, y) \leq 1 . \quad (3.5.21)$$

We now investigate the limiting cases  $\pm 1$ . For  $\rho(u, v) = 1$  the variance is  $\sigma(u - v) = 0$ , i.e., the random variable  $(u - v)$  is a constant. Expressed in terms of  $x, y$  one has therefore

$$u - v = \frac{x - \hat{x}}{\sigma(x)} - \frac{y - \hat{y}}{\sigma(y)} = \text{const} . \quad (3.5.22)$$

The equation is always fulfilled if

$$y = a + bx , \quad (3.5.23)$$

where  $b$  is positive. Therefore in the case of a linear dependence ( $b$  positive) between  $x$  and  $y$  the correlation coefficient takes the value  $\rho(x, y) = +1$ . Correspondingly one finds  $\rho(x, y) = -1$  for a negative linear dependence ( $b$  negative). We would expect the covariance to vanish for two independent variables  $x$  and  $y$ , i.e., for which the probability density obeys Eq. (3.4.6). Indeed with (3.5.9) and (3.5.1) we find

$$\begin{aligned} \text{cov}(x, y) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \hat{x})(y - \hat{y}) g(x) h(y) dx dy \\ &= \left( \int_{-\infty}^{\infty} (x - \hat{x}) g(x) dx \right) \left( \int_{-\infty}^{\infty} (y - \hat{y}) h(y) dy \right) \\ &= 0 . \end{aligned}$$

## 3.6 More than Two Variables: Vector and Matrix Notation

In analogy to (3.4.1) we now define a *distribution function of  $n$  variables*  $x_1, x_2, \dots, x_n$ :

$$F(x_1, x_2, \dots, x_n) = P(x_1 < x_1, x_2 < x_2, \dots, x_n < x_n) . \quad (3.6.1)$$

If the function  $F$  is differentiable with respect to the  $x_i$ , then the *joint probability density* is given by

$$f(x_1, x_2, \dots, x_n) = \frac{\partial^n}{\partial x_1 \partial x_2 \cdots \partial x_n} F(x_1, x_2, \dots, x_n) \quad . \quad (3.6.2)$$

The probability density of one of the variables  $x_r$ , the *marginal probability density*, is given by

$$g_r(x_r) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, x_2, \dots, x_n) dx_1 \cdots dx_{r-1} dx_{r+1} \cdots dx_n \quad . \quad (3.6.3)$$

If  $H(x_1, x_2, \dots, x_n)$  is a function of  $n$  variables, then the *expectation value* of  $H$  is

$$\begin{aligned} & E\{H(x_1, x_2, \dots, x_n)\} \\ &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} H(x_1, x_2, \dots, x_n) f(x_1, x_2, \dots, x_n) dx_1 \cdots dx_n \quad . \end{aligned} \quad (3.6.4)$$

With  $H(x) = x_r$  one obtains

$$\begin{aligned} E(x_r) &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} x_r f(x_1, x_2, \dots, x_n) dx_1 \cdots dx_n \quad , \\ E(x_r) &= \int_{-\infty}^{\infty} x_r g_r(x_r) dx_r \quad . \end{aligned} \quad (3.6.5)$$

The variables are *independent* if

$$f(x_1, x_2, \dots, x_n) = g_1(x_1) g_2(x_2) \cdots g_n(x_n) \quad . \quad (3.6.6)$$

In analogy to Eq. (3.6.3) one can define the joint marginal probability density of  $\ell$  out of the  $n$  variables,\* by integrating (3.6.3) over only the  $n - \ell$  remaining variables,

$$g(x_1, x_2, \dots, x_\ell) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, x_2, \dots, x_n) dx_{\ell+1} \cdots dx_n \quad . \quad (3.6.7)$$

These  $\ell$  variables are independent if

$$g(x_1, x_2, \dots, x_\ell) = g_1(x_1) g_2(x_2) \cdots g_\ell(x_\ell) \quad . \quad (3.6.8)$$

The *moments of order*  $\ell_1, \ell_2, \dots, \ell_n$  *about the origin* are the expectation values of the functions

$$H = x_1^{\ell_1} x_2^{\ell_2} \cdots x_n^{\ell_n} \quad ,$$

---

\*Without loss of generality we can take these to be the variables  $x_1, x_2, \dots, x_\ell$ .

that is,

$$\lambda_{\ell_1 \ell_2 \dots \ell_n} = E(\mathbf{x}_1^{\ell_1} \mathbf{x}_2^{\ell_2} \cdots \mathbf{x}_n^{\ell_n}) \quad .$$

In particular one has

$$\begin{aligned}\lambda_{100\dots0} &= E(\mathbf{x}_1) = \hat{x}_1 \quad , \\ \lambda_{010\dots0} &= E(\mathbf{x}_2) = \hat{x}_2 \quad , \\ &\vdots \\ \lambda_{000\dots1} &= E(\mathbf{x}_n) = \hat{x}_n \quad .\end{aligned}\tag{3.6.9}$$

The moments about  $(\hat{x}_1, \hat{x}_2, \dots, \hat{x}_n)$  are correspondingly

$$\mu_{\ell_1 \ell_2 \dots \ell_n} = E\{(\mathbf{x}_1 - \hat{x}_1)^{\ell_1} (\mathbf{x}_2 - \hat{x}_2)^{\ell_2} \cdots (\mathbf{x}_n - \hat{x}_n)^{\ell_n}\} \quad .\tag{3.6.10}$$

The *variances* of the  $x_i$  are then

$$\begin{aligned}\mu_{200\dots0} &= E\{(\mathbf{x}_1 - \hat{x}_1)^2\} = \sigma^2(\mathbf{x}_1) \quad , \\ \mu_{020\dots0} &= E\{(\mathbf{x}_2 - \hat{x}_2)^2\} = \sigma^2(\mathbf{x}_2) \quad , \\ &\vdots \\ \mu_{000\dots2} &= E\{(\mathbf{x}_n - \hat{x}_n)^2\} = \sigma^2(\mathbf{x}_n) \quad .\end{aligned}\tag{3.6.11}$$

The moment with  $\ell_i = \ell_j = 1$ ,  $\ell_k = 0$  ( $i \neq k \neq j$ ) is called the *covariance* between the variables  $\mathbf{x}_i$  and  $\mathbf{x}_j$ ,

$$c_{ij} = \text{cov}(\mathbf{x}_i, \mathbf{x}_j) = E\{(\mathbf{x}_i - \hat{x}_i)(\mathbf{x}_j - \hat{x}_j)\} \quad .\tag{3.6.12}$$

It proves useful to represent the  $n$  variables  $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$  as components of a vector  $\mathbf{x}$  in an  $n$ -dimensional space. We can then write the distribution function (3.6.1) as

$$F = F(\mathbf{x}) \quad .\tag{3.6.13}$$

Correspondingly, the probability density (3.6.2) is then

$$f(\mathbf{x}) = \frac{\partial^n}{\partial x_1 \partial x_2 \cdots \partial x_n} F(\mathbf{x}) \quad .\tag{3.6.14}$$

The expectation value of a function  $H(\mathbf{x})$  is then simply

$$E\{H(\mathbf{x})\} = \int H(\mathbf{x}) f(\mathbf{x}) d\mathbf{x} \quad .\tag{3.6.15}$$

We would now like to express the variances and covariances by means of a matrix.<sup>†</sup> This is called the *covariance matrix*

---

<sup>†</sup>For details on matrix notation see Appendix A.

$$C = \begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & c_{22} & \cdots & c_{2n} \\ \vdots & & & \\ c_{n1} & c_{n2} & \cdots & c_{nn} \end{pmatrix} . \quad (3.6.16)$$

The elements  $c_{ij}$  are given by (3.6.12); the diagonal elements are the variances  $c_{ii} = \sigma^2(\mathbf{x}_i)$ . The covariance matrix is clearly symmetric, since

$$c_{ij} = c_{ji} . \quad (3.6.17)$$

If we now also write the expectation values of the  $\mathbf{x}_i$  as a vector,

$$E(\mathbf{x}) = \hat{\mathbf{x}} , \quad (3.6.18)$$

we see that each element of the covariance matrix

$$c_{ij} = E\{(\mathbf{x}_i - \hat{\mathbf{x}}_i)(\mathbf{x}_j - \hat{\mathbf{x}}_j)^T\}$$

is given by the expectation value of the product of the row vector  $(\mathbf{x} - \hat{\mathbf{x}})^T$  and the column vector  $(\mathbf{x} - \hat{\mathbf{x}})$ , where

$$\mathbf{x}^T = (x_1, x_2, \dots, x_n) , \quad \mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} .$$

The covariance matrix can therefore be written simply as

$$C = E\{(\mathbf{x} - \hat{\mathbf{x}})(\mathbf{x} - \hat{\mathbf{x}})^T\} . \quad (3.6.19)$$

## 3.7 Transformation of Variables

As already mentioned in Sect. 3.3, a function of a random variable is itself a random variable, e.g.,

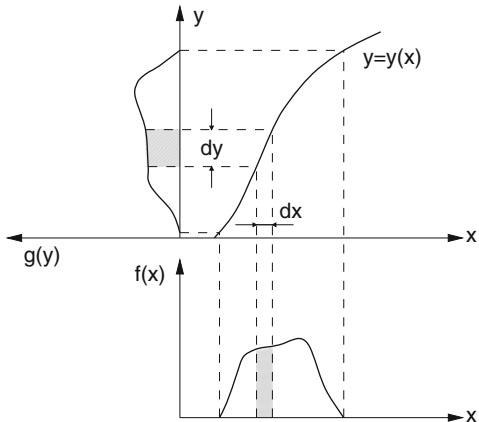
$$\mathbf{y} = \mathbf{y}(\mathbf{x}) .$$

We now ask for the probability density  $g(y)$  for the case where the probability density  $f(x)$  is known.

Clearly the probability

$$g(y) dy$$

that  $y$  falls into a small interval  $dy$  must be equal to the probability  $f(x) dx$  that  $x$  falls into the “corresponding interval”  $dx$ ,  $f(x) dx = g(y) dy$ . This is illustrated in Fig. 3.9. The intervals  $dx$  and  $dy$  are related by



**Fig. 3.9:** Transformation of variables for a probability density of  $x$  to  $y$ .

$$dy = \left| \frac{dy}{dx} \right| dx , \quad \text{i.e.,} \quad dx = \left| \frac{dx}{dy} \right| dy .$$

The absolute value ensures that we consider the values  $dx, dy$  as intervals without a given direction. Only in this way are the probabilities  $f(x) dx$  and  $g(x) dy$  always positive. The probability density is then given by

$$g(y) = \left| \frac{dx}{dy} \right| f(x) . \quad (3.7.1)$$

We see immediately that  $g(y)$  is defined only in the case of a single-valued function  $y(x)$  since only then is the derivative in (3.7.1) uniquely defined. For functions where this is not the case, e.g.,  $y = \sqrt{x}$ , one must consider the individual single-valued parts separately, i.e.,  $y = +\sqrt{x}$ . Equation (3.7.1) also guarantees that the probability distribution of  $y$  is normalized to unity:

$$\int_{-\infty}^{\infty} g(y) dy = \int_{-\infty}^{\infty} f(x) dx = 1 .$$

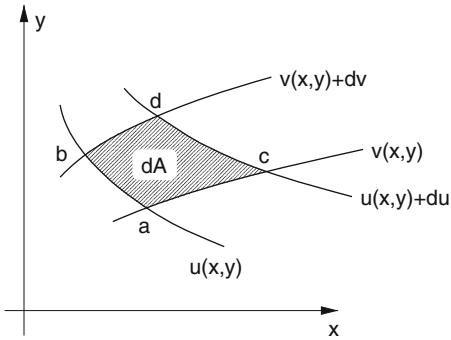
In the case of two independent variables  $x, y$  the transformation to the new variables

$$u = u(x, y) , \quad v = v(x, y) \quad (3.7.2)$$

can be illustrated in a similar way. One must find the quantity  $J$  that relates the probabilities  $f(x, y)$  and  $g(u, v)$ :

$$g(u, v) = f(x, y) \left| J \left( \frac{x, y}{u, v} \right) \right| . \quad (3.7.3)$$

Figure 3.10 shows in the  $(x, y)$  plane two lines each for  $u = \text{const}$  and  $v = \text{const}$ . They bound the surface element  $dA$  of the transformed variables  $u, v$  corresponding to the element  $dx dy$  of the original variables.



**Fig. 3.10:** Variable transformation from  $x, y$  to  $u, v$ .

These curves of course need not be straight lines. Since, however,  $dA$  is an “infinitesimal” surface element, it can be treated as a parallelogram, whose area we will now compute. The coordinates of the corner points  $a, b, c$  are

$$\begin{aligned} x_a &= x(u, v) \quad , & y_a &= y(u, v) \quad , \\ x_b &= x(u, v + dv) \quad , & y_b &= y(u, v + dv) \quad , \\ x_c &= x(u + du, v) \quad , & y_c &= y(u + du, v) \quad . \end{aligned}$$

We can expand the last two lines in a series and obtain

$$\begin{aligned} x_b &= x(u, v) + \frac{\partial x}{\partial v} dv \quad , & y_b &= y(u, v) + \frac{\partial y}{\partial v} dv \quad , \\ x_c &= x(u, v) + \frac{\partial x}{\partial u} du \quad , & y_c &= y(u, v) + \frac{\partial y}{\partial u} du \quad . \end{aligned}$$

The area of the parallelogram is equal to the absolute value of the determinant

$$dA = \begin{vmatrix} 1 & x_a & y_a \\ 1 & x_b & y_b \\ 1 & x_c & y_c \end{vmatrix} = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial y}{\partial u} \\ \frac{\partial x}{\partial v} & \frac{\partial y}{\partial v} \end{vmatrix} du dv = J\left(\frac{x, y}{u, v}\right) du dv \quad , \quad (3.7.4)$$

where the sign is of no consequence because of the absolute value in Eq. (3.7.3). The expression

$$J\left(\frac{x, y}{u, v}\right) = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial y}{\partial u} \\ \frac{\partial x}{\partial v} & \frac{\partial y}{\partial v} \end{vmatrix} \quad (3.7.5)$$

is called the *Jacobian (determinant)* of the transformation (3.7.2).

For the general case of  $n$  variables  $\mathbf{x} = (x_1, x_2, \dots, x_n)$  and the transformation

$$\begin{aligned} y_1 &= y_1(\mathbf{x}) \quad , \\ y_2 &= y_2(\mathbf{x}) \quad , \\ &\vdots \\ y_n &= y_n(\mathbf{x}) \quad , \end{aligned} \quad (3.7.6)$$

the probability density of the transformed variables is given by

$$g(\mathbf{y}) = \left| J\left(\frac{\mathbf{x}}{\mathbf{y}}\right) \right| f(\mathbf{x}) , \quad (3.7.7)$$

where the Jacobian is

$$J\left(\frac{\mathbf{x}}{\mathbf{y}}\right) = J\left(\frac{x_1, x_2, \dots, x_n}{y_1, y_2, \dots, y_n}\right) = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_2}{\partial y_1} & \dots & \frac{\partial x_n}{\partial y_1} \\ \frac{\partial x_1}{\partial y_2} & \frac{\partial x_2}{\partial y_2} & \dots & \frac{\partial x_n}{\partial y_2} \\ \vdots & & & \\ \frac{\partial x_1}{\partial y_n} & \frac{\partial x_2}{\partial y_n} & \dots & \frac{\partial x_n}{\partial y_n} \end{vmatrix} . \quad (3.7.8)$$

A requirement for the existence of  $g(\mathbf{y})$  is of course again the uniqueness of all derivatives occurring in  $J$ .

### 3.8 Linear and Orthogonal Transformations: Error Propagation

In practice we deal frequently with linear transformations of variables. The main reason is that they are particularly easy to handle, and we try therefore to approximate other transformations by linear ones using Taylor series techniques.

Consider  $r$  linear functions of the  $n$  variables  $\mathbf{x} = (x_1, x_2, \dots, x_n)$ :

$$\begin{aligned} y_1 &= a_1 + t_{11}x_1 + t_{12}x_2 + \dots + t_{1n}x_n , \\ y_2 &= a_2 + t_{21}x_1 + t_{22}x_2 + \dots + t_{2n}x_n , \\ &\vdots \\ y_r &= a_r + t_{r1}x_1 + t_{r2}x_2 + \dots + t_{rn}x_n , \end{aligned} \quad (3.8.1)$$

or in matrix notation,

$$\mathbf{y} = T\mathbf{x} + \mathbf{a} . \quad (3.8.2)$$

The expectation value of  $\mathbf{y}$  follows from the generalization of (3.5.3)

$$E(\mathbf{y}) = \widehat{\mathbf{y}} = T\widehat{\mathbf{x}} + \mathbf{a} . \quad (3.8.3)$$

Together with (3.6.19) one obtains the covariance matrix for  $\mathbf{y}$ ,

$$\begin{aligned} C_y &= E\{(\mathbf{y} - \widehat{\mathbf{y}})(\mathbf{y} - \widehat{\mathbf{y}})^T\} \\ &= E\{(T\mathbf{x} + \mathbf{a} - T\widehat{\mathbf{x}} - \mathbf{a})(T\mathbf{x} + \mathbf{a} - T\widehat{\mathbf{x}} - \mathbf{a})^T\} \\ &= E\{T(\mathbf{x} - \widehat{\mathbf{x}})(\mathbf{x} - \widehat{\mathbf{x}})^T T^T\} \\ &= T E\{(\mathbf{x} - \widehat{\mathbf{x}})(\mathbf{x} - \widehat{\mathbf{x}})^T\} T^T , \\ C_y &= T C_x T^T . \end{aligned} \quad (3.8.4)$$

Equation (3.8.4) expresses the well-known law of *error propagation*. Suppose the expectation values  $\hat{x}_i$  have been measured. Suppose as well that the errors (i.e., the standard deviations or variances) and the covariances of  $\mathbf{x}$  are known. One would like to know the errors of an arbitrary function  $\mathbf{y}(\mathbf{x})$ . If the errors are relatively small, then the probability density is only significantly large in a small region (on the order of the standard deviation) around the point  $\hat{\mathbf{x}}$ . One then performs a Taylor expansion of the functions,

$$\begin{aligned} y_i &= y_i(\hat{\mathbf{x}}) + \left( \frac{\partial y_i}{\partial x_1} \right)_{\mathbf{x}=\hat{\mathbf{x}}} (\mathbf{x}_1 - \hat{x}_1) + \cdots + \left( \frac{\partial y_i}{\partial x_n} \right)_{\mathbf{x}=\hat{\mathbf{x}}} (\mathbf{x}_n - \hat{x}_n) \\ &\quad + \text{higher-order terms} , \end{aligned}$$

or in matrix notation,

$$\mathbf{y} = \mathbf{y}(\hat{\mathbf{x}}) + T(\mathbf{x} - \hat{\mathbf{x}}) + \text{higher-order terms} \quad (3.8.5)$$

with

$$T = \begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \dots & \frac{\partial y_1}{\partial x_n} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & \dots & \frac{\partial y_2}{\partial x_n} \\ \vdots & & & \\ \frac{\partial y_r}{\partial x_1} & \frac{\partial y_r}{\partial x_2} & \dots & \frac{\partial y_r}{\partial x_n} \end{pmatrix}_{\mathbf{x}=\hat{\mathbf{x}}} . \quad (3.8.6)$$

If one neglects the higher-order terms and substitutes the first partial derivatives of the matrix  $T$  into Eq. (3.8.4), then one obtains the law of error propagation. We see in particular that not only the errors (i.e., the variances) of  $\mathbf{x}$  but *also the covariances* make a significant contribution to the errors of  $\mathbf{y}$ , that is, to the diagonal elements of  $C_y$ . If these are not taken into account in the error propagation, then the result cannot be trusted.

The covariances can only be neglected when they vanish anyway, i.e., in the case of independent original variables  $\mathbf{x}$ . In this case  $C_x$  simplifies to a diagonal matrix. The diagonal elements of  $C_y$  then have the simple form

$$\sigma^2(y_i) = \sum_{j=1}^n \left( \frac{\partial y_i}{\partial x_j} \right)_{\mathbf{x}=\hat{\mathbf{x}}}^2 \sigma^2(x_j) . \quad (3.8.7)$$

If we now call the standard deviation, i.e., the positive square root of the variance, the error of the corresponding quantity and use for this the symbol  $\Delta$ , Eq. (3.8.7) leads immediately to the formula

$$\Delta y_i = \sqrt{\sum_{j=1}^n \left( \frac{\partial y_i}{\partial x_j} \right)^2 (\Delta x_j)^2} , \quad (3.8.8)$$

known commonly as the law of the propagation of errors. It must be emphasized that this expression is incorrect in cases of non-vanishing covariances. This is illustrated in the following example.

**Example 3.6:** Error propagation and covariance

In a Cartesian coordinate system a point  $(x, y)$  is measured. The measurement is performed with a coordinate measuring device whose error in  $y$  is three times larger than that in  $x$ . The measurements of  $x$  and  $y$  are independent. We therefore can take the covariance matrix to be (up to a factor common to all elements)

$$C_{x,y} = \begin{pmatrix} 1 & 0 \\ 0 & 9 \end{pmatrix} .$$

We now evaluate the errors (i.e., the covariance matrix) in polar coordinates

$$r = \sqrt{x^2 + y^2} , \quad \varphi = \arctan \frac{y}{x} .$$

The transformation matrix (3.8.6) is

$$T = \begin{pmatrix} \frac{x}{r} & \frac{y}{r} \\ -\frac{y}{r^2} & \frac{x}{r^2} \end{pmatrix} .$$

To simplify the numerical calculations we consider only the point  $(1, 1)$ . Then

$$T = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{2} & \frac{1}{2} \end{pmatrix}$$

and therefore

$$C_{r\varphi} = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{2} & \frac{1}{2} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 9 \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{2} \\ \frac{1}{\sqrt{2}} & \frac{1}{2} \end{pmatrix} = \begin{pmatrix} 5 & \frac{4}{\sqrt{2}} \\ \frac{4}{\sqrt{2}} & \frac{5}{2} \end{pmatrix} .$$

We can now return to the original Cartesian coordinate system

$$x = r \cos \varphi , \quad y = r \sin \varphi$$

by use of the transformation

$$T' = \begin{pmatrix} \cos \varphi & -r \sin \varphi \\ \sin \varphi & r \cos \varphi \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{2}} & -1 \\ \frac{1}{\sqrt{2}} & 1 \end{pmatrix} .$$

As expected we obtain

$$C_{xy} = \begin{pmatrix} \frac{1}{\sqrt{2}} & -1 \\ \frac{1}{\sqrt{2}} & 1 \end{pmatrix} \begin{pmatrix} 5 & \frac{4}{\sqrt{2}} \\ \frac{4}{\sqrt{2}} & \frac{5}{2} \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -1 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 9 \end{pmatrix} .$$

If instead we had used the formula (3.8.8), i.e., if we had neglected the covariances in the transformation of  $r, \varphi$  to  $x, y$ , then we would have obtained

$$C'_{xy} = \begin{pmatrix} \frac{1}{\sqrt{2}} & -1 \\ \frac{1}{\sqrt{2}} & 1 \end{pmatrix} \begin{pmatrix} 5 & 0 \\ 0 & \frac{5}{2} \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -1 & 1 \end{pmatrix} = \begin{pmatrix} 5 & 0 \\ 0 & 5 \end{pmatrix} ,$$

which is different from the original covariance matrix. This example stresses the importance of covariances, since it is obviously not possible to change errors of measurements by simply transforming back and forth between coordinate systems. ■

Finally we discuss a special type of linear transformation. We consider the case of exactly  $n$  functions  $y$  of the  $n$  variables  $x$ . In particular we take  $\mathbf{a} = 0$  in (3.8.2). One then has

$$\mathbf{y} = R \mathbf{x} , \quad (3.8.9)$$

where  $R$  is a square matrix. We now require that the transformation (3.8.9) leaves the modulus of a vector invariant

$$\mathbf{y}^2 = \sum_{i=1}^n y_i^2 = \mathbf{x}^2 = \sum_{i=1}^n x_i^2 . \quad (3.8.10)$$

Using Eq. (A.1.9) we can write

$$\mathbf{y}^T \mathbf{y} = (R \mathbf{x})^T (R \mathbf{x}) = \mathbf{x}^T R^T R \mathbf{x} = \mathbf{x}^T \mathbf{x} .$$

This means

$$R^T R = I ,$$

or written in terms of components,

$$\sum_{i=1}^n r_{ik} r_{i\ell} = \delta_{k\ell} = \begin{cases} 0, & k \neq \ell \\ 1, & k = \ell \end{cases} . \quad (3.8.11)$$

A transformation of the type (3.8.9) that fulfills condition (3.8.11) is said to be *orthogonal*. We now consider the determinant of the transformation matrix

$$D = \begin{vmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ r_{21} & r_{22} & \cdots & r_{2n} \\ \vdots & & & \\ r_{n1} & r_{n2} & \cdots & r_{nn} \end{vmatrix}$$

and form its square. According to rules for computing determinants we obtain from Eq. (3.8.11)

$$D^2 = \begin{vmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \cdots & & & \\ 0 & 0 & \cdots & 1 \end{vmatrix},$$

i.e.,  $D = \pm 1$ . The determinant  $D$ , however, is the Jacobian determinant of the transformation (3.8.9),

$$J\left(\frac{\mathbf{y}}{\mathbf{x}}\right) = \pm 1 \quad . \quad (3.8.12)$$

We multiply the system of equations (3.8.9) on the left with  $R^T$  and obtain

$$R^T \mathbf{y} = R^T R \mathbf{x} \quad .$$

Because of Eq. (3.8.11) this expression reduces to

$$\mathbf{x} = R^T \mathbf{y} \quad . \quad (3.8.13)$$

The inverse transformation of an orthogonal transformation is described simply by the transposed transformation matrix. It is itself orthogonal.

An important property of any linear transformation of the type

$$y_1 = r_{11}x_1 + r_{12}x_2 + \cdots + r_{1n}x_n$$

is the following. By constructing additional functions  $y_2, y_3, \dots, y_n$  of equivalent form it can be extended to yield an orthogonal transformation as long as the condition

$$\sum_{i=1}^n r_{1i}^2 = 1$$

is fulfilled.

## 4. Computer Generated Random Numbers: The Monte Carlo Method

### 4.1 Random Numbers

Up to now in this book we have considered the observation of random variables, but not prescriptions for producing them. It is in many applications useful, however, to have a sequence of values of a randomly distributed variable  $x$ . Since operations must often be carried out with a large number of such *random numbers*, it is particularly convenient to have them directly available on a computer. The correct procedure to create such random numbers would be to use a statistical process, e.g., the measurement of the time between two decays from a radioactive source, and to transfer the measured results into the computer. In practical applications, however, the random numbers are almost always calculated directly by the computer. Since this works in a strictly deterministic way, the resulting values are not really random, but rather can be exactly predicted. They are therefore called *pseudorandom*.

Computations with random numbers currently make up a large part of all computer calculations in the planning and evaluation of experiments. The statistical behavior which stems either from the nature of the experiment or from the presence of measurement errors can be simulated on the computer. The use of random numbers in computer programs is often called *the Monte Carlo method*.

We begin this chapter with a discussion of the representation of numbers in a computer (Sect. 4.2), which is indispensable for an understanding of what follows. The best studied method for the creation of uniformly distributed random numbers is the subject of Sects. 4.3–4.7. Sections 4.8 and 4.9 cover the creation of random numbers that follow an arbitrary distribution and the especially common case of normally distributed numbers. In the last two sections one finds discussion and examples of the Monte Carlo method in applications of numerical integration and simulation.

In many examples and exercises we will simulate measurements with the Monte Carlo method and then analyze them. We possess in this way a *computer laboratory*, which allows us to study individually the influence of simulated measurement errors on the results of an analysis.

## 4.2 Representation of Numbers in a Computer

For most applications the representation of numbers used in a computation is unimportant. It can be of decisive significance, however, for the properties of computer-generated random numbers. We will restrict ourselves to the binary representation, which is used today in practically all computers. The elementary unit of information is the *bit*,\* which can assume the values of 0 or 1. This is realized physically by two distinguishably different electric or magnetic states of a component in the computer.

If one has  $k$  bits available for the representation of an integer, then 1 bit is sufficient to encode the sign. The remaining  $k - 1$  bits are used for the binary representation of the absolute value in the form

$$a = a^{(k-2)}2^{k-2} + a^{(k-3)}2^{k-3} + \cdots + a^{(1)}2^1 + a^{(0)}2^0 . \quad (4.2.1)$$

Here each of the coefficients  $a^{(j)}$  can assume only the values 0 or 1, and thus can be represented by a single bit.

The binary representation for non-negative integers is

$$\begin{aligned} 00\cdots000 &= 0 \\ 00\cdots001 &= 1 \\ 00\cdots010 &= 2 \\ 00\cdots011 &= 3 \\ &\vdots \end{aligned}$$

One could simply use the first bit to encode the sign and represent the corresponding negative numbers such that in the first bit the 0 is replaced by a 1. That would give, however, two different representations for the number zero, or rather  $+0$  and  $-0$ . In fact, one uses for negative numbers the “complementary representation”

$$\begin{aligned} 11\cdots111 &= -1 \\ 11\cdots110 &= -2 \\ 11\cdots101 &= -3 \\ &\vdots \end{aligned}$$

---

\*Abbreviation of *binary digit*.

Then using  $k$  bits, integers in the interval

$$-2^{k-1} \leq x \leq 2^{k-1} - 1 \quad (4.2.2)$$

can be represented.

In most computers 8 bit are grouped together into one *byte*. Four bytes are generally used for the representation of integers, i.e.,  $k = 32$ ,  $2^{k-1} - 1 = 2\,147\,483\,647$ . In many small computers only two bytes are available,  $k = 16$ ,  $2^{k-1} - 1 = 32\,767$ . This constraint (4.2.2) must be taken into consideration when designing a program to generate random numbers.

Before turning to the representation of fractional numbers in a computer, let us consider a finite decimal fraction, which we can write in various ways, e.g.,

$$x = 17.23 = 0.1723 \cdot 10^2$$

or in general

$$x = M \cdot 10^e \quad .$$

The quantities  $M$  and  $e$  are called the *mantissa* and *exponent*, respectively. One chooses the exponent such that the mantissa's nonzero digits are all to the right of the decimal point, and the first place after the decimal point is not zero. If one has available  $n$  decimal places for the representation of the value  $M$ , then

$$m = M \cdot 10^n$$

is an integer. In our example,  $n = 4$  and  $m = 1723$ . In this way the decimal fraction  $d$  is represented by the two integers  $m$  and  $e$ .

The representation of fractions in the binary system is done in a completely analogous way. One decomposes a number of the form

$$x = M \cdot 2^e \quad (4.2.3)$$

into a mantissa  $M$  and exponent  $e$ . If  $n_m$  bits are available for the representation of the mantissa (including sign), it can be expressed by the integer

$$m = M \cdot 2^{n_m-1} \quad , \quad -2^{n_m-1} \leq m \leq 2^{n_m-1} - 1 \quad . \quad (4.2.4)$$

If the exponent with its sign is represented by  $n_e$  bits, then it can cover the interval

$$-2^{n_e} \leq e \leq 2^{n_e} - 1 \quad . \quad (4.2.5)$$

In our Java classes we use floating-point numbers of the type `double` with 64 bit,  $n_m = 53$  for the mantissa and  $n_e = 11$  for the exponent.

For the interval of values in which a floating point number can be represented in a computer, the constraint (4.2.2) no longer applies but one has rather the weaker condition

$$2^{e_{\min}} < |x| < 2^{e_{\max}} . \quad (4.2.6)$$

Here  $e_{\min}$  and  $e_{\max}$  are given by (4.2.5). If 11 bit are available for representing the exponent (including sign), then one has  $e_{\max} = 2^{10} - 1 = 1023$ . Therefore, one has the constraint  $|x| < 2^{1023} \approx 10^{308}$ .

When computing with floating point numbers, the concept of the *relative precision* of the representation is of considerable significance. There are a fixed number of binary digits corresponding to a fixed number of decimal places available for the representation of the mantissa  $M$ . If we designate by  $\alpha$  the smallest possible mantissa, then two numbers  $x_1$  and  $x_2$  can still be represented as being distinct if

$$x_1 = x = M \cdot 2^e , \quad x_2 = (M + \alpha) \cdot 2^e .$$

The *absolute precision* in the representation of  $x$  is thus

$$\Delta x = x_1 - x_2 = \alpha \cdot 2^e ,$$

which depends on the exponent of  $x$ . The relative precision

$$\frac{\Delta x}{x} = \frac{\alpha}{M}$$

is in contrast independent of  $x$ . If  $n$  binary digits are available for the representation of the mantissa, then one has  $M \approx 2^n$ , since the exponent is chosen such that all  $n$  places for the mantissa are completely used. The smallest possible mantissa is  $\alpha = 2^0$ , so that the *relative precision* in the representation of  $x$  is

$$\frac{\Delta x}{x} = 2^{-n} . \quad (4.2.7)$$

### 4.3 Linear Congruential Generators

Since, as mentioned, computers work in a strictly deterministic way, all (pseudo)"random numbers generated in a computer are in the most general case a function of all of the preceding (pseudo)random numbers<sup>†</sup>

$$x_{j+1} = f(x_j, x_{j-1}, \dots, x_1) . \quad (4.3.1)$$

Programs for creating random numbers are called *random number generators*.

---

<sup>†</sup>Since the numbers are pseudorandom and not strictly random, we use the notation  $x$  in place of  $\mathbf{x}$ .

The best studied algorithm is based on the following rule,

$$x_{j+1} = (ax_j + c) \bmod m \quad . \quad (4.3.2)$$

All of the quantities in (4.3.2) are integer valued. Generators using this rule are called linear congruential generators (LCG). The symbol mod  $m$  or modulo  $m$  in (4.3.2) means that the expression before the symbol is divided by  $m$  and only the remainder of the result is taken, e.g.,  $6 \bmod 5 = 1$ . Each random number made by an LCG according to the rule (4.3.2) depends only on the number immediately preceding it and on the constant  $a$  (the *multiplier*), on  $c$  (the *increment*), and on  $m$  (the *modulus*). When these three constants and one *initial value*  $x_0$  are given, the infinite sequence of random numbers  $x_0, x_1, \dots$  is determined.

The sequence is clearly periodic. The maximum period length is  $m$ . Only partial sequences that are short compared to the period length are useful for computations.

**Theorem on the maximum period of an LCG with  $c \neq 0$ :**

An LCG defined by the values  $m$ ,  $a$ ,  $c$ , and  $x_0$  has the period  $m$  if and only if

- (a)  $c$  and  $m$  have no common factors;
- (b)  $b = a - 1$  is a multiple of  $p$  for every prime number  $p$  that is a factor of  $m$ ;
- (c)  $b$  is a multiple of 4 if  $m$  is a multiple of 4.

The proof of this theorem as well as the theorems of Sect. 4.4 can be found in, e.g., KNUTH [2].

A simple example is  $c = 3$ ,  $a = 5$ ,  $m = 16$ . One can easily compute that  $x_0 = 0$  results in the sequence

$$0, 3, 2, 13, 4, 7, 6, 1, 8, 11, 10, 5, 12, 15, 14, 9, 0, \dots \quad .$$

Since the period  $m$  can only be attained when all  $m$  possible values are actually assumed, the choice of the initial value  $x_0$  is unimportant.

## 4.4 Multiplicative Linear Congruential Generators

If one chooses  $c = 0$  in (4.3.2), then the algorithm simplifies to

$$x_{j+1} = (ax_j) \bmod m \quad . \quad (4.4.1)$$

Generators based on this rule are called multiplicative linear congruential generators (MLCG). The computation becomes somewhat shorter and thus

faster. The exact value zero, however, can no longer be produced (except for the unusable sequence 0, 0, ...). In addition the period becomes shorter. Before giving the theorem on the maximum period length for this case, we introduce the concept of the primitive element modulo  $m$ .

Let  $a$  be an integer having no common factors (except unity) with  $m$ . We consider all  $a$  for which  $a^\lambda \bmod m = 1$  for integer  $\lambda$ . The smallest value of  $\lambda$  for which this relation is valid is called the *order* of  $a$  modulo  $m$ . All values  $a$  having the same largest possible order  $\lambda(m)$  are called *primitive elements* modulo  $m$ .

**Theorem on the order  $\lambda(m)$  of a primitive element modulo  $m$ :**

For every integer  $e$  and prime number  $p$

$$\begin{aligned}\lambda(2) &= 1 ; \\ \lambda(4) &= 2 ; \\ \lambda(2^e) &= 2^{e-2} , \quad e > 2 ; \\ \lambda(p^e) &= p^{e-1}(p-1) , \quad p > 2 .\end{aligned}\tag{4.4.2}$$

**Theorem on primitive elements modulo  $p^e$ :** The number  $a$  is a primitive element modulo  $p^e$  if and only if

$$\begin{aligned}a \text{ odd} , \quad p^e &= 2 ; \\ a \bmod 4 &= 3 , \quad p^e = 4 ; \\ a \bmod 8 &= 3, 5, 7 , \quad p^e = 8 ; \\ a \bmod 8 &= 3, 5 , \quad p = 2 , \quad e > 3 ; \\ a \bmod p &\neq 0 , \quad a^{(p-1)/q} \bmod p \neq 1 , \quad p > 2 , \quad e = 1 , \\ &\quad q \text{ every prime factor of } p-1 ; \\ a \bmod p &\neq 0 , \quad a^{p-1} \bmod p^2 \neq 1 , \quad a^{(p-1)/q} \bmod p \neq 1 , \\ &\quad p > 2 , \quad e > 1 , \quad q \text{ every prime factor of } p-1 .\end{aligned}\tag{4.4.3}$$

For large values of  $p$  the primitive elements must be determined with computer programs with the aid of this theorem.

**Theorem on the maximum period of an MLCG:** The maximum period of an MLCG defined by the quantities  $m, a, c = 0$ ,  $x_0$  is equal to the order  $\lambda(m)$ . This is attained if the multiplier  $a$  is a primitive element modulo  $m$  and when the initial value  $x_0$  and the multiplier  $m$  have no common factors (except unity).

In fact, MLC generators with  $c = 0$  are frequently used in practice. There are two cases of practical significance in choosing the multiplier  $m$ .

- (i)  $m = 2^e$ : Here  $m - 1$  can be the largest integer that can be represented on the computer. According to (4.4.2) the maximum attainable period length is  $m/4$ .
- (ii)  $m = p$ : If  $m$  is a prime number, the period of  $m - 1$  can be attained according to (4.4.2).

## 4.5 Quality of an MLCG: Spectral Test

When producing random numbers, the main goal is naturally not just to attain the longest possible period. This could be achieved very simply with the sequence  $0, 1, 2, \dots, m - 1, 0, 1, \dots$ . Much more importantly, the individual elements within a period should follow each other “randomly”. First the modulus  $m$  is chosen, and then one chooses various multipliers  $a$  corresponding to (4.4.3) that guarantee a maximum period. One then constructs generators with the constants  $a$ ,  $m$ , and  $c = 0$  in the form of computer programs and checks with statistical tests the randomness of the resulting numbers. General tests, also applicable to this particular question, will be discussed in Sect. 8. The spectral test was especially developed for investigating random numbers, in particular for detecting non-random dependencies between neighboring elements in a sequence.

In a simple example we first consider the case  $a = 3$ ,  $m = 7$ ,  $c = 0$ ,  $x_0 = 1$  and obtain the sequence

$$1, 3, 2, 6, 4, 5, 1, \dots .$$

We now form pairs of neighboring numbers

$$(x_j, x_{j+1}) \quad , \quad j = 0, 1, \dots, n-1 \quad . \quad (4.5.1)$$

Here  $n$  is the period, which in our example is  $n = m - 1 = 6$ . In Fig. 4.1 the number pairs (4.5.1) are represented as points in a two-dimensional Cartesian coordinate system. We note – possibly with surprise – that they form a regular lattice. The surprise is somewhat less, however, when we consider two features of the algorithm (4.3.2):

- (i) All coordinate values  $x_j$  are integers. In the accessible range of values  $1 \leq x_j \leq n$  there are, however, only  $n^2$  number pairs (4.5.1) for which both elements are integer. They lie on a lattice of horizontal and vertical lines. Two neighboring lines have a separation of one.
- (ii) There are only  $n$  different pairs (4.5.1), so that only a fraction of the  $n^2$  points mentioned in (i) are actually occupied.

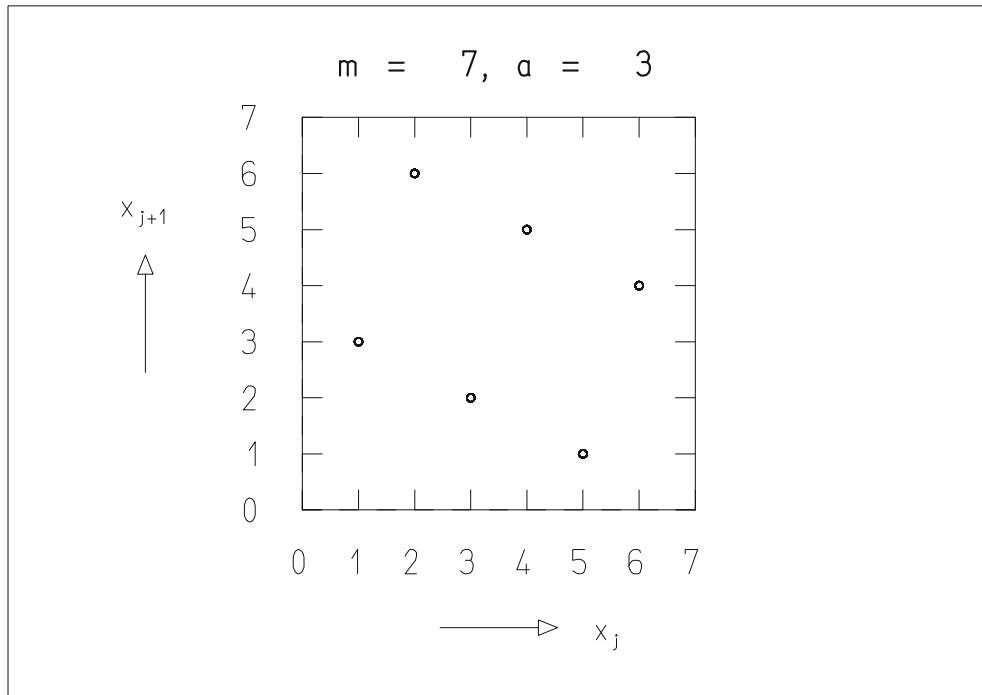
We now go from integer numbers  $x_j$  to transformed numbers

$$u_j = x_j/m \quad (4.5.2)$$

with the property

$$0 < u_j < 1 \quad . \quad (4.5.3)$$

For simplicity we assume that the sequence  $x_0, x_1, \dots$  has the maximum possible period  $m$  for an MLC generator. The pairs



**Fig. 4.1:** Diagram of number pairs (4.5.1) for  $a = 3, m = 7$ .

$$(u_j, u_{j+1}) \quad , \quad j = 0, 1, \dots, m - 1 \quad , \quad (4.5.4)$$

lie in a square whose side has unit length. Because the  $x_j$  are integers, the spacing between the horizontal or vertical lattice lines on which the points (4.5.4) must lie is  $1/m$ . By far not all of these points, however, are occupied. A finite family of lines can be constructed which pass through those points that are actually occupied. We consider now the spacing of neighboring lines within a family, look for the family for which this distance is a maximum, and call this  $d_2$ .

If the distances between neighboring lattice lines for all families are approximately equal, we can then be certain of having a maximally uniform distribution of the occupied lattice points on the unit square. Since this distance is  $1/m$  for a completely occupied lattice ( $m^2$  points), we obtain for a uniformly occupied lattice with  $m$  points a distance of  $d_2 \approx m^{-1/2}$ . With a very nonuniform lattice one obtains the considerably larger value  $d_2 \gg m^{-1/2}$ .

If one now considers not only pairs (4.5.4), but  $t$ -tuples of numbers

$$(u_j, u_{j+1}, \dots, u_{j+t-1}), \quad (4.5.5)$$

one sees that the corresponding points lie on families of  $(t - 1)$ -dimensional hyperplanes in a  $t$ -dimensional cube whose side has unit length. Let us investigate as before the distance between neighboring hyperplanes of a family. We determine the family with the largest spacing and designate this by  $d_t$ . One expects for a uniform distribution of points (4.5.5) a distance

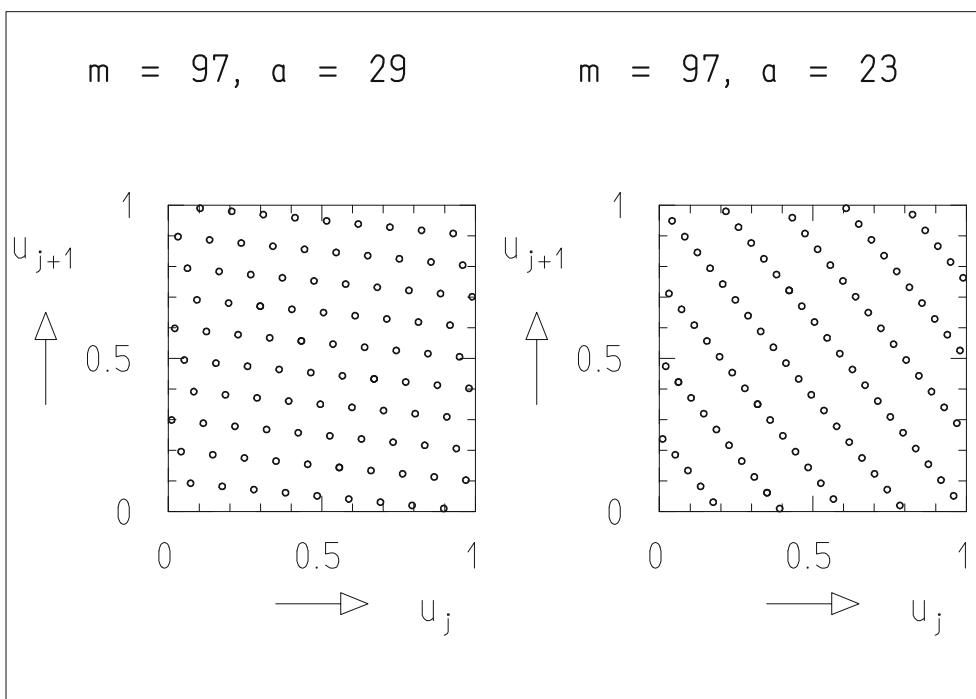
$$d_t \approx m^{-1/t} . \quad (4.5.6)$$

If the lattice is nonuniform, however, we expect

$$d_t \gg m^{-1/t} . \quad (4.5.7)$$

The situations (4.5.6) and (4.5.7) are shown in Fig. 4.2. Naturally one tries to achieve as uniform a lattice as possible. One should note that there is at least a distance (4.5.6) between the lattice points. The lowest decimal places of random numbers are therefore not random, but rather reflect the structure of the lattice.

Theoretical considerations give an upper limit on the smallest possible lattice spacing,



**Fig. 4.2:** Diagram of number pairs (4.5.4) for various small values of  $a$  and  $m$ .

**Table 4.1:** Suitable moduli  $m$  and multipliers  $a$  for portable MLC generators for computers with 32-bit (16-bit) integer arithmetic.

| 32 bit        |        | 16 bit |     |
|---------------|--------|--------|-----|
| $m$           | $a$    | $m$    | $a$ |
| 2 147 483 647 | 39 373 | 32 749 | 162 |
| 2 147 483 563 | 40 014 | 32 363 | 157 |
| 2 147 483 399 | 40 692 | 32 143 | 160 |
| 2 147 482 811 | 41 546 | 32 119 | 172 |
| 2 147 482 801 | 42 024 | 31 727 | 146 |
| 2 147 482 739 | 45 742 | 31 657 | 142 |

$$d_t \geq d_t^* = c_t m^{-1/t} . \quad (4.5.8)$$

The constants  $c_t$  are of order unity. They have the numerical values [2]

$$\begin{aligned} c_2 &= (4/3)^{-1/4}, & c_3 &= 2^{-1/6}, & c_4 &= 2^{-1/4}, & c_5 &= 2^{-3/10}, \\ c_6 &= (64/3)^{-1/12}, & c_7 &= 2^{-3/7}, & c_8 &= 2^{-1/2}. \end{aligned} \quad (4.5.9)$$

The *spectral test* can now be carried out as follows. For given values  $(m, a)$  of the modulus and multiplier of an MLCG one determines with a computer algorithm [2] the values  $d_t(m, a)$  for small  $t$ , e.g.,  $t = 2, 3, \dots, 6$ . One constructs the test quantities

$$S_t(m, a) = \frac{d_t^*(m)}{d_t(m, a)} \quad (4.5.10)$$

and accepts the generator as usable if the  $S_t(m, a)$  do not exceed a given limit. Table 4.1 gives the results of extensive investigations by L'ECUYER [3]. The moduli  $m$  are prime numbers close to the maximum integer values representable by 16 or 32 bit. The multipliers are primitive elements modulo  $m$ . They fulfill the requirement  $a < \sqrt{m}$  (see Sect. 4.6). The prime numbers were chosen such that  $a$  does not have to be much smaller than  $\sqrt{m}$ , but the condition (m, a) in Table 4.1  $S_t(m, a) > 0.65$ ,  $t = 2, 3, \dots, 6$ , still applies.

## 4.6 Implementation and Portability of an MLCG

By *implementation* of an algorithm one means its realization as a computer program for a specific type of computer. If the program can be easily transferred to other computer types and gives there (essentially) the same results, then the program is said to be *portable*. In this section we will give a portable implementation of an MLCG, as realized by WICHMANN and HILL [4] and L'ECUYER [3].

A program that implements the rule (4.4.1) is certain to be portable if the computations are carried out exclusively with integers. If the computer has  $k$  bits for the representation of an integer, then all numbers between  $-m - 1$  and  $m$  for  $m < 2^{k-1}$  are available.

We now choose a multiplier  $a$  with

$$a^2 < m \quad (4.6.1)$$

and define

$$q = m \text{ div } a \quad , \quad r = m \text{ mod } a \quad , \quad (4.6.2)$$

so that

$$m = aq + r \quad . \quad (4.6.3)$$

The expression  $m \text{ div } a$  defined by (4.6.2) and (4.6.3) is the integer part of the quotient  $m/a$ . We now compute the right-hand side of (4.4.1), where we leave off the index  $j$  and note that  $[(x \text{ div } q)m] \text{ mod } m = 0$ , since  $x \text{ div } q$  is an integer:

$$\begin{aligned} [ax] \text{ mod } m &= [ax - (x \text{ div } q)m] \text{ mod } m \\ &= [ax - (x \text{ div } q)(aq + r)] \text{ mod } m \\ &= [a\{x - (x \text{ div } q)q\} - (x \text{ div } q)r] \text{ mod } m \\ &= [a(x \text{ mod } q) - (x \text{ div } q)r] \text{ mod } m \quad . \end{aligned} \quad (4.6.4)$$

Since one always has  $0 < x < m$ , it follows that

$$a(x \text{ mod } q) < aq \leq m \quad , \quad (4.6.5)$$

$$(x \text{ div } q)r < [(aq + r) \text{ div } q]r = ar < a^2 < m \quad . \quad (4.6.6)$$

In this way both terms in square brackets in the last line of (4.6.4) are less than  $m$ , so that the bracketed expression remains in the interval between  $-m$  and  $m$ .

In the Java class DatanRandom we have implemented the expression (4.6.4) in the following three lines, in which all variables are integer:

```
k = x / Q;
x = A * (x - k * Q) - k * R;
if(x < 0) x = x + M;
```

One should note that division of two integer variables results directly in the integer part of the quotient. The first line therefore yields  $x \text{ div } q$  and the last line  $ax \text{ mod } m$ , respectively.

The method DatanRandom.mlcg yields a partial sequence of random numbers of length  $N$ . Each time the subroutine is called, an additional partial sequence is produced. The period of the entire sequence is

$m - 1 = 2147483562$ . The computation is carried out entirely with integer arithmetic, ensuring portability. The output values are, however, floating point valued because of the division by  $m$ , and therefore correspond to a uniform distribution between 0 and 1.

Often one would like to interrupt a computation requiring many random numbers and continue it later starting from the same place. In this case one can read out and store the last computed (integer) random number directly before the interruption, and use it later for producing the next random number. In the technical terminology one calls such a number the *seed* of the generator.

It is sometimes desirable to be able to produce non-overlapping partial sequences of random numbers not one after the other but rather independently. In this way one can, for example, carry out parts of larger simulation problems simultaneously on several computers. As seeds for such partial sequences one uses elements of the total sequence separated by an amount greater than the length of each partial sequence. Such seeds can be determined without having to run through the entire sequence. From (4.4.1) it follows that

$$x_{j+n} = (a^n x_j) \bmod m = [(a^n \bmod m)x_j] \bmod m \quad . \quad (4.6.7)$$

L'ECUYER [3] suggests setting  $n = 2^d$  and choosing some seed  $x_0$ . The expression  $a^{2^d} \bmod m$  can be computed by beginning with  $a$  and squaring it  $d$  times modulo  $m$ . Then one computes  $x_n$  using (4.6.7) and obtains correspondingly  $x_{2n}, x_{3n}, \dots$ .

## 4.7 Combination of Several MLCGs

Since the period of an MLCG is at most  $m - 1$ , and since  $m$  is restricted to  $m < 2^{k-1} - 1$  where  $k$  is the number of bits available in the computer for the representation of an integer, only a relatively short period can be attained with a single MLCG. WICHMANN and HILL [4] and L'ECUYER [3] have given a procedure for combining several MLCGs, which allows for very long periods. The technique is based on the following two theorems.

**Theorem on the sum of discrete random variables, one of which comes from a discrete uniform distribution:** If  $x_1, \dots, x_\ell$  are independent random variables that can only assume integer values, and if  $x_1$  follows a discrete uniform distribution, so that

$$P(x_1 = n) = \frac{1}{d} \quad , \quad n = 0, 1, \dots, d - 1 \quad ,$$

then

$$\mathbf{x} = \left( \sum_{j=1}^{\ell} \mathbf{x}_j \right) \bmod d \quad (4.7.1)$$

also follows this distribution.

We first demonstrate the proof for  $\ell = 2$ , using the abbreviations  $\min(\mathbf{x}_2) = a$ ,  $\max(\mathbf{x}_2) = b$ . One has

$$\begin{aligned} P(\mathbf{x} = n) &= \sum_{k=0}^{\infty} P(\mathbf{x}_1 + \mathbf{x}_2 = n + kd) \\ &= \sum_{i=a}^b P(\mathbf{x}_2 = i) P(\mathbf{x}_1 = (n - i) \bmod d) \\ &= \frac{1}{d} \sum_{i=a}^b P(\mathbf{x}_2 = i) = \frac{1}{d} \quad . \end{aligned}$$

For  $\ell = 3$  we first construct the variable  $\mathbf{x}'_1 = \mathbf{x}_1 + \mathbf{x}_2$ , which follows a discrete uniform distribution between 0 and  $d - 1$ , and then the sum  $\mathbf{x}'_1 + \mathbf{x}_3$ , which has only two terms and therefore possesses the same property. The generalization for  $\ell > 3$  is obvious.

**Theorem on the period of a family of generators:** Consider the random variables  $\mathbf{x}_{j,i}$  coming from a generator  $j$  with a period  $p_j$ , so that the generator gives a sequence  $\mathbf{x}_{j,0}, \mathbf{x}_{j,1}, \dots, \mathbf{x}_{j,p_j-1}$ . We consider now  $\ell$  generators  $j = 1, 2, \dots, \ell$  and the sequence of  $\ell$ -tuples

$$\mathbf{x}_i = \{\mathbf{x}_{1,i}, \mathbf{x}_{2,i}, \dots, \mathbf{x}_{\ell,i}\} \quad , \quad i = 0, 1, \dots \quad . \quad (4.7.2)$$

Its period  $p$  is the smallest common multiple of the periods  $p_1, p_2, \dots, p_\ell$  of the individual generators. The proof is obtained directly from the fact that  $p$  is clearly a multiple of each  $p_j$ .

We now determine the maximum value of the period  $p$ . If the  $\ell$  individual MLCGs have prime numbers  $m_j$  as moduli, then their periods are  $p_j = m_j - 1$  and are therefore even. Therefore one has

$$p \leq \frac{\prod_{j=1}^{\ell} (m_j - 1)}{2^{\ell-1}} \quad . \quad (4.7.3)$$

Equality results if the quantities  $(m_j - 1)/2$  possess no common factors.

The first theorem of this section can now be used to construct a sequence with period given by (4.7.3). One forms first the integer quantity

$$z_i = \left( \sum_{j=1}^{\ell} (-1)^{j-1} x_{j,i} \right) \bmod (m_1 - 1) \quad . \quad (4.7.4)$$

The alternating sign in (4.7.4), which simplifies the construction of the modulus function, does not contradict the prescription of (4.7.1), since one could also use in place of  $x_2, x_4, \dots$ , the variables  $x'_2 = -x_2, x'_4 = -x_4, \dots$ . The quantity  $z_i$  can take on the values

$$z_i \in \{0, 1, \dots, m_1 - 2\} \quad . \quad (4.7.5)$$

The transformation to floating point numbers

$$u_i = \begin{cases} z_i/m_1, & z_i > 0 \\ (m_1 - 1)/m_1, & z_i = 0 \end{cases} \quad (4.7.6)$$

gives values in the range  $0 < u_i < 1$ .

In the method DatanRandom.ecuy we use the techniques, assembled above, to produce uniformly distributed random numbers with a long period. We combine two MLCGs with  $m_1 = 2147483563, a_1 = 40014, m_2 = 2147483399, a_2 = 40692$ . The numbers  $(m_1 - 1)/2$  and  $(m_2 - 1)/2$  have no common factor. Therefore the period of the combined generator is, according to (4.7.3),

$$p = (m_1 - 1)(m_2 - 1)/2 \approx 2.3 \cdot 10^{18} \quad .$$

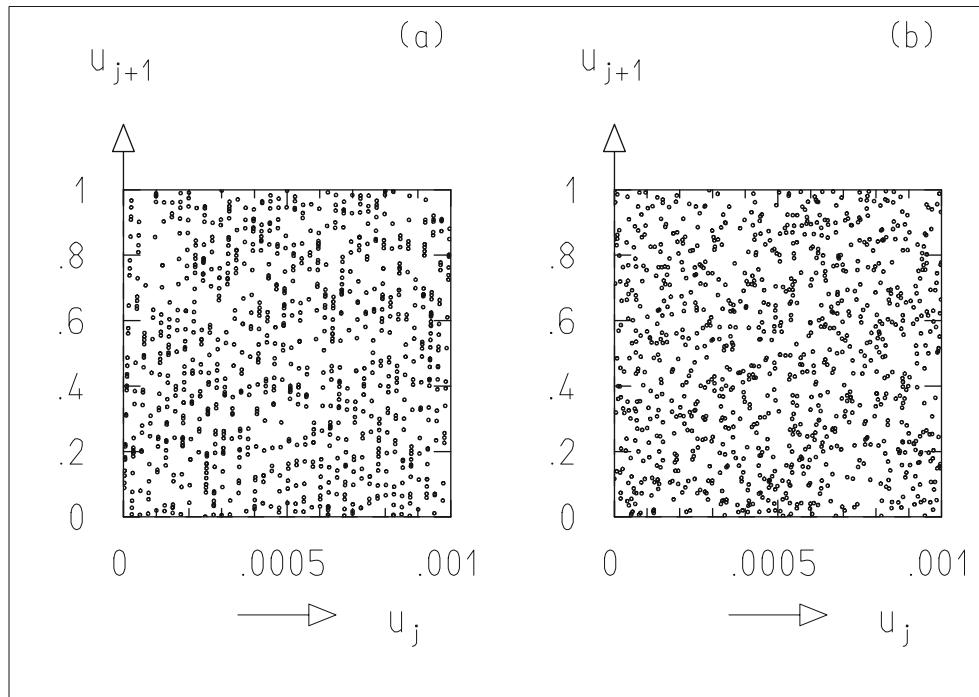
The absolute values of all integers occurring during the computation remain in the range  $\leq 2^{31} - 85$ . The resulting floating point values  $u$  are in the range  $0 < u < 1$ . One does not obtain the values 0 or 1, at least if 23 or more bits are available for the mantissa, which is almost always the case when representing floating point numbers with 32 bit. The program with the given values of  $m_1, m_2, a_1, a_2$  has been subjected to the spectral test and to many other tests by L'ECUYER [3], who has provided a PASCAL version. He determined that it satisfied all of the requirements of the tests.

Figure 4.3 illustrates the difference between the simple MLCG and the combined generator. For the simple MLCG one can still recognize a structure in a scatter plot of the number pairs (4.5.4), although with an expansion of the abscissa by a factor of 1000. The corresponding diagram for the combined generator appears, in contrast, to be completely without structure. For each diagram one million pairs of random numbers were generated. The plots correspond only to a narrow strip on the left-hand edge of the unit square.

In order to initialize non-overlapping partial sequences one can use two methods:

- (i) One applies the procedure discussed in connection with (4.6.7) to both MLCGs, naturally with the same value  $n$ , in order to construct pairs of seeds for each partial sequence.

- (ii) It is considerably easier to use the same seed for the first MLCG for every partial sequence. For the second MLCG one uses an arbitrary seed for the first partial sequence, the following random number from the second MLCG for the second partial sequence, etc. In this way one obtains partial sequences that can reach a length of  $(m_1 - 1)$  without overlapping.



**Fig. 4.3:** Scatter plots of number pairs (4.5.4) from (a) a MLC generator and (b) a combined generator. The methods DatanRandom.mclg and DatanRandom.ecuy, respectively, were used in the generation.

## 4.8 Generation of Arbitrarily Distributed Random Numbers

### 4.8.1 Generation by Transformation of the Uniform Distribution

If  $x$  is a random variable following the uniform distribution,

$$f(x) = 1 \quad , \quad 0 \leq x < 1 \quad ; \quad f(x) = 0 \quad , \quad x < 0 \quad , \quad x \geq 1 \quad , \quad (4.8.1)$$

and  $y$  is a random variable described by the probability density  $g(y)$ , the transformation (3.7.1) simplifies to

$$g(y) dy = dx \quad . \quad (4.8.2)$$

We use the distribution function  $G(y)$ , which is related to  $g(y)$  through  $dG(y)/dy = g(y)$ , and write (4.8.2) in the form

$$dx = g(y) dy = dG(y) , \quad (4.8.3)$$

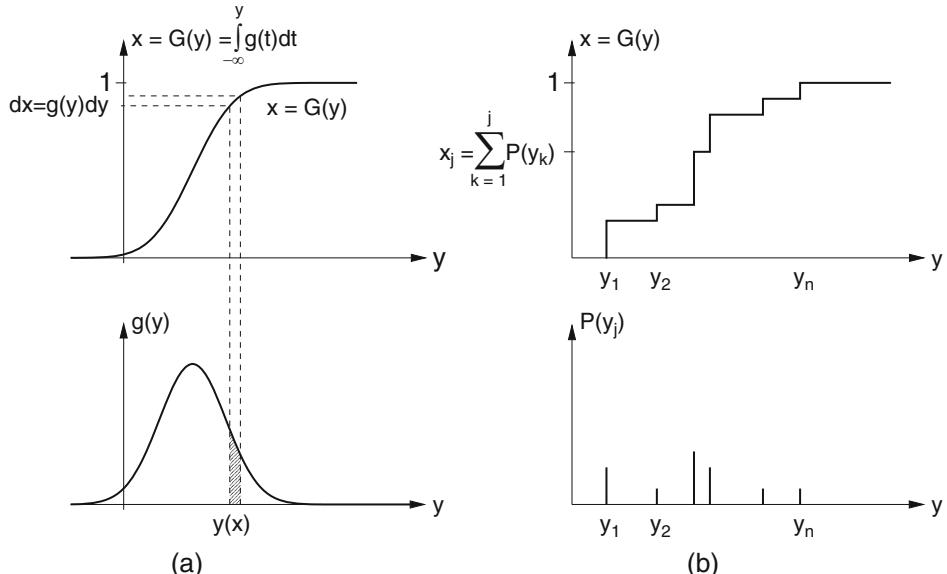
or after integration,

$$x = G(y) = \int_{-\infty}^y g(t) dt . \quad (4.8.4)$$

This equation has the following meaning. If a random number  $x$  is taken from a uniform distribution between 0 and 1 and the function  $x = G(y)$  is inverted,

$$y = G^{-1}(x) , \quad (4.8.5)$$

then one obtains a random number  $y$  described by the probability density  $g(y)$ . The relationship is depicted in Fig. 4.4a. The probability to obtain a random number  $x$  between  $x$  and  $x + dx$  is equal to the probability to have a value  $y(x)$  between  $y$  and  $y + dy$ .



**Fig. 4.4:** Transformation from a uniformly distributed variable  $x$  to a variable  $y$  with the distribution function  $G(y)$ . The variable  $y$  can be continuous (a) or discrete (b).

The relationship (4.8.4) can be also be used to produce discrete probability distributions. An example is shown in Fig. 4.4b. The random variable  $y$  can take on the values  $y_1, y_2, \dots, y_n$  with the probabilities  $P(y_1), P(y_2), \dots, P(y_n)$ . The distribution function as given by (3.2.1) is  $G(y) = P(y < y)$ . The construction of a step function  $x = G(y)$  according to this equation gives the values

$$x_j = G(y_j) = \sum_{k=1}^i P(y_k) , \quad (4.8.6)$$

which lie in the range between 0 and 1. From this one can produce random numbers according to a discrete distribution  $G(y)$  by first producing random numbers  $x$  uniformly distributed between 0 and 1. Depending on the interval  $j$  in which  $x$  falls,  $x_{j-1} < x < x_j$ , the number  $y_j$  is then produced.

**Example 4.1:** Exponentially distributed random numbers

We would like to generate random numbers according to the probability density

$$g(t) = \begin{cases} \frac{1}{\tau} e^{-t/\tau}, & t \geq 0 \\ 0, & t < 0 \end{cases} . \quad (4.8.7)$$

This is the probability density describing the time  $t$  of the decay of a radioactive nucleus that exists at time  $t = 0$  and has a mean lifetime  $\tau$ . The distribution function is

$$x = G(t) = \frac{1}{\tau} \int_{t'=0}^t g(t') dt' = 1 - e^{-t/\tau} . \quad (4.8.8)$$

According to (4.8.4) and (4.8.5) we can obtain exponentially distributed random numbers  $t$  by first generating random numbers uniformly distributed between 0 and 1 and then finding the inverse function  $t = G^{-1}(x)$ , i.e.,

$$t = -\tau \ln(1-x) .$$

Since  $1-x$  is also uniformly distributed between 0 and 1, it is sufficient to compute

$$t = -\tau \ln x . \blacksquare \quad (4.8.9)$$

**Example 4.2:** Generation of random numbers following a Breit–Wigner distribution

To generate random numbers  $y$  which follow a Breit–Wigner distribution (3.3.32),

$$g(y) = \frac{2}{\pi \Gamma} \frac{\Gamma^2}{4(y-a)^2 + \Gamma^2} ,$$

we proceed as discussed in Sect. 4.8.1. We form the distribution function

$$x = G(y) = \int_{-\infty}^y g(y) dy = \frac{2}{\pi \Gamma} \int_{-\infty}^y \frac{\Gamma^2}{4(y-a)^2 + \Gamma^2} dy$$

and perform the integration using the substitution

$$u = \frac{2(y-a)}{\Gamma} , \quad du = \frac{2}{\Gamma} dy .$$

Thus we obtain

$$\begin{aligned} x &= G(y) = \frac{1}{\pi} \int_{\theta=-\infty}^{\theta=2(y-a)/\Gamma} \frac{1}{1+u^2} du = \frac{1}{\pi} [\arctan u]_{-\infty}^{2(y-a)/\Gamma} \\ &= \frac{\arctan 2(y-a)/\Gamma}{\pi} + \frac{1}{2} . \end{aligned}$$

By inversion we obtain

$$\begin{aligned} 2(y-a)/\Gamma &= \tan \left\{ \pi \left( x - \frac{1}{2} \right) \right\} , \\ y &= a + \frac{\Gamma}{2} \tan \left\{ \pi \left( x - \frac{1}{2} \right) \right\} . \end{aligned} \quad (4.8.10)$$

If  $x$  are random numbers uniformly distributed in the interval  $0 < x < 1$ , then  $y$  follows a Breit–Wigner distribution. ■

**Example 4.3:** Generation of random numbers with a triangular distribution

In order to generate random numbers  $y$  following a triangular distribution as in Problem 3.2 we form the distribution function

$$F(y) = \begin{cases} 0, & y < a , \\ \frac{(y-a)^2}{(b-a)(c-a)}, & a \leq y < c , \\ 1 - \frac{(y-b)^2}{(b-a)(b-c)}, & c \leq y < b , \\ 1, & b \leq y . \end{cases}$$

In particular we have

$$F(c) = \frac{c-a}{b-a} .$$

Inverting  $x = F(y)$  gives

$$\begin{aligned} y &= a + \sqrt{(b-a)(c-a)x} , & x < (c-a)/(b-a) , \\ y &= b - \sqrt{(b-a)(b-c)(1-x)} , & x \geq (c-a)/(b-a) . \end{aligned}$$

If  $x$  is uniformly distributed with  $0 < x < 1$ , then  $y$  follows a triangular distribution. ■

## 4.8.2 Generation with the von Neumann Acceptance–Rejection Technique

The elegant technique of the previous section requires that the distribution function  $x = G(y)$  be known and that the inverse function  $y = G^{-1}(x)$  exists and be known as well.

Often one only knows the probability density  $g(y)$ . One can then use the VON NEUMANN acceptance–rejection technique, which we introduce with a simple example before discussing it in its general form.

**Example 4.4:** Semicircle distribution with the simple acceptance–rejection method

As a simple example we generate random numbers following a semicircular probability density,

$$g(y) = \begin{cases} (2/\pi R^2)\sqrt{R^2 - y^2}, & |y| \leq R \\ 0, & |y| > R \end{cases} . \quad (4.8.11)$$

Instead of trying to find and invert the distribution function  $G(y)$ , we generate pairs of random numbers  $(y_i, u_i)$ . Here  $y_i$  is uniformly distributed in the interval available to  $y$ ,  $-R \leq y \leq R$ , and  $u_i$  is uniformly distributed in the range of values assumed by the function  $g(y)$ ,  $0 \leq u \leq R$ . For each pair we test if

$$u_i \geq g(y_i) . \quad (4.8.12)$$

If this inequality is fulfilled, we reject the random number  $y_i$ . The set of random numbers  $y_i$  that are not rejected then follow a probability density  $g(y)$ , since each was accepted with a probability proportional to  $g(y_i)$ . ■

The technique of Example 4.4 can easily be described geometrically. To generate random numbers in the interval  $a \leq y \leq b$  according to the probability density  $g(y)$ , one must consider in the region  $a \leq y \leq b$  the curve

$$u = g(y) \quad (4.8.13)$$

and a constant

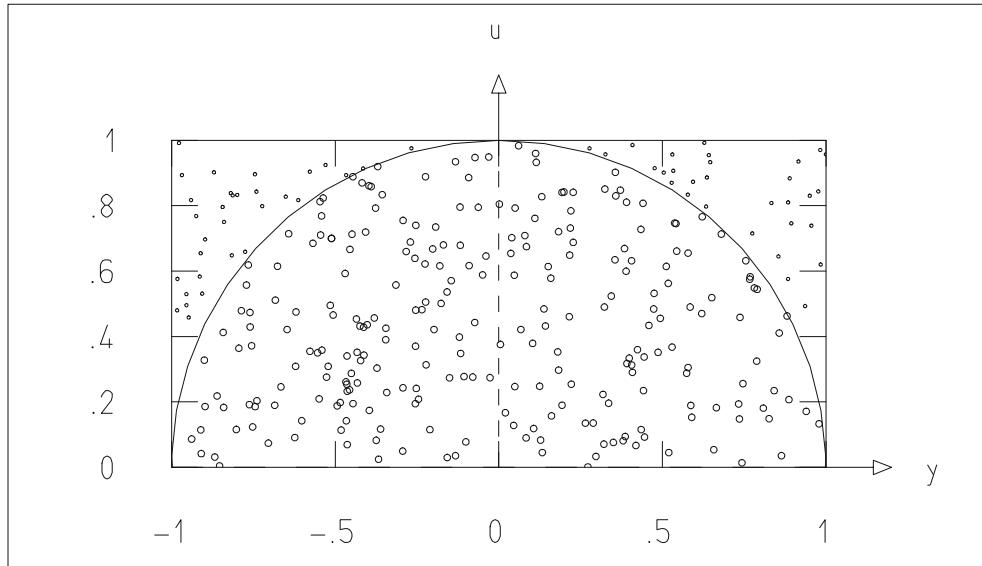
$$u = d , \quad d \geq g_{\max} , \quad (4.8.14)$$

which is greater than or equal to the maximum value of  $g(y)$  in that region. In the  $(y, u)$  plane this constant is described by the line  $u = d$ . Pairs of random numbers  $(y_i, u_i)$  uniformly distributed in the interval  $a \leq y_i \leq b$ ,  $0 \leq u_i \leq d$  correspond to a uniform distribution of points in the corresponding rectangle of the  $(y, u)$ -plane. If all of the points for which (4.8.12) holds are rejected, then only points under the curve  $u = g(y)$  remain. Figure 4.5 shows this situation for the Example 4.4. [It is clear that the technique also gives meaningful results if the function is not normalized to one. In Fig. 4.5 we have simply set  $g(y) = \sqrt{R^2 - y^2}$  and  $R = 1$ .]

For the transformation technique of Sect. 4.8.1, each random number  $y_i$  required only that exactly one random number  $x_i$  be generated from a uniform

distribution and that it be transformed according to (4.8.5). In the acceptance–rejection technique, pairs  $y_i, u_i$  must always be generated, and a considerable fraction of the numbers  $y_i$  – depending on the value of  $u_i$  according to (4.8.12) – are rejected. The probability for  $y_i$  to be accepted is

$$E = \frac{\int_a^b g(y) dy}{(b-a)d} . \quad (4.8.15)$$



**Fig. 4.5:** All the pairs  $(y_i, u_i)$  produced are marked as points in the  $(y, u)$ -plane. Points above the curve  $u = g(y)$  (*small points*) are rejected.

We can call  $E$  the *efficiency* of the procedure. If the interval  $a \leq y \leq b$  includes the entire allowed range of  $y$ , then the numerator of (4.8.15) is equal to unity, and one obtains

$$E = \frac{1}{(b-a)d} . \quad (4.8.16)$$

The numerator and denominator of (4.8.15) are simply the areas contained in the region  $a \leq y \leq b$  under the curves (4.8.13) and (4.8.14), respectively. One distributes points  $(y_i, u_i)$  uniformly under the curve (4.8.14) and rejects the random numbers  $y_i$  if the inequality (4.8.12) holds. The efficiency of the procedure is certainly higher if one uses as the upper curve not the constant (4.8.14) but rather a curve that is closer to  $g(y)$ .

With this in mind the acceptance–rejection technique can be stated in its general form:

- (i) One finds a probability density  $s(y)$  that is sufficiently simple that random numbers can be generated from it using the transformation method, and a constant  $c$  such that

$$g(y) \leq c \cdot s(y) \quad , \quad a < y < b \quad , \quad (4.8.17)$$

holds.

- (ii) One generates one random number  $y$  uniformly distributed in the interval  $a < y < b$  and a second random number  $u$  uniformly distributed in the interval  $0 < u < 1$ .
- (iii) One rejects  $y$

$$u \geq \frac{g(y)}{c \cdot s(y)} \quad . \quad (4.8.18)$$

After the points (ii) and (iii) have been repeated enough times, the resulting set of accepted random numbers  $y$  follows the probability density  $g(y)$ , since

$$P(y < y) = \int_a^y s(t) \frac{g(t)}{c \cdot s(t)} dt = \frac{1}{c} \int_a^y g(t) dt = \frac{1}{c} [G(y) - G(a)] \quad .$$

If the interval  $a \leq y \leq b$  includes the entire range of  $y$  for both  $g(y)$  as well as for  $s(y)$ , then one obtains an efficiency

$$E = \frac{1}{c} \quad . \quad (4.8.19)$$

**Example 4.5:** Semicircle distribution with the general acceptance–rejection method

One chooses for  $c \cdot s(y)$  the polygon

$$c \cdot s(y) = \begin{cases} 0, & y < -R \quad , \\ 3R/2 + y, & -R \leq y < -R/2 \quad , \\ R, & -R/2 \leq y < R/2 \quad , \\ 3R/2 - y, & R/2 \leq y < R \quad , \\ 0, & R \leq y \quad . \end{cases}$$

The efficiency is clearly

$$E = \frac{\pi R^2}{2} \cdot \frac{1}{2R^2 - R^2/4} = \frac{2\pi}{7}$$

in comparison to

$$E = \frac{\pi R^2}{2} \cdot \frac{1}{2R^2} = \frac{\pi}{4}$$

as in Example 4.4. ■

## 4.9 Generation of Normally Distributed Random Numbers

By far the most important distribution for data analysis is the normal distribution, which we will discuss in Sect. 5.7. We present here a program that can produce random numbers  $x_i$  following the standard normal distribution with the probability density

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} . \quad (4.9.1)$$

The corresponding distribution function  $F(x)$  can only be computed and inverted numerically (Appendix C). Therefore the simple transformation method of Sect. 4.8.1 cannot be used. The *polar method* by BOX and MULLER [5] described here combines in an elegant way acceptance–rejection with transformation. The algorithm consists of the following steps:

- (i) Generate two independent random numbers  $u_1, u_2$  from a uniform distribution between 0 and 1. Transform  $v_1 = 2u_1 - 1$ ,  $v_2 = 2u_2 - 1$ .
- (ii) Compute  $s = v_1^2 + v_2^2$ .
- (iii) If  $s \geq 1$ , return to step (i).
- (iv)  $x_1 = v_1 \sqrt{-(2/s) \ln s}$  and  $x_2 = v_2 \sqrt{-(2/s) \ln s}$  are two independent random numbers following the standard normal distribution.

The number pairs  $(v_1, v_2)$  obtained from step (i) are the Cartesian coordinates of a set of points uniformly distributed inside the unit circle. We can write them as  $v_1 = r \cos \theta$ ,  $v_2 = r \sin \theta$  using the polar coordinates  $r = \sqrt{s}$ ,  $\theta = \arctan(v_2/v_1)$ . The point  $(x_1, x_2)$  then has the Cartesian coordinates

$$x_1 = \cos \theta \sqrt{-2 \ln s} , \quad x_2 = \sin \theta \sqrt{-2 \ln s} .$$

We now ask for the probability

$$\begin{aligned} F(r) &= P(\sqrt{-2 \ln s} \leq r) = P(-2 \ln s \leq r^2) \\ &= P(s > e^{-r^2/2}) . \end{aligned}$$

Since  $s = r^2$  is by construction uniformly distributed between 0 and 1, one has

$$F(r) = P(s > e^{-r^2/2}) = 1 - e^{-r^2/2} .$$

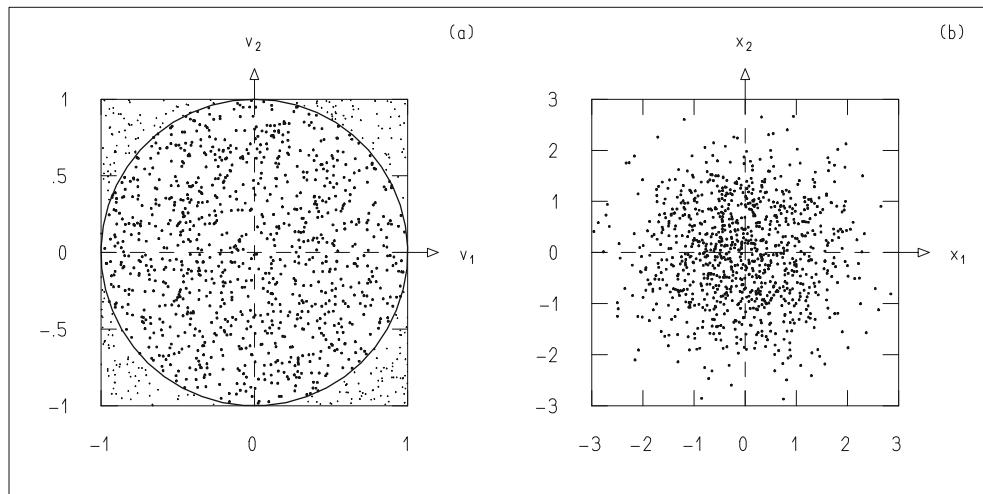
The probability density of  $r$  is

$$f(r) = \frac{dF(r)}{dr} = r e^{-r^2/2} .$$

The joint distribution function of  $x_1$  and  $x_2$ ,

$$\begin{aligned}
 F(x_1, x_2) &= P(x_1 \leq x_1, x_2 \leq x_2) = P(r \cos \theta \leq x_1, r \sin \theta \leq x_2) \\
 &= \frac{1}{2\pi} \int \int_{(x_1 < x_1, x_2 < x_2)} r e^{-r^2/2} dr d\varphi \\
 &= \frac{1}{2\pi} \int \int_{(x_1 < x_1, x_2 < x_2)} e^{-(x_1^2 + x_2^2)/2} dx dy \\
 &= \left( \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x_1} e^{-x_1^2/2} dx_1 \right) \left( \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x_2} e^{-x_2^2/2} dx_2 \right) ,
 \end{aligned}$$

is the product of two distribution functions of the standard normal distribution. The procedure is implemented in the method `DatanRandom.standard-Normal` and illustrated in Fig. 4.6.



**Fig. 4.6:** Illustration of the Box–Muller procedure. (a) Number pairs  $(v_1, v_2)$  are generated that uniformly populate the square. Those pairs are then rejected that do not lie inside the unit circle (marked by *small points*). (b) This is followed by the transformation  $(v_1, v_2) \rightarrow (x_1, x_2)$ .

Many other procedures are described in the literature for the generation of normally distributed random numbers. They are to a certain extent more efficient, but are generally more difficult to program than the BOX–MULLER procedure.

## 4.10 Generation of Random Numbers According to a Multivariate Normal Distribution

The probability density of a multivariate normal distribution of  $n$  variables  $\mathbf{x} = (x_1, x_2, \dots, x_n)$  is according to (5.10.1)

$$\phi(\mathbf{x}) = k \exp \left\{ -\frac{1}{2} (\mathbf{x} - \mathbf{a})^T B (\mathbf{x} - \mathbf{a}) \right\} .$$

Here  $\mathbf{a}$  is the vector of expectation values and  $B = C^{-1}$  is the inverse of the positive-definite symmetric covariance matrix. With the Cholesky decomposition  $B = D^T D$  and the substitution  $\mathbf{u} = D(\mathbf{x} - \mathbf{a})$  the exponent takes on the simple form

$$-\frac{1}{2} \mathbf{u}^T \mathbf{u} = -\frac{1}{2} (u_1^2 + u_2^2 + \cdots + u_n^2) .$$

Thus the elements  $u_i$  of the vectors  $\mathbf{u}$  follow independent standard normal distributions [cf. (5.10.9)]. One obtains vectors  $\mathbf{x}$  of random numbers by first forming a vector  $\mathbf{u}$  of elements  $u_i$  which follow the standard normal distribution and then performing the transformation

$$\mathbf{x} = D^{-1} \mathbf{u} + \mathbf{a} .$$

This procedure is implemented in the method `DatanRandom.multivariateNormal`.

## 4.11 The Monte Carlo Method for Integration

It follows directly from its construction that the acceptance–rejection technique, Sect. 4.8.2, provides a very simple method for numerical integration. If  $N$  pairs of random numbers  $(y_i, u_i)$ ,  $i = 1, 2, \dots, N$  are generated according to the prescription of the general acceptance–rejection technique, and if  $N - n$  of them are rejected because they fulfill condition (4.8.18), then the numbers  $N$  (or  $n$ ) are proportional to the areas under the curves  $c \cdot s(y)$  (or  $g(y)$ ), at least in the limit of large  $N$ , i.e.,

$$\frac{\int_a^b g(y) dy}{c \int_a^b s(y) dy} = \lim_{N \rightarrow \infty} \frac{n}{N} . \quad (4.11.1)$$

Since the function  $s(y)$  is chosen to be particularly simple [in the simplest case one has  $s(y) = 1/(b - a)$ ], the ratio  $n/N$  is a direct measure of the value of the integral

$$I = \int_a^b g(y) dy = \left( \lim_{N \rightarrow \infty} \frac{n}{N} \right) c \int_a^b s(y) dy . \quad (4.11.2)$$

Here the integrand  $g(y)$  does not necessarily have to be normalized, i.e., one does not need to require

$$\int_{-\infty}^{\infty} g(y) dy = 1$$

as long as  $c$  is chosen such that (4.8.17) is fulfilled.

**Example 4.6:** Computation of  $\pi$ 

Referring to Example 4.4 we compute the integral using (4.8.11) with  $R = 1$ :

$$I = \int_0^1 g(y) dy = \pi/4 .$$

Choosing  $s(y) = 1$  and  $c = 1$  we obtain

$$I = \lim_{N \rightarrow \infty} \frac{n}{N} .$$

We expect that when  $N$  points are distributed according to a uniform distribution in the square  $0 \leq y \leq 1$ ,  $0 \leq u \leq 1$ , and when  $n$  of them lie inside the unit circle, then the ratio  $n/N$  approaches the value  $I = \pi/4$  in the limit  $N \rightarrow \infty$ . Table 4.2 shows the results for various values of  $n$  and for various sequences of random numbers. The exact value of  $n/N$  clearly depends on the particular sequence. In Sect. 6.8 we will determine that the typical fluctuations of the number  $n$  are approximately  $\Delta n = \sqrt{n}$ . Therefore one has for the relative precision for the determination of the integral (4.11.2)

$$\frac{\Delta I}{I} = \frac{\Delta n}{n} = \frac{1}{\sqrt{n}} . \quad (4.11.3)$$

We expect therefore in the columns of Table 4.2 to find the value of  $\pi$  with precisions of 10, 1, and 0.1 %. We find in fact in the three columns fluctuations in the first, second, and third places after the decimal point. ■

**Table 4.2:** Numerical values of  $4n/N$  for various values of  $n$ . The entries in the columns correspond to various sequences of random numbers.

| 4n/N       |            |            |
|------------|------------|------------|
| $n = 10^2$ | $n = 10^4$ | $n = 10^6$ |
| 3.419      | 3.122      | 3.141      |
| 3.150      | 3.145      | 3.143      |
| 3.279      | 3.159      | 3.144      |
| 3.419      | 3.130      | 3.143      |

The Monte Carlo method of integration can now be implemented by a very simple program. For integration of single variable functions it is usually better to use other numerical techniques for reasons of computing time. For integrals with many variables, however, the Monte Carlo method is more straightforward and often faster as well.

## 4.12 The Monte Carlo Method for Simulation

Many real situations that are determined by statistical processes can be simulated in a computer with the aid of random numbers. Examples are automobile traffic in a given system of streets or the behavior of neutrons in a nuclear reactor. The Monte Carlo method was originally developed for the latter problem by VON NEUMANN and ULAM. A change of the parameters of the distributions corresponds then to a change in the actual situation. In this way the effect of additional streets or changes in the reactor can be investigated without having to undertake costly and time consuming changes in the real system. Not only processes of interest following statistical laws can be simulated with the Monte Carlo method, but also the measurement errors which occur in every measurement.

**Example 4.7:** Simulation of measurement errors of points on a line

We consider a line in the  $(t, y)$ -plane. It is described by the equation

$$y = at + b \quad . \quad (4.12.1)$$

If we choose discrete values of  $t$

$$t_0 \quad , \quad t_1 = t_0 + \Delta t \quad , \quad t_2 = t_0 + 2\Delta t \quad , \quad \dots \quad , \quad (4.12.2)$$

then they correspond to values of  $y$

$$y_i = at_i + b \quad , \quad i = 0, 1, \dots, n - 1 \quad . \quad (4.12.3)$$

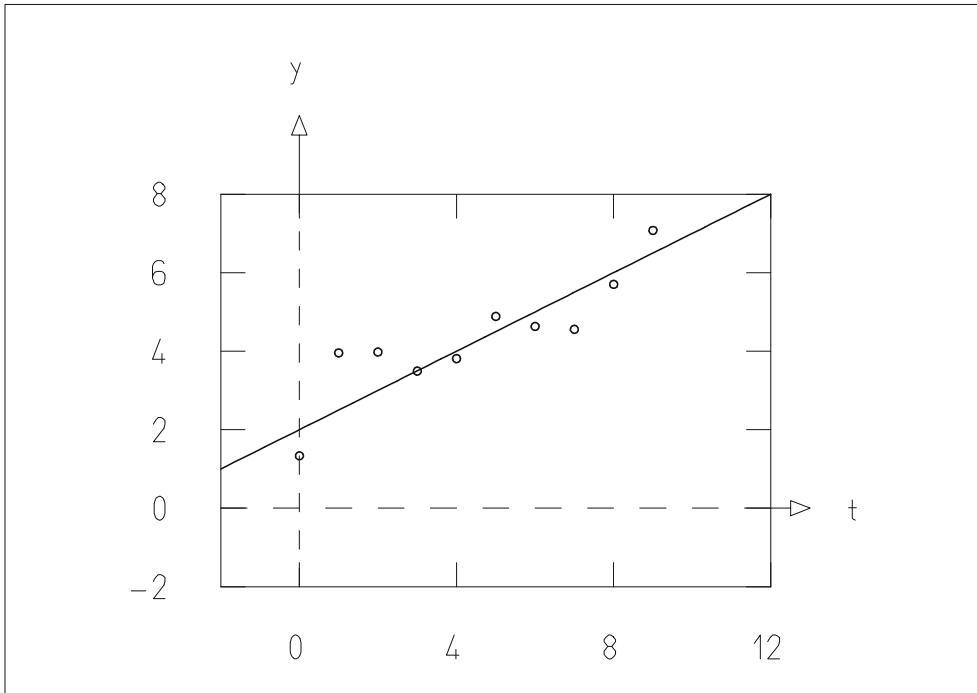
We assume that the values  $t_0, t_1, \dots$  of the “controlled variable”  $t$  can be set without error. Because of measurement errors, however, instead of the values  $y_i$ , one obtains different values

$$y'_i = y_i + \varepsilon_i \quad . \quad (4.12.4)$$

Here  $\varepsilon_i$  are the measurement errors, which follow a normal distribution with mean of zero and standard deviation  $\sigma_y$  (cf. Sect. 5.7). The method DatanRandom.line generates number pairs  $(t_i, y'_i)$ . Figure 4.7 as an example displays 10 simulated points. ■

**Example 4.8:** Generation of decay times for a mixture of two different radioactive substances

At time  $t = 0$  a source consists of  $N$  radioactive nuclei of which  $aN$  decay with a lifetime  $\tau_1$  and  $(a - 1)N$  with a mean lifetime  $\tau_2$ , with  $0 \leq a \leq 1$ . Random numbers for two different problems must be used in the simulation the decay times occurring: for the choice of the type of nucleus and for the determination of the decay time of the nucleus chosen, cf. (4.8.9). The method DatanRandom.radio implements this example. ■



**Fig. 4.7:** Line in the  $(t, y)$ -plane and simulated measured values with errors in  $y$ .

## 4.13 Java Classes and Example Programs

### Java Class for the Generation of Random Numbers

DatanRandom contains methods for the generation of random numbers following various distributions, in particular DatanRandom.ecuy for the uniform, DatanRandom.standard-Normal for the standard normal, and DatanRandom.mvNormal for the multivariate normal Distribution. Further methods are used to illustrate a simple MLC generator or to demonstrate the following examples.

**Example Program 4.1:** The class E1Random demonstrates the generation of random numbers

One can choose interactively between three generators. After clicking on Go 100 random numbers are generated and displayed. The seeds before and after generation are shown and can be changed interactively.

**Example Program 4.2:** The class E2Random demonstrates the generation of measurement points, scattering about a straight line

Example 4.7 is realized. Parameter input is interactive, output both numerical and graphical.

**Example Program 4.3:** The class E3Random demonstrates the simulation of decay times

Example 4.8 is realized. Parameter input is interactive, output in form of a histogram.

**Example Program 4.4:** The class E4Random demonstrates the generation of random numbers from a multivariate normal distribution

The procedure of Sect. 4.10 is realized for the case of two variables. Parameter input is interactive. The generated number pairs are displayed numerically.

## 5. Some Important Distributions and Theorems

We shall now discuss in detail some specific distributions. This chapter could therefore be regarded as a collection of examples. These distributions, however, are of great practical importance and are often encountered in many applications. Moreover, their study will lead us to a number of important theorems.

### 5.1 The Binomial and Multinomial Distributions

Consider an experiment having only two possible outcomes. The sample space can therefore be expressed as

$$E = A + \bar{A} \quad (5.1.1)$$

with the probabilities

$$P(A) = p \quad , \quad P(\bar{A}) = 1 - p = q \quad . \quad (5.1.2)$$

One now performs  $n$  independent trials of the experiment defined by (5.1.1). One wishes to find the probability distribution for the quantity  $x = \sum_{i=1}^n x_i$ , where one has  $x_i = 1$  (or 0) when  $A$  (or  $\bar{A}$ ) occurs as the result of the  $i$ th experiment.

The probability that the first  $k$  trials result in  $A$  and all of the rest in  $\bar{A}$  is, using Eq. (2.3.8),

$$p^k q^{n-k} \quad .$$

Using the rules of combinatorics, the event “outcome  $A$   $k$  times in  $n$  trials” occurs in  $\binom{n}{k} = \frac{n!}{k!(n-k)!}$  different ways, according to the order of the occurrences of  $A$  and  $\bar{A}$  (see Appendix B). The probability of this event is therefore

$$P(k) = W_k^n = \binom{n}{k} p^k q^{n-k} . \quad (5.1.3)$$

We are interested in the mean value and variance of  $\mathbf{x}$ . We first find these quantities for the variable  $x_i$  of an individual event. According to (3.3.2) one has

$$E(x_i) = 1 \cdot p + 0 \cdot q \quad (5.1.4)$$

and

$$\begin{aligned} \sigma^2(x_i) &= E\{(x_i - p)^2\} = (1-p)^2 p + (0-p)^2 q , \\ \sigma^2(x_i) &= pq . \end{aligned} \quad (5.1.5)$$

From the generalization of (3.5.3) for  $\mathbf{x} = \sum x_i$  it follows that

$$E(\mathbf{x}) = \sum_{i=1}^n p = np , \quad (5.1.6)$$

and from (3.5.10), since all of the covariances vanish because the  $x_i$  are independent, one has

$$\sigma^2(\mathbf{x}) = npq . \quad (5.1.7)$$

Figure 5.1 shows the distribution  $W_k^n$  for various  $n$  and for fixed  $p$ , and Fig. 5.2 shows it for fixed  $n$  and various values of  $p$ . Finally in Fig. 5.3  $n$  and  $p$  are both varied but the product  $np$  is held constant. The figures will help us to see relationships between the *binomial distribution* (5.1.3) and other distributions.

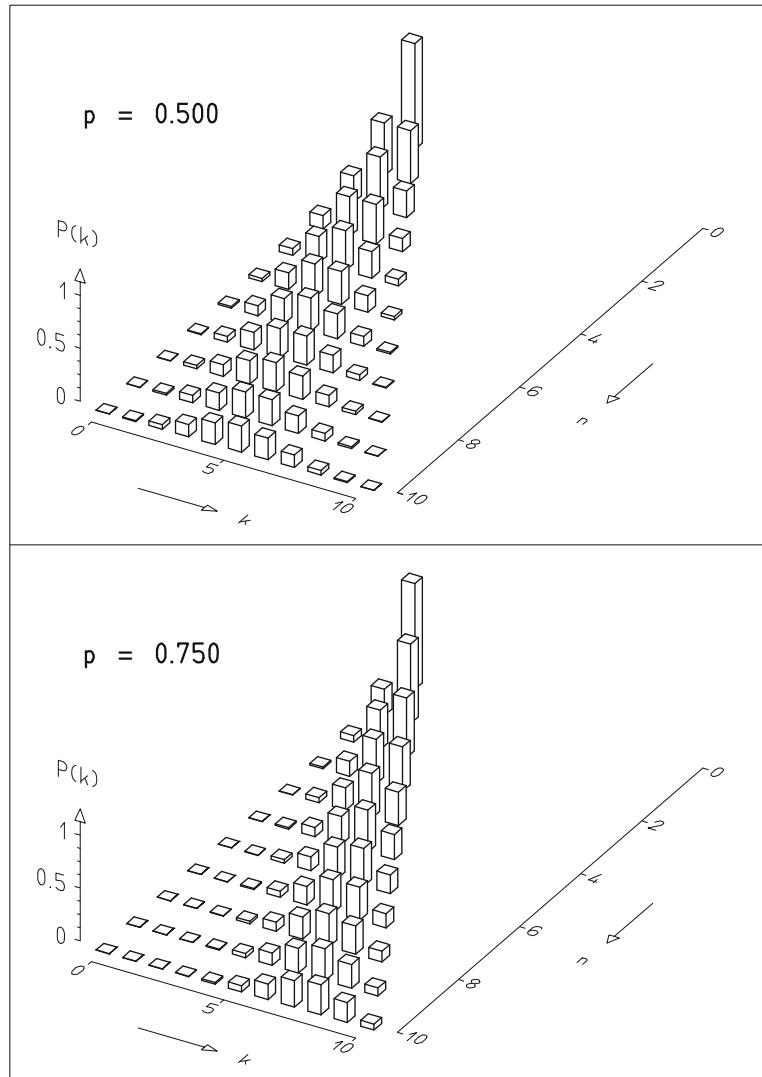
A logical extension of the binomial distribution deals with experiments where more than two different outcomes are possible. Equation (5.1.1) is then replaced by

$$E = A_1 + A_2 + \cdots + A_\ell . \quad (5.1.8)$$

Let the probability for the outcome  $A_j$  be

$$P(A_j) = p_j , \quad \sum_{j=1}^{\ell} p_j = 1 . \quad (5.1.9)$$

We consider again  $n$  trials and ask for the probability that the outcome  $A_j$  occurs  $k_j$  times. This is given by



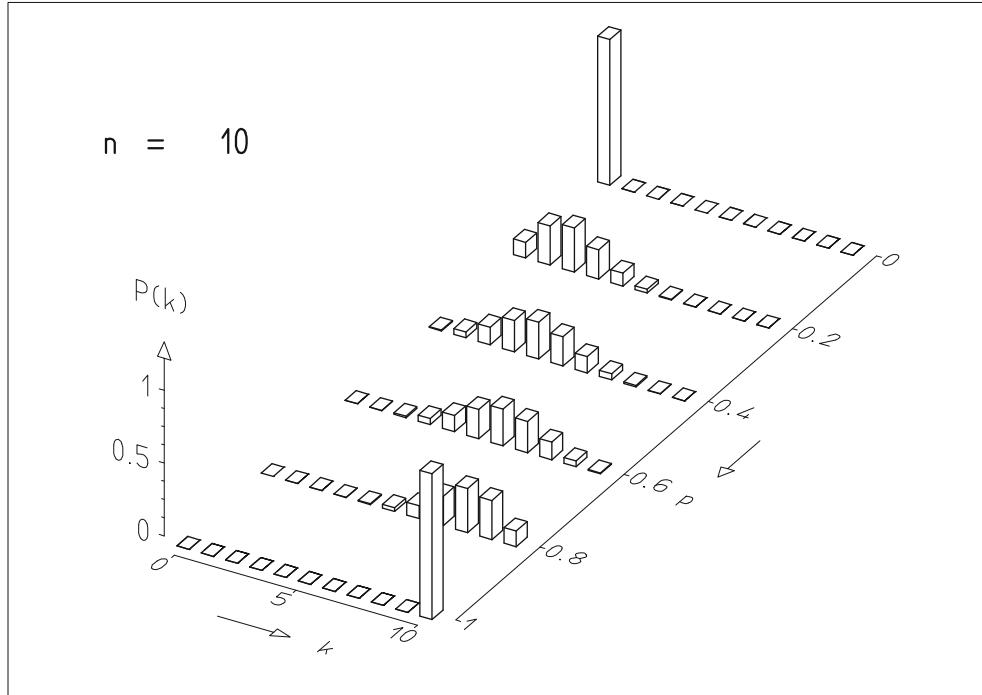
**Fig.5.1:** Binomial distribution for various values of  $n$ , with  $p$  fixed.

$$W_{(k_1, k_2, \dots, k_\ell)}^n = \frac{n!}{\prod_{j=1}^\ell k_j!} \prod_{j=1}^\ell p_j^{k_j} , \quad \sum_{j=1}^\ell k_j = n . \quad (5.1.10)$$

The proof is left to the reader. The probability distribution (5.1.10) is called the *multinomial distribution*.

We can define a random variable  $x_{ij}$  that takes on the value 1 when the  $i$ th trial leads to the outcome  $A_j$ , and is zero otherwise. In addition define  $x_j = \sum_{i=1}^n x_{ij}$ . The expectation value of  $x_j$  is then

$$E(x_j) = \hat{x}_j = np_j . \quad (5.1.11)$$



**Fig. 5.2:** Binomial distribution for various values of  $p$ , with  $n$  fixed.

The elements of the covariance matrix of the  $\mathbf{x}_j$  are

$$c_{ij} = np_i(\delta_{ij} - p_j) \quad . \quad (5.1.12)$$

The off-diagonal elements are clearly not zero. This was to be expected, since from Eq. (5.1.9) the variables  $\mathbf{x}_j$  are not independent.

## 5.2 Frequency: The Law of Large Numbers

Usually the probabilities for the different types of events, e.g.,  $p_j$  in the case of the multinomial distribution, are not known but have to be obtained from experiment. One first measures the *frequency* of the events in  $n$  experiments,

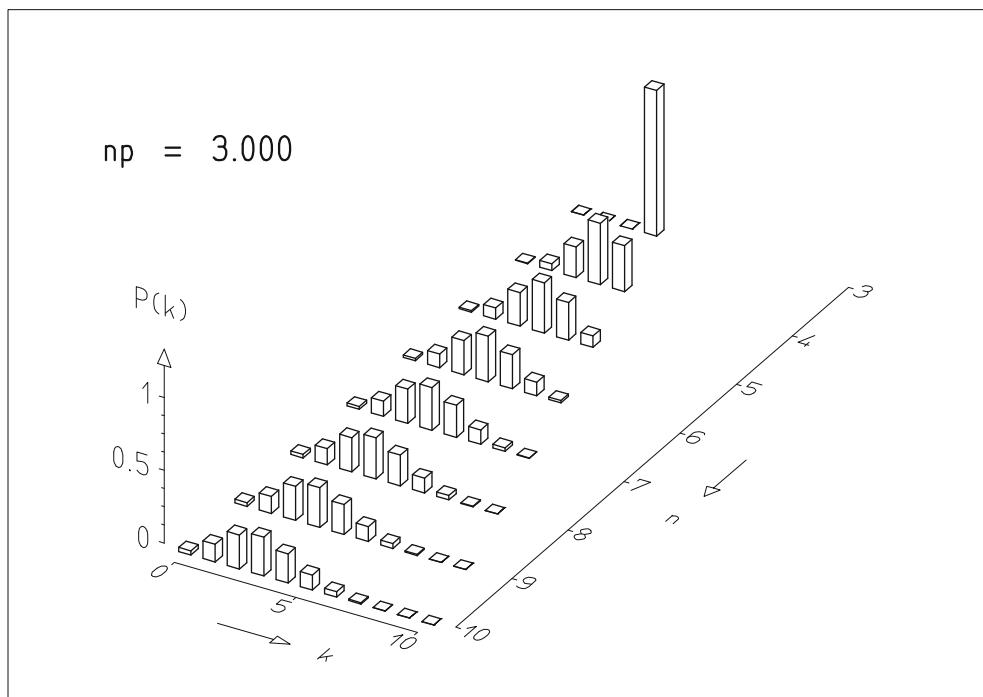
$$\mathbf{h}_j = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_{ij} = \frac{1}{n} \mathbf{x}_j \quad . \quad (5.2.1)$$

Unlike the probability, the frequency is a random quantity, since it depends on the outcomes of the  $n$  individual experiments. By use of (5.1.11), (5.1.12), and (3.3.15) we obtain

$$E(\mathbf{h}_j) = \widehat{\mathbf{h}}_j = E\left(\frac{\mathbf{x}_j}{n}\right) = p_j \quad (5.2.2)$$

and

$$\sigma^2(h_j) = \sigma^2\left(\frac{x_j}{n}\right) = \frac{1}{n^2}\sigma^2(x_j) = \frac{1}{n}p_j(1-p_j) \quad . \quad (5.2.3)$$



**Fig. 5.3:** Binomial distribution for various values of  $n$  but fixed product  $np$ . For higher values of  $n$  the distribution changes very little.

The product  $p_j(1 - p_j)$  in Eq. (5.2.3) is at most  $1/4$ . One sees that the expectation value of the frequency of an event is exactly equal to the probability that the event will occur, and that the variance of frequency about this expectation value can be made arbitrarily small as the number of trials increases. Since  $pq$  is at most  $1/4$ , one can always say that the standard deviation of  $h_j$  is at most  $1/\sqrt{n}$ . This property of the frequency is known as the *law of large numbers*. It is clearly the reason for the frequency definition of probability given by Eq. (2.2.1).

Frequently the purpose of an experimental investigation is to determine the probability for the occurrence of a certain type of event. According to (5.2.2) we can use the frequency as an approximation of the probability. The square of the error of this approximation is then inversely proportional to the number of individual experiments. This kind of error, which originates from the fact that only a finite number of experiments can be performed, is called the *statistical error*. It is of prime importance for applications that are concerned with the counting of individual events, e.g., nuclear particles

passing through a counter, animals with certain traits in heredity experiments, defective items in quality control, and so forth.

**Example 5.1:** Statistical error

Suppose it is known from earlier experiments that a fraction  $R \approx 1/200$  of a sample of fruit flies (*Drosophila*) develop a certain property  $A$  if exposed to a given dose of X-rays. An experiment is planned to determine the fraction  $R$  with an accuracy of 1%. How large must the original sample be in order to achieve this accuracy?

We use Eq. (5.2.3) and find  $p_j = 0.005$ ,  $(1 - p_j) \approx 1$ . We must now choose  $n$  such that  $\sigma(h_j)/h_j = 200\sigma(h_j) = 0.01$ . This gives  $\sigma(h_j) = 0.00005$  and  $\sigma^2(h_j) = 0.25 \times 10^{-8}$ . Equation (5.2.3) gives

$$0.25 \times 10^{-8} = \frac{1}{n} \times 0.005$$

and therefore

$$n = 2 \times 10^6 .$$

A total of two million fruit flies would have to be used. This is practically impossible. To determine the fraction  $R$  with an accuracy of 10% would require 20 000 flies. ■

### 5.3 The Hypergeometric Distribution

Although we shall rigorously introduce the concept of random sampling at a later point, we will now discuss a typical problem of sampling. We consider a container – we shall not break with the habit of mathematicians of calling such a container an urn – with  $K$  white and  $L = N - K$  black balls. We want to determine the probability that in drawing  $n$  balls (without replacing them) we will find exactly  $k$  white and  $l = n - k$  black ones. The problem is rendered difficult by the fact that the drawing of a ball of a particular color changes the ratio of white and black balls and therefore influences the outcome of the next draw. One clearly has  $\binom{N}{n}$  equally likely ways to choose  $n$  out of  $N$  balls. The probability that one of these possibilities will occur is therefore  $1/\binom{N}{n}$ . There are  $\binom{K}{k}$  ways to choose  $k$  of the  $K$  white balls, and  $\binom{L}{\ell}$  ways to choose  $\ell$  of the  $L$  black ones. The required probability is therefore

$$W_k = \frac{\binom{K}{k} \binom{L}{\ell}}{\binom{N}{n}} . \quad (5.3.1)$$

As in Sect. 5.1 we define the random variable  $\mathbf{x} = \sum_{i=1}^n \mathbf{x}_i$  with  $\mathbf{x}_i = 1$  when the  $i$ th draw results in a black ball, and  $\mathbf{x}_i = 0$  otherwise. (In other words, we define  $k$  as the random variable  $\mathbf{x}$ .)

To compute the expectation values of  $\mathbf{x}$  we cannot simply add the expectation values of the  $x_i$ , since these are no longer independent. Instead we must return to the definition (3.3.2),

$$\begin{aligned} E(\mathbf{x}) &= \frac{1}{\binom{N}{n}} \sum_{i=1}^n i \binom{K}{i} \binom{N-K}{n-i} \\ &= \frac{(N-n)!n!}{N!} \sum_{i=1}^n \frac{i K!(N-K)!}{i!(K-i)!(n-i)!(N-K-n+i)!} \\ &= \frac{n(n-1)!(N-n)!}{N(N-1)!} \sum_{i=1}^n \frac{K!}{(i-1)!(K-1-(i-1))!} \\ &\quad \times \frac{(N-K)!}{(n-1-(i-1))!(N-K-(n-1)+(i-1))!} . \end{aligned}$$

If we substitute  $i - 1 = j$ , this gives

$$\begin{aligned} E(\mathbf{x}) &= n \frac{K}{N} \frac{(n-1)!(N-n)!}{(N-1)!} \\ &\quad \times \sum_{j=0}^{n-1} \frac{(K-1)!(N-K)!}{j!(K-1-j)!(n-1-j)!(N-K-(n-1)+j)!} \\ &= n \frac{K}{N} \frac{1}{\binom{N-1}{n-1}} \sum_{j=0}^{n-1} \binom{K-1}{j} \binom{N-K}{n-1-j} . \end{aligned}$$

With Eq. (B.5) we obtain

$$E(\mathbf{x}) = n \frac{K}{N} . \quad (5.3.2)$$

The calculation of the variance follows along the same lines but is rather lengthy. The result is

$$\sigma^2(\mathbf{x}) = \frac{n K (N-K)(N-n)}{N^2(N-1)} . \quad (5.3.3)$$

Figures 5.4 and 5.5 depict several examples of the distribution. If  $n \ll N$ , then drawing a white ball has little influence on the probabilities for the next draw. We therefore expect that in this case  $W_k$  behaves in a manner similar to a binomial distribution with  $p = \frac{K}{N}$  and  $q = \frac{N-K}{N}$ . This is also made clear by the similarity of Figs. 5.5 and 5.1. One obtains in fact the same expectation value,

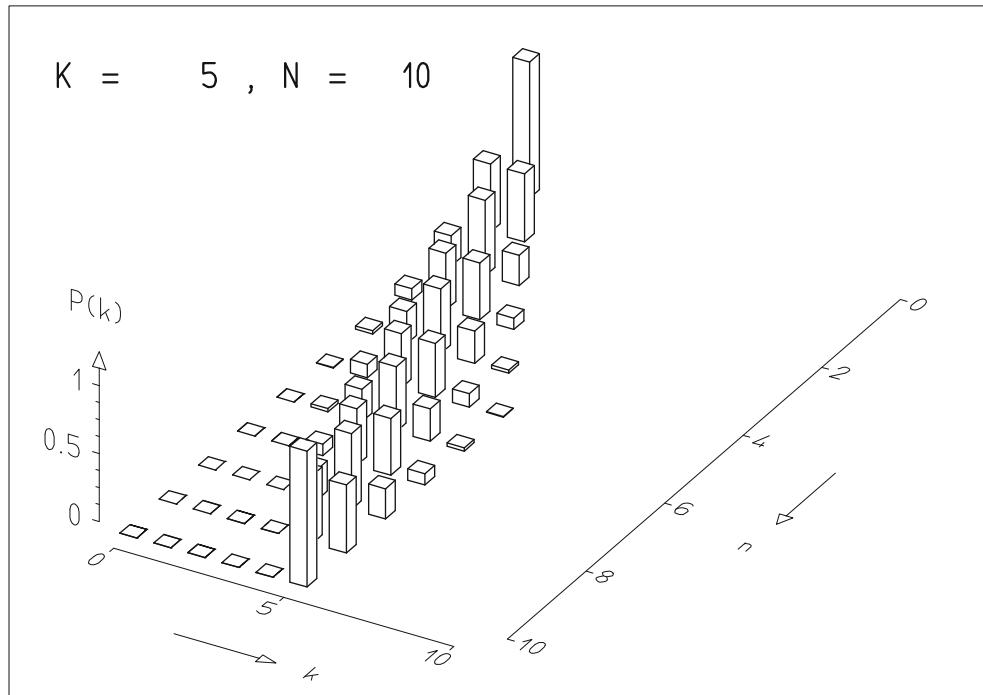
$$E(\mathbf{x}) = n \frac{K}{N} = np ,$$

as for the binomial distribution. The variance is then

$$\sigma^2(x) = \frac{npq(N-n)}{N-1} ,$$

which for the case  $n \ll N$  becomes

$$\sigma^2 = npq .$$



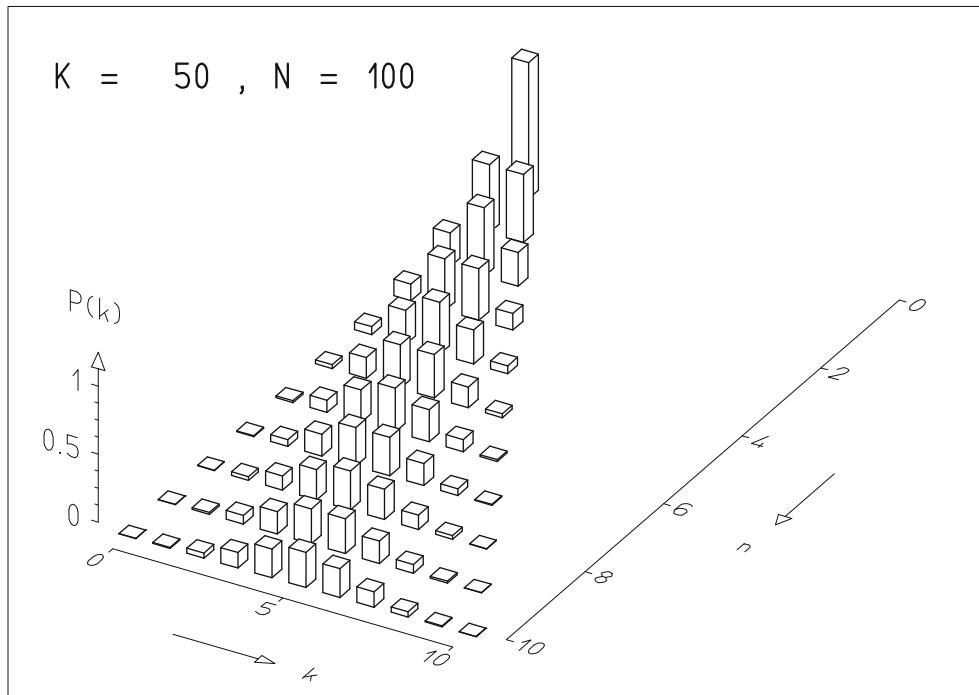
**Fig. 5.4:** Hypergeometric distribution for various values of  $n$  and small values of  $K$  and  $N$ .

There are many applications of the hypergeometric distribution. Opinion polls, quality controls, and so forth are all based on the experimental scheme of taking (polling) an object without replacement back into the original sample or *population*. The distribution can be generalized in two ways. First we can of course consider more properties instead of just two (white and black balls). This leads us to a similar transition as the one from the binomial to the multinomial distribution. The original sample (population) contains  $N$  elements each of which possesses one of  $l$  properties,

$$N = N_1 + N_2 + \cdots + N_\ell .$$

The probability that  $n$  draws (without replacement) will be composed as

$$n = n_1 + n_2 + \cdots + n_\ell$$



**Fig. 5.5:** Hypergeometric distribution for various values of  $n$  and for large values of  $K$  and  $N$ .

is, in analogy to Eq. (5.3.1),

$$W_{n_1, n_2, \dots, n_\ell} = \frac{\binom{N_1}{n_1} \binom{N_2}{n_2} \cdots \binom{N_\ell}{n_\ell}}{\binom{N}{n}} . \quad (5.3.4)$$

Another extension of the hypergeometric distribution is obtained in the following way. We saw earlier that consecutive drawings ceased to be independent because the balls were not replaced. If now each time we draw a ball of one type we place more balls of that type back in the urn, this dependence can be enhanced. One then obtains the *Polya distribution*. It is of importance in the study of epidemic diseases, where the appearance of a case of the disease enhances the probability of future cases.

**Example 5.2:** Application of the hypergeometric distribution for determination of zoological populations

From a pond  $K$  fish are taken and marked. They are then returned to the pond. After a short while  $n$  fish are caught,  $k$  of which are found to be marked. Before the second time that the fish are taken, the pond contains a total of  $N$  fish, of which  $K$  are marked. The probability of finding  $k$  marked out of  $n$  removed fish is given by Eq. (5.3.1). We will return to this problem in Example 7.3. ■

## 5.4 The Poisson Distribution

Looking at Fig. 5.3 it appears that if  $n$  tends to infinity, but at the same time  $np = \lambda$  is kept constant, the binomial distribution approaches a certain fixed distribution. We rewrite Eq. (5.1.3) as

$$\begin{aligned} W_k^n &= \binom{n}{k} p^k q^{n-k} \\ &= \frac{n!}{k!(n-k)!} \left(\frac{\lambda}{n}\right)^k \frac{\left(1 - \frac{\lambda}{n}\right)^n}{\left(1 - \frac{\lambda}{n}\right)^k} \\ &= \frac{\lambda^k}{k!} \frac{n(n-1)(n-2)\cdots(n-k+1)}{n^k} \frac{\left(1 - \frac{\lambda}{n}\right)^n}{\left(1 - \frac{\lambda}{n}\right)^k} \\ &= \frac{\lambda^k}{k!} \left(1 - \frac{\lambda}{n}\right)^n \frac{\left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right)\cdots\left(1 - \frac{k-1}{n}\right)}{\left(1 - \frac{\lambda}{n}\right)^k}. \end{aligned}$$

In the limiting case all of the many individual factors of the term on the right approach unity. In addition one has

$$\lim_{n \rightarrow \infty} \left(1 - \frac{\lambda}{n}\right)^n = e^{-\lambda},$$

so that in the limit one has

$$\lim_{n \rightarrow \infty} W_k^n = f(k) = \frac{\lambda^k}{k!} e^{-\lambda}. \quad (5.4.1)$$

The quantity  $f(k)$  is the probability of the *Poisson distribution*. It is plotted in Fig. 5.6 for various values of  $\lambda$ . As is the case for the other distributions we have encountered so far, the Poisson distribution is only defined for integer values of  $k$ .

The distribution satisfies the requirement that the total probability is equal to unity,

$$\begin{aligned} \sum_{k=0}^{\infty} f(k) &= \sum_{k=0}^{\infty} \frac{e^{-\lambda} \lambda^k}{k!} \\ &= e^{-\lambda} \left(1 + \lambda + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} + \dots\right) \\ &= e^{-\lambda} e^{\lambda}, \end{aligned}$$

$$\sum_{k=0}^{\infty} f(k) = 1 . \quad (5.4.2)$$

The expression in parentheses is in fact the Taylor expansion of  $e^\lambda$ .

We now want to determine the mean, variance, and skewness of the Poisson distribution. The definition (3.3.2) gives

$$\begin{aligned} E(k) &= \sum_{k=0}^{\infty} k \frac{\lambda^k}{k!} e^{-\lambda} = \sum_{k=1}^{\infty} k \frac{\lambda^k}{k!} e^{-\lambda} \\ &= \sum_{k=1}^{\infty} \frac{\lambda \lambda^{k-1}}{(k-1)!} e^{-\lambda} = \lambda \sum_{j=0}^{\infty} \frac{\lambda^j}{j!} e^{-\lambda} \end{aligned}$$

and using this with (5.4.2),

$$E(k) = \lambda . \quad (5.4.3)$$

We would now like to find  $E(k^2)$ . One obtains in a corresponding way

$$\begin{aligned} E(k^2) &= \sum_{k=1}^{\infty} k^2 \frac{\lambda^k}{k!} e^{-\lambda} = \lambda \sum_{k=1}^{\infty} k \frac{\lambda^{k-1}}{(k-1)!} e^{-\lambda} \\ &= \lambda \sum_{j=0}^{\infty} (j+1) \frac{\lambda^j}{j!} e^{-\lambda} = \lambda \left( \sum_{j=0}^{\infty} j \frac{\lambda^j}{j!} e^{-\lambda} + 1 \right) \end{aligned}$$

and therefore

$$E(k^2) = \lambda(\lambda + 1) . \quad (5.4.4)$$

We will use Eqs. (5.4.3) and (5.4.4) to compute the variance. According to Eq. (3.3.16) one has

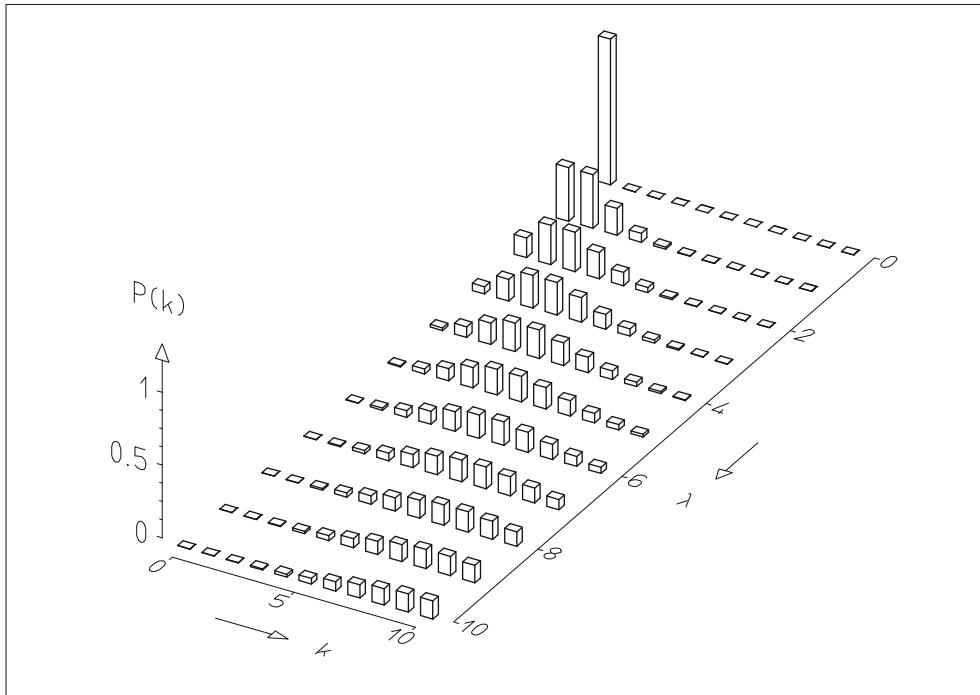
$$\sigma^2(k) = E(k^2) - \{E(k)\}^2 = \lambda(\lambda + 1) - \lambda^2 \quad (5.4.5)$$

or

$$\sigma^2(k) = \lambda . \quad (5.4.6)$$

We now consider the skewness (3.3.13) of the Poisson distribution. Following Sect. 3.3 we easily find that

$$\mu_3 = E\{(k - \hat{k})^3\} = \lambda .$$



**Fig. 5.6:** Poisson distribution for various values of  $\lambda$ .

The skewness (3.3.13) is then

$$\gamma = \frac{\mu_3}{\sigma^3} = \frac{\lambda}{\lambda^{\frac{3}{2}}} = \lambda^{-\frac{1}{2}} , \quad (5.4.7)$$

that is, the Poisson distribution becomes increasingly symmetric as  $\lambda$  increases. Figure 5.6 shows the distribution for various values of  $\lambda$ . In particular the distribution with  $\lambda = 3$  should be compared with Fig. 5.3.

We have obtained the Poisson distribution from the binomial distribution with large  $n$  but constant  $\lambda = np$ , i.e., small  $p$ . We therefore expect it to apply to processes in which a large number of events occur but of which only very few have a certain property of interest to us (i.e., a large number of “trials” but few “successes”).

### Example 5.3: Poisson distribution and independence of radioactive decays

We consider a radioactive nucleus with mean lifetime  $\tau$  and observe it for a time  $T \ll \tau$ . The probability that it decays within this time interval is  $W \ll 1$ . We break the observation time  $T$  into  $n$  smaller time intervals of length  $t$ , so that  $T = nt$ . The probability for the nucleus to decay in a particular time interval is  $p \approx W/n$ . We now observe a radioactive source containing  $N$  nuclei which decay independently from each other for a total time  $T$ , and detect  $a_1$

decays in time interval 1,  $a_2$  decays in interval 2, etc. Let  $h(k)$  be the frequency of decays observed in the interval  $k$ , with  $(k = 0, 1, \dots)$ . That is, if  $n_k$  is the number of intervals with  $k$  decays, then  $h(k) = n_k/n$ . In the limit  $N \rightarrow \infty$  and for large  $n$  the frequency distribution  $h(k)$  becomes the probability distribution (5.4.1). The statistical nature of radioactive decay was established in this way in a famous experiment by RUTHERFORD and GEIGER. ■

Similarly, the frequency of finding  $k$  stars per element of the celestial sphere or  $k$  raisins per volume element of a fruit cake is distributed according to the Poisson law, but not, however, the frequency of finding  $k$  animals of a given species per element of area, at least if these animals live in herds, since in this case the assumption of independence is not fulfilled.

As a quantitative example of the Poisson distribution many textbooks discuss the number of Prussian cavalrymen killed during a period of 20 years by horse kicks, an example originally due to VON BORTKIEWICZ [6]. We prefer to turn our attention to a somewhat less macabre example taken from a lecture of DE SOLLA PRICE [7].

#### **Example 5.4:** Poisson distribution and the independence of scientific discoveries

The author first constructs the model of an apple tree with 1000 apples and 1000 pickers with blindfolded eyes who each try at the same time to pick an apple. Since we are dealing with a model, they do not hinder each other but it can happen that two or several of them will attempt to pick the same apple at the same time. The number of apples grabbed simultaneously by  $k$  people ( $k = 0, 1, 2, \dots$ ) follows a Poisson distribution. It was determined by DE SOLLA PRICE that the number of scientific discoveries made independently twice, three times, etc. is also distributed according to the Poisson law, in a way similar to the principle of the blindfolded apple pickers (Table 5.1). One gets the impression that scientists are not concerned with the activities of their colleagues. DE SOLLA PRICE believes that this can be explained by the assumption that scientists have a strong urge write papers, but feel only a relatively mild need to read them. ■

## 5.5 The Characteristic Function of a Distribution

So far we have only considered real random variables. In fact, in Sect. 3.1 we have introduced the concept of a random quantity as a real number associated with an event. Without changing this concept we can formally construct a *complex random variable* from two real ones by writing

$$z = x + iy . \quad (5.5.1)$$

**Table 5.1:** Simultaneous discovery and the Poisson distribution.

| Number of simultaneous discoveries | Cases of simultaneous discovery | Prediction of Poisson distribution |
|------------------------------------|---------------------------------|------------------------------------|
| 0                                  | Not defined                     | 368                                |
| 1                                  | Not known                       | 368                                |
| 2                                  | 179                             | 184                                |
| 3                                  | 51                              | 61                                 |
| 4                                  | 17                              | 15                                 |
| 5                                  | 6                               | 3                                  |
| $\geq 6$                           | 8                               | 1                                  |

As its expectation value we define

$$E(\mathbf{z}) = E(\mathbf{x}) + i E(\mathbf{y}) \quad . \quad (5.5.2)$$

By analogy with real variables, complex random variables are independent if the real and imaginary parts are independent among themselves.

If  $\mathbf{x}$  is a real random variable with distribution function  $F(x) = P(\mathbf{x} < x)$  and probability density  $f(x)$ , we define its characteristic function to be the expectation value of the quantity  $\exp(it\mathbf{x})$ :

$$\varphi(t) = E\{\exp(it\mathbf{x})\} \quad . \quad (5.5.3)$$

That is, in the case of a continuous variable the characteristic function is a Fourier integral with its known transformation properties:

$$\varphi(t) = \int_{-\infty}^{\infty} \exp(itx) f(x) dx \quad . \quad (5.5.4)$$

For a discrete variable we obtain instead from (3.3.2)

$$\varphi(t) = \sum_i \exp(itx_i) P(\mathbf{x} = x_i) \quad . \quad (5.5.5)$$

We now consider the moments of  $\mathbf{x}$  about the origin,

$$\lambda_n = E(\mathbf{x}^n) = \int_{-\infty}^{\infty} x^n f(x) dx \quad , \quad (5.5.6)$$

and find that  $\lambda_n$  can be obtained simply by differentiating the characteristic function  $n$  times at the point  $t = 0$ :

$$\varphi^{(n)}(t) = \frac{d^n \varphi(t)}{dt^n} = i^n \int_{-\infty}^{\infty} x^n \exp(itx) f(x) dx$$

and therefore

$$\varphi^{(n)}(0) = i^n \lambda_n . \quad (5.5.7)$$

If we now introduce the simple coordinate translation

$$y = x - \hat{x} \quad (5.5.8)$$

and construct the characteristic function

$$\varphi_y(t) = \int_{-\infty}^{\infty} \exp\{it(x - \hat{x})\} f(x) dx = \varphi(t) \exp(-it\hat{x}) , \quad (5.5.9)$$

then its  $n$ th derivative is (up to a power of  $i$ ) equal to the  $n$ th moment of  $x$  about the expectation value [cf. (3.3.8)]:

$$\varphi_y^{(n)}(0) = i^n \mu_n = i^n E\{(x - \hat{x})^n\} , \quad (5.5.10)$$

and in particular

$$\sigma^2(x) = -\varphi_y''(0) . \quad (5.5.11)$$

Inverting the Fourier transform (5.5.4) we see that it is possible to obtain the probability density from the characteristic function,

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp(-itx) \varphi(t) dt . \quad (5.5.12)$$

It is possible to show that a distribution is determined *uniquely* by its characteristic function. This is the case even for discrete variables where one has

$$F(b) - F(a) = \frac{i}{2\pi} \int_{-\infty}^{\infty} \frac{\exp(itb) - \exp(ita)}{t} \varphi(t) dt , \quad (5.5.13)$$

since in this case the probability density is not defined. Often it is more convenient to use the characteristic function rather than the original distribution. Because of the unique relation between the two it is possible to switch back and forth at any place in the course of a calculation.

We now consider the sum of two independent random variables

$$w = x + y .$$

Its characteristic function is

$$\varphi_w(t) = E[\exp\{it(x+y)\}] = E\{\exp(itx)\}\exp(ity) \quad .$$

Generalizing relation (3.5.13) to complex variables we obtain

$$\varphi_w(t) = E\{\exp(itx)\}E\{\exp(ity)\} = \varphi_x(t)\varphi_y(t) \quad , \quad (5.5.14)$$

i.e., the characteristic function of a sum of independent random variables is equal to the product of their respective characteristic functions.

**Example 5.5:** Addition of two Poisson distributed variables with use of the characteristic function

From Eqs. (5.5.5) and (5.4.1) one obtains for the characteristic function of the Poisson distribution

$$\begin{aligned} \varphi(t) &= \sum_{k=0}^{\infty} \exp(itk) \frac{\lambda^k}{k!} \exp(-\lambda) = \exp(-\lambda) \sum_{k=0}^{\infty} \frac{(\lambda \exp(it))^k}{k!} \\ &= \exp(-\lambda) \exp(\lambda e^{it}) = \exp\{\lambda(e^{it} - 1)\} \quad . \end{aligned} \quad (5.5.15)$$

We now form the characteristic function of the sum of two independent Poisson distributed variables with mean values  $\lambda_1$  and  $\lambda_2$ ,

$$\begin{aligned} \varphi_{\text{sum}}(t) &= \exp\{\lambda_1(e^{it} - 1)\} \exp\{\lambda_2(e^{it} - 1)\} \\ &= \exp\{(\lambda_1 + \lambda_2)(e^{it} - 1)\} \quad . \end{aligned} \quad (5.5.16)$$

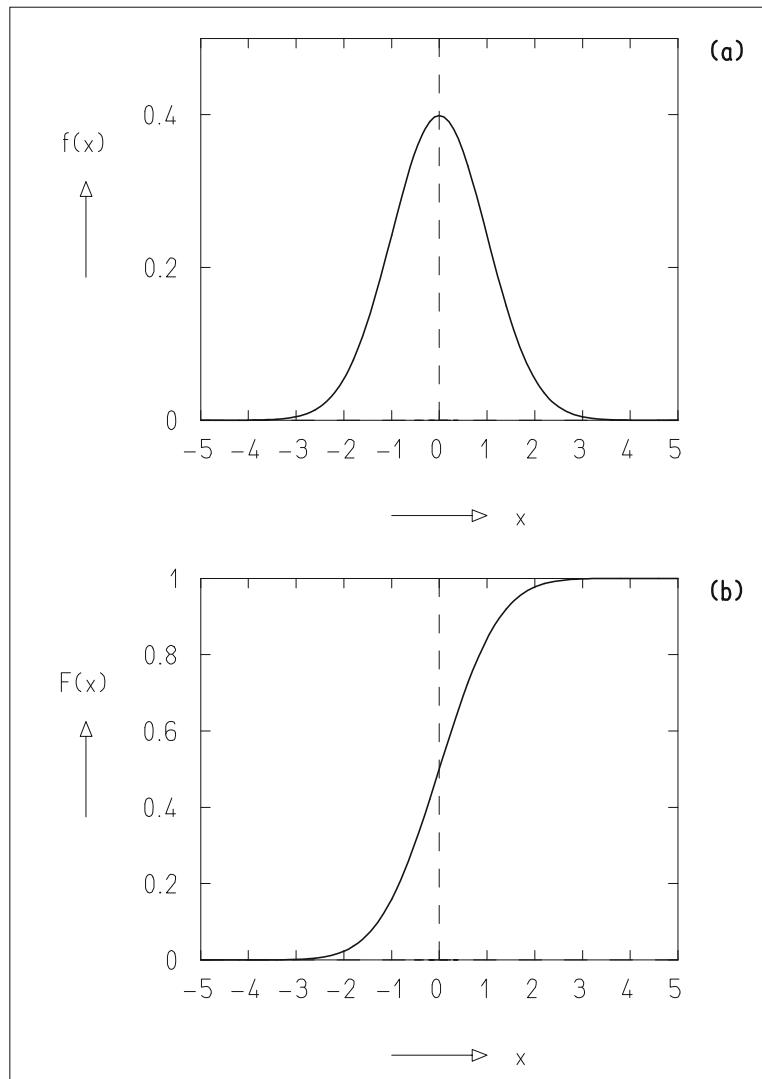
This is again of the form of Eq. (5.5.15). Therefore the distribution of the sum of two independent Poisson distributed variables is itself a Poisson variable. Its mean is the sum of the means of the individual distributions. ■

## 5.6 The Standard Normal Distribution

The probability density of the *standard normal distribution* is defined as

$$f(x) = \phi_0(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \quad . \quad (5.6.1)$$

This function is depicted in Fig. 5.7a. It has a bell shape with the maximum at  $x = 0$ . From Appendix D.1 we have



**Fig. 5.7:** Probability density (a) and distribution function (b) of the standard normal distribution.

$$\int_{-\infty}^{\infty} e^{-x^2/2} dx = \sqrt{2\pi} , \quad (5.6.2)$$

so that  $\phi_0(x)$  is normalized to one as required. Using the symmetry of Fig. 5.7a, or alternatively, using the antisymmetry of the integrand we conclude that the expectation value is

$$\hat{x} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x e^{-x^2/2} dx = 0 . \quad (5.6.3)$$

By integrating by parts we can compute the variance to be

$$\begin{aligned}\sigma^2 &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^2 e^{-x^2/2} dx \\ &= \frac{1}{\sqrt{2\pi}} \left\{ \left[ -x e^{-x^2/2} \right]_{-\infty}^{\infty} + \int_{-\infty}^{\infty} e^{-x^2/2} dx \right\} = 1 \quad , \quad (5.6.4)\end{aligned}$$

since the expression in the square brackets vanishes at the integral's boundaries and the integral in curly brackets is given by Eq. (5.6.2).

The distribution function of the standard normal distribution

$$F(x) = \psi_0(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-t^2/2} dt \quad (5.6.5)$$

is shown in Fig. 5.7b. It cannot be expressed in analytic form. It is tabulated numerically in Appendix C.4.

## 5.7 The Normal or Gaussian Distribution

The standardized distribution of the last section had the properties  $\hat{x} = E(\mathbf{x}) = 0$ ,  $\sigma^2(\mathbf{x}) = 1$ , i.e., the variable  $\mathbf{x}$  had the properties of the standardized variable  $\mathbf{u}$  in Eq. (3.3.17). If we now replace  $\mathbf{x}$  by  $(\mathbf{x} - a)/b$  in (5.6.1), we obtain the probability density of the *normal* or *Gaussian distribution*,

$$f(x) = \phi(x) = \frac{1}{\sqrt{2\pi}b} \exp \left\{ -\frac{(x-a)^2}{2b^2} \right\} \quad , \quad (5.7.1)$$

with

$$\hat{x} = a \quad , \quad \sigma^2(\mathbf{x}) = b^2 \quad . \quad (5.7.2)$$

The characteristic function of the normal distribution (5.7.1) is, using Eq. (5.5.4),

$$\varphi(t) = \frac{1}{\sqrt{2\pi}b} \int_{-\infty}^{\infty} \exp(itx) \exp \left( -\frac{(x-a)^2}{2b^2} \right) dx \quad . \quad (5.7.3)$$

With  $u = (x - a)/b$  one obtains

$$\begin{aligned}\varphi(t) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp \left\{ -\frac{1}{2} u^2 + it(bu + a) \right\} du \\ &= \frac{1}{\sqrt{2\pi}} \exp(ita) \int_{-\infty}^{\infty} \exp \left\{ -\frac{1}{2} u^2 + itbu \right\} du \quad . \quad (5.7.4)\end{aligned}$$

By completing the square the integral can be rewritten as

$$\begin{aligned} & \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2}u^2 + itbu\right\} du \\ &= \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2}(u-itb)^2 - \frac{1}{2}t^2b^2\right\} du \\ &= \exp\left\{-\frac{1}{2}t^2b^2\right\} \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2}(u-itb)^2\right\} du . \end{aligned} \quad (5.7.5)$$

With  $r = u - itb$  the last integral takes on the form

$$\int_{-\infty-itb}^{\infty-itb} \exp\left\{-\frac{1}{2}r^2\right\} dr .$$

The integrand does not have any singularities in the complex  $r$  plane. According to the residue theorem, therefore, the contour integral around any closed path vanishes. Consider a path that runs along the real axis from  $r = -L$  to  $r = L$ , and then parallel to the imaginary axis from  $r = L$  to  $r = L - itb$  and from there antiparallel to the real axis to  $r = -L - itb$ , and finally back to the starting point  $r = L$ . In the limit  $L \rightarrow \infty$  the integrand vanishes along the parts of the path that run parallel to the imaginary axis. One then has

$$\int_{-\infty-itb}^{\infty-itb} \exp\left\{-\frac{1}{2}r^2\right\} dr = \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2}r^2\right\} dr ,$$

i.e., we can extend the integral to cover the entire real axis. The integral is computed in Appendix D.1 and has the value

$$\int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2}r^2\right\} dr = \sqrt{2\pi} . \quad (5.7.6)$$

Substituting this into Eqs. (5.7.5) and (5.7.4) we obtain finally the characteristic function of the normal distribution

$$\varphi(t) = \exp(it\bar{a}) \exp\left(-\frac{1}{2}\bar{b}^2 t^2\right) . \quad (5.7.7)$$

For the case  $\bar{a} = 0$  one obtains from this the following interesting **theorem**:

A normal distribution with mean value zero has a characteristic function that has itself (up to normalization) the form of a normal distribution. The product of the variances of both functions is one.

If we now consider the sum of two independent normal distributions, then by applying Eq. (5.5.14) one immediately sees that the characteristic function of the sum is again of the form of Eq. (5.7.7). The sum of independent normally distributed quantities is therefore itself normally distributed. The Poisson distribution behaves in a similar way (cf. Example 5.5).

## 5.8 Quantitative Properties of the Normal Distribution

Figure 5.7a shows the probability density of the standard Gaussian distribution  $\phi_0(x)$  and the corresponding distribution function. By simple computation one can determine that the points of inflection of (5.6.1) are at  $x = \pm 1$ . [In the case of a general Gaussian distribution (5.7.1) they are at  $x = a \pm b$ .] The distribution function  $\psi_0(x)$  gives the probability for the random variable to take on a value smaller than  $x$ :

$$\psi_0(x) = P(\mathbf{x} < x) . \quad (5.8.1)$$

By symmetry one has

$$P(|\mathbf{x}| > x) = 2\psi_0(-|x|) = 2\{1 - \psi_0(|x|)\} \quad (5.8.2)$$

or conversely, the probability to obtain a random value within an interval of width  $2x$  about zero (the expectation value) is

$$P(|\mathbf{x}| \leq x) = 2\psi_0(|x|) - 1 . \quad (5.8.3)$$

Since the integral (5.6.5) is not easy to evaluate, one typically finds the values of (5.8.1) and (5.8.3) from statistical tables, e.g., in Tables I.2 and I.3 of the appendix.

One can now extend this relation to the general Gaussian distribution given by Eq. (5.7.1). Its distribution function is

$$\psi(x) = \psi_0\left(\frac{x-a}{b}\right) . \quad (5.8.4)$$

We are interested in finding the probability to obtain a random value inside (or outside) of a given multiple of  $\sigma = b$  about the mean value:

$$P(|\mathbf{x}-a| \leq n\sigma) = 2\psi_0\left(\frac{nb}{b}\right) - 1 = 2\psi_0(n) - 1 . \quad (5.8.5)$$

From Table I.3 we find

$$\begin{aligned} P(|\mathbf{x}-a| \leq \sigma) &= 68.3\% , & P(|\mathbf{x}-a| > \sigma) &= 31.7\% , \\ P(|\mathbf{x}-a| \leq 2\sigma) &= 95.4\% , & P(|\mathbf{x}-a| > 2\sigma) &= 4.6\% , \\ P(|\mathbf{x}-a| \leq 3\sigma) &= 99.8\% , & P(|\mathbf{x}-a| > 3\sigma) &= 0.2\% . \end{aligned} \quad (5.8.6)$$

As we will see later in more detail, one can often assume that the measurement errors of a quantity are distributed according to a Gaussian distribution about zero. This means that the probability to obtain a value between  $x$  and  $x + dx$  is given by

$$P(x \leq x < x + dx) = \phi(x) dx .$$

The dispersion  $\sigma$  of the distribution  $\phi(x)$  is called the *standard deviation* or *standard error*. If the standard error of an instrument is known and one carries out a single measurement, then Eq. (5.8.6) tells us that the probability that the true value is within an interval given by plus or minus the standard error about the measured value is 68.3%. It is therefore a common practice to multiply the standard error with a more or less arbitrary factor in order to improve this percentage. (One obtains around 99.8% for the factor 3.) This procedure is, however, misleading and often harmful. If this factor is not explicitly stated, a comparison of different measurements of the same quantity and especially the calculation of a weighted average (cf. Example 9.1) is rendered impossible or is liable to be erroneous.

The quantiles [see Eq. (3.3.25)] of the standard normal distribution are of considerable interest. For the distribution function (5.6.5) one obtains by definition

$$P(x_p) = P(x < x_p) = \psi_0(x_p) . \quad (5.8.7)$$

The quantile  $x_p$  is therefore given by the inverse function

$$x_p = \Omega(P) \quad (5.8.8)$$

of the distribution function  $\psi_0(x_p)$ . This is computed numerically in Appendix C.4 and is given in Table I.4. Figure 5.8 shows a graphical representation.

We now consider the probability

$$P'(x) = P(|x| < x) , \quad x > 0 , \quad (5.8.9)$$

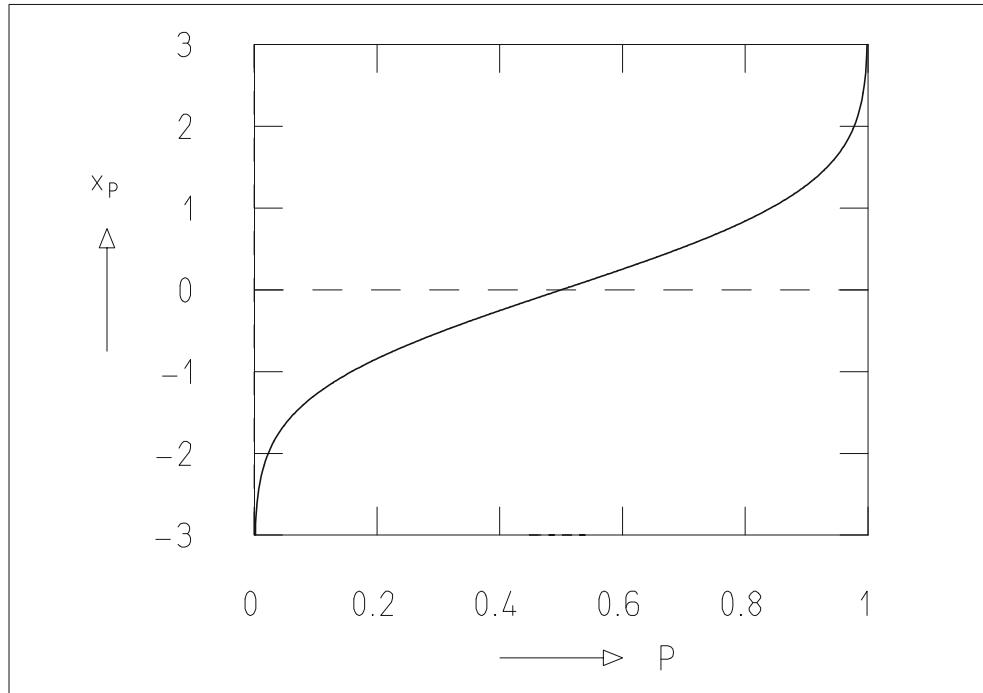
for a quantity distributed according to the standard normal distribution to differ from zero in absolute value by less than  $x$ . Since

$$\begin{aligned} P(x) &= P(x < x) = \psi_0(x) = \int_{-\infty}^x \phi_0(x) dx \\ &= \frac{1}{2} + \int_0^x \phi_0(x) dx = \frac{1}{2} + \frac{1}{2} P'(x) = \frac{1}{2}(P'(x) + 1) , \end{aligned}$$

it is the inverse function and with it the quantiles of the distribution function  $P'(x)$  are obtained by substituting  $(P'(x) + 1)/2$  for the argument of the inverse function of  $P$ :

$$x_p = \Omega'(P') = \Omega((P' + 1)/2) . \quad (5.8.10)$$

This function is tabulated in Table I.5.



**Fig. 5.8:** Quantile of the standard normal distribution.

## 5.9 The Central Limit Theorem

We will now prove the following important theorem. If  $\mathbf{x}_i$  are independent random variables with mean values  $a$  and variances  $b^2$ , then the variable

$$\mathbf{x} = \lim_{n \rightarrow \infty} \sum_{i=1}^n \mathbf{x}_i \quad (5.9.1)$$

is normally distributed with

$$E(\mathbf{x}) = na \quad , \quad \sigma^2(\mathbf{x}) = nb^2 \quad . \quad (5.9.2)$$

From Eq. (3.3.15) one then has that the variable

$$\xi = \frac{1}{n} \mathbf{x} = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \quad (5.9.3)$$

is normally distributed with

$$E(\xi) = a \quad , \quad \sigma^2(\xi) = b^2/n \quad . \quad (5.9.4)$$

To prove this we assume for simplicity that all of the  $x_i$  have the same distribution. If we denote the characteristic function of the  $x_i$  by  $\varphi(t)$ , then the sum of  $n$  variables has the characteristic function  $\{\varphi(t)\}^n$ . We now assume that  $a = 0$ . (The general case can be related to this by a simple coordinate translation  $x'_i = x_i - a$ .) From (5.5.10) we have the first two derivatives of  $\varphi(t)$  at  $t = 0$ ,

$$\varphi'(0) = 0 \quad , \quad \varphi''(0) = -\sigma^2 \quad .$$

We can therefore expand,

$$\varphi_{x'}(t) = 1 - \frac{1}{2}\sigma^2 t^2 + \dots \quad .$$

Instead of  $x_i$  let us now choose

$$u_i = \frac{x'_i}{b\sqrt{n}} = \frac{x_i - a}{b\sqrt{n}}$$

as the variable. If we consider  $n$  to be fixed for the moment, then this implies a simple translation and a change of scale. The corresponding characteristic function is

$$\varphi_{u_i}(t) = E\{\exp(itu_i)\} = E\left\{\exp\left(it\frac{x_i - a}{b\sqrt{n}}\right)\right\} = \varphi_{x'_i}\left(\frac{t}{b\sqrt{n}}\right)$$

or

$$\varphi_{u_i}(t) = 1 - \frac{t^2}{2n} + \dots \quad .$$

The higher-order terms are at most of the order  $n^{-2}$ . If we now consider the limiting case and use

$$u = \lim_{n \rightarrow \infty} \sum_{i=1}^n u_i = \lim_{n \rightarrow \infty} \sum_{i=1}^n \frac{x_i - a}{b\sqrt{n}} = \lim_{n \rightarrow \infty} \frac{(x - na)}{b\sqrt{n}} \quad (5.9.5)$$

we obtain

$$\varphi_u(t) = \lim_{n \rightarrow \infty} \{\varphi_{u_i}(t)\}^n = \lim_{n \rightarrow \infty} \left(1 - \frac{t^2}{2n} + \dots\right)^n$$

or

$$\varphi_u(t) = \exp\left(-\frac{1}{2}t^2\right) \quad . \quad (5.9.6)$$

This, however, is exactly the characteristic function of the standard normal distribution  $\phi_0(u)$ . One therefore has  $E(u) = 0$ ,  $\sigma^2(u) = 1$ . Using Eqs. (5.9.5) and (3.3.15) leads directly to the theorem.

**Example 5.6:** Normal distribution as the limiting case of the binomial distribution

Suppose that the individual variables  $x_i$  in (5.9.1) are described by the simple distribution given by (5.1.1) and (5.1.2), i.e., they can only take on the value 1 (with probability  $p$ ) or 0 (with probability  $1 - p$ ). One then has  $E(x_i) = p$ ,  $\sigma^2(x_i) = p(1 - p)$ . The variable

$$x^{(n)} = \sum_{i=1}^n x_i \quad (5.9.7)$$

then follows the binomial distribution,  $P(x^{(n)} = k) = W_k^n$  [see Eqs. (5.1.3), (5.1.6), (5.1.7)]. As done for (5.9.5) let us consider the distribution of

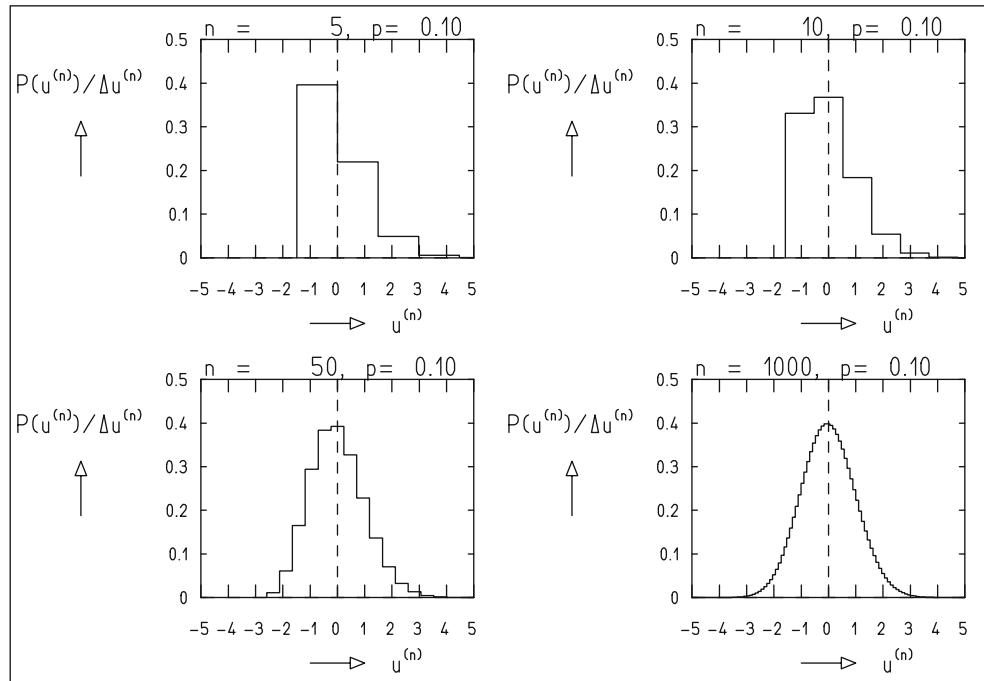
$$u^{(n)} = \sum_{i=1}^n \frac{x_i - p}{\sqrt{np(1-p)}} = \frac{1}{\sqrt{np(1-p)}} \left( \sum_{i=1}^n x_i - np \right) \quad . \quad (5.9.8)$$

One clearly has  $P(x = k) = P(u^{(n)} = (k - np)/\sqrt{np(1-p)}) = W_k^n$ . These values, however, lie increasingly closer to each other on the  $u^{(n)}$  axis as  $n$  increases. Let us denote the distance between two neighboring values of  $u^{(n)}$  by  $\Delta u^{(n)}$ . Then the distribution of a discrete variable  $P(u^{(n)})/\Delta u^{(n)}$  finally becomes the probability density of a continuous variable. According to the Central Limit Theorem this must be a standard normal distribution. This is illustrated in Fig. 5.9, where  $P(u^{(n)})/\Delta u^{(n)}$  is shown for various possible values of  $u^{(n)}$ . ■

**Example 5.7:** Error model of Laplace

In 1783 LAPLACE made the following remarks concerning the origin of errors of an observation. Suppose the true value of a quantity to be measured is  $m_0$ . Now let the measurement be disturbed by a large number  $n$  of independent causes, each resulting in a disturbance of magnitude  $\varepsilon$ . For each disturbance there exists an equal probability for a variation of the measured value in either direction, i.e.,  $+\varepsilon$  or  $-\varepsilon$ . The measurement error is then composed of the sum of the individual disturbances. It is clear that in this model the probability distribution of measurement errors will be given by the binomial distribution. It is interesting nevertheless to follow the model somewhat further, since it leads directly to the famous Pascal triangle.

Figure 5.10 shows how the probability distribution is derived from the model. The starting point is with no disturbance where the probability of measuring  $m_0$  is equal to one. With one disturbance this probability is split equally between the neighboring possibilities  $m_0 + \varepsilon$  and  $m_0 - \varepsilon$ . The same happens with every further disturbance. Of course the individual probabilities leading to the same measured value must be added.



**Fig. 5.9:** The quantity  $P(u^{(n)})/\Delta u^{(n)}$  for various values of the discrete variable  $u^{(n)}$  for increasing  $n$ .

| Number of disturbances<br>$n$ | Deviation from true value |               |                            |                            |                            |               |               |
|-------------------------------|---------------------------|---------------|----------------------------|----------------------------|----------------------------|---------------|---------------|
|                               | $-3\epsilon$              | $-2\epsilon$  | $-\epsilon$                | 0                          | $+\epsilon$                | $+2\epsilon$  | $+3\epsilon$  |
| 0                             |                           |               |                            | 1                          |                            |               |               |
| 1                             |                           |               | $\frac{1}{2}$              |                            | $\frac{1}{2}$              |               |               |
| 2                             |                           | $\frac{1}{4}$ |                            | $\frac{1}{4}, \frac{1}{4}$ |                            | $\frac{1}{4}$ |               |
| 3                             | $\frac{1}{8}$             |               | $\frac{1}{8}, \frac{1}{4}$ | $\frac{1}{2}$              | $\frac{1}{4}, \frac{1}{8}$ |               | $\frac{1}{8}$ |

**Fig. 5.10:** Connection between the Laplacian error model and the binomial distribution.

Each line of the resulting triangle contains the distribution  $W_k^n(k = 0, 1, \dots, n)$  of Eq. (5.1.3) for the case  $p = q = 1/2$ . Multiplied by  $1/(p^k q^{n-k}) = 2^n$  it becomes a line of binomial coefficients of Pascal's triangle (cf. Appendix B).

It is easy to relate this to Example 5.6 by extending Eq. (5.9.8) and substituting  $p = 1/2$ . For  $n \rightarrow \infty$  the quantity

$$u^{(n)} = \frac{2(\sum_{i=1}^n \varepsilon x_i - n\varepsilon/2)}{\sqrt{n\varepsilon}}$$

follows a normal distribution with expectation value zero and standard deviation  $\sqrt{n\varepsilon}/2$ . Thus Gaussian measurement errors can result from a large number of small independent disturbances. ■

The identification of the measurement error distribution as Gaussian is of great significance in many computations, particularly for the method of least squares. The normal distribution for measurement errors is, however, not a law of nature. The causes of experimental errors can be individually very complicated. One cannot, therefore, find a distribution function that describes the behavior of measurement errors in all possible experiments. In particular, it is not always possible to guarantee symmetry and independence. One must ask in each individual case whether the measurement errors can be modeled by a Gaussian distribution. This can be done, for example, by means of a  $\chi^2$ -test, applied to the distribution of a measured quantity (see Sect. 8.7). It is always necessary to check the distribution of experimental errors before more lengthy computations can be used whose results are only meaningful for the case of a Gaussian error distribution.

## 5.10 The Multivariate Normal Distribution

Consider a vector  $\mathbf{x}$  of  $n$  variables,

$$\mathbf{x} = (x_1, x_2, \dots, x_n) \quad .$$

We define the probability density of the joint normal distribution of the  $x_i$  to be

$$\phi(\mathbf{x}) = k \exp\left\{-\frac{1}{2}(\mathbf{x} - \mathbf{a})^T B(\mathbf{x} - \mathbf{a})\right\} = k \exp\left\{-\frac{1}{2}g(\mathbf{x})\right\} \quad (5.10.1)$$

with

$$g(\mathbf{x}) = (\mathbf{x} - \mathbf{a})^T B(\mathbf{x} - \mathbf{a}) \quad . \quad (5.10.2)$$

Here  $\mathbf{a}$  is an  $n$ -component vector and  $B$  is an  $n \times n$  matrix, which is symmetric and positive definite. Since  $\phi(\mathbf{x})$  is clearly symmetric about the point  $\mathbf{x} = \mathbf{a}$ , one has

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} (\mathbf{x} - \mathbf{a}) \phi(\mathbf{x}) dx_1 dx_2 \dots dx_n = 0 \quad . \quad (5.10.3)$$

that is,

$$E(\mathbf{x} - \mathbf{a}) = 0$$

or

$$E(\mathbf{x}) = \mathbf{a} \quad . \quad (5.10.4)$$

The vector of expectation values is therefore given directly by  $\mathbf{a}$ .

We now differentiate Eq. (5.10.3) with respect to  $\mathbf{a}$ ,

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} [I - (\mathbf{x} - \mathbf{a})(\mathbf{x} - \mathbf{a})^T B] \phi(\mathbf{x}) dx_1 dx_2 \dots dx_n = 0 \quad .$$

This means that the expectation value of the quantity in square brackets vanishes,

$$E\{(\mathbf{x} - \mathbf{a})(\mathbf{x} - \mathbf{a})^T\} B = I$$

or

$$C = E\{(\mathbf{x} - \mathbf{a})(\mathbf{x} - \mathbf{a})^T\} = B^{-1} \quad . \quad (5.10.5)$$

Comparing with Eq. (3.6.19) one sees that  $C$  is the covariance matrix of the variables  $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)$ .

Because of the practical importance of the normal distribution, we would like to investigate the case of two variables in somewhat more detail. In particular we are interested in the correlation of the variables. One has

$$C = B^{-1} = \begin{pmatrix} \sigma_1^2 & \text{cov}(\mathbf{x}_1, \mathbf{x}_2) \\ \text{cov}(\mathbf{x}_1, \mathbf{x}_2) & \sigma_2^2 \end{pmatrix} \quad . \quad (5.10.6)$$

By inversion one obtains for  $B$

$$B = \frac{1}{\sigma_1^2 \sigma_2^2 - \text{cov}(\mathbf{x}_1, \mathbf{x}_2)^2} \begin{pmatrix} \sigma_2^2 & -\text{cov}(\mathbf{x}_1, \mathbf{x}_2) \\ -\text{cov}(\mathbf{x}_1, \mathbf{x}_2) & \sigma_1^2 \end{pmatrix} \quad . \quad (5.10.7)$$

One sees that  $B$  is a diagonal matrix if the covariances vanish. One then has

$$B_0 = \begin{pmatrix} 1/\sigma_1^2 & 0 \\ 0 & 1/\sigma_2^2 \end{pmatrix} \quad . \quad (5.10.8)$$

If we substitute  $B_0$  into Eq. (5.10.1), we obtain – as expected – the joint probability density of two *independently* normally distributed variables as the product of two normal distributions:

$$\phi = k \exp\left(-\frac{1}{2} \frac{(x_1 - a_1)^2}{\sigma_1^2}\right) \exp\left(-\frac{1}{2} \frac{(x_2 - a_2)^2}{\sigma_2^2}\right) . \quad (5.10.9)$$

In this simple case the constant  $k$  takes on the value

$$k_0 = \frac{1}{2\pi\sigma_1\sigma_2} ,$$

as can be determined by integration of (5.10.9) or simply by comparison with Eq. (5.7.1). In the general case of  $n$  variables with non-vanishing covariances, one has

$$k = \left(\frac{\det B}{(2\pi)^n}\right)^{\frac{1}{2}} . \quad (5.10.10)$$

Here  $\det B$  is the determinant of the matrix  $B$ . If the variables are not independent, i.e., if the covariance does not vanish, then the expression for the normal distribution of two variables is somewhat more complicated.

Let us consider the reduced variables

$$u_i = \frac{x_i - a_i}{\sigma_i} , \quad i = 1, 2 ,$$

and make use of the correlation coefficient

$$\rho = \frac{\text{cov}(x_1, x_2)}{\sigma_1\sigma_2} = \text{cov}(u_1, u_2) .$$

Equation (5.10.1) then takes on the simple form

$$\phi(u_1, u_2) = k \exp\left(-\frac{1}{2} \mathbf{u}^T B \mathbf{u}\right) = k \exp\left(-\frac{1}{2} g(\mathbf{u})\right) \quad (5.10.11)$$

with

$$B = \frac{1}{1-\rho^2} \begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix} . \quad (5.10.12)$$

Contours of equal probability density are characterized by a constant exponent in (5.10.11):

$$-\frac{1}{2} \cdot \frac{1}{(1-\rho^2)} (u_1^2 + u_2^2 - 2u_1u_2\rho) = -\frac{1}{2} g(\mathbf{u}) = \text{const} . \quad (5.10.13)$$

Let us take for the moment  $g(\mathbf{u}) = 1$ .

In the original variables Eq. (5.10.13) becomes

$$\frac{(x_1 - a_1)^2}{\sigma_1^2} - 2\rho \frac{x_1 - a_1}{\sigma_1} \frac{x_2 - a_2}{\sigma_2} + \frac{(x_2 - a_2)^2}{\sigma_2^2} = 1 - \rho^2 . \quad (5.10.14)$$

This is the equation of an ellipse centered around the point  $(a_1, a_2)$ . The principal axes of the ellipse make an angle  $\alpha$  with respect to the axes  $x_1$  and  $x_2$ . This angle and the half-diameters  $p_1$  and  $p_2$  can be determined from Eq. (5.10.14) by using the known properties of conic sections:

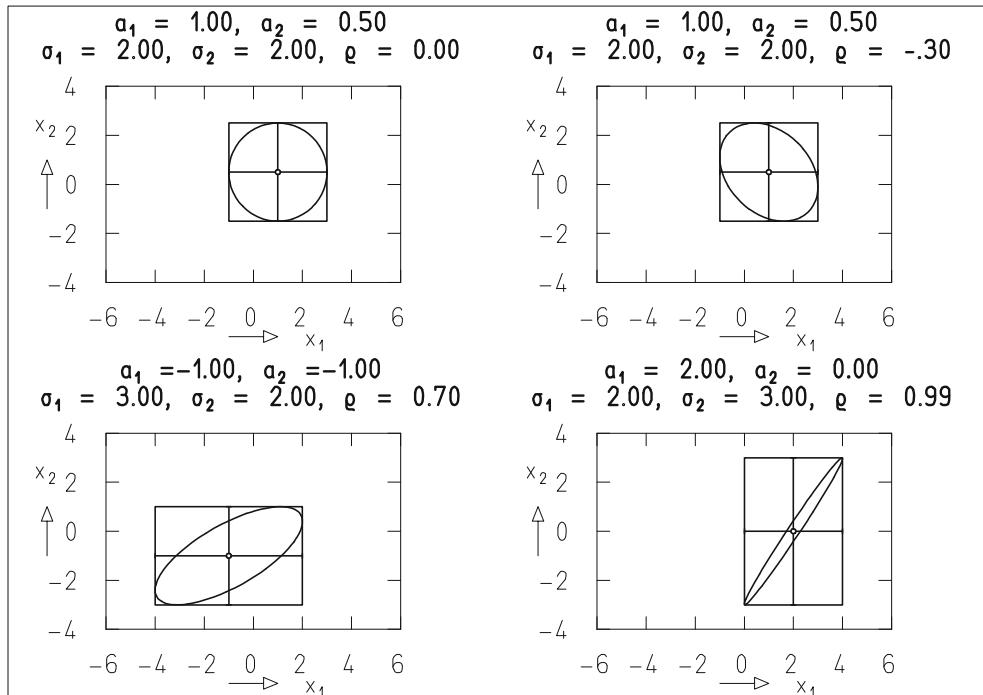
$$\tan 2\alpha = \frac{2\rho\sigma_1\sigma_2}{\sigma_1^2 - \sigma_2^2} , \quad (5.10.15)$$

$$p_1^2 = \frac{\sigma_1^2\sigma_2^2(1-\rho^2)}{\sigma_2^2\cos^2\alpha - 2\rho\sigma_1\sigma_2\sin\alpha\cos\alpha + \sigma_1^2\sin^2\alpha} , \quad (5.10.16)$$

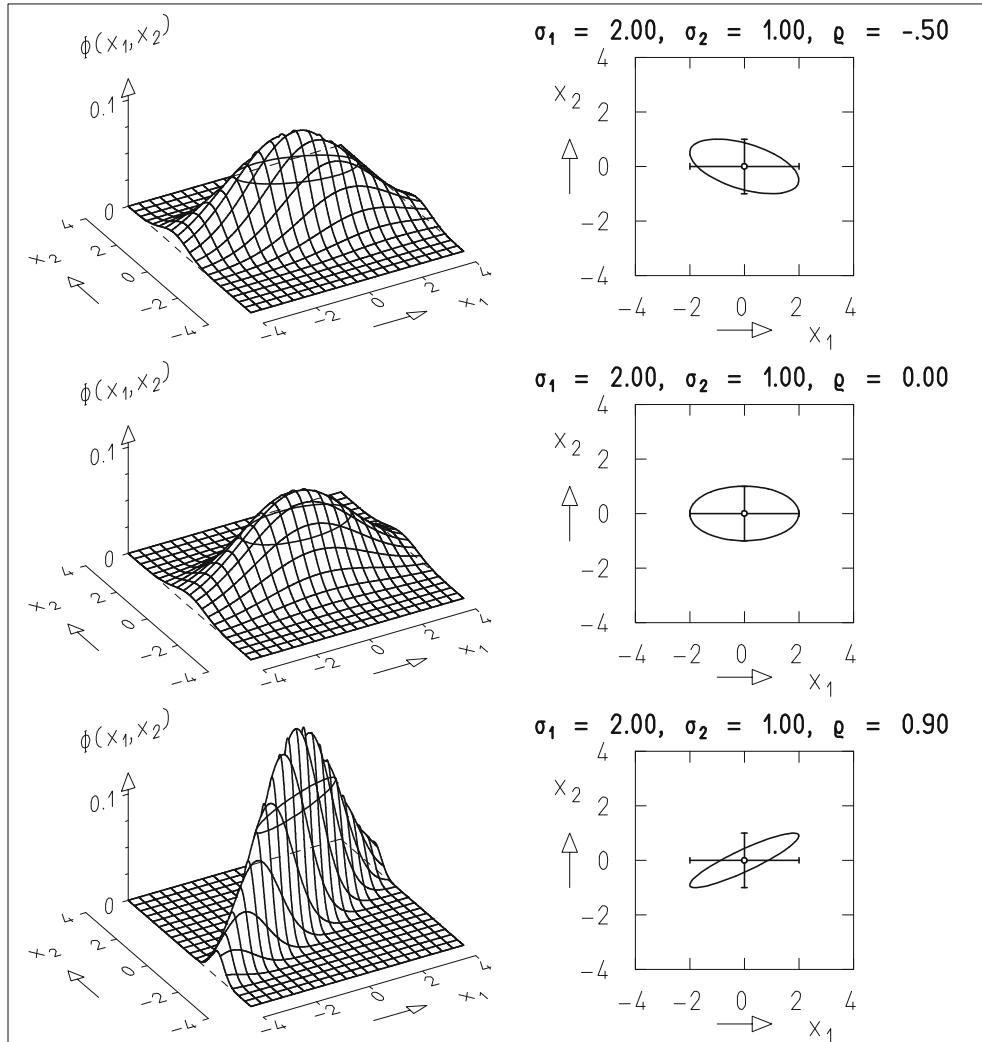
$$p_2^2 = \frac{\sigma_1^2\sigma_2^2(1-\rho^2)}{\sigma_2^2\sin^2\alpha + 2\rho\sigma_1\sigma_2\sin\alpha\cos\alpha + \sigma_1^2\cos^2\alpha} . \quad (5.10.17)$$

The ellipse with these properties is called the *covariance ellipse* of the bivariate normal distribution. Several such ellipses are depicted in Fig. 5.11. The covariance ellipse always lies inside a rectangle determined by the point  $(a_1, a_2)$  and the standard deviations  $\sigma_1$  and  $\sigma_2$ . It touches the rectangle at four points. For the extreme cases  $\rho = \pm 1$  the ellipse becomes one of the two diagonals of this rectangle.

From (5.10.14) it is clear that other lines of constant probability (for  $g \neq 1$ ) are also ellipses, concentric and similar to the covariance ellipse and situated inside (outside) of it for larger (smaller) probability. The bivariate



**Fig. 5.11:** Covariance ellipses.



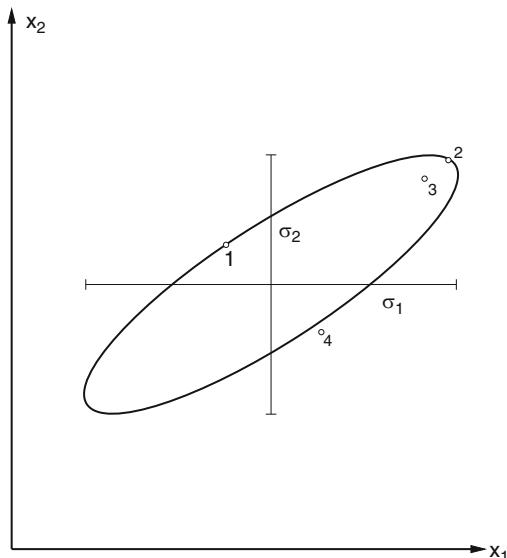
**Fig. 5.12:** Probability density of a bivariate Gaussian distribution (left) and the corresponding covariance ellipse (right). The three rows of the figure differ only in the numerical value of the correlation coefficient  $\rho$ .

normal distribution therefore corresponds to a surface in the three-dimensional space  $(x_1, x_2, \phi)$  (Fig. 5.12), whose horizontal sections are concentric ellipses. For the largest probability this ellipse collapses to the point  $(a_1, a_2)$ . The vertical sections through the center have the form of a Gaussian distribution whose width is directly proportional to the diameter of the covariance ellipse along which the section extends. The probability of observing a pair  $\mathbf{x}_1, \mathbf{x}_2$  of random variables inside the covariance ellipse is equal to the integral

$$\int_A \phi(\mathbf{x}) d\mathbf{x} = 1 - e^{-\frac{1}{2}} = \text{const} \quad , \quad (5.10.18)$$

where the region of integration  $A$  is given by the area within the covariance ellipse (5.10.14). The relation (5.10.18) is obtained by application of the transformation of variables  $\mathbf{y} = T \mathbf{x}$  with  $T = B^{-1}$  to the distribution  $\phi(\mathbf{x})$ . The resulting distribution has the properties  $\sigma(y_1) = \sigma(y_2) = 1$  and  $\text{cov}(y_1, y_2) = 0$ , i.e., it is of the form of (5.10.9). In this way the region of integration is transformed to a unit circle centered about  $(a_1, a_2)$ .

In our consideration of the normal distribution of measurement errors of a variable we found the interval  $a - \sigma \leq x \leq a + \sigma$  to be the region in which the probability density  $f(x)$  exceeded a given fraction, namely  $e^{-1/2}$  of its maximum value. The integral over this region was independent of  $\sigma$ . In the case of two variables, the role of this region is taken over by the covariance ellipse determined by  $\sigma_1$ ,  $\sigma_2$ , and  $\rho$ , and not – as is sometimes incorrectly assumed – by the rectangle that circumscribes the ellipse in Fig. 5.11. The meaning of the covariance ellipse can also be seen from Fig. 5.13. Points 1 and 2, which lie on the covariance ellipse, correspond to equal probabilities ( $P(1) = P(2) = P_e$ ), although the distance of point 1 from the middle is less in both coordinate directions. In addition, point 3 is more probable, and point 4 less probable ( $P(4) < P_e$ ,  $P(3) > P_e$ ), even though point 4 even closer is to  $(a_1, a_2)$  than point 3.



**Fig. 5.13:** Relative probability for various points from a bivariate Gaussian distribution ( $P_1 = P_2 = P_e$ ,  $P_3 > P_e$ ,  $P_4 < P_e$ ).

For three variables one obtains instead of the covariance ellipse a *covariance ellipsoid*, for  $n$  variables a hyperellipsoid in an  $n$ -dimensional space (see also Sect. A.11). According to our construction, the covariance ellipsoid is the hypersurface in the  $n$ -dimensional space on which the function  $g(\mathbf{x})$  in the exponent of the normal distribution (5.10.1) has the constant value  $g(\mathbf{x}) = 1$ . For other values  $g(\mathbf{x}) = \text{const}$  one obtains similar ellipsoids which lie inside ( $g < 1$ ) or outside ( $g > 1$ ) of the covariance ellipsoid. In Sect. 6.6 it

will be shown that the function  $g(\mathbf{x})$  follows a  $\chi^2$ -distribution with  $n$  degrees of freedom if  $\mathbf{x}$  follows the normal distribution (5.10.1). The probability to find  $\mathbf{x}$  inside the ellipsoid  $g = \text{const}$  is therefore

$$W = \int_0^g f(\chi^2; n) d\chi^2 = P\left(\frac{n}{2}, \frac{g}{2}\right) . \quad (5.10.19)$$

Here  $P$  is the incomplete gamma function given in Sect. D.5. For  $g = 1$ , that is, for the covariance ellipsoid in  $n$  dimensions, this probability is

$$W_n = P\left(\frac{n}{2}, \frac{1}{2}\right) . \quad (5.10.20)$$

Numerical values for small  $n$  are

$$\begin{aligned} W_1 &= 0.68269 , & W_2 &= 0.39347 , & W_3 &= 0.19875 , \\ W_4 &= 0.09020 , & W_5 &= 0.03734 , & W_6 &= 0.01439 . \end{aligned}$$

The probability decreases rapidly as  $n$  increases. In order to be able to give regions for various  $n$  which correspond to equal probability content, one specifies a value  $W$  on the left-hand side of (5.10.19) and determines the corresponding value of  $g$ . Then  $g$  is the quantile with probability  $W$  of the  $\chi^2$ -distribution with  $n$  degrees of freedom (see also Appendix C.5),

$$g = \chi_W^2(n) . \quad (5.10.21)$$

The ellipsoid that corresponds to the value of  $g$  that contains  $\mathbf{x}$  with the probability  $W$  is called the *confidence ellipsoid* of probability  $W$ . This expression can be understood to mean that, e.g., for  $W = 0.9$  one should have 90% confidence that  $\mathbf{x}$  lies within the confidence ellipsoid.

The variances  $\sigma_i^2$  or the standard deviations  $\Delta_i = \sigma_i$  also have a certain meaning for  $n$  variables. The probability to observe the variable  $x_i$  in the region  $a_i - \sigma_i < x_i < a_i + \sigma_i$  is, as before, 68.3%, independent of the number  $n$  of the variables. This only holds, however, when one places no requirements on the positions of any of the other variables  $x_j$ ,  $j \neq i$ .

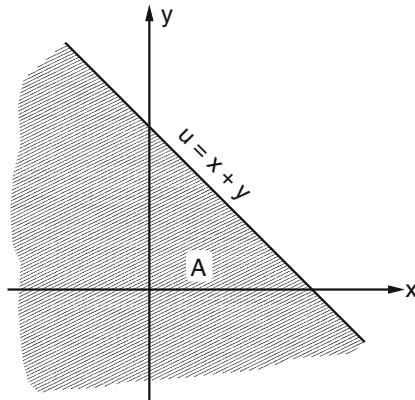
## 5.11 Convolutions of Distributions

### 5.11.1 Folding Integrals

On various occasions we have already discussed sums of random variables, and in the derivation of the Central Limit Theorem, for example, we found the characteristic function to be a useful tool in such considerations. We would

now like to discuss the distribution of the sum of two quantities, but for greater clarity we will not make use of the characteristic function.

A sum of two distributions is often observed in experiments. One could be interested, for example, in the angular distribution of secondary particles from the decay of an elementary particle. This can often be used to determine the spin of the particle. The observed angle is the distribution of a sum of random quantities, namely the decay angle and its measurement error. One speaks of the *convolution* of two distributions.



**Fig. 5.14:** Integration region for (5.11.4).

Let the original quantities be  $x$  and  $y$  and the sum

$$u = x + y \quad . \quad (5.11.1)$$

A requirement for further treatment is that the original variables must be independent. In this case, the joint probability density is the product of simple densities,

$$f(x, y) = f_x(x) f_y(y) \quad . \quad (5.11.2)$$

If we now ask for the distribution function of  $u$ , i.e., for

$$F(u) = P(u < u) = P(x + y < u) \quad , \quad (5.11.3)$$

then this is obtained by integration of (5.11.2) over the hatched region  $A$  in Fig. 5.14;

$$\begin{aligned} F(u) &= \iint_A f_x(x) f_y(y) dx dy = \int_{-\infty}^{\infty} f_x(x) dx \int_{-\infty}^{u-x} f_y(y) dy \\ &= \int_{-\infty}^{\infty} f_y(y) dy \int_{-\infty}^{u-y} f_x(x) dx \quad . \end{aligned} \quad (5.11.4)$$

By differentiation one obtains the probability density for  $u$ ,

$$f(u) = \frac{dF(u)}{du} = \int_{-\infty}^{\infty} f_x(x) f_y(u - x) dx = \int_{-\infty}^{\infty} f_y(y) f_x(u - y) dy \quad . \quad (5.11.5)$$

If  $x$  or  $y$  or both are only defined in a restricted region, then (5.11.5) is still true. The limits of integration, however, may be limited. We will consider various cases:

(a)  $0 \leq x < \infty, -\infty < y < \infty$  :

$$f(u) = \int_{-\infty}^u f_x(u-y) f_y(y) dy . \quad (5.11.6)$$

(Because  $y = u - x$  and since for  $x_{\min} = 0$  one has  $y_{\max} = u$ .)

(b)  $0 \leq x < \infty, 0 \leq y < \infty$  :

$$f(u) = \int_0^u f_x(u-y) f_y(y) dy . \quad (5.11.7)$$

(c)  $a \leq x < b, -\infty < y < \infty$  :

$$f(u) = \int_a^b f_x(x) f_y(u-x) dx . \quad (5.11.8)$$

We will demonstrate case (d) in the following example, in which both  $x$  and  $y$  are bounded from below and from above.

### Example 5.8: Convolution of uniform distributions

With

$$f_x(x) = \begin{cases} 1, & 0 \leq x < 1 \\ 0, & \text{otherwise} \end{cases} \quad \text{and} \quad f_y(y) = \begin{cases} 1, & 0 \leq y < 1 \\ 0, & \text{otherwise} \end{cases}$$

and Eq. (5.11.8) we obtain

$$f(u) = \int_0^1 f_y(u-x) dx .$$

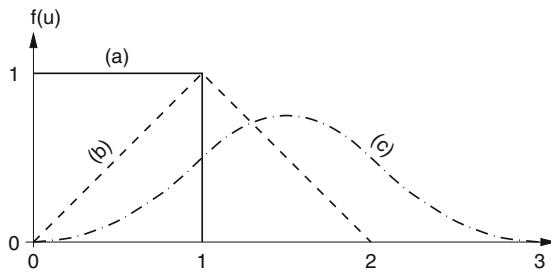
We substitute  $v = u - x, dv = -dx$  and obtain

$$f(u) = - \int_u^{u-1} f_y(v) dv = \int_{u-1}^u f_y(v) dv . \quad (5.11.9)$$

Clearly one has  $0 < u < 2$ . We now consider separately the two cases

$$(a) \quad 0 \leq u < 1 : \quad f_1(u) = \int_0^u f_y(v) dv = \int_0^u 1 dv = u ,$$

$$(b) \quad 1 \leq u < 2 : \quad f_2(u) = \int_{u-1}^1 f_y(v) dv = \int_{u-1}^1 1 dv = 2 - u . \quad (5.11.10)$$



**Fig. 5.15:** Convolution of uniform distributions. Probability density of the sum  $u$  of uniformly distributed random variables  $x$  (a)  $u = x$ , (b)  $u = x + x$ , (c)  $u = x + x + x$ .

Note that the lower (upper) limit of integration is not lower (higher) than the value 0 (1). The result is a triangular distribution (Fig. 5.15).

If this result is folded again with a uniform distribution, i.e., if  $u$  is the sum of three independent uniformly distributed variables, then one obtains

$$f(u) = \begin{cases} \frac{1}{2}u^2, & 0 \leq u < 1 \\ \frac{1}{2}(-2u^2 + 6u - 3), & 1 \leq u < 2 \\ \frac{1}{2}(u - 3)^2, & 2 \leq u < 3 \end{cases} . \quad (5.11.11)$$

The proof is left to the reader. The distribution consists of three parabolic sections (Fig. 5.15) and is similar already to the Gaussian distribution predicted by the Central Limit Theorem. ■

## 5.11.2 Convolutions with the Normal Distribution

Suppose a quantity of experimental interest  $x$  can be considered to be a random variable with probability density  $f_x(x)$ . It is measured with a measurement error  $y$ , which follows a normal distribution with a mean of zero and a variance of  $\sigma^2$ . The result of the measurement is then the sum

$$u = x + y . \quad (5.11.12)$$

Its probability density is [see also (5.11.4)]

$$f(u) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} f_x(x) \exp[-(u-x)^2/2\sigma^2] dx . \quad (5.11.13)$$

By carrying out many measurements,  $f(u)$  can be experimentally determined. The experimenter is interested, however, in the function  $f_x(x)$ . Unfortunately, Eq. (5.11.13) cannot in general be solved for  $f_x(x)$ . This is only possible for a restricted class of functions  $f(u)$ . Therefore one usually approaches the problem in a different way. From earlier measurements or theoretical considerations one possesses knowledge about the form of  $f_x(x)$ , e.g., one might assume that  $f_x(x)$  is described by a uniform distribution, without,

however, knowing its boundaries  $a$  and  $b$ . One then carries out the convolution (5.11.13), compares the resulting function  $f(u)$  with the experiment and in this way determines the unknown parameters (in our example  $a$  and  $b$ ).

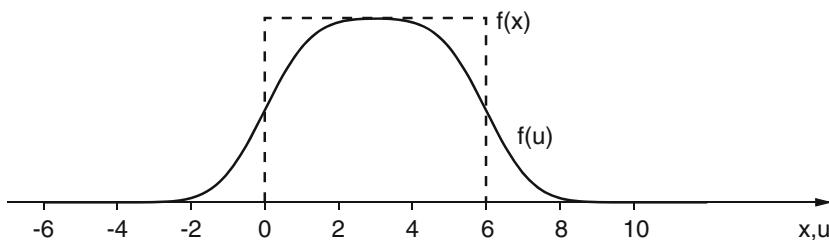
In many cases it is not even possible to perform the integration (5.11.13) analytically. Numerical procedures, e.g., the Monte Carlo method, then have to be used. Sometimes approximations (cf. Example 5.11) give useful results. Because of the importance in many experiments of convolution with the normal distribution we will study some examples.

**Example 5.9:** Convolution of uniform and normal distributions

Using Eqs. (3.3.26) and (5.11.8) and substituting  $v = (x - u)/\sigma$  we obtain

$$\begin{aligned} f(u) &= \frac{1}{b-a} \frac{1}{\sqrt{2\pi}\sigma} \int_a^b \exp[-(u-x)^2/2\sigma^2] dx \\ &= \frac{1}{b-a} \frac{1}{\sqrt{2\pi}} \int_{(a-u)/\sigma}^{(b-u)/\sigma} \exp(-\frac{1}{2}v^2) dv , \\ f(u) &= \frac{1}{b-a} \left\{ \psi_0\left(\frac{b-u}{\sigma}\right) - \psi_0\left(\frac{a-u}{\sigma}\right) \right\} . \end{aligned} \quad (5.11.14)$$

The function  $\psi$  has already been defined in (5.6.5). Figure 5.16 shows the result for  $a = 0$ ,  $b = 6$ ,  $\sigma = 1$ . If one has  $|b-a| \gg \sigma$  (as is the case in Fig. 5.16), one of the terms in parentheses in (5.11.14) is either 0 or 1. The rising edge of the uniform distribution at  $u = a$  is replaced by the distribution function of the normal distribution with standard deviation  $\sigma$  (see also Fig. 5.7). The falling edge at  $u = b$  is its “mirror image”. ■



**Fig. 5.16:** Convolution of a uniform and Gaussian distribution.

**Example 5.10:** Convolution of two normal distributions. “Quadratic addition of errors”

If one convolutes two normal distributions with mean values 0 and variances  $\sigma_x^2$  and  $\sigma_y^2$ , one obtains

$$f(u) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-u^2/2\sigma^2) , \quad \sigma^2 = \sigma_x^2 + \sigma_y^2 . \quad (5.11.15)$$

The proof has already been shown in Sect. 5.7 with the help of the characteristic function. It can also be obtained by computation of the folding integral (5.11.5). If the distributions  $f_x(x)$  and  $f_y(y)$  describe two independent sources of measurement errors, the result (5.11.15) is known as the “*quadratic addition of errors*”. ■

**Example 5.11:** Convolution of exponential and normal distributions

With

$$\begin{aligned} f_x(x) &= \frac{1}{\tau} \exp(-x/\tau) \quad , \quad x > 0 \quad , \\ f_y(y) &= \frac{1}{\sqrt{2\pi}\sigma} \exp(-y^2/2\sigma^2) \quad , \end{aligned}$$

Eq. (5.11.6) takes on the following form:

$$f(u) = \frac{1}{\sqrt{2\pi}\sigma\tau} \int_{-\infty}^u \exp[-(u-y)/\tau] \exp(-y^2/2\sigma^2) dy \quad .$$

We can rewrite the exponent

$$\begin{aligned} -\frac{1}{2\sigma^2\tau}[2\sigma^2(u-y)+\tau y^2] \\ &= -\frac{1}{2\sigma^2\tau} \left[ 2\sigma^2u - 2\sigma^2y + \tau y^2 + \frac{\sigma^4}{\tau} - \frac{\sigma^4}{\tau} \right] \\ &= -\frac{u}{\tau} + \frac{\sigma^2}{2\tau^2} - \frac{1}{2\sigma^2} \left( y - \frac{\sigma^2}{\tau} \right)^2 \end{aligned}$$

and obtain

$$f(u) = \frac{1}{\sqrt{2\pi}\sigma\tau} \exp \left\{ \frac{\sigma^2}{2\tau^2} - \frac{u}{\tau} \right\} \int_{-\infty}^{u-\sigma^2/\tau} \exp \left( \frac{-v^2}{2\sigma^2} \right) dv \quad .$$

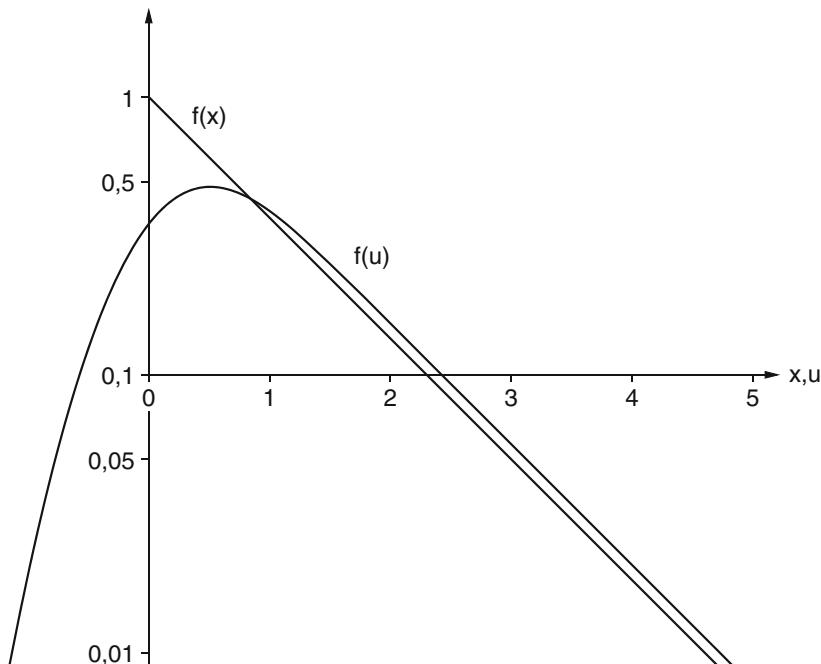
We now require that  $\sigma \ll \tau$ , i.e., that the measurement error is much smaller than the typical value (width) of the exponential distribution. In addition, we only consider values of  $u$  for which  $u - \sigma^2/\tau \gg \sigma$ , i.e.,  $u \gg \sigma$ . The integral is then approximately equal to  $\sqrt{2\pi}\sigma$  or

$$f(u) \approx \frac{1}{\tau} \exp \left\{ -\frac{u}{\tau} + \frac{\sigma^2}{2\tau^2} \right\} \quad .$$

In a semi-logarithmic representation, i.e., in a plot of  $\ln f(u)$  versus  $u$ , the curve  $f(u)$  lies above the curve  $f_x(x)$ , by an amount  $\sigma^2/2\tau^2$ , since

$$\ln f(u) = \ln \frac{1}{\tau} + \frac{\sigma^2}{2\tau^2} - \frac{u}{\tau} = \ln f_x(x) + \frac{\sigma^2}{2\tau^2} \quad .$$

This is plotted in Fig. 5.17. The result can be qualitatively understood in the following way. For each small  $x$  interval of the exponential distribution, the convolution leads with equal probability to a shift to the left or to the right. Since, however, the exponential distribution for a given  $u$  is greater for small values of  $x$ , contributions to the convolution  $f(u)$  originate with greater probability from the left than from the right. This leads to an overall shift to the right of  $f(u)$  with respect to  $f_x(x)$ . ■



**Fig. 5.17:** Convolution of exponential and normal distributions.

## 5.12 Example Programs

**Example Program 5.1:** Class E1Distrib to simulate empirical frequency and demonstrate statistical fluctuations

The program simulates the problem of Example 5.1. It allows input of values for  $n_{\text{exp}}$ ,  $n_{\text{fly}}$ , and  $P(A)$ , and then consecutively performs  $n_{\text{exp}}$  simulated experiments. In each experiment  $n_{\text{fly}}$  objects are analyzed. Each object has a probability  $P(A)$  to have the property  $A$ . For each experiment one line of output is produced containing the current number  $i_{\text{exp}}$  of the experiment, the number  $N_A$  of objects with the property  $A$  and the frequency  $h_A = N_A/n_{\text{fly}}$  with which the property  $A$  was found. The fluctuation of  $h_A$  around the known input value  $P(A)$  in the individual experiments gives a good impression of the statistical error of an experiment.

**Example Program 5.2:** Class E2Distrib to simulate the experiment of Rutherford and Geiger

The principle of the experiment of Rutherford and Geiger is described in Example 5.3. It is simulated as follows. Input quantities are the number  $N$  of decays observed and the number  $n_{\text{int}}$  of partial intervals  $\Delta T$  of the total observation time  $T$ . For simplicity the length of each partial interval is set equal to one. A total of  $N$  random events are simulated by simply generating  $N$  random numbers uniformly distributed between 0 and  $T$ . They are entered into a histogram with  $n_{\text{int}}$  intervals. The histogram is first displayed graphically and then analyzed numerically. For each number  $k = 0, 1, \dots, N_{\text{int}}$  the program determines how many intervals  $N(k)$  of the histogram have  $k$  entries. The numbers  $N(k)$  themselves are presented in the form of another histogram.

Show that for the process simulated in this example program one obtains in the limit  $N \rightarrow \infty$

$$N(k) = n_{\text{int}} W_k^N (p = 1/n_{\text{int}}) .$$

If  $N$  is increased step by step and at the same time  $\lambda = N_p = N/n_{\text{int}}$  is kept constant, then for large  $N$  one has

$$W_k^N (p = \lambda/N) \rightarrow \frac{\lambda^k}{k!} e^{-\lambda}$$

and, in the limit  $N \rightarrow \infty$ ,

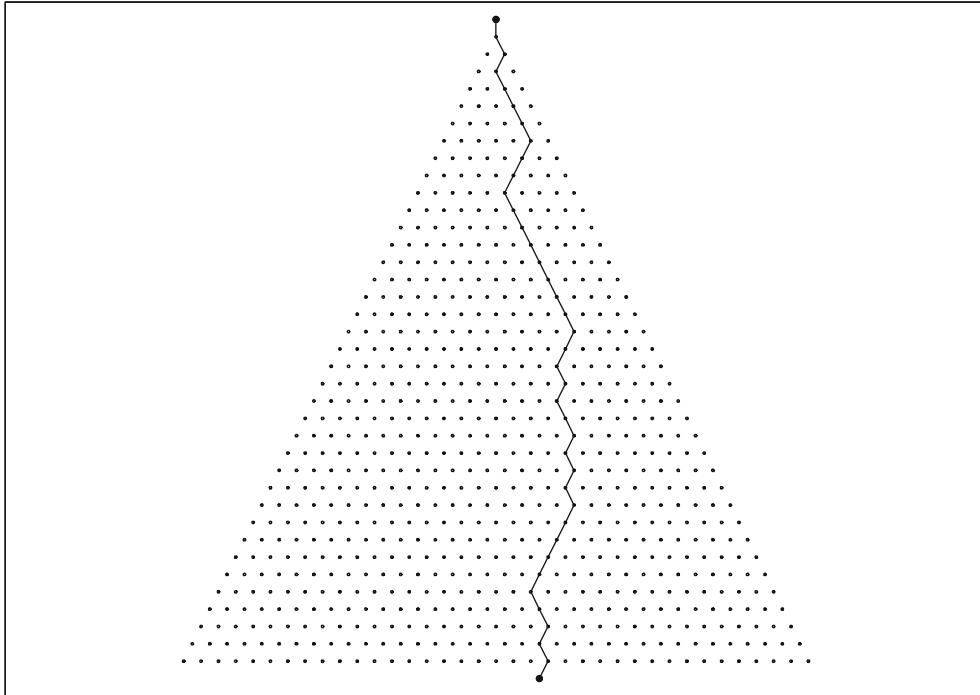
$$N(k) = n_{\text{int}} \frac{\lambda^k}{k!} e^{-\lambda} .$$

Check the above statements by running the program with suitable pairs of numbers, e.g.,  $(N, n_{\text{int}}) = (4, 2), (40, 20), \dots, (2000, 1000)$ , by reading the numbers  $N(k)$  from the graphics display and by comparing them with the statements above.

**Example Program 5.3:** Class E3Distrib to simulate Galton's board

Galton's board is a simple implementation of Laplace's model described in Example 5.7. The vertical board contains rows of horizontally oriented nails as shown in Fig. 5.18. The rows of nails are labeled  $j = 1, 2, \dots, n$ , and row  $j$  has  $j$  nails. One by one a total of  $N_{\text{exp}}$  balls fall onto the nail in row 1. There each ball is deflected with probability  $p$  to the right and with probability  $(1 - p)$  to the left. (In a realistic board one has  $p = 1/2$ .) The distance between the nails is chosen in such a way that in each case the ball hits one of the two nails in row 2 and there again it is deflected with the probability  $p$  to the right. After falling through  $n$  rows each ball assumes one of  $n + 1$  places, which we denote by  $k = 0$  (on the left),  $k = 1, \dots, k = n$  (on the right). After a total of  $N_{\text{exp}}$  experiments (i.e., balls) one finds  $N(k)$  balls for each value  $k$ .

The program allows input of numerical values for  $N_{\text{exp}}$ ,  $n$ , and  $p$ . For each experiment the number  $k$  is first set to zero and  $n$  random numbers  $r_j$  are generated from a uniform distribution and analyzed. For each  $r_j < p$  (corresponding to a deflection to the right in row  $j$ ) the number  $k$  is increased by 1. For each experiment the value of  $k$  is entered into a histogram. After all experiments are simulated the histogram is displayed.



**Fig. 5.18:** Arrangement of nails in Galton's board and one possible trajectory of a ball.

Show that in the limit  $N \rightarrow \infty$  one has

$$N(k) = N W_k^n(p) \quad .$$

By choosing and entering suitable pairs of numbers  $(n, p)$ , e.g.,  $(n, p) = (1, 0.5)$ ,  $(2, 0.25)$ ,  $(10, 0.05)$ ,  $(100, 0.005)$  approximate the Poisson limit

$$N(k) = N \frac{\lambda^k}{k!} e^{-\lambda} \quad , \quad \lambda = np \quad ,$$

and compare these predictions with the results of your simulations.

# 6. Samples

In the last chapter we discussed a number of distributions, but we have not specified how they are realized in a particular case. We have only given the probability that a random variable  $\mathbf{x}$  will lie within an interval with boundaries  $x$  and  $x + dx$ . This probability depends on certain parameters describing its distribution (like  $\lambda$  in the case of the Poisson distribution) which are usually unknown. We therefore have no direct knowledge of the probability distribution and have to approximate it by a *frequency distribution* obtained experimentally. The number of measurements performed for this purpose, called a *sample*, is necessarily finite. To discuss the elements of sampling theory we first have to introduce a number of new definitions.

## 6.1 Random Samples. Distribution of a Sample. Estimators

Every sample is taken from a set of elements which correspond to the possible results of an individual observation. Such a set, which usually has infinitely many elements, is called a *population*. If a sample of  $n$  elements is taken from it, then we say that the sample has the *size*  $n$ . Let the distribution of the random variable  $\mathbf{x}$  in the population be given by the probability density  $f(x)$ . We are interested in the values of  $\mathbf{x}$  assumed by the individual elements of the sample. Suppose that we take  $\ell$  samples of size  $n$  and find the following values for  $\mathbf{x}$ :

$$\begin{aligned} \text{1st sample: } & \mathbf{x}_1^{(1)}, \mathbf{x}_2^{(1)}, \dots, \mathbf{x}_n^{(1)}, \\ & \vdots \\ \text{jth sample: } & \mathbf{x}_1^{(j)}, \mathbf{x}_2^{(j)}, \dots, \mathbf{x}_n^{(j)}, \\ & \vdots \\ \text{\ellth sample: } & \mathbf{x}_1^{(\ell)}, \mathbf{x}_2^{(\ell)}, \dots, \mathbf{x}_n^{(\ell)}. \end{aligned}$$

We group the result of one sample into an  $n$ -dimensional vector

$$\mathbf{x}^{(j)} = (x_1^{(j)}, x_2^{(j)}, \dots, x_n^{(j)}) \quad , \quad (6.1.1)$$

which can be considered the position vector in an  $n$ -dimensional sample space (Sect. 2.1). Its probability density is

$$g(\mathbf{x}) = g(x_1, x_2, \dots, x_n) \quad . \quad (6.1.2)$$

This function must fulfill two conditions in order for the sample to be *random*:

- (a) The individual  $x_i$  must be independent, e.g., one must have

$$g(\mathbf{x}) = g_1(x_1)g_2(x_2)\dots g_n(x_n) \quad . \quad (6.1.3)$$

- (b) The individual marginal distributions must be identical and equal to the probability density  $f(x)$  of the population,

$$g_1(x) = g_2(x) = \dots = g_n(x) = f(x) \quad . \quad (6.1.4)$$

Comparing with (6.1.2) it is clear that there is a simple relation between a population and a sample only if these conditions are fulfilled. In the following we will mean by the word sample a random sample unless otherwise stated.

It should be emphasized that in the actual process of sampling it is often quite difficult to ensure randomness. Because of the large variety of applications, a general prescription cannot be given. In order to obtain reliable results from sampling, we have to take the utmost precautions to meet the requirements (6.1.3) and (6.1.4). Independence (6.1.3) can be checked to a certain extent by comparing the frequency distributions of the first, second, ... elements of a large number of samples. It is very difficult, however, to ensure that the samples in fact come from a population with the probability density  $f(x)$ . If the elements of the population can be numbered, it is often useful to use random numbers to select the elements for the sample.

We now suppose that the  $n$  elements of a sample are ordered according to the value of the variable, e.g., marked on the  $x$  axis, and we ask for the number of elements of the sample  $n_x$  for which  $\mathbf{x} < x$ , for arbitrary  $x$ . The function

$$W_n(x) = n_x/n \quad (6.1.5)$$

takes on the role of an empirical distribution function. It is a step function that increases by  $1/n$  as soon as  $x$  is equal to one of the values  $x$  of an element of the sample. It is called the *sample distribution function*. It is clearly an approximation for  $F(x)$ , the distribution function of the population, which it approaches in the limit  $n \rightarrow \infty$ .

A function of the elements of a sample (6.1.1) is called a *statistic*. Since  $\mathbf{x}$  is a random variable, a statistic is itself a random variable. The most important example is the *sample mean*,

$$\bar{\mathbf{x}} = \frac{1}{n}(\mathbf{x}_1 + \mathbf{x}_2 + \dots + \mathbf{x}_n) \quad . \quad (6.1.6)$$

A typical problem of data analysis is the following. The general mathematical form of the probability density of the population is known. In radioactive decay, for example, the number of nuclei which decay before the time  $t = \tau$  is  $N_\tau = N_0(1 - \exp(-\lambda\tau))$ , if  $N_0$  nuclei existed at time  $t = 0$ . Here, however, the decay constant  $\lambda$  is, in general, not known. By taking a finite sample (measuring a finite number of decay times of individual nuclei) we want to determine the parameter  $\lambda$  as accurately as possible. Since such a task cannot be exactly solved, because the sample is finite, one speaks of the *estimation of parameters*. To estimate a parameter  $\lambda$  of a distribution function one uses an *estimator*

$$\mathbf{S} = \mathbf{S}(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n) \quad . \quad (6.1.7)$$

An estimator is said to be *unbiased* if for arbitrary sample size the expectation value of the (random) quantity  $\mathbf{S}$  is equal to the parameter to be estimated:

$$E\{\mathbf{S}(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)\} = \lambda \quad \text{for all } n. \quad (6.1.8)$$

An estimator is said to be *consistent* if its variance vanishes for arbitrarily large sample size, i.e., if

$$\lim_{n \rightarrow \infty} \sigma(\mathbf{S}) = 0 \quad . \quad (6.1.9)$$

Often one can give a lower limit for the variance of an estimator of a parameter. If one finds an estimator  $\mathbf{S}_0$  whose variance is equal to this limit, then one apparently has the “best possible” estimator.  $\mathbf{S}_0$  is then said to be an *efficient estimator* for  $\lambda$ .

## 6.2 Samples from Continuous Populations: Mean and Variance of a Sample

The case of greatest interest in applications concerns a sample from an infinitely large continuous population described by the probability density  $f(x)$ . The sample mean (6.1.6) is a random variable, as are all statistics. Let us consider its expectation value

$$E(\bar{\mathbf{x}}) = \frac{1}{n}\{E(\mathbf{x}_1) + E(\mathbf{x}_2) + \dots + E(\mathbf{x}_n)\} = \hat{x} \quad . \quad (6.2.1)$$

This expectation value is equal to the expectation value of  $\mathbf{x}$ . Since Eq. (6.2.1) holds for all values of  $n$ , the arithmetic mean of a sample is, as one would expect, an unbiased estimator for the mean of the population. The characteristic function of the random variable  $\bar{\mathbf{x}}$  is

$$\varphi_{\bar{\mathbf{x}}}(t) = \left\{ \varphi_{\frac{\mathbf{x}}{n}}(t) \right\}^n = \left\{ \varphi_{\mathbf{x}}\left(\frac{t}{n}\right) \right\}^n . \quad (6.2.2)$$

Next we are interested in the variance of  $\bar{\mathbf{x}}$ ,

$$\begin{aligned} \sigma^2(\bar{\mathbf{x}}) &= E\{(\bar{\mathbf{x}} - E(\bar{\mathbf{x}}))^2\} = E\left\{ \left( \frac{\mathbf{x}_1 + \mathbf{x}_2 + \dots + \mathbf{x}_n}{n} - \hat{x} \right)^2 \right\} \\ &= \frac{1}{n^2} E\{[(\mathbf{x}_1 - \hat{x}) + (\mathbf{x}_2 - \hat{x}) + \dots + (\mathbf{x}_n - \hat{x})]^2\} . \end{aligned}$$

Since all of the  $\mathbf{x}_i$  are independent, all of the cross terms of the type  $E\{(\mathbf{x}_i - \hat{x})(\mathbf{x}_j - \hat{x})\}$ ,  $i \neq j$  (i.e., all of the covariances) vanish, and we obtain

$$\sigma^2(\bar{\mathbf{x}}) = \frac{1}{n} \sigma^2(\mathbf{x}) . \quad (6.2.3)$$

One thus shows that  $\bar{\mathbf{x}}$  is a consistent estimator for  $\hat{x}$ . The variance (6.2.3) is itself, however, not a random variable, and is therefore not directly obtainable by experiment. As a definition for the *sample variance* we could try using the arithmetic mean of squared differences

$$\mathbf{s}'^2 = \frac{1}{n} \{(\mathbf{x}_1 - \bar{\mathbf{x}})^2 + (\mathbf{x}_2 - \bar{\mathbf{x}})^2 + \dots + (\mathbf{x}_n - \bar{\mathbf{x}})^2\} . \quad (6.2.4)$$

Its expectation value is

$$\begin{aligned} E(\mathbf{s}'^2) &= \frac{1}{n} E\left\{ \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})^2 \right\} \\ &= \frac{1}{n} E\left\{ \sum_{i=1}^n (\mathbf{x}_i - \hat{x} + \hat{x} - \bar{\mathbf{x}})^2 \right\} \\ &= \frac{1}{n} E\left\{ \sum_{i=1}^n (\mathbf{x}_i - \hat{x})^2 + \sum_{i=1}^n (\hat{x} - \bar{\mathbf{x}})^2 + 2 \sum_{i=1}^n (\mathbf{x}_i - \hat{x})(\hat{x} - \bar{\mathbf{x}}) \right\} \\ &= \frac{1}{n} \sum_{i=1}^n \{E((\mathbf{x}_i - \hat{x})^2) - E((\bar{\mathbf{x}} - \hat{x})^2)\} \\ &= \frac{1}{n} \left\{ n\sigma^2(\mathbf{x}) - n\left(\frac{1}{n}\sigma^2(\mathbf{x})\right) \right\} , \\ E(\mathbf{s}'^2) &= \frac{n-1}{n} \sigma^2(\mathbf{x}) . \end{aligned} \quad (6.2.5)$$

Hence one sees that the sample variance defined in this way is a biased estimator for the population variance, having an expectation value smaller than  $\sigma^2(x)$ . We can see directly from (6.2.5), however, the size of the bias. We therefore change our definition (6.2.4) and write for the sample variance

$$s^2 = \frac{1}{n-1} \{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2\} . \quad (6.2.6)$$

This is now an unbiased estimator for  $\sigma^2(x)$ . The value  $(n-1)$  in the denominator appears at first to be somewhat strange. One must consider, however, that for  $n=1$  the sample mean is equal to the value  $x$  of the sole element of the sample ( $x = \bar{x}$ ) and that therefore the quantity (6.2.4) would vanish. That is related to the fact that in (6.2.4) – and also in (6.2.6) – the sample mean  $\bar{x}$  was used instead of the population mean  $\hat{x}$ , since the latter was not known. Part of the information contained in the sample first had to be used and was not available for the calculation of the variance. The effective number of elements available for calculating the variance is therefore reduced. This is taken into consideration by reducing the denominator of the arithmetic mean (6.2.4). The same line of reasoning is repeated quantitatively in Sect. 6.5.

If we substitute the estimator for the population variance (6.2.6) into (6.2.3), we obtain an estimator for the *variance of the mean*

$$s^2(\bar{x}) = \frac{1}{n} s^2(x) = \frac{1}{n(n-1)} \sum_{i=1}^n (x_i - \bar{x})^2 . \quad (6.2.7)$$

The corresponding standard deviation can be considered to be the *error of the mean*

$$\Delta \bar{x} = \sqrt{s^2(\bar{x})} = s(\bar{x}) = \frac{1}{\sqrt{n}} s(x) . \quad (6.2.8)$$

Of course we are also interested in the error of the sample variance (6.2.6). In Sect. 6.6 we will show that this quantity can be determined under the assumption that the population follows a normal distribution. We will use the result here ahead of time. The variance of  $s^2$  is

$$\text{var}(s^2) = \left( \frac{\sigma^2}{n-1} \right)^2 2(n-1) . \quad (6.2.9)$$

If we substitute the estimator (6.2.6) into the right-hand side for  $\sigma^2$  and take the square root, we obtain for the error of the sample variance

$$\Delta s^2 = s^2 \sqrt{\frac{2}{(n-1)}} . \quad (6.2.10)$$

Finally we give explicit expressions for estimators of the *sample standard deviation* and its error. The first is simply the square root of the sample variance

$$\mathbf{s} = \sqrt{\mathbf{s}^2} = \frac{1}{\sqrt{n-1}} \sqrt{\sum (\mathbf{x}_i - \bar{\mathbf{x}})^2} . \quad (6.2.11)$$

The error of the sample standard deviation is obtained from (6.2.10) by error propagation, which gives

$$\Delta \mathbf{s} = \frac{\mathbf{s}}{\sqrt{2(n-1)}} . \quad (6.2.12)$$

**Example 6.1:** Computation of the sample mean and variance from data

Suppose one has  $n = 7$  measurements of a certain quantity (e.g., the length of an object). Their values are 10.5, 10.9, 9.2, 9.8, 9.0, 10.4, 10.7. The computation is made easier if one uses the fact that all of the measured values are near  $a = 10$ , i.e., they are of the form  $\mathbf{x}_i = a + \delta_i$ . The relation (6.1.6) then gives

$$\bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i = \frac{1}{n} \sum_{i=1}^n (a + \delta_i) = a + \frac{1}{n} \sum_{i=1}^n \delta_i = a + \Delta$$

with

$$\begin{aligned} \Delta &= \frac{1}{n} \sum_{i=1}^n \delta_i = \frac{1}{7} (0.5 + 0.9 - 0.8 - 0.2 - 1.0 + 0.4 + 0.7) \\ &= 0.5/7 = 0.07 . \end{aligned}$$

We thus have  $\bar{\mathbf{x}} = 10 + \Delta = 10.07$ .

The sample variance is computed according to (6.2.6) to be

$$\begin{aligned} \mathbf{s}^2 &= \frac{1}{n-1} \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})^2 \\ &= \frac{1}{n-1} \sum_{i=1}^n (\mathbf{x}_i^2 - 2\mathbf{x}_i \bar{\mathbf{x}} + \bar{\mathbf{x}}^2) \\ &= \frac{1}{n-1} \left\{ \sum_{i=1}^n \mathbf{x}_i^2 - n\bar{\mathbf{x}}^2 \right\} . \end{aligned}$$

The result can be obtained either by the first or last line of the relation above. In the last line only one difference is computed, not  $n$ . The numbers to be squared, however, are usually considerably larger, and one must consider the problem of rounding errors. We therefore use the original expression

$$\begin{aligned} \mathbf{s}^2 &= \frac{1}{6} \{0.43^2 + 0.83^2 + 0.87^2 + 0.27^2 + 1.07^2 + 0.33^2 + 0.63^2\} \\ &= \frac{1}{6} \{0.1849 + 0.6889 + 0.7569 + 0.0729 + 1.1449 + 0.1089 \\ &\quad + 0.3969\} \\ &= 3.3543/6 \approx 0.56 . \end{aligned}$$

The sample standard deviation is  $s \approx 0.75$ . From (6.2.8), (6.2.10), and (6.2.12) we obtain finally  $\Delta\bar{x} = 0.28$ ,  $\Delta s^2 = 0.32$ , and  $\Delta s = 0.21$ . ■

Naturally one does not usually compute the sample mean and variance by hand, but rather by the class `Sample` and its methods.

## 6.3 Graphical Representation of Samples: Histograms and Scatter Plots

After the theoretical considerations of the last sections we now turn to some simple practical aspects of the analysis of sample data. An important tool for this is the representation of the data in graphical form.

A sample

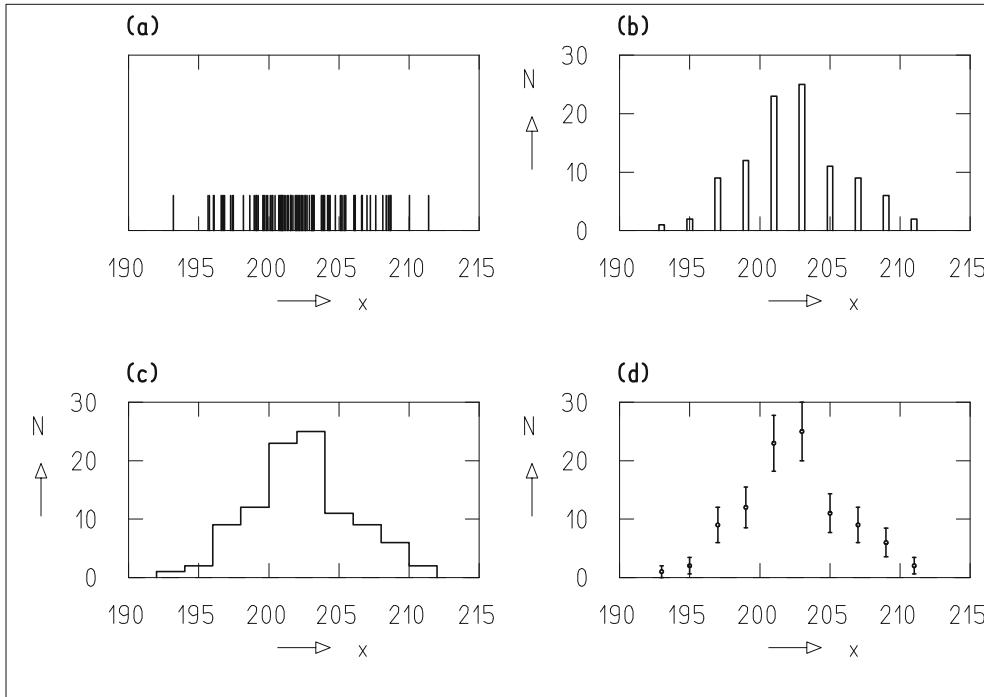
$$x_1, x_2, \dots, x_n ,$$

which depends on a single variable  $x$  can be represented simply by means of tick marks on an  $x$  axis. We will call such a representation a *one-dimensional*

**Table 6.1:** Values of resistance  $R$  of 100 individual resistors of nominal value  $200\text{ k}\Omega$ . The data are graphically represented in Fig. 6.1.

|         |         |         |         |         |
|---------|---------|---------|---------|---------|
| 193.199 | 195.673 | 195.757 | 196.051 | 196.092 |
| 196.596 | 196.679 | 196.763 | 196.847 | 197.267 |
| 197.392 | 197.477 | 198.189 | 198.650 | 198.944 |
| 199.070 | 199.111 | 199.153 | 199.237 | 199.698 |
| 199.572 | 199.614 | 199.824 | 199.908 | 200.118 |
| 200.160 | 200.243 | 200.285 | 200.453 | 200.704 |
| 200.746 | 200.830 | 200.872 | 200.914 | 200.956 |
| 200.998 | 200.998 | 201.123 | 201.208 | 201.333 |
| 201.375 | 201.543 | 201.543 | 201.584 | 201.711 |
| 201.878 | 201.919 | 202.004 | 202.004 | 202.088 |
| 202.172 | 202.172 | 202.297 | 202.339 | 202.381 |
| 202.507 | 202.591 | 202.633 | 202.716 | 202.884 |
| 203.051 | 203.052 | 203.094 | 203.094 | 203.177 |
| 203.178 | 203.219 | 203.764 | 203.765 | 203.848 |
| 203.890 | 203.974 | 204.184 | 204.267 | 204.352 |
| 204.352 | 204.729 | 205.106 | 205.148 | 205.231 |
| 205.357 | 205.400 | 205.483 | 206.070 | 206.112 |
| 206.154 | 206.155 | 206.615 | 206.657 | 206.993 |
| 207.243 | 207.621 | 208.124 | 208.375 | 208.502 |
| 208.628 | 208.670 | 208.711 | 210.012 | 211.394 |

*scatter plot.* It contains all the information about the sample. Table 6.1 contains the values  $x_1, x_2, \dots, x_n$  of a sample of size 100, obtained from measuring the resistance  $R$  of 100 individual resistors of nominal value  $200\text{k}\Omega$ . After obtaining the sample the measurements were ordered.



**Fig. 6.1:** Representation of the data from Table 6.1 as (a) a one-dimensional scatter plot, (b) a bar diagram, (c) a step diagram, and (d) a diagram of measured points with error bars.

Figure 6.1a shows the corresponding scatter plot. Qualitatively one can estimate the mean and variance from the position and width of the clustering of tick marks.

Another graphical representation is usually better suited to visualize the sample by using the second dimension available on the paper. The  $x$  axis is used as abscissa and divided into  $r$  intervals

$$\xi_1, \xi_2, \dots, \xi_r$$

of equal width  $\Delta x$ . These intervals are called *bins*. The centers of the bins have the  $x$ -values

$$x_1, x_2, \dots, x_r \quad .$$

On the vertical axis one plots the corresponding numbers of sample elements

$$n_1, n_2, \dots, n_r$$

that fall into bins  $\xi_1, \xi_2, \dots, \xi_r$ . The diagram obtained in this way is called a *histogram* of the sample. This can be interpreted as a frequency distribution, since  $h_k = n_k/n$  is a frequency, i.e., a measure of the probability  $p_k$  to observe a sample element in the interval  $\xi_k$ . For the graphical form of histograms, various methods are used. In a *bar diagram* the values  $n_k$  are represented as bars perpendicular to the  $x$  axis on the  $x_k$  values (Fig. 6.1b). In a *step diagram* the  $n_k$  are represented as horizontal lines that cover the entire width  $\xi_k$  of the interval. Neighboring horizontal lines are connected by perpendicular lines (Fig. 6.1c). The fraction of the area covering each interval  $\xi_k$  of the  $x$  axis is then proportional to the number  $n_k$  of the sample elements in the interval. (If one uses the area in the interval for the graphical representation of  $n_k$ , then the bins can also have different widths.) In economics bar diagrams are most commonly used. (Sometimes one also sees diagrams in which, instead of bars, line segments are used to connect the tips of the bars. In contrast to the step diagram, the resulting figure does not have an area proportional to the sample size  $n$ .) In the natural sciences, step diagrams are more common.

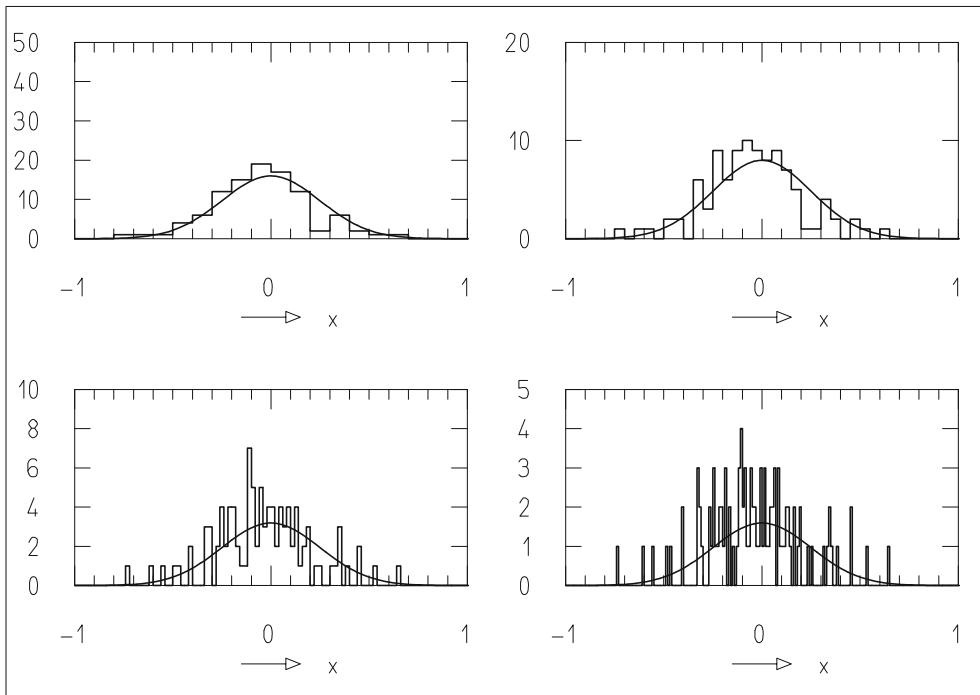
In Sect. 6.8 we will determine that as long as the values  $n_k$  are not too small, their statistical errors are given by  $\Delta n_k = \sqrt{n_k}$ . In order to plot them on a graph, the observed values  $n_k$  can be drawn as points with vertical *error bars* ending at the points  $n_k \pm \sqrt{n_k}$  (Fig. 6.1d).

It is clear that the relative errors  $\Delta n_k / n_k = 1/\sqrt{n_k}$  decrease for increasing  $n_k$ , i.e., for a sample of fixed size  $n$  they decrease for increasing bin width of the histogram. On the other hand, by choosing a larger interval width, one loses any finer structure of the data with respect to the variable  $x$ . The ability of a histogram to convey information therefore depends crucially on the appropriate choice of the bin width, usually found only after several attempts.

### Example 6.2: Histograms of the same sample with various choices of bin width

In Fig. 6.2 four histograms of the same sample are shown. The population is a Gaussian distribution, which is also shown as a continuous curve. This was scaled in such a way that the area under the histogram is equal to the area under the Gaussian curve. Although the information contained in the plot is greater for a smaller bin width – for vanishing bin width the histogram becomes a one dimensional scatter plot – one notices the similarity between the histogram and Gaussian curve much more easily for the larger bin width. This is because for the larger bin width the relative statistical fluctuations of the contents of individual bins are smaller. The individual steps of the histogram differ less from the curve. ■

Constructing a histogram from a sample is a simple programming task. Suppose the histogram has  $n_x$  bins of width  $\Delta x$ , with the first interval extending from  $x = x_0$  to  $x = x_0 + \Delta x$ . The contents of the histogram is put into an



**Fig. 6.2:** Histogram of the same sample from a Gaussian distribution represented with four different bin widths.

array hist, with the first bin in hist[0] the second bin in hist[1], etc.

The histogram is then specified in the computer by the array hist and the three values  $x_0$ ,  $\Delta x$ , and  $n_x$ . The class Histogram permits construction and administration of a histogram.

Graphical display of histograms can be accomplished using methods of the class DatanGraphics (Appendix F). With them the user can freely adjust all of the parameters that determine the appearance of the plot, such as the page format, scale factors, colors, line thickness, etc. Often it is convenient to use the class GraphicsWithHistogram which does not allow for this freedom but by a single call gives rise to the graphical output of a histogram, stored in the computer.

A histogram allows a first direct look at the nature of the data. It answers questions such as “Are the data more or less distributed according to a Gaussian?” or “Do there exist points exceptionally far away from the average value?”. If the histogram leads one to conclude that the population is distributed according to a Gaussian, then the mean and standard deviation can be estimated directly from the plot. The mean is the center of gravity of the histogram. The standard deviation is obtained as in the following example.

**Example 6.3:** Full width at half maximum (FWHM)

If the form of histogram allows one to assume that the represented sample originates from a Gaussian distribution, then one can draw by hand a Gaussian bell curve that follows the histogram as closely as possible. The position of the maximum is a good estimator for the mean of the sample. One then draws a horizontal line at half of the height of the maximum. This crosses the bell curve at the points  $x_a$  and  $x_b$ . The quantity

$$f = x_b - x_a$$

is called the *full width at half maximum* (FWHM). One can easily determine for a Gaussian distribution the simple relation

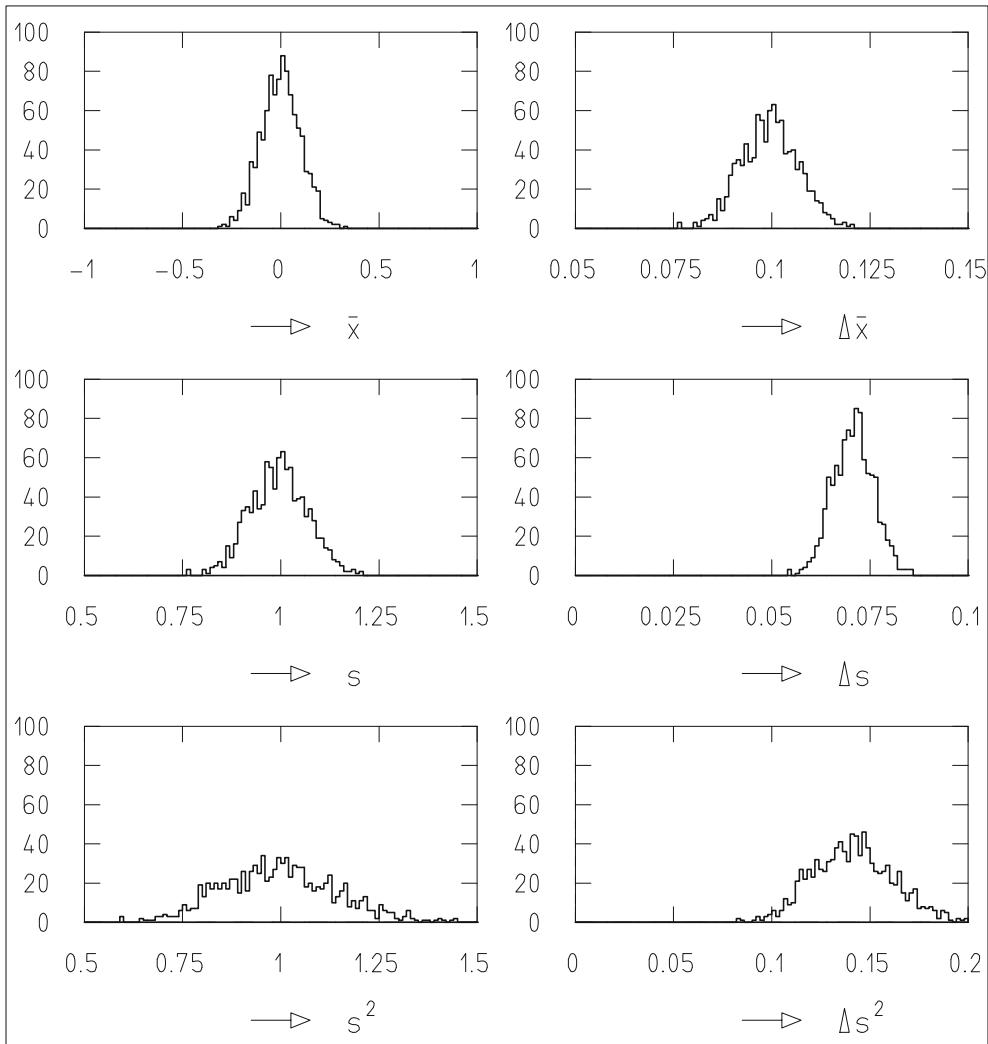
$$\sigma = \frac{f}{\sqrt{-8 \ln \frac{1}{2}}} \approx 0.4247 f \quad (6.3.1)$$

between the standard deviation and FWHM. This expression can be used to estimate the standard deviation of a sample when  $f$  is obtained from a histogram. ■

We now use the Monte Carlo method (Chap. 4) together with histograms in order to illustrate the concepts of mean, standard deviation, and variance of a sample, and their errors, as introduced in Sect. 6.2.

**Example 6.4:** Investigation of characteristic quantities of samples from a Gaussian distribution with the Monte Carlo method

We generate successively 1000 samples of size  $N = 100$  from the standard normal distribution, e.g., compute the mean  $\bar{x}$ , variance  $s^2$ , and standard deviation  $s$  for each sample as well as the errors  $\Delta\bar{x}$ ,  $\Delta s^2$ , and  $\Delta s$ , with the methods of the classSample. We then produce for each of the six quantities a histogram (Fig. 6.3), containing 1000 entries. Since each of the quantities is defined as the sum of many random quantities, we expect in all cases that the histograms should resemble Gaussian distributions. From the histogram for  $\bar{x}$ , we obtain a full width at half maximum of about 0.25, and hence a standard deviation of approximately 0.1. Indeed the histogram for  $\Delta\bar{x}$  shows an approximately Gaussian distribution with mean value  $\Delta\bar{x} = 0.1$ . (From the width of this histogram one could determine the error of the error  $\Delta\bar{x}$  of the mean  $\bar{x}$ !) From both histograms one obtains a very clear impression of the meaning of the error  $\Delta\bar{x}$  of the mean  $\bar{x}$  for a single sample, as computed according to (6.2.8). It gives (within its error) the standard deviation of the population, from which the sample mean  $\bar{x}$  comes. If many samples are successively taken (i.e., if the experiment is repeated many times) then the frequency distribution of the values  $\bar{x}$  follows a Gaussian distribution about the population mean with standard deviation  $\Delta\bar{x}$ . The corresponding considerations also hold for the quantities  $s^2$ ,  $s$ , and their errors  $\Delta s$  and  $\Delta s^2$ . ■



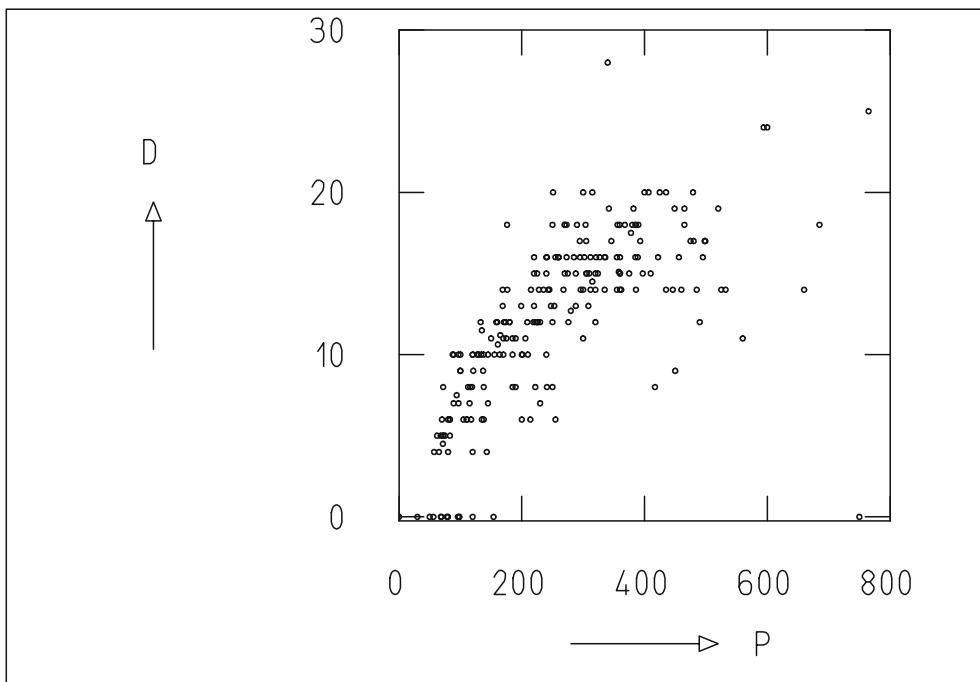
**Fig. 6.3:** Histograms of the quantities  $\bar{x}$ ,  $\Delta\bar{x}$ ,  $s$ ,  $\Delta s$ ,  $s^2$ , and  $\Delta s^2$  from 1000 samples of size 100 from the standard normal distribution.

If the elements of the sample depend on two random variables  $x$  and  $y$ , then one can construct a scatter plot, where each element is represented as a point in a Cartesian coordinate system for the variables  $x$  and  $y$ . Such a *two-dimensional scatter plot* provides useful qualitative information about the relationship between the two variables.

The class `GraphicsWith2DScatterDiagram` generates such a diagram by a single call. (A plot in the format A5 landscape is generated, into which the scatter diagram, itself in square format, is fitted. If another plot format or edge ratio of the diagram is desired, the class has to be adapted accordingly.)

**Example 6.5:** Two-dimensional scatter plot: Dividend versus price for industrial stocks

Table 6.2 contains a list of the first 10 of 226 data sets, in which the dividend in 1967 (first column) and share price on December 31, 1967 (second column) are given for a number of industrial stocks. The third column shows the company name for all German corporations worth more than 10 million marks. The scatter plot of the number pairs (share price, dividend) is shown in Fig. 6.4.



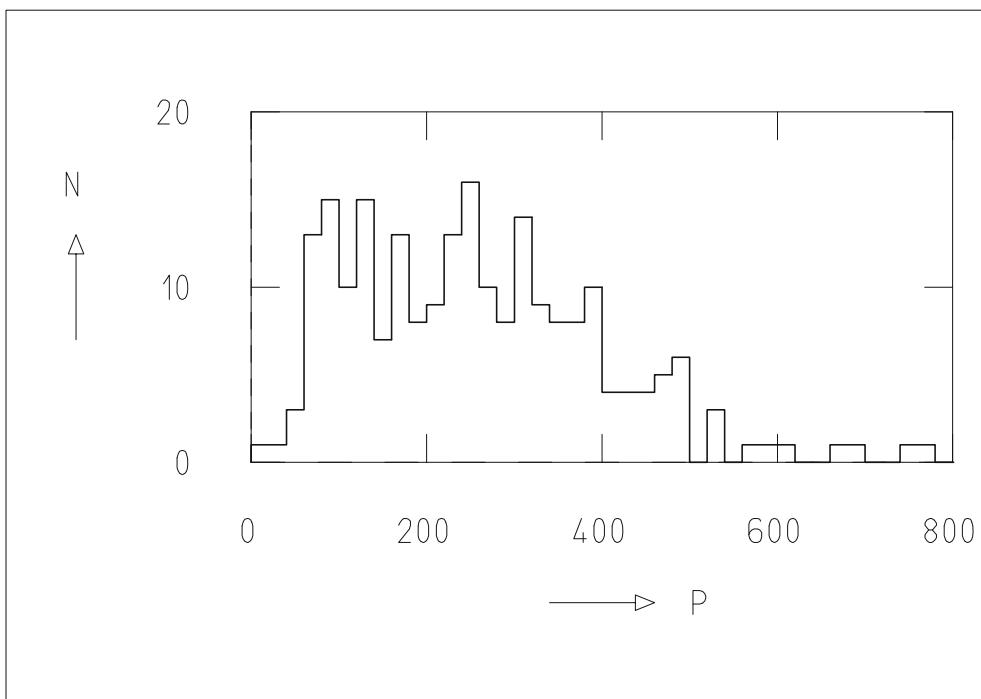
**Fig. 6.4:** Scatter plot of dividend  $D$  versus share price  $P$  for industrial stocks.

As expected we see a strong correlation between dividend and share price. One can see, however, that the dividend does not grow linearly with the price. It appears that factors other than immediate profit determine the price of a stock.

Also shown are histograms for the share price (Fig. 6.5) and dividend (Fig. 6.6) which can be obtained as projections of the scatter plot onto the abscissa and ordinate. One clearly observes a non-statistical behavior for the dividends. It is given as a percent of the nominal value and is therefore almost always integer. One sees that even numbers are considerably more frequent than odd numbers. ■

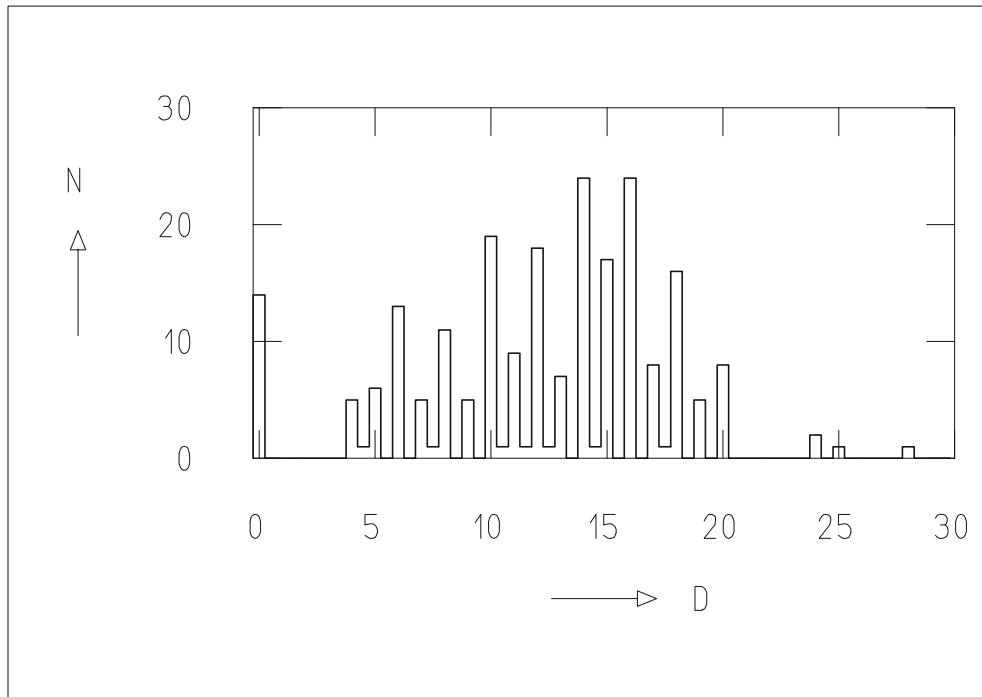
**Table 6.2:** Dividend, share price of a stock, and company name.

- |     |      |                                   |
|-----|------|-----------------------------------|
| 12. | 133. | ACKERMANN-GOEGGINGEN              |
| 08. | 417. | ADLERWERKE KLEYER                 |
| 17. | 346. | AGROB AG FUER GROB U. FEINKERAMIK |
| 25. | 765. | AG.F.ENERGIEWIRTSCHAFT            |
| 16. | 355. | AG F. LICHT- U. KRAFTVERS.,MCHN.  |
| 20. | 315. | AG.F. IND.U.VERKEHRSW.            |
| 08. | 138. | AG. WESER                         |
| 16. | 295. | AEG ALLG.ELEKTR.-GES.             |
| 20. | 479. | ANDREAE-NORIS ZAHN                |
| 10. | 201. | ANKERWERKE                        |

**Fig. 6.5:** Histogram of the price of industrial stocks.

## 6.4 Samples from Partitioned Populations

It is often advantageous to divide a population  $G$  (e.g., all of the students in Europe) into various *subpopulations*  $G_1, G_2, \dots, G_t$  (students at university 1, 2, ...,  $t$ ). Suppose a quantity of interest  $x$  follows in the various subpopulations the probability densities  $f_1(x), f_2(x), \dots, f_t(x)$ . The distribution function corresponding to  $f_i(x)$  is then



**Fig. 6.6:** Histogram of the dividend of industrial stocks.

$$F_i(x) = \int_{-\infty}^x f_i(x) dx = P(\mathbf{x} < x | \mathbf{x} \in G_i) . \quad (6.4.1)$$

This is equal to the conditional probability for  $\mathbf{x} < x$  given that  $\mathbf{x}$  is contained in the subpopulation  $G_i$ . The rule of total probability (2.3.4) provides the relationship between the various  $F_i(x)$  and the distribution function  $F(x)$  for  $G$ ,

$$F(x) = P(\mathbf{x} < x | \mathbf{x} \in G) = \sum_{i=1}^t P(\mathbf{x} < x | \mathbf{x} \in G_i) P(\mathbf{x} \in G_i) ,$$

i.e.,

$$F(x) = \sum_{i=1}^t P(\mathbf{x} \in G_i) F_i(x) . \quad (6.4.2)$$

Correspondingly one has for the probability density

$$f(x) = \sum_{i=1}^t P(\mathbf{x} \in G_i) f_i(x) . \quad (6.4.3)$$

If we now abbreviate  $P(\mathbf{x} \in G_i)$  by  $p_i$ , then one has

$$\begin{aligned}\hat{x} &= E(\mathbf{x}) = \int_{-\infty}^{\infty} xf(x) dx = \sum_{i=1}^t p_i \int_{-\infty}^{\infty} xf_i(x) dx \quad , \\ \hat{x} &= \sum_{i=1}^t p_i \hat{x}_i \quad .\end{aligned}\tag{6.4.4}$$

The population mean is thus the mean of the individual means of the subpopulations, each weighted by probability of its subpopulation. For the population variance one obtains

$$\begin{aligned}\sigma^2(\mathbf{x}) &= \int_{-\infty}^{\infty} (x - \hat{x})^2 f(x) dx \\ &= \int_{-\infty}^{\infty} (x - \hat{x})^2 \sum_{i=1}^t p_i f_i(x) dx \\ &= \sum_{i=1}^t p_i \int_{-\infty}^{\infty} \{(x - \hat{x}_i) + (\hat{x}_i - \hat{x})\}^2 f_i(x) dx \quad .\end{aligned}$$

All cross terms vanish since the  $\mathbf{x}_i$  are independent, leading to

$$\sigma^2(\mathbf{x}) = \sum_{i=1}^t p_i \left\{ \int_{-\infty}^{\infty} (x - \hat{x}_i)^2 f_i(x) dx + (\hat{x}_i - \hat{x})^2 \int_{-\infty}^{\infty} f_i(x) dx \right\}$$

or

$$\sigma^2(\mathbf{x}) = \sum_{i=1}^t p_i \{\sigma_i^2 + (\hat{x}_i - \hat{x})^2\} \quad .\tag{6.4.5}$$

One thus obtains the weighted mean of a sum of two terms. The first gives the dispersion of a subpopulation, the second gives the quadratic deviation of the mean of this subpopulation from the mean of the whole population.

Having discussed separating a population into parts, we now take from each subpopulation  $G_i$  a sample of size  $n_i$  (with  $\sum_{i=1}^t n_i = n$ ) and examine the arithmetic mean of the total partitioned sample

$$\bar{\mathbf{x}}_p = \frac{1}{n} \sum_{i=1}^t \sum_{j=1}^{n_i} \mathbf{x}_{ij} = \frac{1}{n} \sum_{i=1}^t n_i \bar{\mathbf{x}}_i \tag{6.4.6}$$

with the expectation value and variance

$$E(\bar{x}_p) = \frac{1}{n} \sum_{i=1}^n n_i \hat{x}_i , \quad (6.4.7)$$

$$\begin{aligned} \sigma^2(\bar{x}_p) &= E\{(\bar{x}_p - E(\bar{x}_p))^2\} \\ &= E\left\{\left(\sum_{i=1}^t \frac{n_i}{n} (\bar{x}_i - \hat{x}_i)\right)^2\right\} \\ &= \frac{1}{n^2} \sum_{i=1}^t n_i^2 E\{(\bar{x}_i - \hat{x}_i)^2\} , \\ \sigma^2(\bar{x}_p) &= \frac{1}{n^2} \sum_{i=1}^t n_i^2 \sigma^2(\bar{x}_i) . \end{aligned} \quad (6.4.8)$$

Using (6.2.3) this is finally

$$\sigma^2(\bar{x}_p) = \frac{1}{n} \sum_{i=1}^t \frac{n_i}{n} \sigma_i^2 . \quad (6.4.9)$$

One would obtain the same result by application of the law of error propagation (3.8.7) to Eq. (6.4.6).

It is clear that the arithmetic mean  $\bar{x}_p$  cannot in general be an estimator for the sample mean  $\hat{x}$ , since it depends on the arbitrary choice of the size  $n_i$  of the samples from the subpopulations. A comparison of (6.4.7) with (6.4.4) shows that this is only true for the special case  $p_i = n_i/n$ .

The population mean  $\hat{x}$  can be estimated in the following way. One first determines the means  $\bar{x}_i$  for the subpopulations, and constructs then the expression

$$\tilde{x} = \sum_{i=1}^t p_i \bar{x}_i , \quad (6.4.10)$$

in analogy to Eq. (6.4.4). By error propagation one obtains for the variance of  $\tilde{x}$

$$\sigma^2(\tilde{x}) = \sum_{i=1}^t p_i^2 \sigma^2(\bar{x}_i) = \sum_{i=1}^t \frac{p_i^2}{n_i} \sigma_i^2 . \quad (6.4.11)$$

**Example 6.6:** Optimal choice of the sample size for subpopulations

In order to minimize the variance  $\sigma^2(\tilde{x})$ , we cannot simply differentiate the relation (6.4.11) with respect to all  $n_i$ , since the  $n_i$  must satisfy a constraint, namely

$$\sum_{i=1}^t n_i - n = 0 \quad . \quad (6.4.12)$$

We must therefore use the method of *Lagrange multipliers*, by multiplying Eq. (6.4.12) with a factor  $\mu$ , adding this to Eq. (6.4.11), and finally setting the partial derivatives of the  $n_i$  with respect to  $\mu$  equal to zero:

$$L = \sigma^2(\tilde{x}) + \mu(\sum n_i - n) = \sum(p_i^2/n_i)\sigma_i^2 + \mu(\sum n_i - n) \quad , \\ \frac{\partial L}{\partial n_i} = -\frac{p_i^2\sigma_i^2}{n_i^2} + \mu = 0 \quad , \quad (6.4.13)$$

$$\frac{\partial L}{\partial \mu} = \sum n_i - n = 0 \quad . \quad (6.4.14)$$

From (6.4.13) we obtain

$$n_i = p_i\sigma_i/\sqrt{\mu} \quad .$$

Together with (6.4.14) this gives

$$1/\sqrt{\mu} = n / \sum p_i\sigma_i$$

and therefore

$$n_i = np_i\sigma_i / \sum p_i\sigma_i \quad . \quad (6.4.15)$$

The result (6.4.15) states that the sizes  $n_i$  of the samples from the subpopulations  $i$  should be chosen in such a way that they are proportional to the probability  $p_i$  of subpopulation  $i$ , weighted with the corresponding standard deviation.

As an example assume that a scientific publishing company wants to estimate the total amount spent for scientific books by two subpopulations: (1) students and (2) scientific libraries. Further, we will assume that there are 1000 libraries and  $10^6$  students in the population and that the standard deviation of the money spent by students is \$ 100, and for libraries (which are of greatly differing sizes) \$  $3 \cdot 10^5$ . We then have

$$p_1 \approx 1 \quad , \quad p_2 \approx 10^{-3} \quad , \quad \sigma_1 = 100 \quad , \quad \sigma_2 = 3 \times 10^5$$

and from (6.4.15)

$$n_1 = \text{const} \cdot 100 \quad , \quad n_2 = \text{const} \cdot 300 \quad , \quad n_2 = 3n_1 \quad .$$

Note that the result does not depend on the means of the partial populations. The quantities  $p_i$ ,  $x_i$ , and  $\sigma_i$  are in general unknown. They must first be estimated from preliminary samples. ■

The discussion of subpopulations will be taken up again in Chap. 11.

## 6.5 Samples Without Replacement from Finite Discrete Populations. Mean Square Deviation. Degrees of Freedom

We first encountered the concept a sample in connection with the hypergeometric distribution (Sect. 5.3). There we determined that the independence of the individual sample elements was lost by the process of taking elements without replacing them from a finite (and hence discrete) population. We are therefore no longer dealing with genuine random sampling, even if no particular choice among the remaining elements is made.

To discuss this further let us introduce the following notation. Suppose the population consists of  $N$  elements  $y_1, y_2, \dots, y_N$ . From it we take a sample of size  $n$  with the elements  $x_1, x_2, \dots, x_n$ . (In the hypergeometric distribution, the  $y_j$  and hence the  $x_i$  could only take on the values 0 and 1.)

Since it is equally probable for each of the remaining elements  $y_j$  to be chosen, we obtain for the expectation value

$$E(y) = \hat{y} = \bar{y} = \frac{1}{N} \sum_{j=1}^N y_j . \quad (6.5.1)$$

Although  $\hat{y}$  is not a random variable, this expression is the arithmetic mean of a finite number of elements of the population. A definition of the population variance encounters the difficulties discussed at the end of Sect. 6.2. We define it in analogy to (6.2.6) as

$$\begin{aligned} \sigma^2(y) &= \frac{1}{N-1} \sum_{j=1}^N (y_j - \bar{y})^2 \\ &= \frac{1}{N-1} \left\{ \sum_{j=1}^N y_j^2 - \frac{1}{N} \left( \sum_{j=1}^N y_j \right)^2 \right\} . \end{aligned} \quad (6.5.2)$$

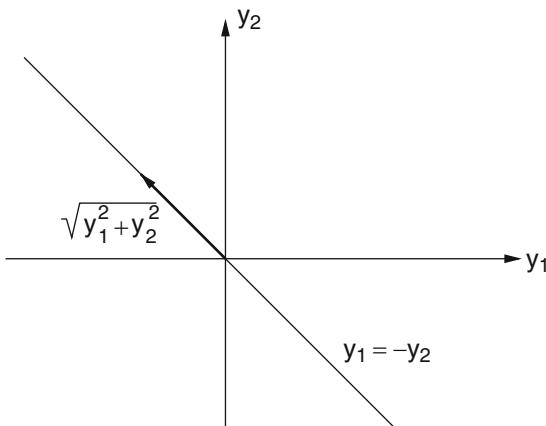
Let us now consider the *sum of squares*,

$$\sum_{j=1}^N (y_j - \bar{y})^2 . \quad (6.5.3)$$

Since we have not constrained the population in any way, the  $y_j$  can take on all possible values. Therefore the first element in the sum in (6.5.2) can also take on any of the possible values. The same holds for the 2nd, 3rd,  $\dots$ ,  $(N - 1)$ th terms summed. The  $N$ th term in the sum is then, however, fixed, since

$$\sum_{j=1}^N (y_j - \bar{y}) = 0 \quad . \quad (6.5.4)$$

We say that the *number of degrees of freedom* of the sum of squares (6.5.3) is  $N - 1$ . One can illustrate this connection geometrically. We consider the case  $\bar{y} = 0$  and construct an  $N$ -dimensional vector space with the  $y_j$ . The quadratic sum (6.5.3) is then the square of the absolute value of the position vector in this space. Because of the equation of constraint (6.5.4) the tip of the position vector can only move in a space of dimension  $(N - 1)$ . In mechanics the dimension of such a constrained space is called the number of degrees of freedom. This is sketched in Fig. 6.7 for the case  $N = 2$ . Here the position vector is constrained to lie on the line  $y_2 = -y_1$ .



**Fig. 6.7:** A sample of size two gives a sum of squares with one degree of freedom.

A sum of squares divided by the number of degrees of freedom, i.e., an expression of the form (6.5.2) is called a *mean square* or to be more complete, since we are dealing with the differences of the individual values from the expectation or mean value, *mean square deviation*. The square root of this expression, which is then a measure of the dispersion, is called the root mean square (RMS) deviation.

We now return to the sample  $x_1, x_2, \dots, x_n$ . For simplicity of notation we will introduce the *Kronecker symbol*, which describes the selection procedure for the sample. It is defined as

$$\delta_i^j = \begin{cases} 1, & \text{if } x_i \text{ is the element } y_j, \\ 0 & \text{otherwise.} \end{cases} \quad (6.5.5)$$

In particular one has

$$\mathbf{x}_i = \sum_{j=1}^N \delta_i^j y_j . \quad (6.5.6)$$

Since the selection of any of the  $y_j$  as the  $i$ th element is equally probable, one has

$$P(\delta_i^j = 1) = 1/N . \quad (6.5.7)$$

Since  $\delta_i^j$  describes a random procedure, it is clearly a random variable itself. Its expectation value is found from Eq. (3.3.2) (where  $n = 2$ ,  $x_1 = 0$ ,  $x_2 = 1$ ) to be

$$E(\delta_i^j) = P(\delta_i^j = 1) = 1/N . \quad (6.5.8)$$

If now one element  $\mathbf{x}_i$  of the sample is determined, one then has only  $(N - 1)$  selection possibilities out of the population for a further element, e.g.,  $\mathbf{x}_k$ . That is,

$$P(\delta_i^j \delta_k^\ell = 1) = \frac{1}{N} \frac{1}{N-1} = E(\delta_i^j \delta_k^\ell) . \quad (6.5.9)$$

Since the sample is taken without replacement, one has  $j \neq \ell$ , i.e.,

$$\delta_i^j \delta_k^\ell = 0 . \quad (6.5.10)$$

Similarly one has

$$\delta_i^j \delta_i^\ell = 0 , \quad (6.5.11)$$

since two different elements of the population cannot simultaneously occur as the  $i$ th element of the sample.

We consider now the expectation value of  $\mathbf{x}_1$ ,

$$E(\mathbf{x}_1) = E \left\{ \sum_{j=1}^N \delta_1^j y_j \right\} = \sum_{j=1}^N y_j E(\delta_1^j) = \frac{1}{N} \sum_{j=1}^N y_j = \bar{y} . \quad (6.5.12)$$

Since  $\mathbf{x}_1$  is in not in any way special, the expectation values of all elements of the sample, and thus also of their arithmetic mean, have the same value

$$E(\bar{\mathbf{x}}) = \frac{1}{n} \sum_{i=1}^n E(\mathbf{x}_i) = \bar{y} . \quad (6.5.13)$$

The arithmetic mean of the sample is thus an unbiased estimator for the population mean.

Next we consider the *sample variance*

$$\mathbf{s}_x^2 = \frac{1}{n-1} \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})^2 . \quad (6.5.14)$$

By means of a somewhat longer calculation it can be shown that the expectation value is

$$E(\mathbf{s}_x^2) = \sigma^2(y) . \quad (6.5.15)$$

The sample variance is thus an unbiased estimator for the population variance.

The *variance of the mean* is also of interest:

$$\sigma^2(\bar{x}) = E\{(\bar{x} - E(\bar{x}))^2\} .$$

$E(\bar{x}) = \bar{y}$  is, however, a fixed, *not* a random quantity, whereas  $\bar{x}$  depends on the individual sample, and is hence a random variable. One therefore has

$$\begin{aligned} \sigma^2(\bar{x}) &= E(\bar{x}^2) - \bar{y}^2 = \frac{1}{n} \left\{ \left(1 - \frac{n}{N}\right) \sigma^2(y) + n\bar{y}^2 \right\} - \bar{y}^2 , \\ \sigma^2(\bar{x}) &= \frac{\sigma^2(y)}{n} \left(1 - \frac{n}{N}\right) . \end{aligned} \quad (6.5.16)$$

Comparing with the case of an infinite continuous sample (6.2.3) one sees the additional factor  $(1 - n/N)$ . This corresponds to the fact that the variance of  $\bar{x}$  vanishes in the case  $n = N$ , where the “sample” contains the entire population and where one has exactly  $\bar{x} = \bar{y}$ .

## 6.6 Samples from Gaussian Distributions: $\chi^2$ -Distribution

We return now to continuously distributed populations and consider in particular a Gaussian distribution with mean  $a$  and variance  $\sigma^2$ . According to (5.7.7), the characteristic function of such a Gaussian distribution is

$$\varphi_x(t) = \exp(it a) \exp\left(-\frac{1}{2}\sigma^2 t^2\right) . \quad (6.6.1)$$

We now take a sample of size  $n$  from the population. The characteristic function of the sample mean was given in (6.2.2) in terms of the characteristic function of the population. From this we have

$$\varphi_{\bar{x}}(t) = \left\{ \exp\left(i\frac{t}{n}a - \frac{\sigma^2}{2}\left(\frac{t}{n}\right)^2\right) \right\}^n . \quad (6.6.2)$$

If we consider  $(\bar{x} - a) = (\bar{x} - \hat{x})$  in place of  $x$ , then one obtains

$$\varphi_{\bar{x}-a}(t) = \exp\left(-\frac{\sigma^2 t^2}{2n}\right) . \quad (6.6.3)$$

This is again the characteristic function of a normal distribution, but with a different variance,

$$\sigma^2(\bar{x}) = \sigma^2(x)/n . \quad (6.6.4)$$

For the simple case of a standard Gaussian distribution ( $a = 0$ ,  $\sigma^2 = 1$ ) we have

$$\varphi_{\bar{x}}(t) = \exp(-t^2/2n) . \quad (6.6.5)$$

We take a sample from this distribution

$$x_1, x_2, \dots, x_n ,$$

but we are interested in particular in the sum of the squares of the sample elements,

$$x^2 = x_1^2 + x_2^2 + \dots + x_n^2 . \quad (6.6.6)$$

We want to show that the quantity  $x^2$  follows the distribution function\*

$$F(\chi^2) = \frac{1}{\Gamma(\lambda)2^\lambda} \int_0^{\chi^2} u^{\lambda-1} e^{-\frac{1}{2}u} du , \quad (6.6.7)$$

where

$$\lambda = \frac{1}{2}n . \quad (6.6.8)$$

The quantity  $n$  is called the *number of degrees of freedom*.

We first introduce the abbreviation

$$\frac{1}{\Gamma(\lambda)2^\lambda} = k \quad (6.6.9)$$

and determine the probability density to be

$$f(\chi^2) = k(\chi^2)^{\lambda-1} e^{-\frac{1}{2}\chi^2} . \quad (6.6.10)$$

For two degrees of freedom, the probability density is clearly an exponential function. First we want to prove what was claimed by (6.6.7) for one degree of freedom ( $\lambda = 1/2$ ). Thus we ask for the probability that  $x^2 < \chi^2$ , or rather, that  $-\sqrt{\chi^2} < x < +\sqrt{\chi^2}$ . This is

$$\begin{aligned} F(\chi^2) &= P(x^2 < \chi^2) = P(-\sqrt{\chi^2} < x < +\sqrt{\chi^2}) \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\sqrt{\chi^2}}^{\sqrt{\chi^2}} e^{-\frac{1}{2}x^2} dx = \frac{2}{\sqrt{2\pi}} \int_0^{\sqrt{\chi^2}} e^{-\frac{1}{2}x^2} dx . \end{aligned}$$

---

\*The symbol  $\chi^2$  (chi squared) was introduced by K. Pearson. Although it is written as something squared, which reminds one of its origin as a sum of squares, it is treated as a usual random variable.

Setting  $x^2 = u$ ,  $du = 2x \, dx$ , we obtain directly

$$F(\chi^2) = \frac{1}{\sqrt{2\pi}} \int_0^{\chi^2} u^{-\frac{1}{2}} e^{-\frac{1}{2}u} du . \quad (6.6.11)$$

To prove the general case we first find the characteristic function of the  $\chi^2$ -distribution to be

$$\varphi_{\chi^2}(t) = \int_0^\infty k(\chi^2)^{\lambda-1} \exp(-\frac{1}{2}\chi^2 + it\chi^2) d\chi^2 \quad (6.6.12)$$

or with  $(1/2 - it)\chi^2 = v$ ,

$$\varphi_{\chi^2}(t) = 2^\lambda (1 - 2it)^{-\lambda} k \int_0^\infty v^{\lambda-1} e^{-v} dv .$$

The integral on the right side is, according to (D.1.1), equal to  $\Gamma(\lambda)$ . One therefore has

$$\varphi_{\chi^2}(t) = (1 - 2it)^{-\lambda} . \quad (6.6.13)$$

If we now consider the case of a second distribution with  $\lambda'$ , then one has

$$\varphi'_{\chi^2}(t) = (1 - 2it)^{-\lambda'} .$$

Since the characteristic function of a sum is equal to the product of the characteristic functions, one has the following important theorem:

The sum of two independent  $\chi^2$  variables with  
 $n_1, n_2$  degrees of freedom follows itself a  $\chi^2$ -distribution  
with  $n = n_1 + n_2$  degrees of freedom.

This theorem can now be used to easily generalize the claim (6.6.7), proven up to now only for  $n = 1$ . The proof follows from the fact that the individual terms of the sum of squares are independent and therefore (6.6.6) can be treated as the sum of  $n$  different  $\chi^2$  variables, each with one degree of freedom.

In order to obtain the expectation value and variance of the  $\chi^2$ -distribution, we use the characteristic function, whose derivatives (5.5.7) give the central moments. We obtain

$$\begin{aligned} E(x^2) &= -i\varphi'(0) = 2\lambda , \\ E(x^4) &= n \end{aligned} \quad (6.6.14)$$

and

$$\begin{aligned} E\{(x^2)^2\} &= -\varphi''(0) = 4\lambda^2 + 4\lambda , \\ \sigma^2(x^2) &= E\{(x^2)^2\} - \{E(x^2)\}^2 = 4\lambda , \\ \sigma^2(x^4) &= 2n . \end{aligned} \quad (6.6.15)$$

The expectation value of the  $\chi^2$ -distribution is thus equal to the number of degrees of freedom, and the variance is two times larger. Figure 6.8 shows the probability density of the  $\chi^2$  distribution for various values of  $n$ . One sees [as can be directly seen also from (6.6.10)] that for  $\chi^2 = 0$ , the function diverges for  $n = 1$ , is equal to 1/2 for  $n = 2$ , and vanishes for  $n \geq 3$ . A table of the  $\chi^2$ -distribution is provided in the appendix (Table I.6).

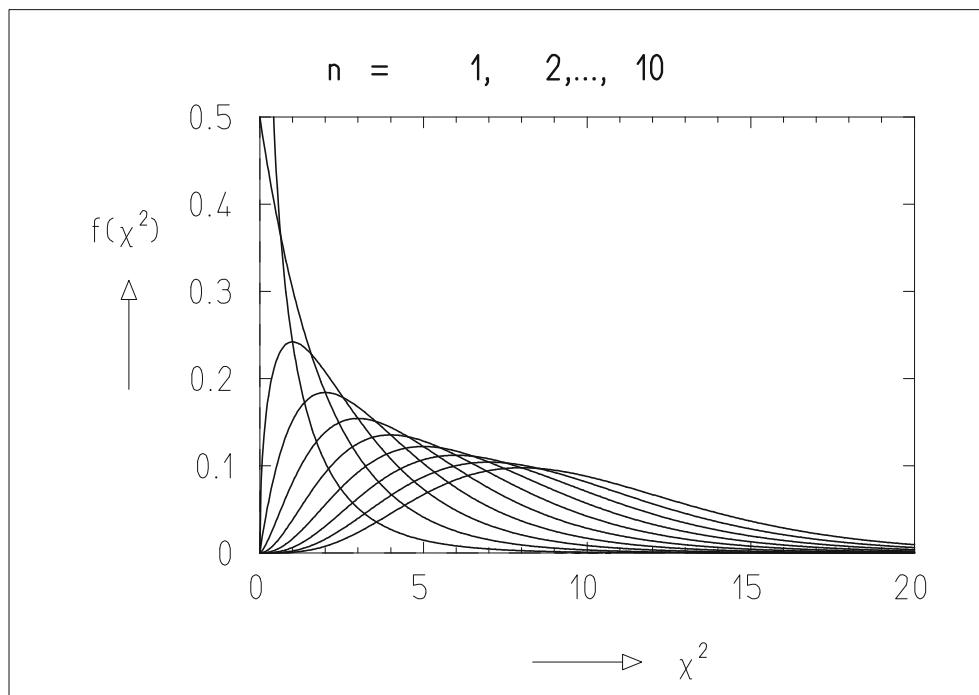
The  $\chi^2$ -distribution is of great significance in many applications, where the quantity  $\chi^2$  is used as a measure of confidence in a certain result. The smaller the value of  $\chi^2$ , the greater is the confidence in the result. (After all,  $\chi^2$  was defined as the sum of squares of deviations of elements of a sample from the population mean. See Sect. 8.7.) The distribution function

$$F(\chi^2) = P(x^2 < \chi^2) \quad (6.6.16)$$

gives the probability that the random variable  $x^2$  is not larger than  $\chi^2$ . In practice, one frequently uses the quantity

$$W(\chi^2) = 1 - F(\chi^2) \quad (6.6.17)$$

as a measure of confidence in a result.  $W(\chi^2)$  is often called the *confidence level*.  $W(\chi^2)$  is large for small values of  $\chi^2$  and falls with increasing  $\chi^2$ . The



**Fig. 6.8:** Probability density of  $\chi^2$  for the number of degrees of freedom  $n = 1, 2, \dots, 10$ . The expectation value  $E(\chi^2) = n$  moves to the right as  $n$  increases.

distribution function (6.6.16) is shown in Fig. 6.9 for various numbers of degrees of freedom  $n$ . The inverse function, which gives the quantiles of the  $\chi^2$ -distribution,

$$\chi_F^2 = \chi^2(F) = \chi^2 \quad (6.6.18)$$

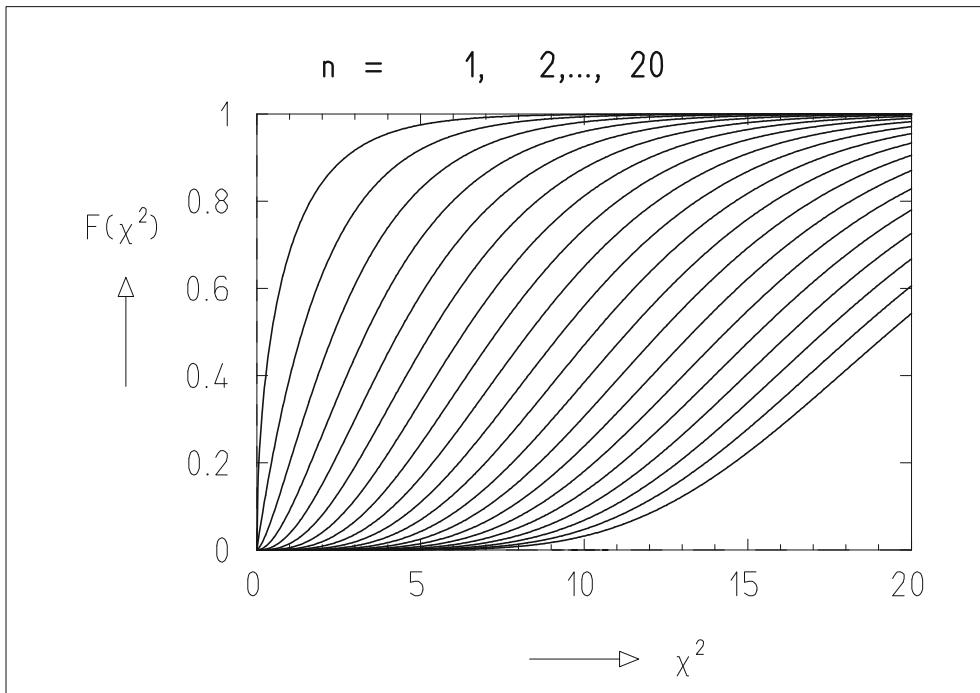
is used especially often in “hypothesis testing” (see Sect. 8.7). It is tabulated in the appendix (Table I.7).

Up to now we have restricted ourselves to the case where the population is described by the standard normal distribution. Usually, however, one has a normal distribution in general form with mean  $a$  and variance  $\sigma^2$ . Then the sum of squares (6.6.6) is clearly no longer distributed according to the  $\chi^2$ -distribution. One immediately obtains, however, a  $\chi^2$ -distribution by considering the quantity

$$\chi^2 = \frac{(x_1 - a)^2 + (x_2 - a)^2 + \cdots + (x_n - a)^2}{\sigma^2} . \quad (6.6.19)$$

This result follows directly from Eq. (5.8.4).

If the expectation values  $a_i$  and variances  $\sigma_i$  of the individual variables are different, then one has



**Fig. 6.9:** Distribution function for  $\chi^2$  for the number of degrees of freedom  $n = 1, 2, \dots, 20$ . The function for  $n = 1$  is the curve at the far left, and the function for  $n = 20$  is at the far right.

$$\chi^2 = \frac{(\mathbf{x}_1 - a_1)^2}{\sigma_1^2} + \frac{(\mathbf{x}_2 - a_2)^2}{\sigma_2^2} + \cdots + \frac{(\mathbf{x}_n - a_n)^2}{\sigma_n^2} . \quad (6.6.20)$$

Finally if the  $n$  variables are not independent, but are described by a joint normal distribution (5.10.1) with the expectation values given by the vector  $\mathbf{a}$  and the covariance matrix  $C = B^{-1}$ , then one has

$$\chi^2 = (\mathbf{x} - \mathbf{a})^T B (\mathbf{x} - \mathbf{a}) . \quad (6.6.21)$$

## 6.7 $\chi^2$ and Empirical Variance

In Eq. (6.2.6) we found that

$$\mathbf{s}^2 = \frac{1}{n-1} \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})^2 \quad (6.7.1)$$

is a consistent, unbiased estimator for the variance  $\sigma^2$  of a population. Let the  $\mathbf{x}_i$  be independent and normally distributed with standard deviation  $\sigma$ . We want to show that the quantity

$$\frac{n-1}{\sigma^2} \mathbf{s}^2 \quad (6.7.2)$$

follows the  $\chi^2$ -distribution with  $f = n - 1$  degrees of freedom. We first carry out an orthogonal transformation of the  $n$  variables  $\mathbf{x}_i$  (see Sect. 3.8):

$$\begin{aligned} \mathbf{y}_1 &= \frac{1}{\sqrt{1 \cdot 2}} (\mathbf{x}_1 - \mathbf{x}_2) , \\ \mathbf{y}_2 &= \frac{1}{\sqrt{2 \cdot 3}} (\mathbf{x}_1 + \mathbf{x}_2 - 2\mathbf{x}_3) , \\ \mathbf{y}_3 &= \frac{1}{\sqrt{3 \cdot 4}} (\mathbf{x}_1 + \mathbf{x}_2 + \mathbf{x}_3 - 3\mathbf{x}_4) , \\ &\vdots \\ \mathbf{y}_{n-1} &= \frac{1}{\sqrt{(n-1)n}} (\mathbf{x}_1 + \mathbf{x}_2 + \cdots + \mathbf{x}_{n-1} - (n-1)\mathbf{x}_n) , \\ \mathbf{y}_n &= \frac{1}{\sqrt{n}} (\mathbf{x}_1 + \mathbf{x}_2 + \cdots + \mathbf{x}_n) = \sqrt{n}\bar{\mathbf{x}} . \end{aligned} \quad (6.7.3)$$

One can verify that this transformation is in fact orthogonal, i.e., that

$$\sum_{i=1}^n \mathbf{x}_i^2 = \sum_{i=1}^n \mathbf{y}_i^2 . \quad (6.7.4)$$

Since a sum or difference of independent normally distributed quantities is again itself normally distributed, all of the  $y_i$  are normally distributed. The factors in (6.7.3) ensure that the  $y_i$  have mean values of zero and standard deviations  $\sigma$ .

From (6.7.1) and (6.7.2) one then has

$$\begin{aligned}(n-1)s^2 &= \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - 2\bar{x} \sum_{i=1}^n x_i + n\bar{x}^2 \\ &= \sum_{i=1}^n x_i^2 - n\bar{x}^2 = \sum_{i=1}^n y_i^2 - y_n^2 = \sum_{i=1}^{n-1} y_i^2 .\end{aligned}$$

This expression is a sum of only  $(n-1)$  independent squared terms. A comparison with (6.6.19) shows that the quantity (6.7.2) in fact follows a  $\chi^2$ -distribution with  $(n-1)$  degrees of freedom.

The squared terms  $(x_i - \bar{x})^2$  are not linearly independent. One has the following relation between them:

$$\sum_{i=1}^n (x_i - \bar{x}) = 0 .$$

One can show that every additional relation between the squared terms reduces the number of degrees of freedom by one. Later we will make frequent use of this result, which we only state here without proof.

## 6.8 Sampling by Counting: Small Samples

Samples are often obtained in the following way. One draws  $n$  elements from a population, checks if they possess a given characteristic and accepts only those  $k$  elements into the sample that have the characteristic. The remaining  $n-k$  elements are rejected, i.e., their properties are not recorded. This approach thus becomes the counting of  $k$  out of  $n$  elements drawn.

This approach corresponds exactly to selecting a sample according to a binomial distribution. The parameters  $p$  and  $q$  of this distribution correspond then to the occurrence or non-occurrence of the property in question. As will be shown in Example 7.5,

$$S(p) = \frac{k}{n} \tag{6.8.1}$$

is the maximum likelihood estimator of the parameter  $p$ . The variance of  $S$  is

$$\sigma^2(S(p)) = \frac{p(1-p)}{n} . \tag{6.8.2}$$

By using (6.8.1) it can be estimated from the sample by

$$\mathbf{s}^2(\mathbf{S}(p)) = \frac{1}{n} \frac{\mathbf{k}}{n} \left(1 - \frac{\mathbf{k}}{n}\right) . \quad (6.8.3)$$

We define the error  $\Delta k$  as

$$\Delta k = \sqrt{[\mathbf{s}^2(\mathbf{S}(np))]} . \quad (6.8.4)$$

By using (6.8.3) we obtain

$$\Delta k = \sqrt{\left[k\left(1 - \frac{k}{n}\right)\right]} . \quad (6.8.5)$$

The error  $\Delta k$  only depends on the number of elements counted and on the size of the sample. It is called the *statistical error*. A particularly important case is that of small  $k$ , or more precisely, the case  $k \ll n$ . In this limit we can define  $\lambda = np$  and following Sect. 5.4 consider the counted number  $k$  as a single element of a sample taken from a Poisson distributed population with parameter  $\lambda$ . From (6.8.1) and (6.8.5) we obtain

$$\mathbf{S}(\lambda) = \mathbf{S}(np) = k , \quad (6.8.6)$$

$$\Delta \lambda = \sqrt{k} . \quad (6.8.7)$$

(This can be derived by using the result of Example 7.4 with  $N = 1$ .) The result (6.8.7) is often written in an actually incorrect but easy to remember form,

$$\Delta k = \sqrt{k} ,$$

which is read: The *statistical error* of the *counted number*  $k$  is  $\sqrt{k}$ .

In order to interpret the statistical error  $\Delta \lambda = \sqrt{k}$  we must examine the Poisson distribution somewhat more closely. Let us begin with the case where  $k$  is not too small (say,  $k > 20$ ). For large values of  $\lambda$  the Poisson distribution becomes a Gaussian distribution with mean  $\lambda$  and variance  $\sigma^2 = \lambda$ . This can be seen qualitatively from Fig. 5.6. As long as  $k$  is not too small, i.e.,  $k \gg 1$ , we can then treat the Poisson distribution in  $k$  with parameter  $\lambda$  as a normal distribution in  $x$  with mean  $\lambda$  and variance  $\sigma^2 = \lambda$ . The discrete variable  $k$  is then replaced by the continuous variable  $x$ . The probability density of  $x$  is

$$f(x; \lambda) = \frac{1}{\sigma \sqrt{2\pi}} \exp \left\{ -\frac{(x - \lambda)^2}{2\sigma^2} \right\} = \frac{1}{\sqrt{2\pi\lambda}} \exp \left\{ -\frac{(x - \lambda)^2}{2\lambda} \right\} . \quad (6.8.8)$$

The observation of  $k$  events corresponds to observing once the value of the random variable  $x = k$ .

With the help of the probability density (6.8.8) we now want to determine the *confidence limits* at a given *confidence level*  $\beta = 1 - \alpha$  in such a way that

$$P(\lambda_- \leq \lambda \leq \lambda_+) = 1 - \alpha \quad . \quad (6.8.9)$$

That is, one requires that the probability that the true value of  $\lambda$  is contained within the confidence limits  $\lambda_-$  and  $\lambda_+$  be equal to the confidence level  $1 - \alpha$ . The limiting cases  $\lambda = \lambda_-$  and  $\lambda = \lambda_+$  are depicted in Fig. 6.10. They are determined such that

$$P(x > k | \lambda = \lambda_+) = 1 - \alpha/2 \quad , \quad P(x < k | \lambda = \lambda_-) = 1 - \alpha/2 \quad . \quad (6.8.10)$$

One clearly has

$$\begin{aligned} \alpha/2 &= \int_{x=-\infty}^{x=k} f(x; \lambda_+) dx = \frac{1}{\sigma \sqrt{2\pi}} \int_{-\infty}^{x=k} \exp\left\{-\frac{(x - \lambda_+)^2}{2\sigma^2}\right\} \\ &= \int_{u=-\infty}^{u=(k-\lambda_+)/\sigma} \phi_0(u) du = \psi_0\left(\frac{k - \lambda_+}{\sigma}\right) \quad , \end{aligned} \quad (6.8.11)$$

and correspondingly

$$1 - \alpha/2 = \int_{-\infty}^{x=k} f(x; \lambda_-) dx = \psi_0\left(\frac{k - \lambda_-}{\sigma}\right) \quad . \quad (6.8.12)$$

Here  $\phi_0$  and  $\psi_0$  are the probability density and distribution function of the standard normal distribution introduced in Sect. 5.8. By using the inverse function  $\Omega$  of the distribution function  $\psi_0$  [see Eq. (5.8.8)], one obtains

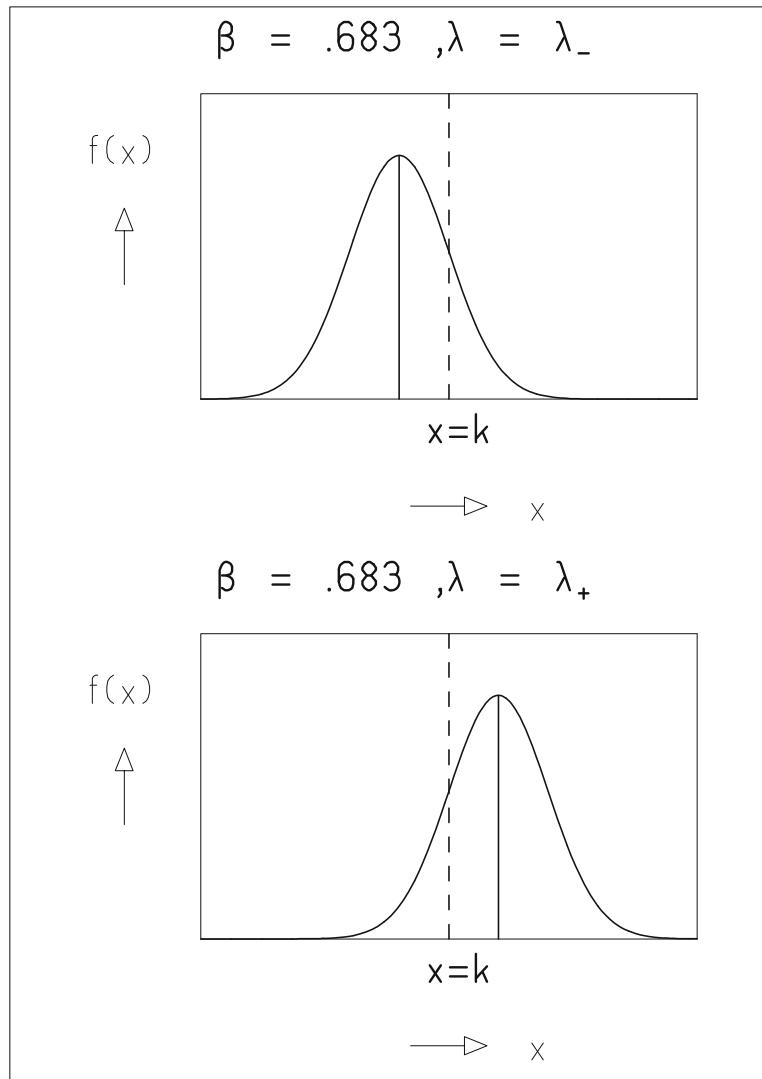
$$\frac{k - \lambda_-}{\sigma} = \Omega(1 - \alpha/2) \quad , \quad \frac{k - \lambda_+}{\sigma} = \Omega(\alpha/2) \quad . \quad (6.8.13)$$

Because of (5.8.10) one has  $\Omega(1 - \alpha/2) = \Omega'(1 - \alpha)$  and because of the symmetry of the function  $\Omega$ ,  $\Omega(1 - \alpha/2) = -\Omega(\alpha/2)$ . Further, since  $\alpha < 1$ , one has  $\Omega(1 - \alpha/2) > 0$ ,  $\Omega(\alpha/2) < 0$ . From this we finally obtain

$$\lambda_- = k - \sigma \Omega'(1 - \alpha) \quad , \quad \lambda_+ = k + \sigma \Omega'(-\alpha) \quad . \quad (6.8.14)$$

According to (6.8.6),  $k$  is the best estimator for  $\lambda$ . Since  $\sigma^2 = \lambda$ , the best estimator for  $\sigma$  is given by  $s = \sqrt{k}$ . Since we have assumed that  $k \gg 1$ , the uncertainty in  $s$  is significantly smaller than the uncertainty in  $k$ . We can therefore substitute  $x = k$  and  $s = \sqrt{k}$  in (6.8.9) and obtain for the confidence interval with confidence level  $1 - \alpha$

$$\lambda_- = k - \sqrt{k} \Omega'(1 - \alpha) \leq \lambda \leq k + \sqrt{k} \Omega'(-\alpha) = \lambda_+ \quad . \quad (6.8.15)$$



**Fig. 6.10:** Normal distribution with mean  $\lambda$  and standard deviation  $\sigma$  for  $\lambda = \lambda_-$  and  $\lambda = \lambda_+$ .

For  $1 - \alpha = 68.3\%$  we find from Sect. 5.8 or Table I.5 that  $\Omega'(\alpha) = 1$ . What is usually reported,

$$\lambda = k \pm \sqrt{k} ,$$

which was already the result of (6.8.6) and (6.8.7), thus gives the confidence limits at the confidence level of 68.3 %, but only for the case  $k \gg 1$ . For the confidence level of 90 %, i.e., for  $\alpha = 0.1$ , we find  $\Omega'(0.1) = 1.65$  and for the confidence level 99 % one has  $\Omega'(0.01) = 2.57$ .

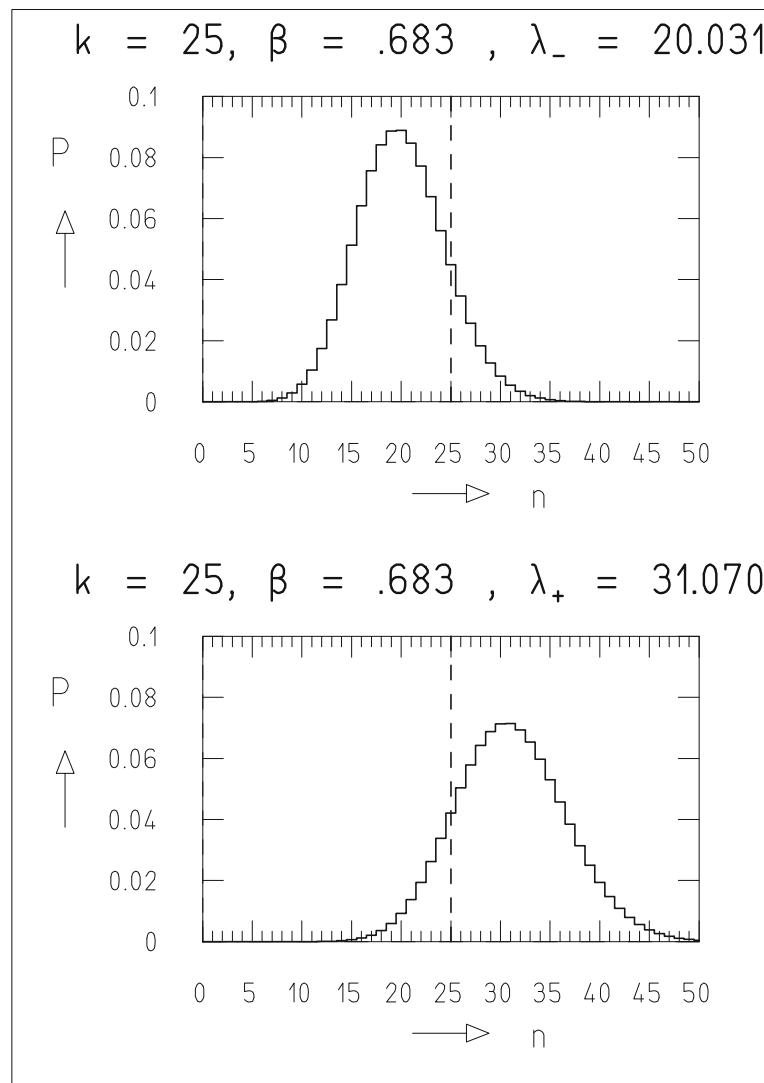
For very small values of  $k$ , one can no longer replace the Poisson distribution with the normal distribution. We follow therefore reference [25]. We start again from Eq. (6.8.10), but use, instead of the probability density (6.8.8) for

the continuous random variable  $x$  with fixed parameter  $\lambda$ , the Poisson probability for observing the discrete random variable  $n$  for a given  $\lambda$ ,

$$f(n; \lambda) = \frac{\lambda^n}{n!} e^{-\lambda} . \quad (6.8.16)$$

For the observation  $k$  we now determine the confidence limits  $\lambda_-$  and  $\lambda_+$ , which fulfill (6.8.10) with  $x = n$  (Fig. 6.11) and obtain in analogy to (6.8.11) and (6.8.12)

$$1 - \alpha/2 = \sum_{n=k+1}^{\infty} f(n; \lambda_+) = 1 - \sum_{n=0}^k f(n; \lambda_+) = 1 - F(k+1; \lambda_+) ,$$



**Fig. 6.11:** Poisson distribution with parameter  $\lambda$  for  $\lambda = \lambda_-$  and  $\lambda = \lambda_+$ .

$$1 - \alpha/2 = \sum_{n=0}^{k-1} f(n; \lambda_-) = F(k; \lambda_-)$$

or

$$\alpha/2 = F(k+1; \lambda_+) , \quad (6.8.17)$$

$$1 - \alpha/2 = F(k; \lambda_-) . \quad (6.8.18)$$

Here,

$$F(k; \lambda) = \sum_{n=0}^{k-1} f(n; \lambda) = P(k < k)$$

is the distribution function of the Poisson distribution.

In order to obtain numerical values for the confidence limits  $\lambda_+$  and  $\lambda_-$ , we solve Eqs. (6.8.17) and (6.8.18). That is, we must construct the inverse function of the Poisson distribution for fixed  $k$  and given probability  $P$  (in our case  $\alpha/2$  and  $1 - \alpha/2$ ),

$$\lambda = \lambda_P(k) . \quad (6.8.19)$$

This is done numerically with the method `StatFunct.quantile` `Poisson`. The function (6.8.19) is given in Table I.1 for frequently occurring values of  $P$ .

For extremely small samples one is often only interested in an *upper confidence limit* at confidence level  $\beta = 1 - \alpha$ . This is obtained by requiring

$$P(n > k | \lambda = \lambda^{(\text{up})}) = \beta = 1 - \alpha \quad (6.8.20)$$

instead of (6.8.10). Thus one has

$$\alpha = \sum_{n=0}^k f(n; \lambda^{(\text{up})}) = F(k+1; \lambda^{(\text{up})}) . \quad (6.8.21)$$

For the extreme case  $k = 0$ , i.e., for a sample in which no event was observed, one obtains the upper limit  $\lambda^{(\text{up})}$  by inverting  $\alpha = F(1; \lambda^{(\text{up})})$ . The upper limit then has the following meaning. If the true value of the parameter were in fact  $\lambda = \lambda^{(\text{up})}$  and if one were to repeat the experiment many times, then the probability of observing at least one event is  $\beta$ . The observation  $k = 0$  is then expressed in the following way: One has  $\lambda < \lambda^{(\text{up})}$  with a confidence level of  $1 - \alpha$ . From Table I.1 one finds that  $k = 0$  corresponds to  $\lambda < 2.996 \approx 3$  at a confidence level of 95 %.

**Example 6.7:** Determination of a lower limit for the lifetime of the proton from the observation of no decays

As already mentioned, the probability for the decay of a radioactive nucleus with the time  $t$  is

$$P(t) = \frac{1}{\tau} \int_0^t e^{-x/\tau} dx .$$

Here  $\tau$  is the mean lifetime of the nucleus. For  $t \ll \tau$  the expression simplifies to

$$P(t) = t/\tau .$$

For a total of  $N$  nuclei one expects that

$$k = NP(t) = N \cdot t/\tau$$

nuclei will decay within the time  $t$ . The mean lifetime  $\tau$  is obtained by counting the number of such decays. If one observes  $k$  decays from a total of  $N$  nuclei in a time  $t$ , then one obtains as the measured value of  $\tau$

$$\tilde{\tau} = \frac{N}{k} t .$$

Of particular interest is the mean lifetime of the proton, one of the primary building blocks of matter. In experiments recently carried out with great effort, one observes large numbers of protons with detectors capable of detecting each individual decay. Up to now, not a single decay has been seen. According to Table I.9, the true expected number of decays  $\lambda$  does not exceed three (at a confidence level of 95 %). One has therefore

$$\tau > \frac{N}{3} t$$

at this confidence level. Typical experimental values are  $t = 0.3$  years,  $N = 10^{33}$ , i.e.,

$$\tau > 10^{32} \text{ years} .$$

The proton can therefore be considered as stable even over cosmological time scales, if one considers that the age of the universe is estimated to be only around  $10^{10}$  years. ■

## 6.9 Small Samples with Background

In many experiments one is faced with the following situation. For the detected events one cannot determine whether they belong to the type that is actually of interest (*signal events*) or to another type (*background events*).

For the expected number of events in the experiment one then has a Poisson distribution with the parameter  $\lambda = \lambda_S + \lambda_B$ . Here  $\lambda_S$  is the sought after parameter of the number of signal events, and  $\lambda_B$  is the parameter for the background events, which must of course be known if one wants to obtain information about  $\lambda_S$ . (In an experiment as in Example 6.7, one might have, for example, an admixture of radioactive nuclei whose decays cannot be distinguished from those of the proton. If the number of such nuclei and their lifetime is known, then  $\lambda_B$  can be computed.)

We are now tempted to simply take the results of the last section, to determine the confidence limits  $\lambda_{\pm}$  and the upper limit  $\lambda^{(\text{up})}$  and to set  $\lambda_{S\pm} = \lambda_{\pm} - \lambda_B$ ,  $\lambda_S^{(\text{up})} = \lambda^{(\text{up})} - \lambda_B$ . This procedure can, however, lead to nonsensical results. (As seen in Example 6.7, one has  $\lambda^{(\text{up})} = 3$  at a confidence level of 95 %, for  $k = 0$ . For  $\lambda_B = 4$ ,  $k = 0$  we would obtain  $\lambda_S^{(\text{up})} = -1$ , although a value  $\lambda_S < 0$  has no meaning.)

The considerations up to now are based on the following. The probability for observing  $n$  events,  $n = n_S + n_B$ , is

$$f(n; \lambda_S + \lambda_B) = \frac{1}{n!} e^{-(\lambda_S + \lambda_B)} (\lambda_S + \lambda_B)^n , \quad (6.9.1)$$

and the probabilities to observe  $n_S$  signal events, and  $n_B$  background events are

$$f(n_S; \lambda_S) = \frac{1}{n_S!} e^{-\lambda_S} \lambda_S^{n_S} , \quad (6.9.2)$$

$$f(n_B; \lambda_B) = \frac{1}{n_B!} e^{-\lambda_B} \lambda_B^{n_B} . \quad (6.9.3)$$

The validity of (6.9.1) was shown in Example 5.5 with the help of the characteristic function, starting from the independence of the two Poisson distributions (6.9.2) and (6.9.3). One can also obtain them directly by summation of all products of the probabilities (6.9.2) and (6.9.3) that lead to  $n = n_S + n_B$ , by application of (B.4) and (B.6),

$$\begin{aligned} & \sum_{n_S=0}^n f(n_S; \lambda_S) f(n - n_S; \lambda_B) \\ &= e^{-(\lambda_S + \lambda_B)} \sum_{n_S=0}^n \frac{1}{n_S!(n - n_S)!} \lambda_S^{n_S} \lambda_B^{n-n_S} \\ &= \frac{1}{n!} e^{-(\lambda_S + \lambda_B)} \sum_{n_S=0}^n \binom{n}{n_S} \lambda_S^{n_S} \lambda_B^{n-n_S} \\ &= \frac{1}{n!} e^{-(\lambda_S + \lambda_B)} (\lambda_S + \lambda_B)^n \\ &= f(n; \lambda_S + \lambda_B) . \end{aligned}$$

The difficulties explained above are overcome with a method developed by ZECH [26]. In an experiment in which  $k$  events are recorded, the number of background events cannot simply be given by (6.9.3), since from the result of the experiment it is known that  $n_B \leq k$ . One must therefore replace (6.9.3) by

$$f'(n_B; \lambda) = f(n_B; \lambda_B) \left/ \sum_{n_B=0}^k f(n_B; \lambda_B) \right. , \quad n_B \leq k . \quad (6.9.4)$$

This distribution is normalized to unity in the region  $0 \leq n_B \leq k$ . In a corresponding way the distribution

$$f'(n; \lambda_S + \lambda_B) = f(n; \lambda_S + \lambda_B) \left/ \sum_{n_B=0}^k f(n_B; \lambda_B) \right. \quad (6.9.5)$$

takes the place of (6.9.1).

In this way one obtains in analogy to (6.8.17) and (6.8.18) for the limits of the confidence interval  $\lambda_{S-} \leq \lambda_S \leq \lambda_{S+}$  at a confidence level of  $1 - \alpha$

$$\alpha/2 = F'(\mathbf{k} + 1, \lambda_{S+} + \lambda_B) , \quad (6.9.6)$$

$$1 - \alpha/2 = F'(\mathbf{k}, \lambda_{S-} + \lambda_B) . \quad (6.9.7)$$

Here,

$$F'(\mathbf{k}; \lambda_S + \lambda_B) = \sum_{n=0}^{k-1} f'(n; \lambda_S + \lambda_B) = P(\mathbf{k} < k) \quad (6.9.8)$$

is the distribution function of the renormalized distribution (6.9.4). If only an upper limit at confidence level  $1 - \alpha$  is desired, then one clearly has in analogy to (6.8.21)

$$\alpha = F'(\mathbf{k} + 1, \lambda_S^{(\text{up})} + \lambda_B) . \quad (6.9.9)$$

Table 6.3 gives some numerical values computed with the methods of the class `SmallSample`. Note that for  $\mathbf{k} = 0$ , Eq. (6.9.7) has no meaning, so that  $\lambda_{S-}$  cannot be defined. In this case Eq. (6.9.6) and also  $\lambda_{S+}$  are not meaningful. (In the table, however, the values for  $\lambda_{S-}$  and  $\lambda_{S+}$  are shown as computed by the program. This sets  $\lambda_{S-} = 0$  and computes  $\lambda_{S+}$  according to (6.9.6).)

## 6.10 Determining a Ratio of Small Numbers of Events

Often a number of *signal* events  $k$  is measured and compared to a number of *reference* events  $d$ . One is interested in the true value  $r$  of the ratio of the

**Table 6.3:** Limits  $\lambda_{S-}$  and  $\lambda_{S+}$  of the confidence interval and upper confidence limits  $\lambda_S^{(up)}$ , for various values of  $\lambda_B$  and various very small sample sizes  $k$  for a fixed confidence level of 90%.

| $\beta = 0.90$ |             |                |                |                    |
|----------------|-------------|----------------|----------------|--------------------|
| $k$            | $\lambda_B$ | $\lambda_{S-}$ | $\lambda_{S+}$ | $\lambda_S^{(up)}$ |
| 0              | 0.0         | 0.000          | 2.996          | 2.303              |
|                | 1.0         | 0.000          | 2.996          | 2.303              |
|                | 2.0         | 0.000          | 2.996          | 2.303              |
| 1              | 0.0         | 0.051          | 4.744          | 3.890              |
|                | 1.0         | 0.051          | 4.113          | 3.272              |
|                | 2.0         | 0.051          | 3.816          | 2.995              |
| 2              | 0.0         | 0.355          | 6.296          | 5.322              |
|                | 1.0         | 0.100          | 5.410          | 4.443              |
|                | 2.0         | 0.076          | 4.824          | 3.877              |
| 3              | 0.0         | 0.818          | 7.754          | 6.681              |
|                | 1.0         | 0.226          | 6.782          | 5.711              |
|                | 2.0         | 0.125          | 5.983          | 4.926              |
| 4              | 0.0         | 1.366          | 9.154          | 7.994              |
|                | 1.0         | 0.519          | 8.159          | 7.000              |
|                | 2.0         | 0.226          | 7.241          | 6.087              |
| 5              | 0.0         | 1.970          | 10.513         | 9.275              |
|                | 1.0         | 1.009          | 9.514          | 8.276              |
|                | 2.0         | 0.433          | 8.542          | 7.306              |

number of signal events to the number of reference events, or more precisely, the ratio of the probability to observe a signal event to that of a reference event. As an estimator for this ratio one clearly uses

$$\tilde{r} = k/d \quad .$$

We now ask for the confidence limits of  $r$ . If  $k$  and  $d$  are sufficiently large, then they may be approximated as Gaussian variables with standard deviations  $\sigma_k = \sqrt{k}$ ,  $\sigma_d = \sqrt{d}$ . Then according to the law of error propagation one has

$$\Delta r = \sqrt{\left(\frac{\partial r}{\partial k}\right)^2 k + \left(\frac{\partial r}{\partial d}\right)^2 d} = r \sqrt{\frac{1}{k} + \frac{1}{d}} \quad . \quad (6.10.1)$$

If in addition one has  $d \gg k$ , then  $\Delta r$  is simply

$$\Delta r = \frac{r}{\sqrt{k}} \quad . \quad (6.10.2)$$

If the requirements for the validity of (6.10.1) or even of (6.10.2) are not fulfilled, i.e., if  $k$  and  $d$  are small numbers, then one must use considerations developed by JAMES and ROOS [23]. Clearly when one observes an individual event, the probability that it is a signal event is given by

$$p = \frac{r}{1+r} , \quad (6.10.3)$$

and the probability that it is a reference event is

$$q = 1 - p = \frac{1}{1+r} . \quad (6.10.4)$$

In an experiment in which a total of  $N = k + d$  are observed, the probability that exactly  $n$  signal events are present is given by a binomial distribution (5.1.3). This is

$$f(n; r) = \binom{N}{n} p^n q^{N-n} = \binom{N}{n} \left(\frac{r}{1+r}\right)^n \left(\frac{1}{1+r}\right)^{N-n} . \quad (6.10.5)$$

The probability to have  $n < k$  is then

$$P(n < k) = \sum_{n=0}^{k-1} f(n; r) = F(k; r) \quad (6.10.6)$$

with

$$F(k; r) = \sum_{n=0}^{k-1} \binom{N}{n} \left(\frac{r}{1+r}\right)^n \left(\frac{1}{1+r}\right)^{N-n} , \quad (6.10.7)$$

i.e., the distribution function of the binomial distribution. To determine the limits  $r_-$  and  $r_+$  of the confidence interval at the confidence level  $\beta = 1 - \alpha$ , we use in analogy to (6.8.17) and (6.8.18)

$$\alpha/2 = F(k+1; r_+) , \quad (6.10.8)$$

$$1 - \alpha/2 = F(k; r_-) . \quad (6.10.9)$$

If one only seeks an upper limit at the confidence level  $\beta = 1 - \alpha$ , it can be obtained from [see Eq. (6.8.21)]

$$\alpha = F(k+1; r^{(up)}) . \quad (6.10.10)$$

The quantities  $r_+$ ,  $r_-$ , and  $r^{(up)}$  can be computed for given values of  $k$ ,  $d$ , and  $\beta$  with the class SmallSample.

## 6.11 Ratio of Small Numbers of Events with Background

By combining as done by SWARTZ [24] the ideas of Sects. 6.9 and 6.10, one can deal with the following situation. In an experiment one has three types of events: *signal*, *background*, and *reference*. Signal and background events cannot be distinguished from each other. Suppose in the experiment one has detected a total of  $k$  signal and background events and  $d$  reference events. Let us label with  $r_S$  and  $r_B$  the true values (in the sense of the definition at the beginning of the previous section) of the ratios of the numbers of signal to reference and background to reference events. Then the probabilities that a randomly selected event is signal or background,  $p_S$  and  $p_B$ , are

$$p_S = \frac{r_S}{1+r_S+r_B} \quad , \quad p_B = \frac{r_B}{1+r_S+r_B} \quad . \quad (6.11.1)$$

The probability that it is a reference event is then

$$p_R = 1 - p_S - p_B = \frac{1}{1+r_S+r_B} \quad . \quad (6.11.2)$$

If one has a total of  $N = k + d$  events in the experiment, then the individual probabilities that one has exactly  $n_S$  signal events,  $n_B$  background events, and  $n_R = N - n_S - n_B$  reference events are

$$f_S(n_S; p_S) = \binom{N}{n_S} p_S^{n_S} (1-p_S)^{N-n_S} \quad , \quad (6.11.3)$$

$$f_B(n_B; p_B) = \binom{N}{n_B} p_B^{n_B} (1-p_B)^{N-n_B} \quad , \quad (6.11.4)$$

$$f_R(n_R; p_R) = \binom{N}{n_R} p_R^{n_R} (1-p_R)^{N-n_R} \quad . \quad (6.11.5)$$

Since there are now three mutually exclusive types of events, one has instead of a binomial distribution (6.10.5) a trinomial distribution, i.e., a multinomial distribution (5.1.10) with  $\ell = 3$ . The probability that in an experiment with a total of  $N$  events one has exactly  $n_S$  signal,  $n_B$  background, and  $N - n_S - n_B$  reference events is therefore

$$\begin{aligned} & f(n_S, n_B; r_S, r_B) \\ &= \frac{N!}{n_S! n_B! (N - n_S - n_B)!} p_S^{n_S} p_B^{n_B} (1 - p_S - p_B)^{N - n_S - n_B} \quad . \end{aligned} \quad (6.11.6)$$

Here, however, we have not taken into consideration that the number of background events cannot be greater than  $k$ . In a manner similar to (6.9.4) one uses

$$f'_B(n_B; p_B) = f_B(n_B; p_B) \left/ \sum_{n_B=0}^k f(n_B; p_B) \right. , \quad n_B \leq k , \quad (6.11.7)$$

in place of  $f_B$ . This replacement must also be made in (6.11.6), which gives

$$f'(n_S, n_B; r_S, r_B) = f(n_S, n_B; r_S, r_B) \left/ \sum_{n_B=0}^k f(n_B; p_B) \right. . \quad (6.11.8)$$

The probability to have  $n_S$  signal events regardless of the number of background events is

$$f'(n_S; r_S, r_B) = \sum_{n_B=0}^k f'(n_S, n_B; r_S, r_B)$$

and finally the probability to have  $n_S \leq k$  is

$$\begin{aligned} F'(k; r_S, r_B) &= \sum_{n_S=0}^{k-n_B-1} f'(n_S; r_S, r_B) \\ &= \frac{\sum_{n_S=0}^{k-n_B-1} \sum_{n_B=0}^{k-1} \frac{N!}{n_S! n_B! (N-n_S-n_B)!} p_S^{n_S} p_B^{n_B} (1-p_S-p_B)^{N-n_S-n_B}}{\sum_{n_B=0}^k \frac{N!}{n_B! (N-n_B)!} p_B^{n_B} (1-p_B)^{N-n_B}} . \end{aligned}$$

Since  $r_B$  was assumed to be known, the quantity  $F'$  for a given  $k$  depends only on  $r_S$ . Similar to (6.9.6) and (6.9.7) one can determine the limits  $r_{S+}$  and  $r_{S-}$  of the confidence region for  $r_S$  with confidence level  $\beta = 1 - \alpha$  from the following requirement:

$$\alpha/2 = F'(k+1; r_{S+}, r_B) , \quad (6.11.9)$$

$$1 - \alpha/2 = F'(k; r_{S-}, r_B) . \quad (6.11.10)$$

If one only wants an upper limit with confidence level  $\beta = 1 - \alpha$ , this can be found according to (6.9.9) from

$$\alpha = F'(k+1; r_S^{(\text{up})}, r_B) . \quad (6.11.11)$$

Table 6.4 contains some numerical values computed with methods of the class SmallSample. For  $k = 0$ , however, (6.11.10) and hence also  $r_{S-}$  have no meaning. Similarly for (6.11.9),  $r_{S+}$  for  $k = 0$  is not meaningful. (In the table, however, values for  $r_{S-}$  and  $r_{S+}$  are given for  $k = 0$  as computed by the program. For  $k = 0$  this sets  $r_{S-} = 0$  and determines  $r_{S+}$  according to (6.11.9).)

**Table 6.4:** Limits  $r_{S-}$  and  $r_{S+}$  of the confidence interval and upper confidence limit  $r_S^{(up)}$  for various values of  $r_B$  and various very small values of  $k$  for a fixed number of reference events  $d$  and fixed confidence level of 90 %.

| $\beta = 0.90, d = 10$ |       |          |          |              |
|------------------------|-------|----------|----------|--------------|
| $k$                    | $r_B$ | $r_{S-}$ | $r_{S+}$ | $r_S^{(up)}$ |
| 0                      | 0.0   | 0.000    | 0.349    | 0.259        |
|                        | 0.1   | 0.000    | 0.349    | 0.259        |
|                        | 0.2   | 0.000    | 0.349    | 0.259        |
| 1                      | 0.0   | 0.005    | 0.573    | 0.450        |
|                        | 0.1   | 0.005    | 0.502    | 0.382        |
|                        | 0.2   | 0.005    | 0.464    | 0.348        |
| 2                      | 0.0   | 0.034    | 0.780    | 0.627        |
|                        | 0.1   | 0.010    | 0.686    | 0.535        |
|                        | 0.2   | 0.007    | 0.613    | 0.467        |
| 3                      | 0.0   | 0.077    | 0.979    | 0.799        |
|                        | 0.1   | 0.020    | 0.880    | 0.701        |
|                        | 0.2   | 0.012    | 0.788    | 0.612        |
| 4                      | 0.0   | 0.127    | 1.174    | 0.968        |
|                        | 0.1   | 0.044    | 1.074    | 0.869        |
|                        | 0.2   | 0.019    | 0.976    | 0.771        |
| 5                      | 0.0   | 0.180    | 1.367    | 1.135        |
|                        | 0.1   | 0.085    | 1.267    | 1.035        |
|                        | 0.2   | 0.034    | 1.167    | 0.936        |

## 6.12 Java Classes and Example Programs

### Java Classes Referring to Samples

Sample contains methods computing characteristic parameters of a sample:  
mean, variance and standard deviation as well as the errors of these  
quantities.

SmallSamle contains methods computing the confidence limits for small  
samples.

Histograms allows the construction and administration of a histogram.

**Example Program 6.1:** The class E1Sample demonstrates the use of the class `Sample`

This short program generates a sample of size  $N$  taken from the standard normal distribution. It computes the six quantities: mean, error of the mean, variance, error of the variance, standard deviation, and error of the standard deviation, and outputs each quantity in a single line.

**Example Program 6.2:** The class E2Sample demonstrates the use of the classes `Histogram` and `GraphicsWithHistogram`

Initialization, filling and graphical representation of a histogram are demonstrated for a sample of  $N$  elements from the standardized normal distribution. Interactive input is provided for  $N$  as well as for the lower boundary  $x_0$ , bin width  $\Delta x$  and the number of bins  $n_x$  of the histogram. The histogram is initialized, the sample elements are generated and entered into the histogram. Finally the histogram graphics is produced.

**Example Program 6.3:** The class E3Sample demonstrates the use of class `GraphicsWith2DScatterDiagrams`

A scatter plot is created and later displayed graphically. The coordinates of the points making up the scatter plot are given as pairs of random numbers from a bivariate normal distribution (cf. Sect. 4.10). The program asks for the parameters of the normal distribution (means  $a_1, a_2$ , standard deviations  $\sigma_1, \sigma_2$ , correlation coefficient  $\rho$ ) and for the number of random number pairs to be generated. It generates the pairs and prepares the caption and the labeling of axes and scales and displays the plot (Fig. 6.12).

**Example Program 6.4:** The class E4Sample demonstrates using the methods of the class `SmallSample` to compute confidence limits

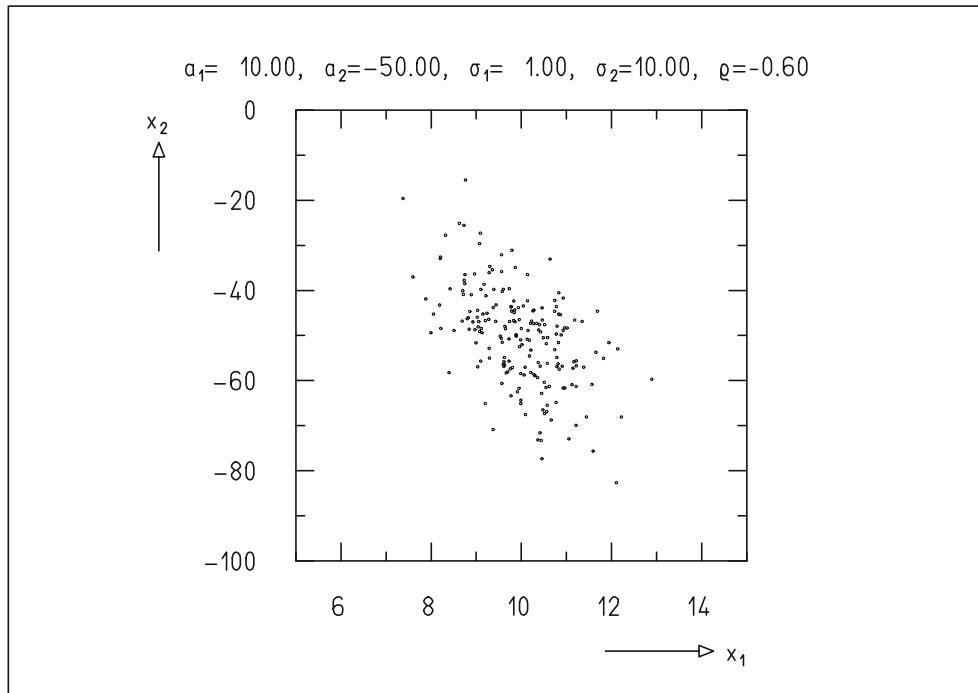
The program computes the limits  $\lambda_{S-}$ ,  $\lambda_{S+}$ , and  $\lambda_S^{(\text{up})}$  for the Poisson parameter of a signal. The user enters interactively the number of observed events  $k$ , the confidence level  $\beta = 1 - \alpha$ , and the Poisson parameter  $\lambda_B$  of the background. **Suggestions:** (a) Verify a few lines from Table 6.3.

(b) Choose  $\beta = 0.683$ ,  $\lambda_B = 0$  and compare for different values  $k$  the values  $\lambda_{S-}$  and  $\lambda_{S+}$  with the naive statement  $\lambda = k \pm \sqrt{k}$ .

**Example Program 6.5:** The class E5Sample demonstrates the use of methods of the class `SmallSample` to compute confidence limits of ratios

The program computes the limits  $r_{S-}$ ,  $r_{S+}$ , and  $r_S^{(\text{up})}$  for the ratio  $r$  of the number of signal to reference events in the limit of a large number of events. More precisely,  $r$  is the ratio of the Poisson parameters  $\lambda_S$  to  $\lambda_R$  of the signal to reference events. The program asks interactively for the number  $k$  of observed (signal plus reference) events, for the number  $d$  of reference events, for the confidence level  $\beta = 1 - \alpha$ , and for the expected ratio  $r_B = \lambda_B / \lambda_R$  for background events.

**Suggestion:** Verify a few lines from Table 6.4.



**Fig. 6.12:** Pairs of random numbers taken from a bivariate normal distribution.

**Example Program 6.6:** The class E6Sample simulates experiments with few events and background

A total of  $n_{\text{exp}}$  experiments are simulated. In each experiment  $N$  objects are analyzed. Each object yields with probability  $p_S = \lambda_S/N$  a signal event and with probability  $p_B = \lambda_B/N$  a background event. The numbers of events found in the simulated experiment are  $k_S$ ,  $k_B$ , and  $k = k_S + k_B$ . In the real experiment only  $k$  is known. The limits  $\lambda_{S-}$ ,  $\lambda_{S+}$ , and  $\lambda_S^{(\text{up})}$  for  $k$  are computed for a given confidence level  $\beta = 1 - \alpha$  and a given value of  $\lambda_B$  and are displayed for each experiment.

**Suggestion:** Choose, e.g.,  $n_{\text{exp}} = 20$ ,  $N = 1000$ ,  $\lambda_S = 5$ ,  $\lambda_B = 2$ ,  $\beta = 0.9$  and find out whether, as expected, this simulation yields for 10% of the experiments an interval  $(\lambda_{S-}, \lambda_{S+})$ , which does not contain the value  $\lambda_S$ . Keep in mind the meaning of the statistical error of your observation when using only 20 experiments.

**Example Program 6.7:** The class E7Sample simulates experiments with few signal events and with reference events

The program asks interactively for the quantities  $n_{\text{exp}}$ ,  $N$ ,  $\lambda_S$ ,  $\lambda_B$ ,  $\lambda_R$ , and the confidence level  $\beta = 1 - \alpha$ . It computes the probabilities  $p_S = \lambda_S/N$ ,  $p_B = \lambda_B/N$ ,  $p_R = \lambda_R/N$ , as well as the ratios  $r_S = p_S/p_R$  and  $r_B = p_B/p_R$  of a total of  $n_{\text{exp}}$  simulated experiments, in each of which  $N$  objects are analyzed. Each object is taken to be a signal event with probability  $p_S$ , a background event with probability  $p_B$ , and a reference event with probability  $p_R$ . (Here  $p_S + p_B + p_R \ll 1$  is assumed.) The simulation yields the numbers  $k_S$ ,  $k_B$ , and  $d$  for signal, background, and refer-

ence events, respectively. In a real experiment only the numbers  $k = k_S + k_B$  and  $d$  are known, since signal and background events cannot be distinguished. For the given values of  $\beta$  and  $r_B$  and for the quantities  $k$  and  $d$  which were found in the simulated experiments, the limits  $r_{S-}$ ,  $r_{S+}$ , and  $r_S^{(\text{up})}$  are computed and displayed.

**Suggestion:** Modify the suggestion accompanying Example Program 6.6 by choosing an additional input parameter  $\lambda_S = 20$ . Find out in how many cases the true value  $r_S$  used in the simulation is not contained in the interval  $(r_{S-}, r_{S+})$ .

# 7. The Method of Maximum Likelihood

## 7.1 Likelihood Ratio: Likelihood Function

In the last chapter we introduced the concept of parameter estimation. We have also described the desirable properties of estimators, though without specifying how such estimators can be constructed in a particular case. We have derived estimators only for the important quantities expectation value and variance. We now take on the general problem.

In order to specify explicitly the parameters

$$\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \dots, \lambda_p) ,$$

we now write the probability density of the random variables

$$\mathbf{x} = (x_1, x_2, \dots, x_n)$$

in the form

$$f = f(\mathbf{x}; \boldsymbol{\lambda}) . \quad (7.1.1)$$

If we now carry out a certain number of experiments, say  $N$ , or we draw a sample of size  $N$  from a population, then we can give a number to each experiment  $j$ :

$$dP^{(j)} = f(\mathbf{x}^{(j)}; \boldsymbol{\lambda}) d\mathbf{x} . \quad (7.1.2)$$

The number  $dP^{(j)}$  has the character of an *a posteriori probability*, i.e., given after the experiment, how probable it was to find the result  $\mathbf{x}^{(j)}$  (within a small interval). The total probability to find exactly all of the events

$$\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(j)}, \dots, \mathbf{x}^{(N)}$$

is then the product

$$dP = \prod_{j=1}^N f(\mathbf{x}^{(j)}; \boldsymbol{\lambda}) d\mathbf{x} . \quad (7.1.3)$$

This probability still clearly depends on  $\lambda$ . There are cases where the population is determined by only two possible sets of parameters,  $\lambda_1$  and  $\lambda_2$ . Such cases occur, for example, in nuclear physics, where the parity of a state is necessarily “even” or “odd”. One can construct the ratio

$$Q = \frac{\prod_{j=1}^N f(\mathbf{x}^{(j)}; \lambda_1)}{\prod_{j=1}^N f(\mathbf{x}^{(j)}; \lambda_2)} \quad (7.1.4)$$

and say that the values  $\lambda_1$  are “ $Q$  times more probable” than the values  $\lambda_2$ . This factor is called the *likelihood ratio*.\*

A product of the form

$$L = \prod_{j=1}^N f(\mathbf{x}^{(j)}; \lambda) \quad (7.1.5)$$

is called a *likelihood function*. One must clearly distinguish between a probability density and a likelihood function, which is a function of a sample and is hence a random variable. In particular, the a posteriori nature of the probability in (7.1.5) is of significance in many discussions.

### Example 7.1: Likelihood ratio

Suppose one wishes to decide whether a coin belongs to type  $A$  or  $B$  by means of a number of tosses. The coins in question are asymmetric in such a way that  $A$  shows heads with a probability of  $1/3$ , and  $B$  shows heads with a probability of  $2/3$ .

|       | $A$   | $B$   |
|-------|-------|-------|
| Heads | $1/3$ | $2/3$ |
| Tails | $2/3$ | $1/3$ |

If an experiment yields heads once and tails four times, then one has

$$L_A = \frac{1}{3} \cdot (\frac{2}{3})^4 \text{ and } L_B = \frac{2}{3} \cdot (\frac{1}{3})^4,$$

$$Q = \frac{L_A}{L_B} = 8.$$

One would therefore tend towards the position that the coin is of type  $A$ . ■

---

\* Although the likelihood ratio  $Q$  and the likelihood functions  $L$  and  $\ell$  introduced below are random variables, since they are functions of a sample, we do not write them here with a special character type.

## 7.2 The Method of Maximum Likelihood

The generalization of the likelihood ratio is now clear. One gives the greatest confidence to that choice of the parameters  $\lambda$  for which the likelihood function (7.1.5) is a maximum. Figure 7.1 illustrates the situation for various forms of the likelihood function for the case of a single parameter  $\lambda$ .

The maximum can be located simply by setting the first derivative of the likelihood function with respect to the parameter  $\lambda_i$  equal to zero. The derivative of a product with many factors is, however, unpleasant to deal with. One first constructs therefore the logarithm of the likelihood function,

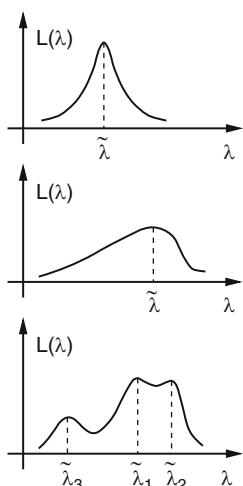
$$\ell = \ln L = \sum_{j=1}^N \ln f(\mathbf{x}^{(j)}; \lambda). \quad (7.2.1)$$

The function  $\ell$  is also often called the likelihood function. Sometimes one says explicitly “*log-likelihood function*”. Clearly the maxima of (7.2.1) are identical with those of (7.1.5). For the case of a single parameter we now construct

$$\ell' = d\ell/d\lambda = 0. \quad (7.2.2)$$

The problem of estimating a parameter is now reduced to solving this *likelihood equation*. By application of (7.2.1) we can write

$$\ell' = \sum_{j=1}^N \frac{d}{d\lambda} \ln f(\mathbf{x}^{(j)}; \lambda) = \sum_{j=1}^N \frac{f'}{f} = \sum_{j=1}^N \varphi(\mathbf{x}^{(j)}; \lambda), \quad (7.2.3)$$



**Fig. 7.1:** Likelihood functions.

where

$$\varphi(\mathbf{x}^{(j)}; \lambda) = \left( \frac{d}{d\lambda} f(\mathbf{x}^{(j)}; \lambda) \right) / f(\mathbf{x}^{(j)}; \lambda) \quad (7.2.4)$$

is the *logarithmic derivative* of the density  $f$  with respect to  $\lambda$ .

In the general case of  $p$  parameters the likelihood equation (7.2.2) is replaced by the system of  $p$  simultaneous equations,

$$\frac{\partial \ell}{\partial \lambda_i} = 0, \quad i = 1, 2, \dots, p. \quad (7.2.5)$$

### Example 7.2: Repeated measurements of differing accuracy

If a quantity is measured with different instruments, then the measurement errors are in general different. The measurements  $\mathbf{x}^{(j)}$  are spread about the true value  $\lambda$ . Suppose the errors are normally distributed, so that a measurement corresponds to obtaining a sample from a Gaussian distribution with mean  $\lambda$  and standard deviation  $\sigma_j$ . The a posteriori probability for a measured value is then

$$f(\mathbf{x}^{(j)}; \lambda) dx = \frac{1}{\sqrt{2\pi}\sigma_j} \exp\left(-\frac{(\mathbf{x}^{(j)} - \lambda)^2}{2\sigma_j^2}\right) dx.$$

From all  $N$  measurements one obtains the likelihood function

$$L = \prod_{j=1}^N \frac{1}{\sqrt{2\pi}\sigma_j} \exp\left(-\frac{(\mathbf{x}^{(j)} - \lambda)^2}{2\sigma_j^2}\right) \quad (7.2.6)$$

with the logarithm

$$\ell = -\frac{1}{2} \sum_{j=1}^N \frac{(\mathbf{x}^{(j)} - \lambda)^2}{\sigma_j^2} + \text{const.} \quad (7.2.7)$$

The likelihood equation thus becomes

$$\frac{d\ell}{d\lambda} = \sum_{j=1}^N \frac{\mathbf{x}^{(j)} - \lambda}{\sigma_j^2} = 0.$$

It has the solution

$$\tilde{\lambda} = \frac{\sum_{j=1}^N \frac{\mathbf{x}^{(j)}}{\sigma_j^2}}{\sum_{j=1}^N \frac{1}{\sigma_j^2}}. \quad (7.2.8)$$

Since  $d^2\ell/d\lambda^2 = -\sum \sigma_j^{-2} < 0$ , the solution is, in fact, a maximum. Thus we see that we obtain the maximum likelihood estimator as the mean of the  $N$  measurements weighted inversely by the variances of the individual measurements. ■

**Example 7.3:** Estimation of the parameter  $N$  of the hypergeometric distribution

As in the example with coins at the beginning of this chapter, sometimes parameters to be estimated can only take on discrete values. In Example 5.2 we indicated the possibility of estimating zoological population densities by means of tagging and recapture. According to (5.3.1), the probability to catch exactly  $n$  fish of which  $k$  are tagged out of a pond with an unknown total of  $N$  fish, out of which  $K$  are tagged, is given by

$$L(k; n, K, N) = \frac{\binom{K}{k} \binom{N-K}{n-k}}{\binom{N}{n}}.$$

We must now find the value of  $N$  for which the function  $L$  is maximum. For this we use the ratio

$$\frac{L(k; n, k, N)}{L(k; n, k, N-1)} = \frac{(N-n)(N-k)}{(N-n-K+k)N} \begin{cases} > 1, & Nk < nK, \\ < 1, & Nk > nK. \end{cases}$$

The function  $L$  is thus maximum when  $N$  is the integer closest to  $nK/k$ . ■

### 7.3 Information Inequality. Minimum Variance Estimators. Sufficient Estimators

We now want to discuss once more the quality of an estimator. In Sect. 6.1 we called an estimator unbiased if for every sample the *bias* vanished,

$$B(\lambda) = E(\mathbf{S}) - \lambda = 0. \quad (7.3.1)$$

Lack of bias is, however, not the only characteristic required of a “good” estimator. More importantly one should require that the variance

$$\sigma^2(\mathbf{S})$$

is small. Here one must often find a compromise, since there is a connection between  $B$  and  $\sigma^2$ , described by the *information inequality*.<sup>†</sup>

One immediately sees that it is easy to achieve  $\sigma^2(\mathbf{S}) = 0$  simply by using a constant for  $\mathbf{S}$ . We consider an estimator  $\mathbf{S}(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(N)})$  that is

---

<sup>†</sup>This inequality was independently found by H. Cramer, M. Fréchet, and C. R. Rao as well as by other authors. It is also called the Cramer–Rao or Fréchet inequality.

a function of the sample  $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(N)}$ . According to (6.1.3) and (6.1.4) the joint probability density of the elements of the sample is

$$f(x^{(1)}, x^{(2)}, \dots, x^{(N)}; \lambda) = f(x^{(1)}; \lambda) f(x^{(2)}; \lambda) \cdots f(x^{(N)}; \lambda).$$

The expectation value of  $\mathbf{S}$  is thus

$$\begin{aligned} E(\mathbf{S}) &= \int \mathbf{S}(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N)}) f(x^{(1)}; \lambda) \cdots f(x^{(N)}; \lambda) \\ &\quad \times dx^{(1)} dx^{(2)} \cdots dx^{(N)}. \end{aligned} \quad (7.3.2)$$

According to (7.3.1), however, one also has

$$E(\mathbf{S}) = B(\lambda) + \lambda.$$

We now assume that we can differentiate with respect to  $\lambda$  in the integral. We then obtain

$$\begin{aligned} 1 + B'(\lambda) &= \\ &\int \mathbf{S} \left( \sum_{j=1}^N \frac{f'(x^{(j)}; \lambda)}{f(x^{(j)}; \lambda)} \right) f(x^{(1)}; \lambda) \cdots f(x^{(N)}; \lambda) dx^{(1)} \cdots dx^{(N)}, \end{aligned}$$

which is equivalent to

$$1 + B'(\lambda) = E \left\{ \mathbf{S} \sum_{j=1}^N \frac{f'(x^{(j)}; \lambda)}{f(x^{(j)}; \lambda)} \right\} = E \left\{ \mathbf{S} \sum_{j=1}^N \varphi(x^{(j)}; \lambda) \right\}.$$

From (7.2.3) we have

$$\ell' = \sum_{j=1}^N \varphi(x^{(j)}; \lambda)$$

and therefore

$$1 + B'(\lambda) = E\{\mathbf{S}\ell'\}. \quad (7.3.3)$$

One clearly has

$$\int f(x^{(1)}; \lambda) \cdots f(x^{(N)}; \lambda) dx^{(1)} \cdots dx^{(N)} = 1.$$

If we also compute the derivative with respect to  $\lambda$ , we obtain

$$\int \sum_{j=1}^N \frac{f'(x^{(j)}; \lambda)}{f(x^{(j)}; \lambda)} f(x^{(1)}; \lambda) \cdots f(x^{(N)}; \lambda) dx^{(1)} \cdots dx^{(N)} = E(\ell') = 0.$$

By multiplying this equation by  $E(\mathbf{S})$  and subtracting the result of (7.3.3) one obtains

$$1 + B'(\lambda) = E\{\mathbf{S}\ell'\} - E(\mathbf{S})E(\ell') = E\{[\mathbf{S} - E(\mathbf{S})]\ell'\}. \quad (7.3.4)$$

In order to see the significance of this expression, we need to use the Cauchy–Schwarz inequality in the following form:

If  $\mathbf{x}$  and  $\mathbf{y}$  are random variables and if  $\mathbf{x}^2$  and  $\mathbf{y}^2$  have finite expectation values, then

$$\{E(\mathbf{xy})\}^2 \leq E(\mathbf{x}^2)E(\mathbf{y}^2). \quad (7.3.5)$$

To prove this inequality we consider the expression

$$E((a\mathbf{x} + \mathbf{y})^2) = a^2 E(\mathbf{x}^2) + 2a E(\mathbf{xy}) + E(\mathbf{y}^2) \geq 0. \quad (7.3.6)$$

This is a non-negative number for all values of  $a$ . If we consider for the moment the case of equality, then this is a quadratic equation for  $a$  with the solutions

$$a_{1,2} = -\frac{E(\mathbf{xy})}{E(\mathbf{x}^2)} \pm \sqrt{\left(\frac{E(\mathbf{xy})}{E(\mathbf{x}^2)}\right)^2 - \frac{E(\mathbf{y}^2)}{E(\mathbf{x}^2)}}. \quad (7.3.7)$$

The inequality (7.3.6) is then valid for all  $a$  if the term under the square root is negative or zero. From this follows the assertion

$$\frac{\{E(\mathbf{xy})\}^2}{\{E(\mathbf{x}^2)\}^2} - \frac{E(\mathbf{y}^2)}{E(\mathbf{x}^2)} \leq 0.$$

If we now apply the inequality (7.3.5) to (7.3.4), one obtains

$$\{1 + B'(\lambda)\}^2 \leq E\{[\mathbf{S} - E(\mathbf{S})]^2\}E(\ell'^2). \quad (7.3.8)$$

We now use (7.2.3) in order to rewrite the expression for  $E(\ell'^2)$ ,

$$\begin{aligned} E(\ell'^2) &= E \left\{ \left( \sum_{j=1}^N \varphi(\mathbf{x}^{(j)}; \lambda) \right)^2 \right\} \\ &= E \left\{ \sum_{j=1}^N (\varphi(\mathbf{x}^{(j)}; \lambda))^2 \right\} + E \left\{ \sum_{i \neq j} \varphi(\mathbf{x}^{(i)}; \lambda) \varphi(\mathbf{x}^{(j)}; \lambda) \right\}. \end{aligned}$$

All terms on the right-hand side vanish, since for  $i \neq j$

$$E\{\varphi(\mathbf{x}^{(i)}; \lambda) \varphi(\mathbf{x}^{(j)}; \lambda)\} = E\{\varphi(\mathbf{x}^{(i)}; \lambda)\} E\{\varphi(\mathbf{x}^{(j)}; \lambda)\},$$

$$E\{\varphi(\mathbf{x}; \lambda)\} = \int_{-\infty}^{\infty} \frac{f'(x; \lambda)}{f(x; \lambda)} f(x; \lambda) dx = \int f'(x; \lambda) dx,$$

and

$$\int_{-\infty}^{\infty} f(x; \lambda) dx = 1.$$

By differentiating the last line with respect to  $\lambda$  one obtains

$$\int_{-\infty}^{\infty} f'(x; \lambda) dx = 0.$$

Thus one has simply

$$E(\ell'^2) = E \left\{ \sum_{j=1}^N (\varphi(\mathbf{x}^{(j)}; \lambda))^2 \right\} = E \left\{ \sum_{j=1}^N \left( \frac{f'(\mathbf{x}^{(j)}; \lambda)}{f(\mathbf{x}^{(j)}; \lambda)} \right)^2 \right\}.$$

Since the individual terms of the sum are independent, the expectation value of the sum is simply the sum of the expectation values. The individual expectation values do not depend on the elements of the sample. Therefore one has

$$I(\lambda) = E(\ell'^2) = N E \left\{ \left( \frac{f'(x; \lambda)}{f(x; \lambda)} \right)^2 \right\}.$$

This expression is called the *information of the sample with respect to  $\lambda$* . It is a non-negative number, which vanishes if the likelihood function does not depend on the parameter  $\lambda$ .

It is sometimes useful to write the information in a somewhat different form. To do this we differentiate the expression

$$E \left( \frac{f'(x; \lambda)}{f(x; \lambda)} \right) = \int_{-\infty}^{\infty} \frac{f'(x; \lambda)}{f(x; \lambda)} f(x; \lambda) dx = 0$$

once more with respect to  $\lambda$  and obtain

$$\begin{aligned} 0 &= \int_{-\infty}^{\infty} \left\{ \frac{f'^2}{f} + f \left( \frac{f'}{f} \right)' \right\} dx = \int_{-\infty}^{\infty} \left\{ \left( \frac{f'}{f} \right)^2 + \left( \frac{f'}{f} \right)' \right\} f dx \\ &= E \left\{ \left( \frac{f'}{f} \right)^2 \right\} + E \left\{ \left( \frac{f'}{f} \right)' \right\}. \end{aligned}$$

The information can then be written as

$$I(\lambda) = N E \left\{ \left( \frac{f'(x; \lambda)}{f(x; \lambda)} \right)^2 \right\} = -N E \left\{ \left( \frac{f'(x; \lambda)}{f(x; \lambda)} \right)' \right\}$$

or

$$I(\lambda) = E(\ell'^2) = -E(\ell''). \quad (7.3.9)$$

The inequality (7.3.8) can now be written in the following way:

$$\{1 + B'(\lambda)\}^2 \leq \sigma^2(\mathbf{S})I(\lambda)$$

or

$$\sigma^2(\mathbf{S}) \geq \frac{\{1 + B'(\lambda)\}^2}{I(\lambda)}. \quad (7.3.10)$$

This is the *information inequality*. It gives the connection between the bias and the variance of an estimator and the information of a sample. It should be noted that in its derivation no assumption about the estimator was made. The right-hand side of the inequality (7.3.10) is therefore a lower bound for the variance of an estimator. It is called the *minimum variance bound* or *Cramer–Rao bound*. In cases where the bias does not depend on  $\lambda$ , i.e., particularly in cases of vanishing bias, the inequality (7.3.10) simplifies to

$$\sigma^2(\mathbf{S}) \geq 1/I(\lambda). \quad (7.3.11)$$

This relation justifies using the name information. As the information of a sample increases, the variance of an estimator can be made smaller.

We now ask under which circumstances the minimum variance bound is attained, or explicitly, when the equals sign in the relation (7.3.10) holds. In the inequality (7.3.6), one has equality if  $(a\mathbf{x} + \mathbf{y})$  vanishes, since only then does one have  $E\{(a\mathbf{x} + \mathbf{y})^2\} = 0$  for all values of  $a$ ,  $\mathbf{x}$ , and  $\mathbf{y}$ . Applied to (7.3.8), this means that

$$\ell' + a(\mathbf{S} - E(\mathbf{S})) = 0$$

or

$$\ell' = A(\lambda)(\mathbf{S} - E(\mathbf{S})). \quad (7.3.12)$$

Here  $A$  means an arbitrary quantity that does not depend on the sample  $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(N)}$ , but may be, however, a function of  $\lambda$ . By integration we obtain

$$\ell = \int \ell' d\lambda = B(\lambda)\mathbf{S} + C(\lambda) + D \quad (7.3.13)$$

and finally

$$L = d \exp\{B(\lambda)\mathbf{S} + C(\lambda)\}. \quad (7.3.14)$$

The quantities  $d$  and  $D$  do not depend on  $\lambda$ .

We thus see that estimators attain the minimum variance bound when the likelihood function is of the special form (7.3.14). They are therefore called *minimum variance estimators*.

For the case of an unbiased minimum variance estimator we obtain from (7.3.11)

$$\sigma^2(\mathbf{S}) = \frac{1}{I(\lambda)} = \frac{1}{E(\ell'^2)}. \quad (7.3.15)$$

By substituting (7.3.12) one obtains

$$\sigma^2(\mathbf{S}) = \frac{1}{(A(\lambda))^2 E\{(\mathbf{S} - E(\mathbf{S}))^2\}} = \frac{1}{(A(\lambda))^2 \sigma^2(\mathbf{S})}$$

or

$$\sigma^2(\mathbf{S}) = \frac{1}{|A(\lambda)|}. \quad (7.3.16)$$

If instead of (7.3.14) only the weaker requirement

$$L = g(\mathbf{S}, \lambda) c(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(N)}) \quad (7.3.17)$$

holds, then the estimator  $\mathbf{S}$  is said to be *sufficient* for  $\lambda$ . One can show [see, e.g., Kendall and Stuart, Vol. 2 (1967)], that no other estimator can contribute information about  $\lambda$  that is not already contained in  $S$ , if the requirement (7.3.17) is fulfilled. Hence the name “sufficient estimator” (or statistic).

#### **Example 7.4:** Estimator for the parameter of the Poisson distribution

Consider the *Poisson distribution* (5.4.1)

$$f(k) = \frac{\lambda^k}{k!} e^{-\lambda}.$$

The likelihood function of a sample  $k^{(1)}, k^{(2)}, \dots, k^{(N)}$  is

$$\ell = \sum_{j=1}^N \{k^{(j)} \ln \lambda - \ln(k^{(j)}!) - \lambda\}$$

and its derivative with respect to  $\lambda$  is

$$\begin{aligned} \frac{d\ell}{d\lambda} &= \ell' = \sum_{j=1}^N \left\{ \frac{k^{(j)}}{\lambda} - 1 \right\} = \frac{1}{\lambda} \sum_{j=1}^N \{k^{(j)} - \lambda\}, \\ \ell' &= \frac{N}{\lambda} (\bar{k} - \lambda). \end{aligned} \quad (7.3.18)$$

Comparing with (7.3.12) and (7.3.16) shows that the arithmetic mean  $\bar{k}$  is an unbiased minimum variance estimator with variance  $\lambda/N$ . ■

**Example 7.5:** Estimator for the parameter of the binomial distribution

The likelihood function of a sample from a *binomial distribution* with the parameters  $p = \lambda$ ,  $q = 1 - \lambda$  is given directly by (5.1.3),

$$L(k, \lambda) = \binom{n}{k} \lambda^k (1 - \lambda)^{n-k}.$$

(The result of the sample can be summarized by the statement that in  $n$  experiments, the event  $A$  occurred  $k$  times; see Sect. 5.1.) One then has

$$\begin{aligned}\ell &= \ln L = k \ln \lambda + (n - k) \ln(1 - \lambda) + \ln \binom{n}{k}, \\ \ell' &= \frac{k}{\lambda} - \frac{n - k}{1 - \lambda} = \frac{n}{\lambda(1 - \lambda)} \left( \frac{k}{n} - \lambda \right).\end{aligned}$$

By comparing with (7.3.12) and (7.3.16) one finds  $k/n$  to be a minimum variance estimator with variance  $\lambda(1 - \lambda)/n$ . ■

**Example 7.6:** Law of error combination (“Quadratic averaging of individual errors”)

We now return to the problem of Example 7.2 of repeated measurements of the same quantity with varying uncertainties, or expressed in another way, to the problem of drawing a sample from normal distributions with the same mean  $\lambda$  and different but known variances  $\sigma_j^2$ . From (7.2.7) we obtain

$$\frac{d\ell}{d\lambda} = \ell' = \sum_{j=1}^N \frac{x^{(j)} - \lambda}{\sigma_j^2}.$$

We can rewrite this expression as

$$\begin{aligned}\ell' &= \sum \frac{x^{(j)}}{\sigma_j^2} - \sum \frac{\lambda}{\sigma_j^2} \\ &= \sum_{j=1}^N \frac{1}{\sigma_j^2} \left\{ \frac{\sum x^{(j)}}{\sum \frac{1}{\sigma_j^2}} - \lambda \right\}.\end{aligned}$$

As in Example 7.2 we recognize

$$\mathbf{S} = \tilde{\lambda} = \frac{\sum \frac{x^{(j)}}{\sigma_j^2}}{\sum \frac{1}{\sigma_j^2}} \quad (7.3.19)$$

as an unbiased estimator for  $\lambda$ . Comparing with (7.3.12) shows that it is also a minimum variance estimator. From (7.3.16) one determines that its variance is

$$\sigma^2(\tilde{\lambda}) = \left( \sum_{j=1}^N \frac{1}{\sigma_j^2} \right)^{-1}. \quad (7.3.20)$$

The relation (7.3.20) often goes by the name of the *law of error combination* or *quadratic averaging of individual errors*. It could have been obtained by application of the rule of error propagation (3.8.7) to (7.3.19). If we identify  $\sigma(\tilde{\lambda})$  as the error of the estimator  $\tilde{\lambda}$  and  $\sigma_j$  as the error of the  $j$ th measurement, then we can write it in its usual form

$$\Delta\tilde{\lambda} = \left( \frac{1}{(\Delta x_1)^2} + \frac{1}{(\Delta x_2)^2} + \cdots + \frac{1}{(\Delta x_n)^2} \right)^{-\frac{1}{2}}. \quad (7.3.21)$$

If all of the measurements have the same error  $\sigma = \sigma_j$ , Eqs. (7.3.19), (7.3.20) simplify to

$$\tilde{\lambda} = \bar{x}, \quad \sigma^2(\tilde{\lambda}) = \sigma^2/n,$$

which we have already found in Sect. 6.2. ■

## 7.4 Asymptotic Properties of the Likelihood Function and Maximum-Likelihood Estimators

We can now show heuristically several important properties of the likelihood function and maximum-likelihood estimators for very large data samples, that is, for the limit  $N \rightarrow \infty$ . The estimator  $S = \tilde{\lambda}$  was defined as the solution to the likelihood equation

$$\ell'(\lambda) = \sum_{j=1}^N \left( \frac{f'(\mathbf{x}^{(j)}; \lambda)}{f(\mathbf{x}^{(j)}; \lambda)} \right)'_{\tilde{\lambda}} = 0. \quad (7.4.1)$$

Let us assume that  $\ell'(\lambda)$  can be differentiated with respect to  $\lambda$  one more time, so that we can expand it in a series around the point  $\lambda = \tilde{\lambda}$ ,

$$\ell'(\lambda) = \ell'(\tilde{\lambda}) + (\lambda - \tilde{\lambda})\ell''(\tilde{\lambda}) + \cdots \quad (7.4.2)$$

The first term on the right side vanishes because of Eq. (7.4.1). In the second term one can write explicitly

$$\ell''(\tilde{\lambda}) = \sum_{j=1}^N \left( \frac{f'(\mathbf{x}^{(j)}; \lambda)}{f(\mathbf{x}^{(j)}; \lambda)} \right)''_{\tilde{\lambda}}.$$

This expression has the form of the mean value of a sample. For very large  $N$  it can be replaced by the expectation value of the population (Sect. 6.2),

$$\ell''(\tilde{\lambda}) = NE \left\{ \left( \frac{f'(x; \lambda)}{f(x; \lambda)} \right)'_{\tilde{\lambda}} \right\}. \quad (7.4.3)$$

Using Eq. (7.3.9) we can now write

$$\ell''(\tilde{\lambda}) = E(\ell''(\tilde{\lambda})) = -E(\ell'^2(\tilde{\lambda})) = -I(\tilde{\lambda}) = -1/b^2. \quad (7.4.4)$$

In this way we can replace the expression for  $\ell''(\tilde{\lambda})$ , which is a function the sample  $x^{(1)}, x^{(2)}, \dots, x^{(N)}$ , by the quantity  $-1/b^2$ , which only depends on the probability density  $f$  and the estimator  $\tilde{\lambda}$ . If one neglects higher-order terms, Eq. (7.4.2) can be expressed as

$$\ell'(\lambda) = -\frac{1}{b^2}(\lambda - \tilde{\lambda}). \quad (7.4.5)$$

By integration one obtains

$$\ell(\lambda) = -\frac{1}{2b^2}(\lambda - \tilde{\lambda})^2 + c.$$

Inserting  $\lambda = \tilde{\lambda}$  gives  $c = \ell(\tilde{\lambda})$ , or

$$\ell(\lambda) - \ell(\tilde{\lambda}) = -\frac{1}{2} \frac{(\lambda - \tilde{\lambda})^2}{b^2}. \quad (7.4.6)$$

By exponentiation one obtains

$$L(\lambda) = k \exp\{-(\lambda - \tilde{\lambda})^2/2b^2\}, \quad (7.4.7)$$

where  $k$  is a constant. The likelihood function  $L(\lambda)$  has the form of a normal distribution with mean  $\tilde{\lambda}$  and variance  $b^2$ . At the values  $\lambda = \tilde{\lambda} \pm b$ , where  $\lambda$  is one standard deviation from  $\tilde{\lambda}$ , one has

$$-(\ell(\lambda) - \ell(\tilde{\lambda})) = \frac{1}{2}. \quad (7.4.8)$$

We can now compare (7.4.7) with Eqs. (7.3.12) and (7.3.16). Since we are estimating the parameter  $\lambda$ , we must write  $S = \tilde{\lambda}$  and thus  $E(S) = \lambda$ . The estimator  $\tilde{\lambda}$  is therefore an unbiased minimum variance estimator with variance

$$\sigma^2(\tilde{\lambda}) = b^2 = \frac{1}{I(\tilde{\lambda})} = \frac{1}{E(\ell'^2(\tilde{\lambda}))} = -\frac{1}{E(\ell''(\tilde{\lambda}))}. \quad (7.4.9)$$

Since the estimator  $\tilde{\lambda}$  only possesses this property for the limiting case  $N \rightarrow \infty$ , we call it *asymptotically unbiased*. This is equivalent to the statement

that the maximum likelihood estimator is consistent (Sect. 6.1). For the same reason the likelihood function is called *asymptotically normal*.

In Sect. 7.2 we interpreted the likelihood function  $L(\lambda)$  as a measure of the probability that the true value  $\lambda_0$  of a parameter is equal to  $\lambda$ . The result of an estimator is often represented in abbreviated form,

$$\lambda = \tilde{\lambda} \pm \sigma(\tilde{\lambda}) = \tilde{\lambda} \pm \Delta\tilde{\lambda}.$$

Since the likelihood function is asymptotically normal, at least in the case of large samples, i.e., many measurements, this can be interpreted by saying that the probability that the true value  $\lambda_0$  lies in the interval

$$\tilde{\lambda} - \Delta\tilde{\lambda} < \lambda_0 < \tilde{\lambda} + \Delta\tilde{\lambda}$$

is 68.3 % (Sect. 5.8). In practice the relation above is used for large but finite samples. Unfortunately one cannot construct any general rule for determining when a sample is large enough for this procedure to be reliable. Clearly if  $N$  is finite, (7.4.3) is only an approximation, whose accuracy depends not only on  $N$ , but also on the particular probability density  $f(x; \lambda)$ .

**Example 7.7:** Determination of the mean lifetime from a small number of decays

The probability that a radioactive nucleus, which exists at time  $t = 0$ , decays in the time interval between  $t$  and  $t + dt$  is

$$f(t) dt = \frac{1}{\tau} \exp(-t/\tau) dt.$$

For observed decay times  $t_1, t_2, \dots, t_N$  the likelihood function is

$$L = \frac{1}{\tau^N} \exp \left\{ -\frac{1}{\tau} \sum_{i=1}^N t_i \right\} = \frac{1}{\tau^N} \exp \left\{ -\frac{N}{\tau} \bar{t} \right\}.$$

Its logarithm is

$$\ell = \ln L = -\frac{N}{\tau} \bar{t} - N \ln \tau$$

with the derivative

$$\ell' = \frac{N}{\tau} \left( \frac{\bar{t}}{\tau} - 1 \right) = \frac{N}{\tau^2} (\bar{t} - \tau).$$

Comparing with (7.3.12) we see that  $\tilde{\tau} = \bar{t}$  is the maximum likelihood solution, which has a variance of  $\sigma^2(\tau) = \tau^2/N$ . For  $\tau = \tilde{\tau} = \bar{t}$  one obtains  $\Delta\tilde{\tau} = \bar{t}/\sqrt{N}$ .

For  $\tilde{\tau} = \bar{t}$  one has

$$\ell(\tilde{\tau}) = \ell_{\max} = -N(1 + \ln \bar{t}).$$

We can write

$$-(\ell(\tau) - \ell(\tilde{\tau})) = N \left( \frac{\bar{t}}{\tau} + \ln \frac{\tau}{\bar{t}} - 1 \right).$$

From this expression for the log-likelihood function one cannot so easily recognize the asymptotic form (7.4.6) for  $N \rightarrow \infty$ . For small values of  $N$  it clearly does not have this form. Corresponding to (7.4.8), we want to use the values  $\tau_+ = \tilde{\tau} + \Delta_+$  and  $\tau_- = \tilde{\tau} - \Delta_-$ , where one has

$$-(\ell(\tau_{\pm}) - \ell(\tilde{\tau})) = \frac{1}{2}$$

for the *asymmetric errors*  $\Delta_+$ ,  $\Delta_-$ . Clearly we expect for  $N \rightarrow \infty$  that  $\Delta_+$ ,  $\Delta_- \rightarrow \Delta\tilde{\tau} = \sigma(\tilde{\tau})$ .

In Fig. 7.2 the  $N$  observed decay times  $t_i$  are marked as vertical tick marks on the abscissa for various small values of  $N$ . In addition the function  $-(\ell - \ell_{\max}) = -(\ell(\tau) - \ell(\tilde{\tau}))$  is plotted. The points  $\tau_+$  and  $\tau_-$  are found where a horizontal line intersects  $-(\ell - \ell_{\max}) = 1/2$ . The point  $\tilde{\tau}$  is indicated by an additional mark on the horizontal line. One sees that with increasing  $N$  the function  $-(\ell - \ell_{\max})$  approaches more and more the symmetric parabolic form and that the errors  $\Delta_+$ ,  $\Delta_-$ , and  $\Delta\tilde{\tau}$  become closer to each other. ■

## 7.5 Simultaneous Estimation of Several Parameters: Confidence Intervals

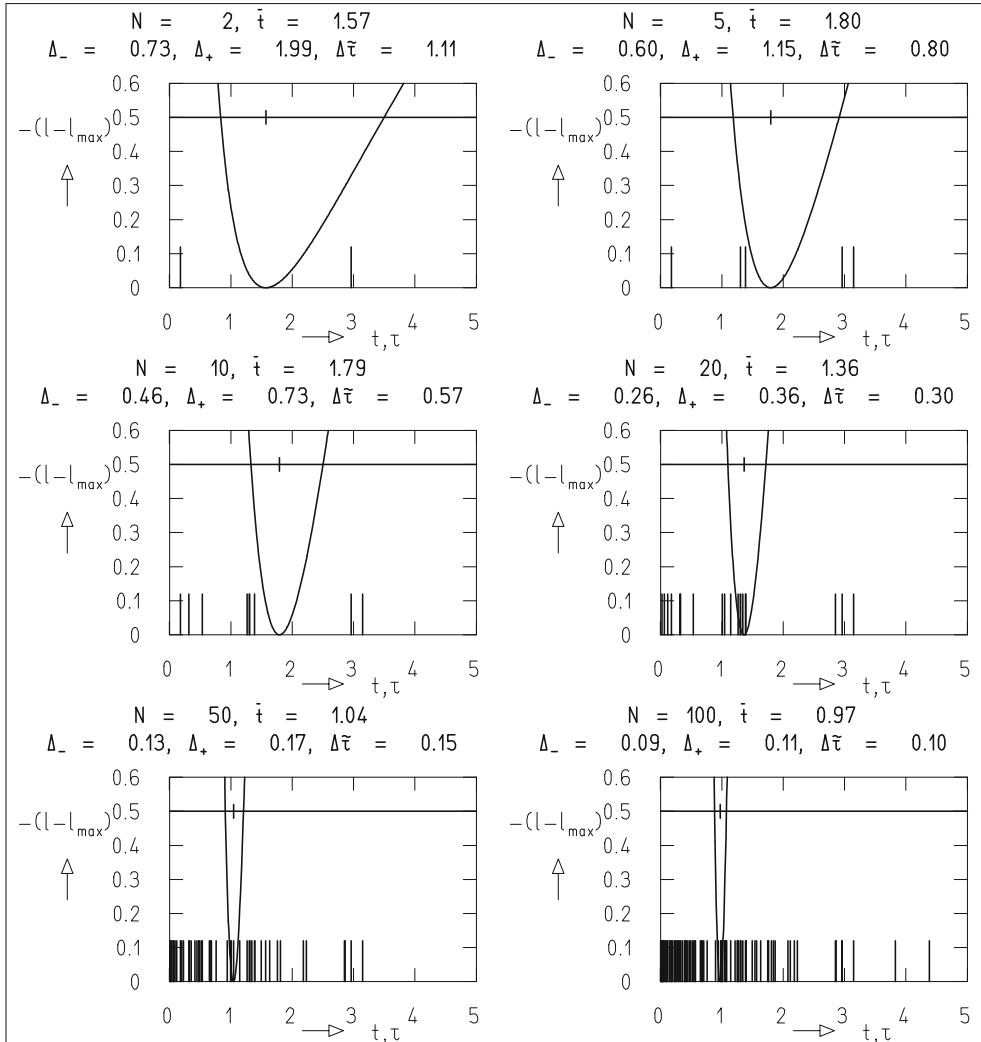
We have already given a system of equations (7.2.5) allowing the simultaneous determination of  $p$  parameters  $\lambda = (\lambda_1, \lambda_2, \dots, \lambda_p)$ . It turns out that it is not the parameter determination but rather the estimation of their errors that becomes significantly more complicated in the case of several parameters. In particular we will need to consider correlations as well as errors of the parameters.

We extend our considerations from Sect. 7.4 on the properties of the likelihood function to the case of several parameters. The log-likelihood function

$$\ell(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(N)}; \boldsymbol{\lambda}) = \sum_{j=1}^N \ln f(\mathbf{x}^{(j)}; \boldsymbol{\lambda}) \quad (7.5.1)$$

can be expanded in a series about the point

$$\tilde{\boldsymbol{\lambda}} = (\tilde{\lambda}_1, \tilde{\lambda}_2, \dots, \tilde{\lambda}_p) \quad (7.5.2)$$



**Fig. 7.2:** Data and log-likelihood function for Example 7.7.

to give

$$\begin{aligned} \ell(\boldsymbol{\lambda}) &= \ell(\tilde{\boldsymbol{\lambda}}) + \sum_{k=1}^p \left( \frac{\partial \ell}{\partial \lambda_k} \right)_{\tilde{\boldsymbol{\lambda}}} (\lambda_k - \tilde{\lambda}_k) \\ &\quad + \frac{1}{2} \sum_{\ell=1}^p \sum_{m=1}^p \left( \frac{\partial^2 \ell}{\partial \lambda_\ell \partial \lambda_m} \right)_{\tilde{\boldsymbol{\lambda}}} (\lambda_\ell - \tilde{\lambda}_\ell)(\lambda_m - \tilde{\lambda}_m) + \dots \end{aligned} \quad (7.5.3)$$

Since by the definition of  $\tilde{\boldsymbol{\lambda}}$  one has

$$\left( \frac{\partial \ell}{\partial \lambda_k} \right)_{\tilde{\boldsymbol{\lambda}}} = 0, \quad k = 1, 2, \dots, p, \quad (7.5.4)$$

which holds for all  $k$ , the series simplifies to

$$-\left(\ell(\boldsymbol{\lambda}) - \ell(\tilde{\boldsymbol{\lambda}})\right) = \frac{1}{2}(\boldsymbol{\lambda} - \tilde{\boldsymbol{\lambda}})^T A(\boldsymbol{\lambda} - \tilde{\boldsymbol{\lambda}}) + \dots \quad (7.5.5)$$

with

$$-A = \begin{pmatrix} \frac{\partial^2 \ell}{\partial \lambda_1^2} & \frac{\partial^2 \ell}{\partial \lambda_1 \partial \lambda_2} & \cdots & \frac{\partial^2 \ell}{\partial \lambda_1 \partial \lambda_p} \\ \frac{\partial^2 \ell}{\partial \lambda_1 \partial \lambda_2} & \frac{\partial^2 \ell}{\partial \lambda_2^2} & \cdots & \frac{\partial^2 \ell}{\partial \lambda_2 \partial \lambda_p} \\ \vdots & & & \\ \frac{\partial^2 \ell}{\partial \lambda_1 \partial \lambda_p} & \frac{\partial^2 \ell}{\partial \lambda_2 \partial \lambda_p} & \cdots & \frac{\partial^2 \ell}{\partial \lambda_p^2} \end{pmatrix}_{\boldsymbol{\lambda}=\tilde{\boldsymbol{\lambda}}}. \quad (7.5.6)$$

In the limit  $N \rightarrow \infty$  we can replace the elements of  $A$ , which still depend on the specific sample, by the corresponding expectation values,

$$B = E(A) = - \begin{pmatrix} E\left(\frac{\partial^2 \ell}{\partial \lambda_1^2}\right) & E\left(\frac{\partial^2 \ell}{\partial \lambda_1 \partial \lambda_2}\right) & \cdots & E\left(\frac{\partial^2 \ell}{\partial \lambda_1 \partial \lambda_p}\right) \\ E\left(\frac{\partial^2 \ell}{\partial \lambda_1 \partial \lambda_2}\right) & E\left(\frac{\partial^2 \ell}{\partial \lambda_2^2}\right) & \cdots & E\left(\frac{\partial^2 \ell}{\partial \lambda_2 \partial \lambda_p}\right) \\ \vdots & & & \\ E\left(\frac{\partial^2 \ell}{\partial \lambda_1 \partial \lambda_p}\right) & E\left(\frac{\partial^2 \ell}{\partial \lambda_2 \partial \lambda_p}\right) & \cdots & E\left(\frac{\partial^2 \ell}{\partial \lambda_p^2}\right) \end{pmatrix}_{\boldsymbol{\lambda}=\tilde{\boldsymbol{\lambda}}}. \quad (7.5.7)$$

If we neglect higher-order terms, we can give the likelihood function as

$$L = k \exp\left\{-\frac{1}{2}(\boldsymbol{\lambda} - \tilde{\boldsymbol{\lambda}})^T B(\boldsymbol{\lambda} - \tilde{\boldsymbol{\lambda}})\right\}. \quad (7.5.8)$$

Comparing with (5.10.1) shows that this is a  $p$ -dimensional normal distribution with mean  $\tilde{\boldsymbol{\lambda}}$  and covariance matrix

$$C = B^{-1}. \quad (7.5.9)$$

The variances of the maximum likelihood estimators  $\tilde{\lambda}_1, \tilde{\lambda}_2, \dots, \tilde{\lambda}_p$  are given by the diagonal elements of the matrix (7.5.9). The off-diagonal elements are the covariances between all possible pairs of estimators,

$$\sigma^2(\tilde{\lambda}_i) = c_{ii}, \quad (7.5.10)$$

$$\text{cov}(\tilde{\lambda}_j, \tilde{\lambda}_k) = c_{jk}. \quad (7.5.11)$$

For the correlation coefficient between the estimators  $\tilde{\lambda}_j, \tilde{\lambda}_k$  we can define

$$\varrho(\tilde{\lambda}_j, \tilde{\lambda}_k) = \frac{\text{cov}(\tilde{\lambda}_j, \tilde{\lambda}_k)}{\sigma(\tilde{\lambda}_j)\sigma(\tilde{\lambda}_k)}. \quad (7.5.12)$$

As in the case of a single parameter, the square roots of the variances are given as the error or standard deviations of the estimators,

$$\Delta\tilde{\lambda}_i = \sigma(\tilde{\lambda}_i) = \sqrt{c_{ii}}. \quad (7.5.13)$$

In Sect. 7.4 we determined that by giving the maximum-likelihood estimator and its error one defines a region that contains the true value of the parameter with a probability of 68.3 %. Since the likelihood function in the several parameter case is asymptotically a Gaussian distribution of several variables, this region is not determined only by the errors, but rather by the entire covariance matrix. In the special case of two parameters this is the covariance ellipse, which we introduced in Sect. 5.10.

The expression (7.5.8) has (with the replacement  $\mathbf{x} = \boldsymbol{\lambda}$ ) exactly the form of (5.10.1). We can therefore use it for all of the results of Sect. 5.10. For the exponent one has

$$-\frac{1}{2}(\boldsymbol{\lambda} - \tilde{\boldsymbol{\lambda}})^T B(\boldsymbol{\lambda} - \tilde{\boldsymbol{\lambda}}) = -\frac{1}{2}g(\boldsymbol{\lambda}) = -\left\{\ell(\boldsymbol{\lambda}) - \ell(\tilde{\boldsymbol{\lambda}})\right\}. \quad (7.5.14)$$

In the parameter space spanned by  $\lambda_1, \dots, \lambda_p$ , the covariance ellipsoid of the distribution (7.5.8) is then determined by the condition

$$g(\boldsymbol{\lambda}) = 1 = 2\left\{\ell(\boldsymbol{\lambda}) - \ell(\tilde{\boldsymbol{\lambda}})\right\}. \quad (7.5.15)$$

For other values of  $g(\boldsymbol{\lambda})$  one obtains the confidence ellipsoids introduced in Sect. 5.10. For smaller values of  $N$ , the series (7.5.3) cannot be truncated and the approximation (7.5.7) is not valid. Nevertheless, the solution (7.5.4) can clearly still be computed. For a given probability  $W$  one obtains instead of a confidence ellipsoid a *confidence region*, contained within the hypersurface

$$g(\boldsymbol{\lambda}) = 2\left\{\ell(\boldsymbol{\lambda}) - \ell(\tilde{\boldsymbol{\lambda}})\right\} = \text{const.} \quad (7.5.16)$$

The value of  $g$  is determined in the same way as for the confidence ellipsoid as in Sect. 5.10.

In Example 7.7 we computed the region  $\tilde{\lambda} - \Delta_- < \lambda < \tilde{\lambda} + \Delta_+$  for the case of a single variable. This corresponds to a confidence region with the probability content 68.3 %.

**Example 7.8:** Estimation of the mean and variance of a normal distribution

We want to determine the mean  $\lambda_1$  and standard deviation  $\lambda_2$  of a normal distribution using a sample of size  $N$ . This problem occurs, for example, in the measurement of the range of  $\alpha$ -particles in matter. Because of the statistical nature of the energy loss through a large number of independent individual collisions, the range of the individual particles is Gaussian distributed about some mean value. By measuring the range  $\mathbf{x}^{(j)}$  of  $N$  different particles, the mean  $\lambda_1$  and “straggling constant”  $\lambda_2 = \sigma$  can be estimated. We obtain the likelihood function

$$L = \prod_{j=1}^N \frac{1}{\lambda_2 \sqrt{2\pi}} \exp\left(-\frac{(\mathbf{x}^{(j)} - \lambda_1)^2}{2\lambda_2^2}\right)$$

and

$$\ell = -\frac{1}{2} \sum_{j=1}^N \frac{(\mathbf{x}^{(j)} - \lambda_1)^2}{\lambda_2^2} - N \ln \lambda_2 - \text{const.}$$

The system of likelihood equations is

$$\begin{aligned} \frac{\partial \ell}{\partial \lambda_1} &= \sum_{j=1}^N \frac{\mathbf{x}^{(j)} - \lambda_1}{\lambda_2^2} = 0, \\ \frac{\partial \ell}{\partial \lambda_2} &= \frac{1}{\lambda_2^3} \sum_{j=1}^N (\mathbf{x}^{(j)} - \lambda_1)^2 - \frac{N}{\lambda_2} = 0. \end{aligned}$$

Its solution is

$$\begin{aligned} \tilde{\lambda}_1 &= \frac{1}{N} \sum_{j=1}^N \mathbf{x}^{(j)}, \\ \tilde{\lambda}_2 &= \sqrt{\frac{\sum_{j=1}^N (\mathbf{x}^{(j)} - \tilde{\lambda}_1)^2}{N}}. \end{aligned}$$

For the estimator of the mean, the maximum-likelihood method leads to the arithmetic mean of the individual measurements. For the variance it gives the quantity  $s'^2$  (6.2.4), which has a small bias, and not  $s^2$ , the unbiased estimator (6.2.6).

Let us now determine the matrix  $B$ . The second derivatives are

$$\frac{\partial^2 \ell}{\partial \lambda_1^2} = -\frac{N}{\lambda_2^2},$$

$$\begin{aligned}\frac{\partial^2 \ell}{\partial \lambda_1 \partial \lambda_2} &= -\frac{2 \sum (\mathbf{x}^{(j)} - \lambda_1)}{\lambda_2^3}, \\ \frac{\partial^2 \ell}{\partial \lambda_2^2} &= -\frac{3 \sum (\mathbf{x}^{(j)} - \lambda_1)^2}{\lambda_2^4} + \frac{N}{\lambda_2^2}.\end{aligned}$$

We use the procedure of (7.5.7), substitute  $\lambda_1, \lambda_2$  by  $\tilde{\lambda}_1, \tilde{\lambda}_2$  and find

$$B = \begin{pmatrix} N/\tilde{\lambda}_2^2 & 0 \\ 0 & 2N/\tilde{\lambda}_2^2 \end{pmatrix}$$

or for the covariance matrix

$$C = B^{-1} = \begin{pmatrix} \tilde{\lambda}_2^2/N & 0 \\ 0 & \tilde{\lambda}_2^2/2N \end{pmatrix}.$$

We interpret the diagonal elements as the errors of the corresponding parameters, i.e.,

$$\Delta \tilde{\lambda}_1 = \tilde{\lambda}_2 / \sqrt{N}, \quad \Delta \tilde{\lambda}_2 = \tilde{\lambda}_2 / \sqrt{2N}.$$

The estimators for  $\lambda_1$  and  $\lambda_2$  are not correlated. ■

**Example 7.9:** Estimators for the parameters of a two-dimensional normal distribution

To conclude we consider a population described by a two-dimensional normal distribution (Sect. 5.10)

$$\begin{aligned}f(x_1, x_2) &= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\varrho^2}} \exp\left[-\frac{1}{2(1-\varrho^2)}\right. \\ &\quad \times \left. \left\{ \frac{(x_1 - a_1)^2}{\sigma_1^2} + \frac{(x_2 - a_2)^2}{\sigma_2^2} - 2\varrho \frac{(x_1 - a_1)(x_2 - a_2)}{\sigma_1\sigma_2} \right\} \right].\end{aligned}$$

By constructing and solving a system of five simultaneous likelihood equations for the five parameters  $a_1, a_2, \sigma_1^2, \sigma_2^2, \varrho$  we obtain for the maximum-likelihood estimators

$$\begin{aligned}\bar{\mathbf{x}}_1 &= \frac{1}{N} \sum_{j=1}^N \mathbf{x}_1^{(j)}, & \bar{\mathbf{x}}_2 &= \frac{1}{N} \sum_{j=1}^N \mathbf{x}_2^{(j)}, \\ \mathbf{s}'_1^2 &= \frac{1}{N} \sum_{j=1}^N (\mathbf{x}_1^{(j)} - \bar{\mathbf{x}}_1)^2, & \mathbf{s}'_2^2 &= \frac{1}{N} \sum_{j=1}^N (\mathbf{x}_2^{(j)} - \bar{\mathbf{x}}_2)^2,\end{aligned}$$

$$r = \frac{\sum_{j=1}^N (\mathbf{x}_1^{(j)} - \bar{\mathbf{x}}_1)(\mathbf{x}_2^{(j)} - \bar{\mathbf{x}}_2)}{N \mathbf{s}'_1 \mathbf{s}'_2}. \quad (7.5.17)$$

Exactly as in Example 7.8, the estimators for the variances  $s'_1{}^2$  and  $s'_2{}^2$  are biased. This also holds for the expression (7.5.17), the *sample correlation coefficient r*. Like all maximum likelihood estimators, r is consistent, i.e., it provides a good estimation of  $\rho$  for very large samples. For  $N \rightarrow \infty$  the probability density of the random variable r becomes a normal distribution with mean  $\rho$  and variance

$$\sigma^2(r) = (1 - \rho^2)^2 / N. \quad (7.5.18)$$

For finite samples the distribution is asymmetric. It is therefore important to have a sufficiently large sample before applying Eq. (7.5.17). As a rule of thumb,  $N \geq 500$  is usually recommended. ■

## 7.6 Example Programs

**Example Program 7.1:** The class E1MaxLife computes the mean lifetime and its asymmetric errors from a small number of radioactive decays

The program performs the computations and the graphical display for the problem described in Example 7.7. First by Monte Carlo method a total of  $N$  decay times  $t_i$  of radioactive nuclei with a mean lifetime of  $\tau = 1$  are simulated. The number  $N$  of decays and also the seeds for the random number generator are entered interactively.

**Example Program 7.2:** The class E2MaxLife computes the maximum-likelihood estimates of the parameters of a bivariate normal distribution from a simulated sample

The program asks interactively for the number  $n_{\text{exp}}$  of experiments to simulate (i.e., of the samples to be treated consecutively), for the size  $n_{\text{pt}}$  of each sample and for the means  $a_1, a_2$ , the standard deviations  $\sigma_1, \sigma_2$ , and the correlation coefficient  $\rho$  of a bivariate Gaussian distribution.

The covariance matrix  $C$  of the normal distribution is calculated and the generator of random numbers from a multivariate Gaussian distribution is initialized. Each sample is generated and then analyzed, i.e., the quantities  $\bar{x}_1, \bar{x}_2, s'_1, s'_2$ , and  $r$  are computed, which are estimates of  $a_1, a_2, \sigma_1, \sigma_2$ , and  $\rho$  [cf. (7.5.17)]. The quantities are displayed for each sample.

**Suggestions:** Choose  $n_{\text{exp}} = 20$ , keep all other parameters fixed, and study the statistical fluctuations of  $r$  for  $n_{\text{pt}} = 5, 50, 500$ . Use the values  $\rho = 0, 0.5, 0.95$ .

# 8. Testing Statistical Hypotheses

## 8.1 Introduction

Often the problem of a statistical analysis does not involve determining an originally unknown parameter, but rather, one already has a preconceived opinion about the value of the parameter, i.e., a *hypothesis*. In a sample taken for quality control, for example, one might initially assume that certain critical values are normally distributed within tolerance levels around their nominal values. One would now like to test this hypothesis. To elucidate such test procedures, called *statistical tests*, we will consider such an example and for simplicity make the hypothesis that a sample of size 10 originates from a standard normal distribution.

Suppose the analysis of the sample resulted in the arithmetic mean  $\bar{x} = 0.154$ . Under the assumption that our hypothesis is correct, the random variable  $\bar{x}$  is normally distributed with mean 0 and standard deviation  $\frac{1}{\sqrt{10}}$ . We now ask for the probability to observe a value  $|\bar{x}| \geq 0.154$  from such a distribution. From (5.8.5) and Table I.3 this is

$$P(|\bar{x}| \geq 0.154) = 2\{1 - \psi_0(0.154\sqrt{10})\} = 0.62 \quad .$$

Thus we see that even if our hypothesis is correct, there is a probability of 62% that a sample of size 10 will lead to a sample mean that differs from the population mean by 0.154 or more.

We now find ourselves in the difficult situation of having to answer the simple question: “Is our hypothesis true or false?” A solution to this problem is provided by the concept of the *significance level*: One specifies before the test a (small) test probability  $\alpha$ . Staying with our previous example, if  $P(|\bar{x}| \geq 0.154) < \alpha$ , then one would regard the occurrence of  $\bar{x} = 0.154$  as improbable. That is, one would say that  $\bar{x}$  differs significantly from the hypothesized value

and the hypothesis would be rejected. The converse is, however, not true. If  $P$  does not fall below  $\alpha$ , we cannot say that “the hypothesis is true”, but rather “it is not contradicted by the result of the sample”. The choice of the significance level depends on the problem being considered. For quality control of pencils one might be satisfied with 1%. If, however, one wishes to determine insurance premiums such that the probability for the company to go bankrupt is less than  $\alpha$ , then one would probably still regard 0.01% as too high. In the analysis of scientific data  $\alpha$  values of 5, 1, or 0.1% are typically used. From Table I.3 we can obtain limiting values for  $|\bar{x}|$  such that a deviation in excess of these values corresponds to the given probabilities. These are

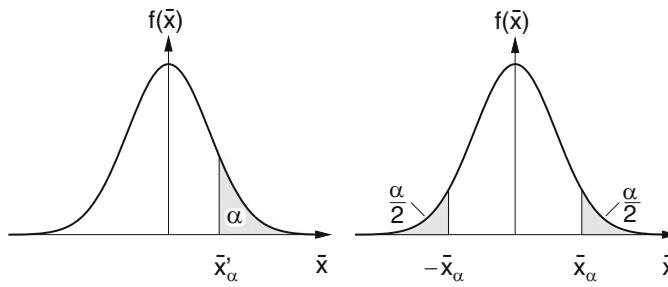
$$\begin{aligned} 0.05 &= 2\{1 - \psi_0(1.96)\} = 2\{1 - \psi_0(0.62\sqrt{10})\} , \\ 0.01 &= 2\{1 - \psi_0(2.58)\} = 2\{1 - \psi_0(0.82\sqrt{10})\} , \\ 0.001 &= 2\{1 - \psi_0(3.29)\} = 2\{1 - \psi_0(1.04\sqrt{10})\} . \end{aligned}$$

At these significance levels the value  $|\bar{x}|$  would have to exceed the values 0.62, 0.82, 1.04 before we could reject the hypothesis.

In some cases the sign of  $\bar{x}$  is important. In many production processes, deviations in a positive and negative direction are of different importance. (If a baker's rolls are too heavy, this reduces profits; if they are too light they cost the baker his license.) If one sets, e.g.,

$$P(\bar{x} \geq x'_\alpha) < \alpha ,$$

then this is referred to as a *one-sided test* in contrast to the *two-sided test*, which we have already considered (Fig. 8.1).



**Fig. 8.1:** One-sided and two-sided tests

For many tests one does not construct the sample mean but rather a certain function of the sample called a *test statistic*, which is particularly suited for tests of certain hypotheses. As above one specifies a certain significance level  $\alpha$  and determines a region  $U$  in the space of possible values of the test statistic  $T$  in which

$$P_H(T \in U) = \alpha .$$

The index  $H$  means that the probability was computed under the assumption that the hypothesis  $H$  is valid. One then obtains a sample, which results in a

particular value  $T'$  for the test statistic. If  $T'$  is in the region  $U$ , the *critical region* of the test, then the hypothesis is rejected.

In the next sections we will discuss some important tests in detail and then turn to a more rigorous treatment of test theory.

## 8.2 *F*-Test on Equality of Variances

The problem of comparing two variances occurs frequently in the development of measurement techniques or production procedures. Consider two populations with the same expectation value; e.g., one measures the same quantity with two different devices without systematic error. One may then ask if they also have the same variance.

To test this hypothesis we take samples of size  $N_1$  and  $N_2$  from both populations, which we assume to be normally distributed. We construct the sample variance (6.2.6) and consider the ratio

$$F = s_1^2/s_2^2 \quad . \quad (8.2.1)$$

If the hypothesis is true, then  $F$  will be near unity. It is known from Sect. 6.6 that for every sample we can construct a quantity that follows a  $\chi^2$ -distribution:

$$\begin{aligned} X_1^2 &= \frac{(N_1 - 1)s_1^2}{\sigma_1^2} = \frac{f_1 s_1^2}{\sigma_1^2} \quad , \\ X_2^2 &= \frac{(N_2 - 1)s_2^2}{\sigma_2^2} = \frac{f_2 s_2^2}{\sigma_2^2} \quad . \end{aligned}$$

The two distributions have  $f_1 = (N_1 - 1)$  and  $f_2 = (N_2 - 1)$  degrees of freedom.

Under the assumption that the hypothesis ( $\sigma_1^2 = \sigma_2^2$ ) is true, one has

$$F = \frac{f_2}{f_1} \frac{X_1^2}{X_2^2} \quad .$$

The probability density of a  $\chi^2$ -distribution with  $f$  degrees of freedom is [see (6.6.10)]

$$f(\chi^2) = \frac{1}{\Gamma(\frac{1}{2}f)2^{\frac{1}{2}f}} (\chi^2)^{\frac{1}{2}(f-2)} e^{-\frac{1}{2}\chi^2} \quad .$$

We now compute the probability\*

---

\*We use here the symbol  $W$  for a distribution function in order to avoid confusion with the ratio  $F$ .

$$W(Q) = P\left(\frac{X_1^2}{X_2^2} < Q\right)$$

that the ratio  $X_1^2/X_2^2$  is smaller than  $Q$ :

$$W(Q) = \frac{1}{\Gamma(\frac{1}{2}f_1)\Gamma(\frac{1}{2}f_2)2^{\frac{1}{2}(f_1+f_2)}} \int \int_{\substack{x>0 \\ y>0 \\ x/y>Q}} x^{\frac{1}{2}f_1-1} e^{-\frac{1}{2}x} y^{\frac{1}{2}f_2-1} e^{-\frac{1}{2}y} dx dy .$$

Calculating the integral gives

$$W(Q) = \frac{\Gamma(\frac{1}{2}f)}{\Gamma(\frac{1}{2}f_1)\Gamma(\frac{1}{2}f_2)} \int_0^Q t^{\frac{1}{2}f_1-1} (t+1)^{-\frac{1}{2}f} dt , \quad (8.2.2)$$

where

$$f = f_1 + f_2 .$$

Finally if one sets

$$F = Q f_2/f_1 ,$$

then the distribution function of the ratio  $F$  can be obtained from (8.2.2),

$$W(F) = P\left(\frac{s_1^2}{s_2^2} < F\right) .$$

This is called the Fisher  $F$ -distribution.<sup>†</sup> It depends on the parameters  $f_1$  and  $f_2$ . The probability density for the  $F$ -distribution is

$$f(F) = \left(\frac{f_1}{f_2}\right)^{\frac{1}{2}f_1} \frac{\Gamma(\frac{1}{2}(f_1+f_2))}{\Gamma(\frac{1}{2}f_1)\Gamma(\frac{1}{2}f_2)} F^{\frac{1}{2}f_1-1} \left(1 + \frac{f_1}{f_2}F\right)^{-\frac{1}{2}(f_1+f_2)} . \quad (8.2.3)$$

This is shown in Fig. 8.2 for fixed values of  $f_1$  and  $f_2$ . The distribution is reminiscent of the  $\chi^2$ -distribution; it is only non-vanishing for  $F \geq 0$ , and has a long tail for  $F \rightarrow \infty$ . Therefore it cannot be symmetric. One can easily show that for  $f_2 > 2$  the expectation value is simply

$$E(F) = f_2/(f_2 - 2) .$$

We can now determine a limit  $F'_\alpha$  with the requirement

$$P\left(\frac{s_1^2}{s_2^2} > F'_\alpha\right) = \alpha . \quad (8.2.4)$$

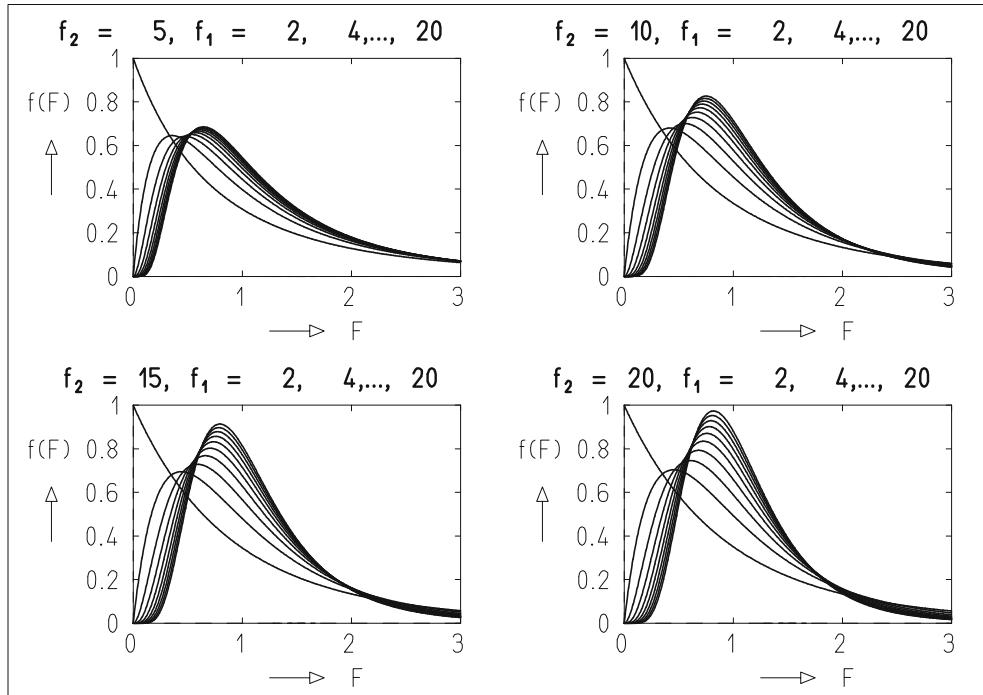
---

<sup>†</sup>This is also often called the  $v^2$ -distribution,  $\omega^2$ -distribution, or Snedecor distribution.

This expression means that the limit  $F'_\alpha$  is equal to the *quantile*  $F_{1-\alpha}$  of the  $F$ -distribution (see Sect. 3.3) since

$$P\left(\frac{s_1^2}{s_2^2} < F'_\alpha\right) = P\left(\frac{s_1^2}{s_2^2} < F_{1-\alpha}\right) = 1 - \alpha . \quad (8.2.5)$$

If this limit is exceeded, then we say that  $\sigma_1^2 > \sigma_2^2$  with the significance level  $\alpha$ . The quantiles  $F_{1-\alpha}$  for various pairs of values  $(f_1, f_2)$  are given in Table I.8. In general one applies a two-sided test, i.e., one tests whether the ratio  $F$  is between two limits  $F''_\alpha$  and  $F'''_\alpha$ , which are determined by



**Fig. 8.2:** Probability density of the  $F$ -distribution for fixed values of  $f_2 = 5, 10, 15, 20$ . For  $f_1 = 2$  one has  $f(F) = e^{-F}$ . For  $f_1 > 2$  the function has a maximum which increases for increasing  $f_1$ .

$$P\left(\frac{s_1^2}{s_2^2} > F''_\alpha\right) = \frac{1}{2}\alpha, \quad P\left(\frac{s_1^2}{s_2^2} < F'''_\alpha\right) = \frac{1}{2}\alpha . \quad (8.2.6)$$

Because of the definition of  $F$  as a ratio, the inequality

$$s_1^2/s_2^2 < F'''_\alpha(f_1, f_2)$$

clearly has the same meaning as

$$\mathbf{s}_2^2 / \mathbf{s}_1^2 > F''_\alpha(f_2, f_1) \quad .$$

Here the first argument gives the number of degrees of freedom in the numerator, and the second in the denominator. The requirement (8.2.6) can also be written:

$$P\left(\frac{\mathbf{s}_1^2}{\mathbf{s}_2^2} > F''_\alpha(f_1, f_2)\right) = \frac{1}{2}\alpha, \quad P\left(\frac{\mathbf{s}_2^2}{\mathbf{s}_1^2} > F''_\alpha(f_2, f_1)\right) = \frac{1}{2}\alpha \quad . \quad (8.2.7)$$

Table I.8 can therefore be used for the one-sided as well as the two-sided  $F$ -test.

A glance at Table I.8 also shows that  $F_{1-\alpha/2} > 1$  for all reasonable values of  $\alpha$ . Therefore one needs only to find the limit for the ratio

$$\mathbf{s}_g^2 / \mathbf{s}_k^2 > F_{1-\frac{1}{2}\alpha}(f_g, f_k) \quad . \quad (8.2.8)$$

Here the indices  $g$  and  $k$  give the larger and smaller values of the two variances, i.e.,  $s_g^2 > s_k^2$ . If the inequality (8.2.8) is satisfied, then the hypothesis of equal variances must be rejected.

**Example 8.1:**  $F$ -test of the hypothesis of equal variance of two series of measurements

A standard length ( $100 \mu\text{m}$ ) is measured using two traveling microscopes. The measurements and computations are summarized in Table 8.1. From Table I.8 we find for the two-sided  $F$ -test with a significance level of 10%,

$$F''_{0.1}(6, 9) = F_{0.95}(6, 9) = 3.37 \quad .$$

The hypothesis of equal variances cannot be rejected. ■

### 8.3 Student's Test: Comparison of Means

We now consider a population that follows a standard Gaussian distribution. Let  $\bar{x}$  be the arithmetic mean of a sample of size  $N$ . According to (6.2.3) the variance of  $\bar{x}$  is related to the population variance by

$$\sigma^2(\bar{x}) = \sigma^2(x)/N \quad . \quad (8.3.1)$$

If  $N$  is sufficiently large, then from the Central Limit Theorem,  $\bar{x}$  will be normally distributed with mean  $\hat{x}$  and variance  $\sigma^2(\bar{x})$ . That is,

$$y = (\bar{x} - \hat{x})/\sigma(\bar{x}) \quad (8.3.2)$$

will be described by a standard normal distribution. The quantity  $\sigma(x)$  is, however, not known. We only know the estimate for  $\sigma^2(x)$ ,

$$s_x^2 = \frac{1}{N-1} \sum_{j=1}^N (x_j - \bar{x})^2 \quad . \quad (8.3.3)$$

Then with (8.3.1) we can also estimate  $\sigma^2(\bar{x})$  to be

$$s_{\bar{x}}^2 = \frac{1}{N(N-1)} \sum_{j=1}^N (x_j - \bar{x})^2 \quad . \quad (8.3.4)$$

We now ask to what extent (8.3.2) differs from the standard Gaussian distribution if  $\sigma(\bar{x})$  is replaced by  $s_{\bar{x}}$ . By means of a simple translation of coordinates we can always have  $\hat{x} = 0$ . We therefore only consider the distribution of

$$t = \bar{x}/s_{\bar{x}} = \bar{x}\sqrt{N}/s_x \quad . \quad (8.3.5)$$

Since  $(N-1)s_x^2 = f s_x^2$  follows a  $\chi^2$ -distribution with  $f = N-1$  degrees of freedom, we can write

**Table 8.1:** *F*-test on the equality of variances. Data from Example 8.1.

| Measurement number  | Measurement with     |                      |
|---------------------|----------------------|----------------------|
|                     | Instrument 1<br>[μm] | Instrument 2<br>[μm] |
| 1                   | 100                  | 97                   |
| 2                   | 101                  | 102                  |
| 3                   | 103                  | 103                  |
| 4                   | 98                   | 96                   |
| 5                   | 97                   | 100                  |
| 6                   | 98                   | 101                  |
| 7                   | 102                  | 100                  |
| 8                   | 101                  |                      |
| 9                   | 99                   |                      |
| 10                  | 101                  |                      |
| Mean                | 100                  | 99.8                 |
| Degrees of freedom  | 9                    | 6                    |
| $s^2$               | $34/9 = 3.7$         | $39/6 = 6.5$         |
| $F = 6.5/3.7 = 1.8$ |                      |                      |

$$t = \bar{x}\sqrt{N}\sqrt{f}/\chi \quad .$$

The distribution function of  $t$  is given by

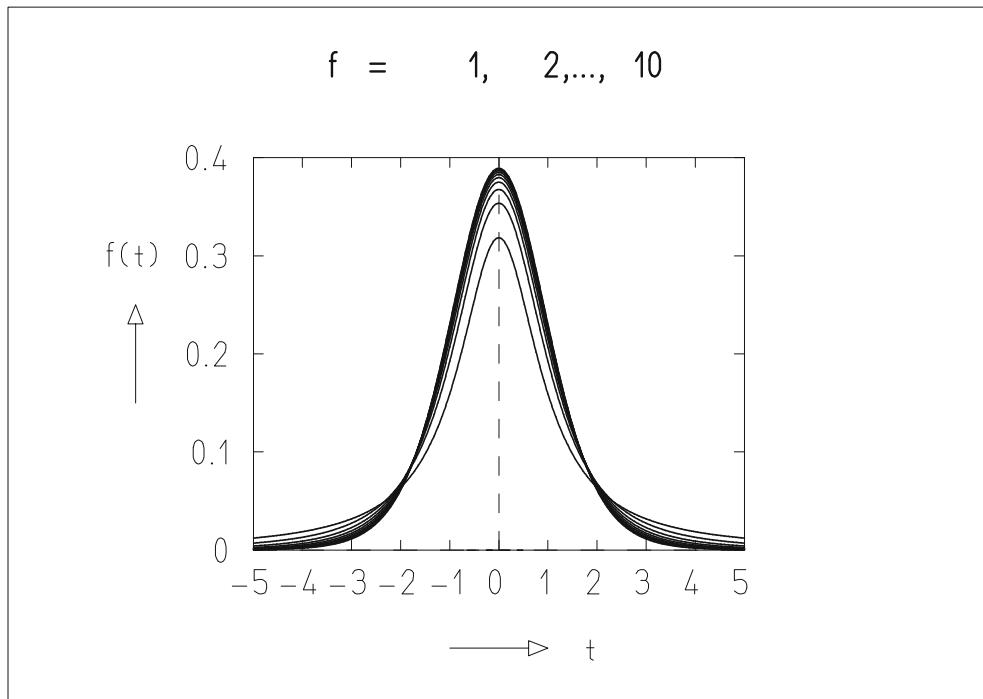
$$F(t) = P(t < t) = P\left(\frac{\bar{x}\sqrt{N}\sqrt{f}}{\chi} < t\right) \quad . \quad (8.3.6)$$

After a somewhat lengthy calculation one finds

$$F(t) = \frac{\Gamma(\frac{1}{2}(f+1))}{\Gamma(\frac{1}{2}f)\sqrt{\pi}\sqrt{f}} \int_{-\infty}^t \left(1 + \frac{t^2}{f}\right)^{-\frac{1}{2}(f+1)} dt \quad .$$

The corresponding probability density is

$$f(t) = \frac{\Gamma(\frac{1}{2}(f+1))}{\Gamma(\frac{1}{2}f)\sqrt{\pi}\sqrt{f}} \left(1 + \frac{t^2}{f}\right)^{-\frac{1}{2}(f+1)} \quad . \quad (8.3.7)$$



**Fig. 8.3:** Student's distribution  $f(t)$  for  $f = 1, 2, \dots, 10$  degrees of freedom. For  $f = 1$  the maximum is lowest and the tails are especially prominent.

Figure 8.3 shows the function  $f(t)$  for various degrees of freedom  $f = N - 1$ . A comparison with Fig. 5.7 shows that for  $f \rightarrow \infty$ , the distribution (8.3.7) becomes the standard normal distribution  $\phi_0(t)$ , as expected.

Like  $\phi_0(t)$ ,  $f(t)$  is symmetric about 0 and has a bell shape. Corresponding to (5.8.3) one has

$$P(|t| \leq t) = 2F(|t|) - 1 \quad . \quad (8.3.8)$$

By requiring

$$\int_0^{t'_\alpha} f(t) dt = \frac{1}{2}(1 - \alpha) \quad (8.3.9)$$

we can again determine limits  $\pm t'_\alpha$  at a given significance level  $\alpha$ , where

$$t'_\alpha = t_{1-\frac{1}{2}\alpha} \quad .$$

The quantiles  $t_{1-\frac{1}{2}\alpha}$  are given in Table I.9 for various values of  $\alpha$  and  $f$ .

The application of Student's test<sup>‡</sup> can be described in the following way: One has a hypothesis  $\lambda_0$  for the population mean of a normal distribution. A sample of size  $N$  yields the sample mean  $\bar{x}$  and sample variance  $s_x^2$ . If the inequality

$$|t| = \frac{|\bar{x} - \lambda_0| \sqrt{N}}{s_x} > t'_\alpha = t_{1-\frac{1}{2}\alpha} \quad (8.3.10)$$

is fulfilled for a given significance level  $\alpha$ , then the hypothesis must be rejected.

This is clearly a two-sided test. If deviations only in one direction are important, then the corresponding test at the significance level  $\alpha$  is

$$t = \frac{(\bar{x} \pm \lambda_0) \sqrt{N}}{s_x} > t'_{2\alpha} = t_{1-\alpha} \quad . \quad (8.3.11)$$

We can make the test more general and apply it to the problem of comparing two mean values. Suppose samples of size  $N_1$  and  $N_2$  have been taken from two populations  $X$  and  $Y$ . We wish to find a measure of correctness for the hypothesis that the expectation values are equal,

$$\hat{x} = \hat{y} \quad .$$

Because of the Central Limit Theorem, the mean values are almost normally distributed. Their variances are

$$\sigma^2(\bar{x}) = \frac{1}{N_1} \sigma^2(x), \quad \sigma^2(\bar{y}) = \frac{1}{N_2} \sigma^2(y) \quad (8.3.12)$$

and the estimates for these quantities are

---

<sup>‡</sup>The  $t$ -distribution was introduced by W. S. Gosset and published under the pseudonym "Student".

$$\begin{aligned}s_x^2 &= \frac{1}{N_1(N_1 - 1)} \sum_{j=1}^{N_1} (\mathbf{x}_j - \bar{\mathbf{x}})^2 , \\ s_y^2 &= \frac{1}{N_2(N_2 - 1)} \sum_{j=1}^{N_2} (\mathbf{y}_j - \bar{\mathbf{y}})^2 .\end{aligned}\quad (8.3.13)$$

According to the discussion in Example 5.10, the difference

$$\Delta = \bar{\mathbf{x}} - \bar{\mathbf{y}} \quad (8.3.14)$$

also has an approximate normal distribution with

$$\sigma^2(\Delta) = \sigma^2(\bar{\mathbf{x}}) + \sigma^2(\bar{\mathbf{y}}) . \quad (8.3.15)$$

If the hypothesis of equal means is true, i.e.,  $\hat{\Delta} = 0$ , then the ratio

$$\Delta / \sigma(\Delta) \quad (8.3.16)$$

follows the standard normal distribution. If  $\sigma(\Delta)$  were known one could immediately give the probability according to (5.8.2) for the hypothesis to be fulfilled. But only  $s_\Delta$  is known. The corresponding ratio

$$\Delta / s_\Delta \quad (8.3.17)$$

will in general be somewhat larger.

Usually the hypothesis  $\hat{\mathbf{x}} = \hat{\mathbf{y}}$  implies that  $\bar{\mathbf{x}}$  and  $\bar{\mathbf{y}}$  come from the same population. Then  $\sigma^2(\mathbf{x})$  and  $\sigma^2(\mathbf{y})$  are equal, and we can use the weighted mean of  $s_x^2$  and  $s_y^2$  as the corresponding estimator. The weights are given by  $(N_1 - 1)$  and  $(N_2 - 1)$ :

$$s^2 = \frac{(N_1 - 1)s_x^2 + (N_2 - 1)s_y^2}{(N_1 - 1) + (N_2 - 1)} . \quad (8.3.18)$$

From this we construct

$$s_{\bar{\mathbf{x}}}^2 = \frac{s^2}{N_1}, \quad s_{\bar{\mathbf{y}}}^2 = \frac{s^2}{N_2} ,$$

and

$$s_\Delta^2 = s_{\bar{\mathbf{x}}}^2 + s_{\bar{\mathbf{y}}}^2 = \frac{N_1 + N_2}{N_1 N_2} s^2 . \quad (8.3.19)$$

It can be shown (see [8]) that the ratio (8.3.17) follows the Student's *t*-distribution with  $f = N_1 + N_2 - 2$  degrees of freedom. With this one can now perform *Student's difference test*:

The quantity (8.3.17) is computed from the results of two samples. This value is compared to a quantile of Student's distribution with  $f = N_1 + N_2 - 2$  degrees of freedom with a significance level  $\alpha$ . If

$$|t| = \frac{|\Delta|}{s_\Delta} = \frac{|\bar{x} - \bar{y}|}{s_\Delta} \geq t'_\alpha = t_{1-\frac{1}{2}\alpha} \quad , \quad (8.3.20)$$

then the hypothesis of equal means must be rejected. Instead one would assume  $\hat{x} > \hat{y}$  or  $\hat{x} < \hat{y}$ , depending on whether one has  $\bar{x} > \bar{y}$  or  $\bar{x} < \bar{y}$ .

**Example 8.2:** Student's test of the hypothesis of equal means of two series of measurements

Column  $x$  of Table 8.2 contains measured values (in arbitrary units) of the concentration of neuraminic acid in the red blood cells of patients suffering from a certain blood disease. Column  $y$  gives the measured values for a group of healthy persons. From the mean values and variances of the two samples one finds

$$\begin{aligned} |\Delta| &= |\bar{x} - \bar{y}| = 1.3 \quad , \\ s^2 &= \frac{15s_x^2 + 6s_y^2}{21} = 9.15 \quad , \\ s_\Delta^2 &= \frac{23}{112}s^2 = 1.88 \quad . \end{aligned}$$

For  $\alpha = 5\%$  and  $f = 21$  we find  $t_{1-\alpha/2} = 2.08$ . We must therefore conclude that the experimental data is not sufficient to determine an influence of the disease on the concentration. ■

## 8.4 Concepts of the General Theory of Tests

The test procedures discussed so far have been obtained more or less intuitively and without rigorous justification. In particular we have not given any specific reasons for the choice of the critical region. We now want to deal with the theory of statistical tests in a somewhat more critical way. A complete treatment of this topic would, however, go beyond the scope of this book.

Each sample of size  $N$  can be characterized by  $N$  points in the sample space of Sect. 2.1. For simplicity we will limit ourselves to a continuous random variable  $x$ , so that the sample can be described by  $N$  points  $(x^{(1)}, x^{(2)}, \dots, x^{(N)})$  on the  $x$  axis. In the case of  $r$  random variables we would have  $N$  points in an  $r$ -dimensional space. The result of such a sample, however, could also be specified by a single point in a space of dimension  $rN$ . A sample of size 2 with a single variable could, for example, be depicted as a point in

**Table 8.2:** Student's difference test on the equality of means. Data from Example 8.2.

| $x$                | $y$              |
|--------------------|------------------|
| 21                 | 16               |
| 24                 | 20               |
| 18                 | 22               |
| 19                 | 19               |
| 25                 | 18               |
| 17                 | 19               |
| 18                 | 19               |
| 22                 |                  |
| 21                 |                  |
| 23                 |                  |
| 18                 |                  |
| 13                 |                  |
| 16                 |                  |
| 23                 |                  |
| 22                 |                  |
| 24                 |                  |
| $N_1 = 16$         | $N_2 = 7$        |
| $\bar{x} = 20.3$   | $\bar{y} = 19.0$ |
| $s_x^2 = 171.8/15$ | $s_y^2 = 20/6$   |

a two-dimensional plane, spanned by the axes  $x^{(1)}, x^{(2)}$ . We will call such a space the  $E$  space. Every *hypothesis*  $H$  consists of an assumption about the probability density

$$f(x; \lambda_1, \lambda_2, \dots, \lambda_p) = f(x; \boldsymbol{\lambda}) \quad . \quad (8.4.1)$$

The hypothesis is said to be *simple* if the function  $f$  is completely specified, i.e., if the hypothesis gives the values of all of the parameters  $\lambda_i$ . It is said to be *composite* if the general mathematical form of  $f$  is known, but the exact value of at least one parameter remains undetermined. A simple hypothesis could, for example, specify a standard Gaussian distribution. A Gaussian distribution with a mean of zero but an undetermined variance, however, is a composite hypothesis. The hypothesis  $H_0$  is called the *null hypothesis*. Sometimes we will write explicitly

$$H_0(\boldsymbol{\lambda} = \boldsymbol{\lambda}_0) = H_0(\lambda_1 = \lambda_{10}, \lambda_2 = \lambda_{20}, \dots, \lambda_p = \lambda_{p0}) \quad . \quad (8.4.2)$$

Other possible hypotheses are called *alternative hypotheses*, e.g.,

$$H_1(\boldsymbol{\lambda} = \boldsymbol{\lambda}_1) = H_1(\lambda_1 = \lambda_{11}, \lambda_2 = \lambda_{21}, \dots, \lambda_p = \lambda_{p1}) \quad . \quad (8.4.3)$$

Often one wants to test a null hypothesis of the type (8.4.2) against a composite alternative hypothesis

$$H_1(\lambda \neq \lambda_0) = H_1(\lambda_1 \neq \lambda_{10}, \lambda_2 \neq \lambda_{20}, \dots, \lambda_p \neq \lambda_{p0}) \quad . \quad (8.4.4)$$

Since the null hypothesis makes a statement about the probability density in the sample space, it also predicts the probability for observing a point  $X = (\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(N)})$  in the  $E$  space.<sup>§</sup> We now define a *critical region*  $S_c$  with the significance level  $\alpha$  by the requirement

$$P(X \in S_c | H_0) = \alpha \quad , \quad (8.4.5)$$

i.e., we determine  $S_c$  such that the probability to observe a point  $X$  within  $S_c$  is  $\alpha$ , under the assumption that  $H_0$  is true. If the point  $X$  from the sample actually falls into the region  $S_c$ , then the hypothesis  $H_0$  is rejected. One must note that the requirement (8.4.5) does not necessarily determine the critical region  $S_c$  uniquely.

Although using the  $E$  space is conceptually elegant, it is usually not very convenient for carrying out tests. Instead one constructs a *test statistic*

$$T = T(X) = T(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(N)}) \quad (8.4.6)$$

and determines a region  $U$  of the variable  $T$  such that it corresponds to the critical region  $S_c$ , i.e., one performs the mapping

$$X \rightarrow T(X), \quad S_c(X) \rightarrow U(X) \quad . \quad (8.4.7)$$

The null hypothesis is rejected if  $T \in U$ .

Because of the statistical nature of the sample, it is clearly possible that the null hypothesis could be true, even though it was rejected since  $X \in S_c$ . The probability for such an error, an *error of the first kind*, is equal to  $\alpha$ . There is in addition another possibility to make a wrong decision, if one does not reject the hypothesis  $H_0$  because  $X$  was not in the critical region  $S_c$ , even though the hypothesis was actually false and an alternative hypothesis was true. This is an *error of the second kind*. The probability for this,

$$P(X \notin S_c | H_1) = \beta \quad , \quad (8.4.8)$$

depends of course on the particular alternative hypotheses  $H_1$ . This connection provides us with a method to specify the critical region  $S_c$ . A test is clearly most reasonable if for a given significance level  $\alpha$  the critical region is chosen such that the probability  $\beta$  for an error of the second kind is a

---

<sup>§</sup>Although  $X$  and the function  $T(x)$  introduced below are random variables, we do not use for them a special character type.

minimum. The critical region and therefore the test itself naturally depend on the alternative hypothesis under consideration.

Once the critical region has been determined, we can consider the probability for rejecting the null hypothesis as a function of the “true” hypothesis, or rather as a function of the parameters that describe it. In analogy to (8.4.5), this is

$$M(S_c, \lambda) = P(X \in S_c | H) = P(X \in S_c | \lambda) . \quad (8.4.9)$$

This probability is a function of  $S_c$  and of the parameters  $\lambda$ . It is called the *power function* of a test. The complementary probability

$$L(S_c, \lambda) = 1 - M(S_c, \lambda) \quad (8.4.10)$$

is called the *acceptance probability* or the *operating characteristic function* of the test. It gives the probability to accept <sup>¶</sup> the null hypothesis. One clearly has

$$\begin{aligned} M(S_c, \lambda_0) &= \alpha, & M(S_c, \lambda_1) &= 1 - \beta, \\ L(S_c, \lambda_0) &= 1 - \alpha, & L(S_c, \lambda_1) &= \beta . \end{aligned} \quad (8.4.11)$$

The *most powerful test* of a simple hypothesis  $H_0$  with respect to the simple alternative hypothesis is defined by the requirement

$$M(S_c, \lambda_1) = 1 - \beta = \max . \quad (8.4.12)$$

Sometimes there exists a *uniformly most powerful test*, for which the requirement (8.4.12) holds for all possible alternative hypotheses.

A test is said to be *unbiased* if its power function is greater than or equal to  $\alpha$  for all alternative hypotheses:

$$M(S_c, \lambda_1) \geq \alpha . \quad (8.4.13)$$

This definition is reasonable, since the probability to reject the null hypothesis is then a minimum if the null hypothesis is true. An *unbiased most powerful test* is the most powerful of all the unbiased tests. Correspondingly one can define a *unbiased uniformly most powerful test*. In the next sections we will learn the rules which sometimes allow one to construct tests with such desirable properties. Before turning to this task, we will first give an example to illustrate the definitions just introduced.

---

<sup>¶</sup>We use here the word “acceptance” of a hypothesis, although more precisely one should say, “There is no evidence to reject the hypothesis.”

**Example 8.3:** Test of the hypothesis that a normal distribution with given variance  $\sigma^2$  has the mean  $\lambda = \lambda_0$

We wish to test the hypothesis  $H_0(\lambda = \lambda_0)$ . As a test statistic we use the arithmetic mean  $\bar{x} = \frac{1}{n}(x_1 + x_2 + \dots + x_n)$ . (We will see in Example 8.4 that this is the most appropriate test statistic for our purposes.) From Sect. 6.2 we know that  $\bar{x}$  is normally distributed with mean  $\lambda$  and variance  $\sigma^2/n$ , i.e., that the probability density for  $\bar{x}$  for the case  $\lambda = \lambda_0$  is given by

$$f(\bar{x}; \lambda_0) = \frac{\sqrt{n}}{\sqrt{2\pi}\sigma} \exp\left(-\frac{n}{2\sigma^2}(\bar{x} - \lambda_0)^2\right) . \quad (8.4.14)$$

This is shown in Fig. 8.4 together with four different critical regions, all of which have the same significance level  $\alpha$ .

These are the regions

$$\begin{aligned} U_1 : \bar{x} < \lambda^I &\text{ and } \bar{x} > \lambda^{II} & \text{with } \int_{-\infty}^{\lambda^I} f(\bar{x}) d\bar{x} = \int_{\lambda^{II}}^{\infty} f(\bar{x}) d\bar{x} = \frac{1}{2}\alpha , \\ U_2 : \bar{x} > \lambda^{III} & & \text{with } \int_{\lambda^{III}}^{\infty} f(\bar{x}) d\bar{x} = \alpha , \\ U_3 : \bar{x} < \lambda^{IV} & & \text{with } \int_{-\infty}^{\lambda^{IV}} f(\bar{x}) d\bar{x} = \alpha , \\ U_4 : \lambda^V \leq \bar{x} < \lambda^{VI} & & \text{with } \int_{\lambda^V}^{\lambda^VI} f(\bar{x}) d\bar{x} = \int_{\lambda_0}^{\lambda^VI} f(\bar{x}) d\bar{x} = \frac{1}{2}\alpha . \end{aligned}$$

In order to obtain the power functions for each of these regions, we must vary the mean value  $\lambda$ . The probability density of  $\bar{x}$  for an arbitrary value of  $\lambda$  is, in analogy to (8.4.14), given by

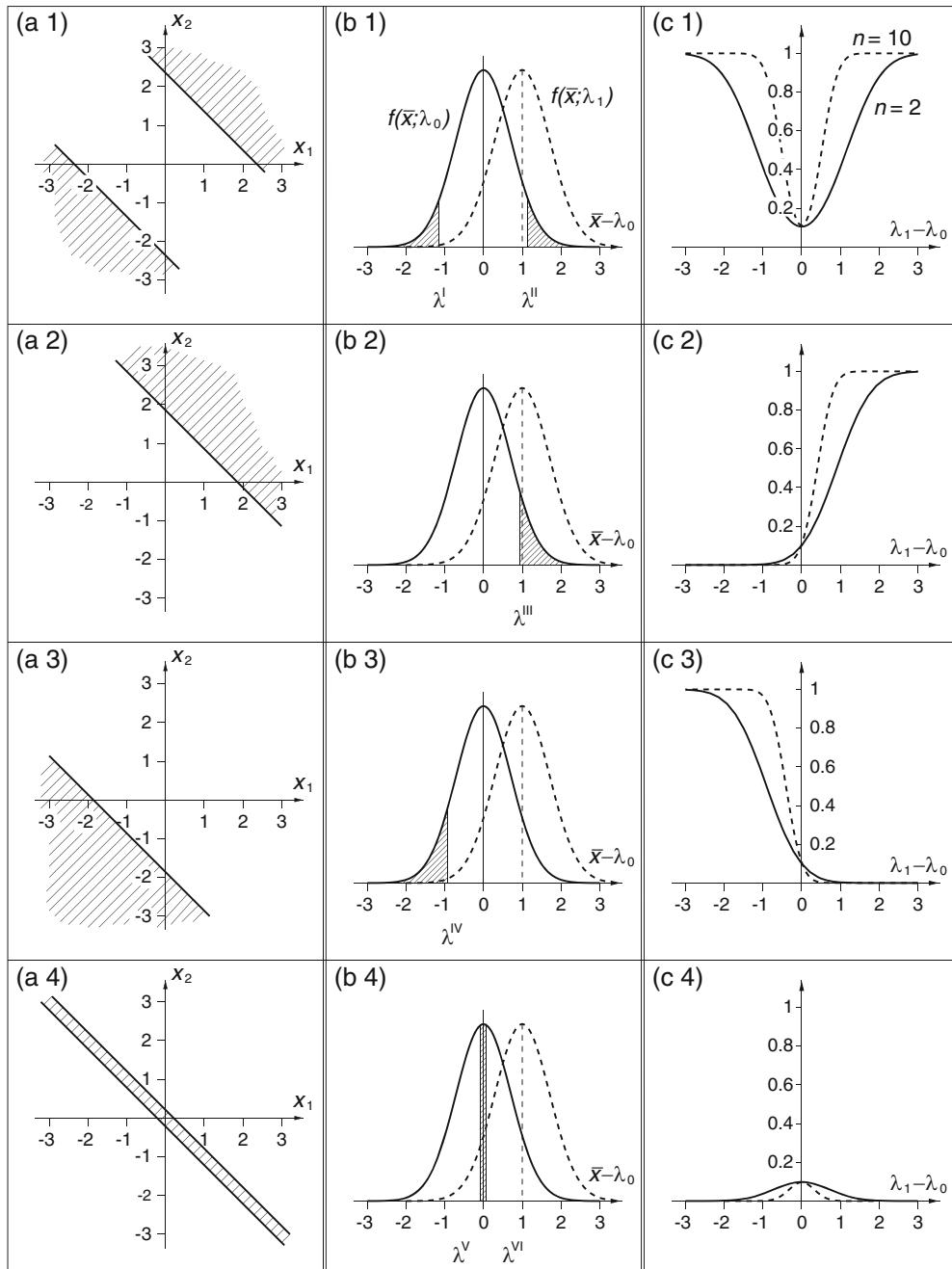
$$f(\bar{x}; \lambda) = \frac{\sqrt{n}}{\sqrt{2\pi}\sigma} \exp\left[-\frac{n}{2\sigma^2}(\bar{x} - \lambda)^2\right] . \quad (8.4.15)$$

The dashed curve in Fig. 8.4b represents the probability density for  $\lambda = \lambda_1 = \lambda_0 + 1$ . The power function (8.4.9) is now simply

$$P(\bar{x} \in U | \lambda) = \int_U f(\bar{x}; \lambda) d\bar{x} . \quad (8.4.16)$$

The power functions obtained in this way for the critical regions  $U_1, U_2, U_3, U_4$  are shown in Fig. 8.4c for  $n = 2$  (solid curve) and  $n = 10$  (dashed curve).

We can now compare the effects of the four tests corresponding to the various critical regions. From Fig. 8.4c we immediately see that  $U_1$  corresponds to an unbiased test, since the requirement (8.4.13) is clearly fulfilled. On the other hand, the test with the critical region  $U_2$  is more powerful for the alternative hypothesis  $H_1(\lambda_1 > \lambda_0)$ , but is not good for  $H_1(\lambda_1 < \lambda_0)$ . For the test with  $U_3$ , the situation is exactly the opposite. Finally, the region  $U_4$  provides a test for which the rejection probability is a maximum if the null hypothesis is true. Clearly this is very undesirable. The test was only constructed for demonstration purposes. If we compare the first three tests, we see that none of them are more powerful than the other two for all values of  $\lambda_1$ . Thus



**Fig. 8.4:** (a) Critical regions in  $E$  space, (b) critical region of the test function, and (c) power function of the test from Example 8.3.

we have not succeeded in finding a uniformly most powerful test. In Example 8.4, where we will continue the discussion of the present example, we will determine that for this problem there does not exist a uniformly most powerful test. ■

## 8.5 The Neyman–Pearson Lemma and Applications

In the last section we introduced the  $E$  space, in which a sample is represented by a single point  $X$ . The probability to observe a point  $X$  within the critical region  $S_c$  – providing the null hypothesis  $H_0$  is true – was defined in (8.4.5),

$$P(X \in S_c | H_0) = \alpha . \quad (8.5.1)$$

We now define a conditional probability in  $E$  space,

$$f(X | H_0).$$

One clearly has

$$\int_{S_c} f(X | H_0) dX = P(X \in S_c | H_0) = \alpha . \quad (8.5.2)$$

The NEYMAN–PEARSON lemma states the following:

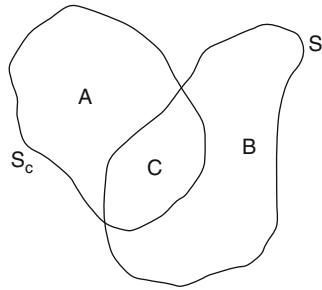
A test of the simple hypothesis  $H_0$  with respect to the simple alternative hypothesis  $H_1$  is a most powerful test if the critical region  $S_c$  in  $E$  space is chosen such that

$$\frac{f(X | H_0)}{f(X | H_1)} \begin{cases} \leq c & \text{for all } X \in S_c \\ \geq c & \text{for all } X \notin S_c \end{cases} . \quad (8.5.3)$$

Here  $c$  is a constant which depends on the significance level.

We will prove this by considering another region  $S$  along with  $S_c$ . It may partially overlap with  $S_c$ , as sketched in Fig. 8.5. We choose the size of the region  $S$  such that it corresponds to the same significance level, i.e.,

$$\int_S f(X | H_0) dX = \int_{S_c} f(X | H_0) dX = \alpha .$$

**Fig. 8.5:** The regions  $S$  and  $S_c$ .

Using the notation of Fig. 8.5, we can write

$$\begin{aligned}\int_A f(X|H_0) dX &= \int_{S_c} f(X|H_0) dX - \int_C f(X|H_0) dX \\ &= \int_S f(X|H_0) dX - \int_C f(X|H_0) dX \\ &= \int_B f(X|H_0) dX .\end{aligned}$$

Since  $A$  is contained in  $S_c$ , we can use (8.5.3), i.e.,

$$\int_A f(X|H_0) dX \leq c \int_A f(X|H_1) dX .$$

Correspondingly, since  $B$  is outside of  $S_c$ , one has

$$\int_B f(X|H_0) dX \geq c \int_B f(X|H_1) dX .$$

We can now express the power function (8.4.9) using these integrals:

$$\begin{aligned}M(S_c, \lambda_1) &= \int_{S_c} f(X|H_1) dX = \int_A f(X|H_1) dX + \int_C f(X|H_1) dX \\ &\geq \frac{1}{c} \int_A f(X|H_0) dX + \int_C f(X|H_1) dX \\ &\geq \int_B f(X|H_1) dX + \int_C f(X|H_1) dX \\ &\geq \int_S f(X|H_1) dX = M(S, \lambda_1)\end{aligned}$$

or directly,

$$M(S_c, \lambda_1) \geq M(S, \lambda_1) . \quad (8.5.4)$$

This is exactly the condition (8.4.12) for a uniformly most powerful test. Since we have not made any assumptions about the alternative hypothesis  $H_1(\lambda = \lambda_1)$  or the region  $S$ , we have proven that the requirement (8.5.3) provides a uniformly most powerful test when it is fulfilled by the alternative hypothesis.

**Example 8.4:** Most powerful test for the problem of Example 8.3

We now continue with the ideas from Example 8.3, i.e., we consider tests with a sample of size  $N$ , obtained from a normal distribution with known variance  $\sigma^2$  and unknown mean  $\lambda$ . The conditional probability density of a point  $X = (\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(N)})$  in  $E$  space is the joint probability density of the  $\mathbf{x}^{(j)}$  for given values of  $\lambda$ , i.e.,

$$f(X|H_0) = \left( \frac{1}{\sqrt{2\pi}\sigma} \right)^N \exp \left[ -\frac{1}{2\sigma^2} \sum_{j=1}^N (\mathbf{x}^{(j)} - \lambda_0)^2 \right] \quad (8.5.5)$$

for the null hypothesis and

$$f(X|H_1) = \left( \frac{1}{\sqrt{2\pi}\sigma} \right)^N \exp \left[ -\frac{1}{2\sigma^2} \sum_{j=1}^N (\mathbf{x}^{(j)} - \lambda_1)^2 \right] \quad (8.5.6)$$

for the alternative hypothesis. The ratio (8.5.3) takes on the form

$$\begin{aligned} Q = \frac{f(X|H_0)}{f(X|H_1)} &= \exp \left[ -\frac{1}{2\sigma^2} \left\{ \sum_{j=1}^N (\mathbf{x}^{(j)} - \lambda_0)^2 - \sum_{j=1}^N (\mathbf{x}^{(j)} - \lambda_1)^2 \right\} \right] \\ &= \exp \left[ -\frac{1}{2\sigma^2} \left\{ N(\lambda_0^2 - \lambda_1^2) - 2(\lambda_0 - \lambda_1) \sum_{j=1}^N \mathbf{x}^{(j)} \right\} \right]. \end{aligned}$$

The expression

$$\exp \left[ -\frac{N}{2\sigma^2} (\lambda_0^2 - \lambda_1^2) \right] = k \geq 0$$

is a non-negative constant. The condition (8.5.3) thus has the form

$$k \exp \left[ \frac{\lambda_0 - \lambda_1}{\sigma^2} \sum_{j=1}^N \mathbf{x}^{(j)} \right] \left\{ \begin{array}{ll} \leq c, & X \in S_c \\ \geq c, & X \notin S_c \end{array} \right. .$$

This is the same as

$$(\lambda_0 - \lambda_1) \bar{\mathbf{x}} \left\{ \begin{array}{ll} \leq c', & X \in S_c \\ \geq c', & X \notin S_c \end{array} \right. . \quad (8.5.7)$$

Here  $c'$  is a constant different from  $c$ . Equation (8.5.7) is, however, not only a condition for  $S_c$ , but also specifies directly that  $\bar{\mathbf{x}}$  should be used as the test variable. For each given  $\lambda_1$ , i.e., for every simple alternative hypothesis  $H_1(\lambda = \lambda_1)$ , (8.5.7) gives a clear prescription for the choice of  $S_c$  or  $U$ , i.e., for the critical region and the test variable  $\bar{\mathbf{x}}$ .

For the case  $\lambda_1 < \lambda_0$ , the relation (8.5.7) becomes

$$\bar{x} \begin{cases} \leq c'', & X \in S_c \\ \geq c'', & X \notin S_c \end{cases},$$

This corresponds to the situation in Fig. 8.4b3 with  $c'' = \lambda^{IV}$ . Similarly, for every alternative hypothesis with  $\lambda_1 > \lambda_0$ , the critical region of the most powerful test is given by

$$\bar{x} \geq c'''$$

(see Fig. 8.4b2 with  $c''' = \lambda'''$ ). There does not exist a uniformly most powerful test, since the factor  $(\lambda_0 - \lambda_1)$  in Eq. (8.5.7) changes sign at  $\lambda_1 = \lambda_0$ . ■

## 8.6 The Likelihood-Ratio Method

The Neyman–Pearson lemma gave the condition for a uniformly most powerful test. Such a test did not exist if the alternative hypothesis included parameter values that could be both greater and less than that of the null hypothesis. We determined this in Example 8.4; it can be shown, however, that it is true in general. The question thus arises as to what test should be used when no uniformly most powerful test exists. Clearly this question is not formulated precisely enough to allow a unique answer. We would like in the following to give a prescription that allows us to construct tests that have desirable properties and that have the advantage of being relatively easy to use.

We consider from the beginning the general case with  $p$  parameters  $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \dots, \lambda_p)$ . The result of a sample, i.e., the point  $X = (x^{(1)}, x^{(2)}, \dots, x^{(N)})$  in  $E$  space is to be used to test a given hypothesis. The (composite) null hypothesis is characterized by a given region for each parameter. We can use a  $p$ -dimensional space, with the  $\lambda_1, \lambda_2, \dots, \lambda_p$  as coordinate axes, and consider the region allowed by the null hypothesis as a region in this parameter space, called  $\omega$ . We denote the region in this space representing all possible values of the parameters by  $\Omega$ . The most general alternative hypothesis is then the part of  $\Omega$  that does not contain  $\omega$ . We denote this by  $\Omega - \omega$ . Recall now from Chap. 7 the maximum-likelihood estimator  $\tilde{\lambda}$  for a parameter  $\lambda$ . It is that value of  $\lambda$  for which the likelihood function is a maximum. In Chap. 7 we tacitly assumed that one searched for the maximum in the entire allowable parameter space. In the following we will consider maxima in a restricted region (e.g., in  $\omega$ ). We write in this case  $\tilde{\lambda}^{(\omega)}$ . The *likelihood-ratio test* defines a test statistic

$$T = \frac{f(x^{(1)}, x^{(2)}, \dots, x^{(N)}; \tilde{\lambda}_1^{(\Omega)}, \tilde{\lambda}_2^{(\Omega)}, \dots, \tilde{\lambda}_p^{(\Omega)})}{f(x^{(1)}, x^{(2)}, \dots, x^{(N)}; \tilde{\lambda}_1^{(\omega)}, \tilde{\lambda}_2^{(\omega)}, \dots, \tilde{\lambda}_p^{(\omega)})} . \quad (8.6.1)$$

Here  $f(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(N)}; \lambda_1, \lambda_2, \dots, \lambda_p)$  is the joint probability density of the  $\mathbf{x}^{(j)}$  ( $j = 1, 2, \dots, N$ ), i.e., the likelihood function (7.1.5). The procedure of the likelihood ratio test prescribes that we reject the null hypothesis if

$$T > T_{1-\alpha} . \quad (8.6.2)$$

Here  $T_{1-\alpha}$  is defined by

$$P(T > T_{1-\alpha} | H_0) = \int_{T_{1-\alpha}}^{\infty} g(T | H_0) dT , \quad (8.6.3)$$

and  $g(T | H_0)$  is the conditional probability density for the test statistic  $T$ . The following theorem by WILKS [9] concerns the distribution function of  $T$  (or actually  $-2 \ln T$ ) in the limiting case of very large samples:

If a population is described by the probability density  $f(x; \lambda_1, \lambda_2, \dots, \lambda_p)$  that satisfies reasonable requirements of continuity, and if  $p - r$  of the  $p$  parameters are fixed by the null hypothesis, while  $r$  parameters remain free, then the statistic  $-2 \ln T$  follows a  $\chi^2$ -distribution with  $p - r$  degrees of freedom for very large samples, i.e., for  $N \rightarrow \infty$ .

We now apply this method to the problem of Examples 8.3 and 8.4, i.e., we consider tests with samples from a normally distributed population with known variance and unknown mean.

### **Example 8.5:** Power function for the test from Example 8.3

For the simple hypothesis  $H_0(\lambda = \lambda_0)$ , the region  $\omega$  of the parameter space is reduced to the point  $\lambda = \lambda_0$ . We have thus

$$\tilde{\lambda}^{(\omega)} = \lambda_0 . \quad (8.6.4)$$

If we consider the most general alternative hypothesis  $H_1(\lambda = \lambda_1 \neq \lambda_0)$ , then we obtain as the maximum-likelihood estimator of  $\lambda$  the sample mean  $\bar{x}$ . The likelihood ratio (8.6.1) thus becomes

$$T = \frac{f(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(N)}; \bar{x})}{f(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(N)}; \lambda_0)} . \quad (8.6.5)$$

The joint probability density is given by (7.2.6),

$$f(x^{(1)}, x^{(2)}, \dots, x^{(N)}) = \left( \frac{1}{\sqrt{2\pi}\sigma} \right)^N \exp \left[ -\frac{1}{2\sigma^2} \sum_{j=1}^N (x^{(j)} - \lambda)^2 \right] . \quad (8.6.6)$$

Therefore,

$$\begin{aligned}
T &= \exp \left[ \frac{1}{2\sigma^2} \left\{ - \sum_{j=1}^N (\mathbf{x}^{(j)} - \bar{\mathbf{x}})^2 + \sum_{j=1}^N (\mathbf{x}^{(j)} - \lambda_0)^2 \right\} \right] \\
&= \exp \left[ \frac{1}{2\sigma^2} \sum_{j=1}^N (\bar{\mathbf{x}} - \lambda_0)^2 \right] \\
&= \exp \left[ \frac{N}{2\sigma^2} (\bar{\mathbf{x}} - \lambda_0)^2 \right] \quad .
\end{aligned}$$

We must now calculate  $T_{1-\alpha}$  and reject the hypothesis  $H_0$  if the inequality (8.6.2) is fulfilled. Since the logarithm of  $T$  is a monotonic function of  $T$ , we can use

$$T' = 2 \ln T = \frac{N}{\sigma^2} (\bar{\mathbf{x}} - \lambda_0)^2 \quad (8.6.7)$$

as the test statistic and reject  $H_0$  if

$$T' > T'_{1-\alpha}$$

with

$$\int_{T'_{1-\alpha}}^{\infty} h(T'|H_0) dT' = \alpha \quad .$$

In order to calculate the probability density  $h(T'|H_0)$  of  $T'$ , we start with the density  $f(\bar{x})$  for the sample mean with the condition  $\lambda = \lambda_0$ ,

$$f(\bar{x}|H_0) = \sqrt{\frac{N}{2\pi\sigma^2}} \exp\left(-\frac{N}{2\sigma^2}(\bar{x} - \lambda_0)^2\right) \quad .$$

In order to carry out the transformation of variables (3.7.1), we need in addition the derivative,

$$\left| \frac{d\bar{x}}{dT'} \right| = \frac{1}{2} \sqrt{\frac{\sigma^2}{N}} T'^{-1/2} \quad ,$$

which can be easily obtained from (8.6.7). One then has

$$h(T'|H_0) = \left| \frac{d\bar{x}}{dT'} \right| f(\bar{x}|H_0) = \frac{1}{\sqrt{2\pi}} T'^{-1/2} e^{-T'/2} \quad . \quad (8.6.8)$$

This is indeed a  $\chi^2$ -distribution for one degree of freedom. Thus in our example, WILKS' theorem holds even for finite  $N$ . We see, therefore, that the likelihood-ratio test yields the unbiased test of Fig. 8.4b1. The test

$$T' = \frac{N}{\sigma^2} (\bar{x} - \lambda_0)^2 > T'_{1-\alpha}$$

is equivalent to

$$\left(\frac{N}{\sigma^2}\right)^{1/2} |\bar{x} - \lambda_0| < \lambda', \quad \left(\frac{N}{\sigma^2}\right)^{1/2} |\bar{x} - \lambda_0| > \lambda'' \quad (8.6.9)$$

with

$$-\lambda' = \lambda'' = (T'_{1-\alpha})^{1/2} = (\chi^2_{1-\alpha})^{1/2} = \chi_{1-\alpha} \quad .$$

We can use this result to compute explicitly the power function of our test. For a given value of the population mean  $\lambda$ , the probability density for the sample mean is

$$f(\bar{x}; \lambda) = \left(\frac{N}{2\pi\sigma^2}\right)^{1/2} \exp\left[-\frac{N(\bar{x} - \lambda)^2}{2\sigma^2}\right] = \phi_0\left(\frac{\bar{x} - \lambda}{\sigma/\sqrt{N}}\right) \quad .$$

Using (8.4.9) and (8.6.9) we obtain

$$\begin{aligned} M(S_c; \lambda) &= \int_{-\infty}^A f(\bar{x}; \lambda) d\bar{x} + \int_B^\infty f(\bar{x}; \lambda) d\bar{x} \quad (8.6.10) \\ &= \psi_0\left(\chi_{1-\alpha} - \frac{\lambda - \lambda_0}{\sigma/\sqrt{N}}\right) + \psi_0\left(\chi_{1-\alpha} + \frac{\lambda - \lambda_0}{\sigma/\sqrt{N}}\right) \quad , \\ A &= -\chi_{1-\alpha}\sigma/\sqrt{N} - \lambda_0, \quad B = \chi_{1-\alpha}\sigma/\sqrt{N} - \lambda_0 \quad . \end{aligned}$$

Here  $\phi_0$  and  $\psi_0$  are the probability density and distribution function of the standard normal distribution. The power function (8.6.10) is shown in Fig. 8.6 for  $\alpha = 0.05$  and various values of  $N/\sigma^2$ . ■

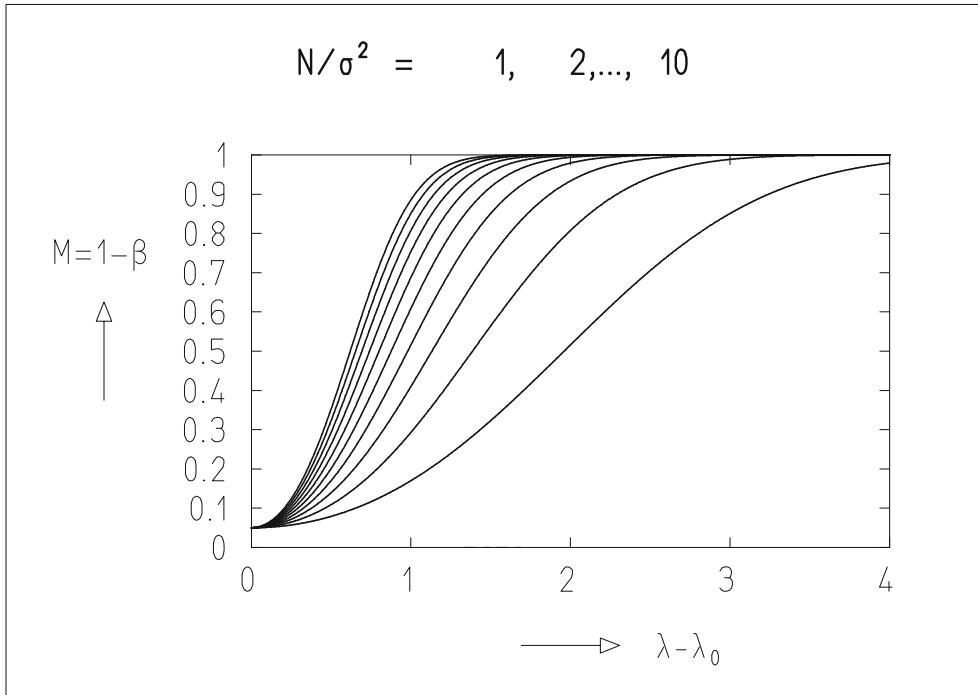
**Example 8.6:** Test of the hypothesis that a normal distribution of unknown variance has the mean value  $\lambda = \lambda_0$

In this case the null hypothesis  $H_0(\lambda = \lambda_0)$  is composite, i.e., it makes no statement about  $\sigma^2$ . From Example 7.8 we know the maximum-likelihood estimator in the full parameter space,

$$\begin{aligned} \tilde{\lambda}^{(\Omega)} &= \bar{x} \quad , \\ \tilde{\sigma}^{2(\Omega)} &= \frac{1}{N} \sum_{j=1}^N (\mathbf{x}^{(j)} - \bar{\mathbf{x}})^2 = \mathbf{s}'^2 \quad . \end{aligned}$$

In the parameter space of the null hypothesis we have

$$\begin{aligned} \tilde{\lambda}^{(\omega)} &= \lambda_0, \\ \tilde{\sigma}^{2(\omega)} &= \frac{1}{N} \sum_{j=1}^N (\mathbf{x}^{(j)} - \lambda_0)^2 \quad . \end{aligned}$$



**Fig. 8.6:** Power function of the test from Example 8.5. The right-most curve corresponds to  $N/\sigma^2 = 1$ .

The likelihood ratio (8.6.1) is then

$$\begin{aligned} T &= \left( \frac{\sum(\mathbf{x}^{(j)} - \lambda_0)^2}{\sum(\mathbf{x}^{(j)} - \bar{\mathbf{x}})^2} \right)^{N/2} \exp \left( -\frac{N}{2} \frac{\sum(\mathbf{x}^{(j)} - \bar{\mathbf{x}})^2}{\sum(\mathbf{x}^{(j)} - \bar{\mathbf{x}})^2} + \frac{N}{2} \frac{\sum(\mathbf{x}^{(j)} - \lambda_0)^2}{\sum(\mathbf{x}^{(j)} - \bar{\mathbf{x}})^2} \right) \\ &= \left( \frac{\sum(\mathbf{x}^{(j)} - \lambda_0)^2}{\sum(\mathbf{x}^{(j)} - \bar{\mathbf{x}})^2} \right)^{N/2}. \end{aligned}$$

We transform again to a different test statistic  $T'$  that is a monotonic function of  $T$ ,

$$\begin{aligned} T' &= T^{2/N} = \frac{\sum(\mathbf{x}^{(j)} - \lambda_0)^2}{\sum(\mathbf{x}^{(j)} - \bar{\mathbf{x}})^2} = \frac{\sum(\mathbf{x}^{(j)} - \bar{\mathbf{x}})^2 + N(\bar{\mathbf{x}} - \lambda_0)^2}{\sum(\mathbf{x}^{(j)} - \bar{\mathbf{x}})^2} , \quad (8.6.11) \\ T' &= 1 + \frac{t^2}{N-1}, \end{aligned}$$

where

$$t = \sqrt{N} \frac{\bar{\mathbf{x}} - \lambda_0}{\left( \frac{\sum(\mathbf{x}^{(j)} - \bar{\mathbf{x}})^2}{N-1} \right)^{1/2}} = \sqrt{N} \frac{\bar{\mathbf{x}} - \lambda_0}{\mathbf{s}_x} = \frac{\bar{\mathbf{x}} - \lambda_0}{\mathbf{s}_{\bar{\mathbf{x}}}} \quad (8.6.12)$$

is Student's test variable introduced in Sect. 8.3. From (8.6.11) we can compute a value of  $t$  for a given sample and reject the null hypothesis if

$$|t| > t_{1-\frac{1}{2}\alpha} \quad .$$

The very generally formulated method of the likelihood ratio has led us to Student's test, which was originally constructed for tests with samples from a normal distribution with known mean and unknown variance. ■

## 8.7 The $\chi^2$ -Test for Goodness-of-Fit

### 8.7.1 $\chi^2$ -Test with Maximal Number of Degrees of Freedom

Suppose one has  $N$  measured values  $g_i$ ,  $i = 1, 2, \dots, N$ , each with a known measurement error  $\sigma_i$ . The meaning of the measurement error is the following:  $g_i$  is a measurement of the (unknown) true quantity  $h_i$ . One has

$$g_i = h_i + \varepsilon_i, \quad i = 1, 2, \dots, N \quad . \quad (8.7.1)$$

Here the deviation  $\varepsilon_i$  is a random variable that follows a normal distribution with mean 0 and standard deviation  $\sigma_i$ .

We now want to test the hypothesis specifying the values  $h_i$  on which the measurement is based,

$$h_i = f_i, \quad i = 1, 2, \dots, N \quad . \quad (8.7.2)$$

If this hypothesis is true, then all of the quantities

$$u_i = \frac{g_i - f_i}{\sigma_i}, \quad i = 1, 2, \dots, N \quad , \quad (8.7.3)$$

follow the standard Gaussian distribution. Therefore,

$$T = \sum_{i=1}^N u_i^2 = \sum_{i=1}^N \left( \frac{g_i - f_i}{\sigma_i} \right)^2 \quad (8.7.4)$$

follows a  $\chi^2$ -distribution for  $N$  degrees of freedom. If the hypothesis (8.7.2) is false, then the individual deviations of the measured values  $g_i$  from the values predicted by the hypothesis  $f_i$ , normalized by the errors  $\sigma_i$ , (8.7.3), will be greater. For a given significance level  $\alpha$ , the hypothesis (8.7.2) is rejected if

$$T > \chi_{1-\alpha}^2 \quad , \quad (8.7.5)$$

i.e., if the quantity (8.7.4) is greater than the quantile  $\chi_{1-\alpha}^2$  of the  $\chi^2$ -distribution for  $N$  degrees of freedom.

### 8.7.2 $\chi^2$ -Test with Reduced Number of Degrees of Freedom

The number of degrees of freedom is reduced when the hypothesis to be tested is less explicit than (8.7.2). For this case we consider the following example. Suppose a quantity  $g$  can be measured as a function of an independent *controlled* variable  $t$ , which itself can be set without error,

$$g = g(t) \quad .$$

The individual measurements  $g_i$  correspond to given fixed values  $t_i$  of the independent variable. The corresponding true quantities  $h_i$  are given by some function

$$h_i = h(t_i) \quad .$$

A particularly simple hypothesis for this function is the linear equation

$$f(t) = h(t) = at + b \quad . \quad (8.7.6)$$

The hypothesis can in fact include specifying the numerical values for the parameters  $a$  and  $b$ . In this case, all values  $f_i$  in (8.7.2) are exactly known, and the quantity (8.7.4) follows – if the hypothesis is true – a  $\chi^2$ -distribution for  $N$  degrees of freedom.

The hypothesis may, however, only state: There exists a linear relationship (8.7.6) between the controlled variable  $t$  and the variable  $h$ . The numerical values of the parameters  $a$  and  $b$  are, however, unknown. In this case one constructs estimators  $\tilde{a}$ ,  $\tilde{b}$  for the parameters, which are functions of the measurements  $g_i$  and the errors  $\sigma_i$ . The hypothesis (8.7.2) is then

$$h_i = h(t_i) = f_i = \tilde{a}t_i + \tilde{b} \quad .$$

Since, however,  $\tilde{a}$  and  $\tilde{b}$  are functions of the measurements  $g_i$ , the normalized deviations  $u_i$  in (8.7.3) are no longer all independent. Therefore the number of degrees of freedom of the  $\chi^2$ -distribution for the sum of squares (8.7.4) is reduced by 2 to  $N - 2$ , since the determination of the two quantities  $\tilde{a}$ ,  $\tilde{b}$  introduces two equations of constraint between the quantities  $u_i$ .

### 8.7.3 $\chi^2$ -Test and Empirical Frequency Distribution

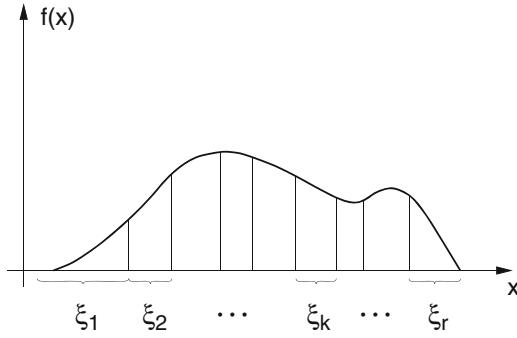
Suppose we have a distribution function  $F(x)$  and its probability density  $f(x)$ . The full region of the random variable  $x$  can be divided into  $r$  intervals

$$\xi_1, \xi_2, \dots, \xi_i, \dots, \xi_r \quad ,$$

as shown in Fig. 8.7. By integrating  $f(x)$  over the individual intervals we obtain the probability to observe  $x$  in  $\xi_i$ ,

$$p_i = P(x \in \xi_i) = \int_{\xi_i} f(x) dx; \quad \sum_{i=1}^r p_i = 1 \quad . \quad (8.7.7)$$

We now take a sample of size  $n$  and denote by  $n_i$  the number of elements of the sample that fall into the interval  $\xi_i$ . An appropriate graphical representation of the sample is a histogram, as described in Sect. 6.3.



**Fig. 8.7:** Dividing the range of the variable  $x$  into the intervals  $\xi_k$ .

One clearly has

$$\sum_{i=1}^r n_i = n \quad . \quad (8.7.8)$$

From the (hypothetical) probability density for the population we would have expected the value

$$n p_i$$

for  $n_i$ . For large values of  $n_i$ , the variance of  $n_i$  is equal to  $n_i$  (Sect. 6.8), and the distribution of the quantity  $u_i$  with

$$u_i^2 = \frac{(n_i - n p_i)^2}{n_i} \quad . \quad (8.7.9)$$

becomes approximately – if the hypothesis is true – a standard Gaussian distribution. This holds also if one uses the expected variances  $n p_i$  instead of the observed quantities  $n_i$  in the denominator of (8.7.9),

$$u_i^2 = \frac{(n_i - n p_i)^2}{n p_i} \quad . \quad (8.7.10)$$

If we now construct the sum of squares of the  $u_i$  for all intervals,

$$X^2 = \sum_{i=1}^r u_i^2 \quad , \quad (8.7.11)$$

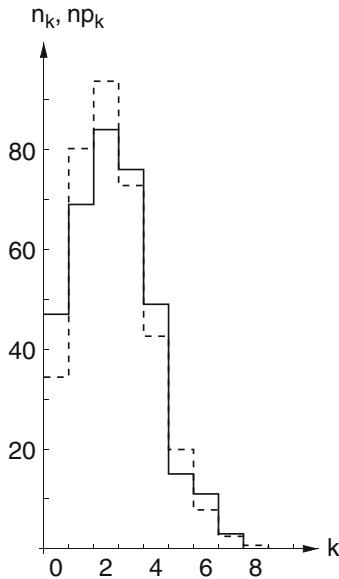
then we expect (for large  $n$ ) that this follows a  $\chi^2$ -distribution if the hypothesis is true. The number of degrees of freedom is  $r - 1$ , since the  $u_i$  are not independent because of (8.7.8). The number of degrees of freedom is reduced to  $r - 1 - p$  if, in addition,  $p$  parameters are determined from the observations.

**Example 8.7:**  $\chi^2$ -test for the fit of a Poisson distribution to an empirical frequency distribution

In an experiment investigating photon-proton interactions, a beam of high energy photons ( $\gamma$ -quanta) impinge on a hydrogen bubble chamber. The processes by which a photon materializes in the chamber, electron-positron pair conversion, are counted in order to obtain a measure of the intensity of the photon beam. The frequency of cases in which 0,1,2,... pairs are observed simultaneously, i.e., in the same bubble-chamber photograph, follows a Poisson distribution (see Example 5.3). Deviations from the Poisson distribution provide information about measurement losses, which are important for uncovering systematic errors. The results of observing  $n = 355$  photographs are given in column 2 of Table 8.3 and in Fig. 8.8. From Example 7.4, we know that the maximum-likelihood estimator of the parameter of the Poisson distribution is given by  $\tilde{\lambda} = \sum_k k n_k / \sum_k n_k$ . We find  $\tilde{\lambda} = 2.33$ . The values  $p_k$  of the Poisson distribution with this parameter multiplied by  $n$  are given in column 3. By summing the squared terms in column 4 one obtains the value  $X^2 = 10.44$ . The problem has six degrees of freedom, since  $r = 8$ ,  $p = 1$ . We chose  $\alpha = 1\%$  and find  $\chi^2_{0.99} = 16.81$  from Table I.7. We therefore have no reason to reject the hypothesis of a Poisson distribution. ■

**Table 8.3:** Data for the  $\chi^2$ -test from Example 8.7.

| Number of<br>electron pairs<br>per photograph | Number of<br>photographs with<br>$k$ electron pairs | Prediction<br>of Poisson<br>distribution | $\frac{(n_k - n p_k)^2}{n p_k}$ |
|---|---|--|---------------------------------|
| $k$   | $n_k$   | $n p_k$                                  |                                 |
| 0   | 47  | 34.4                                     | 4.61                            |
| 1   | 69  | 80.2                                     | 1.56                            |
| 2   | 84  | 93.7                                     | 1.00                            |
| 3   | 76  | 72.8                                     | 0.14                            |
| 4   | 49  | 42.6                                     | 0.96                            |
| 5   | 16  | 19.9                                     | 0.76                            |
| 6   | 11  | 7.8                                      | 1.31                            |
| 7   | 3   | 2.5                                      | 0.10                            |
| 8   | —   | (0.7)                                    |                                 |
| $n = \sum n_k = 355$                          |   | $X^2 = 10.44$                            |                                 |



**Fig. 8.8:** Comparison of the experimental distribution  $n_k$  (histogram with solid line) from Example 8.7 with the Poisson distribution  $np_k$  (dashed line).

## 8.8 Contingency Tables

Suppose  $n$  experiments have been carried out whose results are characterized by the values of two random variables  $x$  and  $y$ . We consider the two variables as discrete, being able to take on the values  $x_1, x_2, \dots, x_k; y_1, y_2, \dots, y_\ell$ . Continuous variables can be approximated by discrete ones by dividing their range into intervals, as shown in Fig. 8.7. Let the number of times the result  $x = x_i$  and  $y = y_j$  is observed be  $n_{ij}$ . One can arrange the numbers  $n_{ij}$  in a matrix, called a *contingency table* (Table 8.4).

**Table 8.4:** Contingency table.

|          | $y_1$    | $y_2$    | $\dots$ | $y_\ell$    |
|----------|----------|----------|---------|-------------|
| $x_1$    | $n_{11}$ | $n_{12}$ | $\dots$ | $n_{1\ell}$ |
| $x_2$    | $n_{21}$ | $n_{22}$ | $\dots$ | $n_{2\ell}$ |
| $\vdots$ | $\vdots$ | $\vdots$ |         | $\vdots$    |
| $x_k$    | $n_{k1}$ | $n_{k2}$ | $\dots$ | $n_{k\ell}$ |

We denote by  $p_i$  the probability for  $x = x_i$  to occur, and by  $q_j$  the probability for  $y = y_j$ . If the variables are independent, then the probability to simultaneously observe  $x = x_i$  and  $y = y_j$  is equal to the product  $p_i q_j$ . The maximum-likelihood estimators for  $p$  and  $q$  are

$$\tilde{p}_i = \frac{1}{n} \sum_{j=1}^{\ell} n_{ij}, \quad \tilde{q}_j = \frac{1}{n} \sum_{i=1}^k n_{ij} .$$

Since

$$\sum_{j=1}^{\ell} \tilde{q}_j = \sum_{i=1}^k \tilde{p}_i = \frac{1}{n} \sum_{j=1}^{\ell} \sum_{i=1}^k n_{ij} = 1 \quad ,$$

one has  $k + \ell - 2$  independent estimators  $\tilde{p}_i, \tilde{q}_j$ . We can now organize the elements of the contingency table into a single line,

$$n_{11}, n_{12}, \dots, n_{1\ell}, n_{21}, n_{22}, \dots, n_{2\ell}, \dots, n_{k\ell} \quad ,$$

and carry out a  $\chi^2$ -test. For this we must compute the quantity

$$X^2 = \sum_{i=1}^k \sum_{j=1}^{\ell} \frac{(n_{ij} - n \tilde{p}_i \tilde{q}_j)^2}{n \tilde{p}_i \tilde{q}_j} \quad (8.8.1)$$

and compare it to the quantile  $\chi^2_{1-\alpha}$  of the  $\chi^2$ -distribution corresponding to a given significance level  $\alpha$ . The number of degrees of freedom is still obtained from the number of intervals minus the number of estimated parameters minus one,

$$f = k\ell - 1 - (k + \ell - 2) = (k - 1)(\ell - 1) \quad .$$

If the variables are not independent, then  $n_{ij}$  will not, in general, be near  $n \tilde{p}_i \tilde{q}_j$ , i.e., one will find

$$X^2 > \chi^2_{1-\alpha} \quad (8.8.2)$$

and the hypothesis will be rejected.

## 8.9 $2 \times 2$ Table Test

The simplest nontrivial contingency table has only two rows and two columns, and is called a  $2 \times 2$  table, as shown in Table 8.5. It is often used in medical studies. (The variables  $x_1$  and  $x_2$  could represent, for example, two different treatment methods, and  $y_1$  and  $y_2$  could represent success and failure of the treatment. One wishes to determine whether success is independent of the treatment.)

**Table 8.5:**  $2 \times 2$  table.

|       | $y_1$        | $y_2$        |
|-------|--------------|--------------|
| $x_1$ | $n_{11} = a$ | $n_{12} = b$ |
| $x_2$ | $n_{21} = c$ | $n_{22} = d$ |

One computes the quantity  $X^2$  either according to (8.8.1) or from the formula

$$X^2 = \frac{n(ad - bc)^2}{(a+b)(c+d)(a+c)(b+d)} ,$$

which is obtained by rearranging (8.8.1). If the variables  $x$  and  $y$  are independent, then  $X^2$  follows a  $\chi^2$ -distribution with one degree of freedom. One rejects the hypothesis of independence at the significance level  $\alpha$  if

$$X^2 > \chi_{1-\alpha}^2 .$$

In order for the quantity  $X^2$  to actually follow a  $\chi^2$ -distribution is again necessary that the individual  $n_{ij}$  are sufficiently large, (and the hypothesis of independence must be true).

## 8.10 Example Programs

**Example Program 8.1:** The class E1Test generates samples and tests the equality of their variances

The program performs a total of  $n_{\text{exp}}$  simulated experiments. Each experiment consists of the simulation of two samples of sizes  $N_1$  and  $N_2$  from normal distributions with standard deviations  $\sigma_1$  and  $\sigma_2$ . The variance of each of the samples is computed using the class Sample. The sample variances are called  $s_g^2$  and  $s_k^2$ , so that

$$s_g^2 > s_k^2 .$$

From the corresponding sample sizes the numbers of degrees of freedom  $f_g = N_g - 1$  and  $f_k = N_k - 1$  are computed. Finally, the ratio  $s_g^2/s_k^2$  is compared with the quantile  $F_{1-\alpha/2}(f_g, f_k)$  at a given confidence level  $\beta = 1 - \alpha$ . If the ratio is larger than the quantile, then the hypothesis of equal variances has to be rejected. The program asks for the quantities  $n_{\text{exp}}$ ,  $N_1$ ,  $N_2$ ,  $\sigma_1$ ,  $\sigma_2$ , and  $\beta$ . For each simulated experiment one line of output is displayed.

**Suggestions:** Choose  $n_{\text{exp}} = 20$  and  $\beta = 0.9$ . (a) For  $\sigma_1 = \sigma_2$  you would expect the hypothesis to be rejected in 2 out of 20 cases because of an error of the first kind. Note the large statistical fluctuations, which obviously depend on  $N_1$  and  $N_2$ , and choose different pairs of values  $N_1$ ,  $N_2$  for  $\sigma_1 = \sigma_2$ . (b) Check the power of the test for different variances  $\sigma_1 \neq \sigma_2$ .

**Example Program 8.2:** The class E2Test generates samples and tests the equality of their means with a given value using Student's Test

This short program performs  $n_{\text{exp}}$  simulation experiments. In each experiment a sample of size  $N$  is drawn from a normal distribution with mean  $x_0$  and width  $\sigma$ . Using the class Sample the sample mean  $\bar{x}$  and the sample variance  $s_x^2$  are determined. If  $\lambda_0$  is the population mean specified by the hypothesis, then the quantity

$$|t| = \frac{|\bar{x} - \lambda_0| \sqrt{N}}{s_x}$$

can be used to test the hypothesis. At a given confidence level  $\beta = 1 - \alpha$  the hypothesis is rejected if

$$|t| > t_{1-\alpha/2} .$$

Here  $t_{1-\alpha/2}$  is the quantile of Student's distribution with  $f = N - 1$  degrees of freedom. The program asks for the quantities  $n_{\text{exp}}$ ,  $N$ ,  $x_0$ ,  $\sigma$ ,  $\lambda_0$ , and  $\beta$ . For each simulated experiment one line of output is displayed.

**Suggestion:** Modify the suggestions at the end of Sect. 8.1 to apply to Student's test.

**Example Program 8.3:** The class E3Test generates samples

and computes the test statistic  $\chi^2$  for the hypothesis that the samples are taken from a normal distribution with known parameters

For samples of size  $N$  the hypothesis  $H_0$  that they stem from a normal distribution with mean  $a_0$  and standard deviation  $\sigma_0$  is tested. A total of  $n_{\text{exp}}$  samples are drawn in simulated experiments from a normally distributed population with mean  $a$  and standard deviation  $\sigma$ . For each sample the quantity

$$\mathbf{X}^2 = \sum_{i=1}^N \left( \frac{\mathbf{x}_i - a_0}{\sigma_0} \right)^2 \quad (8.10.1)$$

is computed. Here  $\mathbf{x}_i$  are the elements of the sample. If  $a = a_0$  and  $\sigma = \sigma_0$ , then the quantity  $\mathbf{X}^2$  follows a  $\chi^2$ -distribution for  $N$  degrees of freedom. This quantity can therefore be used to perform a  $\chi^2$ -test on the hypothesis  $H_0$ . The program does not, however, perform the  $\chi^2$ -test, but rather it displays a histogram of the quantity  $\mathbf{X}^2$  together with the  $\chi^2$ -distribution. One observes that for  $a = a_0$  and  $\sigma = \sigma_0$  the histogram and  $\chi^2$ -distribution indeed coincide within statistical fluctuations. If, however,  $a \neq a_0$  and/or  $\sigma \neq \sigma_0$ , then deviations appear. These deviations become particularly clear if instead of  $\mathbf{X}^2$  the quantity

$$P(\mathbf{X}^2) = 1 - F(\mathbf{X}^2; N) \quad (8.10.2)$$

is displayed. Here  $F(\mathbf{X}^2, N)$  is the distribution function (C.5.2) of the  $\chi^2$ -distribution for  $N$  degrees of freedom.  $F(\mathbf{X}^2, N)$  is equal to the probability that a random variable drawn from a  $\chi^2$ -distribution is smaller than  $\mathbf{X}^2$ . Thus,  $P$  is the probability that it is greater than or equal to  $\mathbf{X}^2$ . If the hypothesis  $H_0$  is true, then  $F$  and therefore also  $P$  follow uniform distributions between 0 and 1. If, however,  $H_0$  is false, then the distribution of the  $\mathbf{X}^2$  is not a  $\chi^2$ -distribution, and the distribution of the  $P$  is not a uniform distribution. The test statistic  $\mathbf{X}^2$  often (not completely correctly) is simply called " $\chi^2$ " and the quantity  $P$  is then called the " $\chi^2$ -probability". Large values of  $\mathbf{X}^2$  obviously signify that the terms in the sum (8.10.1) are on the average large compared to unity, i.e., that the  $\mathbf{x}_i$  are significantly different from  $a_0$ . For large values of  $\mathbf{X}^2$ , however,  $P$  becomes small, cf. (8.10.2). Large values of " $\chi^2$ " therefore correspond to small values of the " $\chi^2$ -probability". The hypothesis  $H_0$  is rejected at the confidence level  $\beta = 1 - \alpha$  if  $\mathbf{X}^2 > \chi^2_{1-\alpha}(N)$ . That is equivalent to  $F(\mathbf{X}^2, N) > \beta$  or  $P < \alpha$ .

The program allows for interactive input of the quantities  $n_{\text{exp}}$ ,  $N$ ,  $a$ ,  $a_0$ ,  $\sigma$ ,  $\sigma_0$  and displays the distributions of both  $X^2$  and  $P(X^2)$  in the form of histograms.

**Suggestions:** (a) Choose  $n_{\text{exp}} = 1000$ ;  $a = a_0 = 0$ ,  $\sigma = \sigma_0 = 1$  and for  $N = 1$ ,  $N = 2$ , and  $N = 10$  display both  $X^2$  and  $P(X^2)$ . (b) Repeat (a), keeping  $a = 0$  and choosing  $a_0 = 1$  and  $a_0 = 5$ . Explain the shift of the histogram for  $P(X^2)$ . (c) Repeat (a), keeping  $\sigma = 1$  fixed and choosing  $\sigma_0 = 0.5$  and  $\sigma_0 = 2$ . Discuss the results. (d) Modify the program so that instead of  $a_0$  and  $\sigma_0^2$ , the sample mean  $\bar{x}$  sample variance  $s^2$  are used for the computation of  $X^2$ . The quantity  $X^2$  can be used for a  $\chi^2$ -test of the hypothesis that the samples were drawn from a normal distribution. Display histograms of  $X^2$  and  $P(X^2)$  and show that  $X^2$  follows a  $\chi^2$ -distribution with  $N - 2$  degrees of freedom.

## 9. The Method of Least Squares

The method of *least squares* was first developed by LEGENDRE and GAUSS. In the simplest case it consists of the following prescription:

The repeated measurements  $y_j$  can be treated as the sum of the (unknown) quantity  $x$  and the measurement error  $\varepsilon_j$ ,

$$y_j = x + \varepsilon_j \quad .$$

The quantity  $x$  should be determined such that the sum of squares of the errors  $\varepsilon_j$  is a minimum,

$$\sum_j \varepsilon_j^2 = \sum_j (x - y_j)^2 = \min \quad .$$

We will see that in many cases this prescription can be derived as a result of the principle of maximum likelihood, which historically was developed much later, but that in other cases as well, it provides results with optimal properties. The method of least squares, which is the most widely used of all statistical methods, can also be used in the case where the measured quantities  $y_j$  are not directly related to the unknown  $x$ , but rather indirectly, i.e., as a linear (or also nonlinear) combination of several unknowns  $x_1, x_2, \dots$ . Because of the great practical significance of the method we will illustrate the various cases individually with examples before turning to the most general case.

### 9.1 Direct Measurements of Equal or Unequal Accuracy

The simplest case of *direct measurements with equal accuracy* has already been mentioned. Suppose one has carried out  $n$  measurements of an unknown quantity  $x$ . The measured quantities  $y_j$  have measurement errors  $\varepsilon_j$ . We now make the additional assumption that these are normally distributed about zero:

$$y_j = x + \varepsilon_j \quad , \quad E(\varepsilon_j) = 0 \quad , \quad E(\varepsilon_j^2) = \sigma^2 \quad . \quad (9.1.1)$$

This assumption can be justified in many cases by the Central Limit Theorem. The probability to observe the value  $y_j$  as the result of a single measurement is thus proportional to

$$f_j dy = \frac{1}{\sigma \sqrt{2\pi}} \exp\left(-\frac{(y_j - x)^2}{2\sigma^2}\right) dy \quad .$$

The log-likelihood function for all  $n$  measurements is thus (see Example 7.2)

$$\ell = -\frac{1}{2\sigma^2} \sum_{j=1}^n (y_j - x)^2 + \text{const} \quad . \quad (9.1.2)$$

The maximum likelihood condition,

$$\ell = \max \quad ,$$

is thus equivalent to

$$M = \sum_{j=1}^n (y_j - x)^2 = \sum_{j=1}^n \varepsilon_j^2 = \min \quad . \quad (9.1.3)$$

This is exactly the least-squares prescription. As we have shown in Examples 7.2 and 7.6, this leads to the result that the best estimator for  $x$  is given by the arithmetic mean of the  $y_j$

$$\tilde{x} = \bar{y} = \frac{1}{n} \sum_{j=1}^n y_j \quad . \quad (9.1.4)$$

The variance of this estimator is

$$\sigma^2(\bar{y}) = \sigma^2/n \quad , \quad (9.1.5)$$

or, if we set the measurement errors and standard deviations equal,

$$\Delta \tilde{x} = \Delta y / \sqrt{n} \quad . \quad (9.1.6)$$

The more general case of *direct measurements of different accuracy* has also already been treated in Example 7.6. Let us assume again a normal distribution centered about zero for the measurement errors, i.e.,

$$y_j = x + \varepsilon_j \quad , \quad E(\varepsilon_j) = 0 \quad , \quad E(\varepsilon_j^2) = \sigma_j^2 = 1/g_j \quad . \quad (9.1.7)$$

Comparing with (7.2.7) gives the requirement of the maximum-likelihood method

$$M = \sum_{j=1}^n \frac{(y_j - x)^2}{\sigma_j^2} = \sum_{j=1}^n g_j (y_j - x)^2 = \sum_{j=1}^n g_j \varepsilon_j^2 = \min \quad . \quad (9.1.8)$$

The individual terms in the sum are now *weighted* with the inverse of the variances. The best estimator for  $x$  is then [cf. (7.2.8)]

$$\tilde{x} = \frac{\sum_{j=1}^n g_j y_j}{\sum_{j=1}^n g_j} \quad , \quad (9.1.9)$$

i.e., the weighted mean of the individual measurements. One sees that an individual measurement thus contributes to the final result less when its measurement error is greater. From (7.3.20) we know the variance of  $\tilde{x}$ ; it is

$$\sigma^2(\tilde{x}) = \left( \sum_{j=1}^n \frac{1}{\sigma_j^2} \right)^{-1} = \left( \sum_{j=1}^n g_j \right)^{-1} \quad . \quad (9.1.10)$$

We can use the result (9.1.9) in order to compute the best estimates  $\tilde{\varepsilon}_j$  of the original measurement errors  $\varepsilon_j$  from (9.1.1) to obtain

$$\tilde{\varepsilon}_j = y_j - \tilde{x} \quad .$$

We expect these quantities to be normally distributed about zero with the variance  $\sigma_j^2$ . That is, the quantities  $\tilde{\varepsilon}_j / \sigma_j$  should follow a standard Gaussian distribution. According to Sect. 6.6, the sum of squares

$$M = \sum_{j=1}^n \left( \frac{\tilde{\varepsilon}_j}{\sigma_j} \right)^2 = \sum_{j=1}^n \frac{(y_j - \tilde{x})^2}{\sigma_j^2} = \sum_{j=1}^n g_j (y_j - \tilde{x})^2 \quad (9.1.11)$$

then follows a  $\chi^2$ -distribution with  $n - 1$  degrees of freedom.

This property of the quantity  $M$  can now be used to carry out a  $\chi^2$ -test on the validity of the assumption (9.1.7). If, for a given significance level  $\alpha$ , the quantity  $M$  exceeds the value  $\chi^2_{1-\alpha}$ , then we would have to recheck the assumption (9.1.7). Usually one does not doubt that the  $y_j$  are in fact measurements of the unknown  $x$ . It may be, however, that the errors  $\varepsilon_j$  are not normally distributed. In particular, the measurements may also be biased, i.e., the expectation value of the errors  $\varepsilon_j$  may be different from zero. The presence of such systematic errors can often be inferred from the failure of the  $\chi^2$ -test.

**Example 9.1:** Weighted mean of measurements of different accuracy

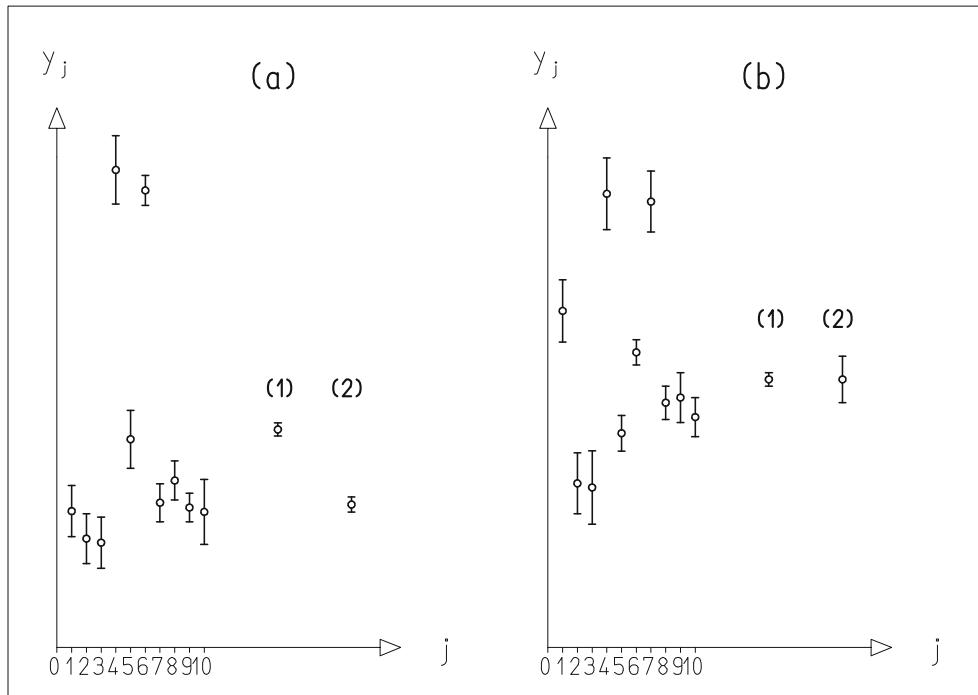
The best values for constants of fundamental significance, such as important constants of nature, are usually obtained as weighted averages of measurements obtained by different experimental groups. For the properties of elementary particles, such mean values are compiled at regular intervals. We will consider as an example somewhat older measurements of the mass of the neutral K-meson ( $K^0$ ), taken from such a compilation from 1967 [10]. An average was computed from the results of four experiments, all carried out with different techniques. The calculation can be carried out following the scheme of Table 9.1. The resulting value of  $M$  is 7.2. If we choose a significance level of 5 %, we find from Table I.7 for three degrees of freedom  $\chi^2_{0.95} = 7.82$ . At the time of the averaging, one could therefore assume that the result  $m_{K^0} = (497.9 \pm 0.2)$  MeV represented the best value for the mass of the K-meson, as long as no further experiments were carried out. (More than 40 years later the weighted mean of all measurements was  $m_{K^0} = (497.614 \pm 0.024)$  MeV [11]). ■

**Table 9.1:** Construction of the weighted mean from four measurements of the mass of the neutral K meson. The  $y_j$  are the measured values in MeV.

| $j$   | $y_j$  | $\sigma_j$ | $1/\sigma_j^2 = g_j$ | $y_j g_j$ | $y_j - \tilde{x}$ | $(y_j - \tilde{x})^2 g_j$ |
|---|--------|------------|----------------------|-----------|-------------------|---------------------------|
| 1   | 498.1  | 0.4        | 6.3                  | 3038.0    | 0.2               | 0.3                       |
| 2   | 497.44 | 0.33       | 10                   | 4974.4    | -0.46             | 2.1                       |
| 3   | 498.9  | 0.5        | 4                    | 1995.6    | 1.0               | 4.0                       |
| 4   | 497.44 | 0.5        | 4                    | 1989.8    | -0.46             | 0.8                       |
| $\sum$  |        |            | 24.3                 | 11 997.8  |                   | 7.2                       |
| $\tilde{x} = \sum y_j g_j / \sum g_j = 497.9$ , $\Delta \tilde{x} = (\sum g_j)^{-\frac{1}{2}} = 0.20$ |        |            |                      |           |                   |                           |

Let us now consider the case where the  $\chi^2$ -test fails. As mentioned above, one usually assumes that at least one of the measurements has a systematic error. By investigation of the individual measurements, one can sometimes determine that one or two measurements deviate from the others by a large amount. Such a case is illustrated in Fig. 9.1a, where several different measurements are shown with their errors. (The measured value is plotted along the vertical axis; the horizontal axis merely distinguishes between the different measurements.) Although the  $\chi^2$ -test would fail if all measurements from Fig. 9.1a are used, this is not the case if measurements 4 and 6 are excluded.

Unfortunately the situation is not always so clear. Figure 9.1b shows a further example where the  $\chi^2$ -test fails. (According to Chap. 8 one would reject the hypothesis that the measurements are determinations of the same quantity.) There is no single measurement, however, that is responsible for



**Fig. 9.1:** Averaging of 10 measurements where the  $\chi^2$ -test fails. **(a)** Anomalous deviation of certain measurements: (1) Averaging of all measurements; (2) averaging without  $y_4$  and  $y_6$ . **(b)** Errors of individual measurements clearly too small: (1) Error of the mean according to (9.1.10); (2) error of the mean according to (9.1.13).

this fact. It would now be mathematically correct to not give any average value at all, and to make no statement about a best value, as long as no further measurements are available. In practice, this is clearly not satisfactory. ROSENFELD et al. [10] have suggested that the individual measurement errors should be increased by a *scale factor*  $\sqrt{M/(n-1)}$ , i.e., one replaces the  $\sigma_j$  by

$$\sigma'_j = \sigma_j \sqrt{\frac{M}{n-1}} \quad . \quad (9.1.12)$$

The weighted mean  $\tilde{x}$  obtained by using these measurement errors does not differ from that of expression (9.1.9). The variance, however, is different from (9.1.10). It becomes

$$\sigma'^2(\tilde{x}) = \frac{M}{n-1} \left( \sum_{j=1}^n \frac{1}{\sigma_j^2} \right)^{-1} \quad . \quad (9.1.13)$$

We now compute the analogous expression to (9.1.11),

$$M' = \frac{n-1}{M} \sum_{j=1}^n \frac{(y_j - \tilde{x})^2}{\sigma_j^2} = \frac{n-1}{M} M = n-1 \quad . \quad (9.1.14)$$

This is the expectation value of  $\chi^2$  for  $n-1$  degrees of freedom. Equation (9.1.14) clearly provided the motivation for relation (9.1.12). We repeat that this relation has no rigorous mathematical basis. It should be used with caution, since it hides the influence of systematic errors. On the other hand it provides reasonable errors for the mean for cases such as those in Fig. 9.1b, while direct application of Eq. (9.1.10) leads to an error that is far to small to reflect the actual dispersion of the individual measurements about the mean. Both solutions for the error of the mean are shown in Fig. 9.1b.

## 9.2 Indirect Measurements: Linear Case

Let us now consider the more general case of several unknown quantities  $x_i$  ( $i = 1, 2, \dots, r$ ). The unknowns are often not measured directly. Instead, only a set of linear functions of the  $x_i$  are measurable,

$$\eta_j = p_{j0} + p_{j1}x_1 + p_{j2}x_2 + \dots + p_{jr}x_r \quad . \quad (9.2.1)$$

We now write this relation in a somewhat different form,

$$f_j = \eta_j + a_{j0} + a_{j1}x_1 + a_{j2}x_2 + \dots + a_{jr}x_r = 0 \quad . \quad (9.2.2)$$

We define a column vector,

$$a_j = \begin{pmatrix} a_{j1} \\ a_{j2} \\ \vdots \\ a_{jr} \end{pmatrix} \quad , \quad (9.2.3)$$

and write (9.2.2) in the more compact form,

$$f_j = \eta_j + a_{j0} + \mathbf{a}_j^T \mathbf{x} = 0 \quad , \quad j = 1, 2, \dots, n \quad . \quad (9.2.4)$$

If we define in addition

$$\boldsymbol{\eta} = \begin{pmatrix} \eta_1 \\ \eta_2 \\ \vdots \\ \eta_n \end{pmatrix}, \quad \mathbf{a}_0 = \begin{pmatrix} a_{10} \\ a_{20} \\ \vdots \\ a_{n0} \end{pmatrix}, \quad A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1r} \\ a_{21} & a_{22} & \cdots & a_{2r} \\ \vdots & & & \\ a_{n1} & a_{n2} & \cdots & a_{nr} \end{pmatrix}, \quad (9.2.5)$$

then the system of equations (9.2.4) can be written as a matrix equation,

$$\mathbf{f} = \boldsymbol{\eta} + \mathbf{a}_0 + A\mathbf{x} = 0 \quad . \quad (9.2.6)$$

Of course the measured quantities still have measurement errors  $\varepsilon_j$ , which we assume to be normally distributed. We then have\*

$$\begin{aligned} y_j &= \eta_j + \varepsilon_j \quad , \\ E(\varepsilon_j) &= 0 \quad , \\ E(\varepsilon_j^2) &= \sigma_j^2 = 1/g_j \quad . \end{aligned} \quad (9.2.7)$$

Since the  $y_j$  are *independent* measurements, we can arrange the variances  $\sigma_j^2$  in a *diagonal* covariance matrix for  $y_j$  or  $\varepsilon_j$ ,

$$C_y = C_\varepsilon = \begin{pmatrix} \sigma_1^2 & & & 0 \\ & \sigma_2^2 & & \\ & & \ddots & \\ 0 & & & \sigma_n^2 \end{pmatrix} \quad . \quad (9.2.8)$$

In analogy to (9.1.7) we call the inverse of the covariance matrix a *weight matrix*,

$$G_y = G_\varepsilon = C_y^{-1} = C_\varepsilon^{-1} = \begin{pmatrix} g_1 & & & 0 \\ & g_2 & & \\ & & \ddots & \\ 0 & & & g_n \end{pmatrix} \quad . \quad (9.2.9)$$

If we now put the measurements and errors together into vectors, we obtain from (9.2.7)

$$\mathbf{y} = \boldsymbol{\eta} + \boldsymbol{\varepsilon} \quad . \quad (9.2.10)$$

From (9.2.6) one then has

$$\mathbf{y} - \boldsymbol{\varepsilon} + \mathbf{a}_0 + A\mathbf{x} = 0 \quad . \quad (9.2.11)$$

We want to solve this system of equations for the unknowns  $\mathbf{x}$  with the maximum-likelihood method. With our assumption (9.2.7), the measurements  $y_j$  are normally distributed with the probability density

$$f(y_j) = \frac{1}{\sigma_j \sqrt{2\pi}} \exp\left(-\frac{(y_j - \eta_j)^2}{2\sigma_j^2}\right) = \frac{1}{\sigma_j \sqrt{2\pi}} \exp\left(-\frac{\varepsilon_j^2}{2\sigma_j^2}\right) \quad . \quad (9.2.12)$$

---

\*For simplicity of notation we no longer write random variables with a special character type. From context it will always be evident which variables are random.

Thus for all measurements one obtains likelihood functions

$$L = \prod_{j=1}^n f(y_j) = (2\pi)^{-\frac{1}{2}n} \left( \prod_{j=1}^n \sigma_j^{-1} \right) \exp \left( -\frac{1}{2} \sum_{j=1}^n \frac{\varepsilon_j^2}{\sigma_j^2} \right) , \quad (9.2.13)$$

$$\ell = \ln L = -\frac{1}{2}n \ln 2\pi + \ln \left( \prod_{j=1}^n \sigma_j^{-1} \right) - \frac{1}{2} \sum_{j=1}^n \frac{\varepsilon_j^2}{\sigma_j^2} . \quad (9.2.14)$$

This expression is clearly a maximum when

$$M = \sum_{j=1}^n \frac{\varepsilon_j^2}{\sigma_j^2} = \sum_{j=1}^n \frac{(y_j + \mathbf{a}_j^T \mathbf{x} + a_{j0})^2}{\sigma_j^2} = \min . \quad (9.2.15)$$

Using (9.2.9) and (9.2.11), we can rewrite this expression as

$$M = \mathbf{\epsilon}^T G_y \mathbf{\epsilon} = \min \quad (9.2.16)$$

or

$$M = (\mathbf{y} + \mathbf{a}_0 + A\mathbf{x})^T G_y (\mathbf{y} + \mathbf{a}_0 + A\mathbf{x}) = \min , \quad (9.2.17)$$

or in abbreviated form

$$\mathbf{c} = \mathbf{y} + \mathbf{a}_0 , \quad (9.2.18)$$

$$M = (\mathbf{c} + A\mathbf{x})^T G_y (\mathbf{c} + A\mathbf{x}) = \min . \quad (9.2.19)$$

We will simplify this expression further by using the Cholesky decomposition (cf. Sect. A.9) of the positive-definite symmetric weight matrix  $G_y$ ,

$$G_y = H^T H . \quad (9.2.20)$$

In the frequently occurring case of uncorrelated measurements (9.2.9) one has

$$H = H^T = \begin{pmatrix} 1/\sigma_1 & & & 0 \\ & 1/\sigma_2 & & \\ & & \ddots & \\ 0 & & & 1/\sigma_n \end{pmatrix} . \quad (9.2.21)$$

Using the notation

$$\mathbf{c}' = H\mathbf{c} , \quad A' = HA \quad (9.2.22)$$

Eq. (9.2.19) takes on the simple form

$$M = (A'\mathbf{x} + \mathbf{c}')^2 = \min . \quad (9.2.23)$$

The method for solving this equation for  $\mathbf{x}$  is described in detail in Appendix A, in particular in Sects. A.5 through A.14. The solution can be written in the form

$$\tilde{\mathbf{x}} = -A'^+ \mathbf{c}' . \quad (9.2.24)$$

[See (A.10.3).] Here  $A'^+$  is the pseudo-inverse of the matrix  $A'$  (see Sect. A.10). In Sect. A.14, a solution of the form of (9.2.24) is indeed found with the help of the singular value decomposition of the matrix  $A'$ . This procedure is particularly accurate numerically.

To compute by hand one uses instead of Eq. (9.2.24) the mathematically equivalent expression for the solution of the normal equations [see (A.5.17)],

$$\tilde{\mathbf{x}} = -(A'^T A')^{-1} A'^T \mathbf{c}' \quad (9.2.25)$$

or, with (9.2.22) in terms of the quantities  $\mathbf{c}$  and  $A$ ,

$$\tilde{\mathbf{x}} = -(A^T G_y A)^{-1} A^T G_y \mathbf{c} \quad . \quad (9.2.26)$$

The solution includes, of course, the special case of Sect. 9.1. In the case of direct measurements of different accuracy,  $\mathbf{x}$  has only one element,  $\mathbf{a}_0$  vanishes, and  $A$  is simply an  $n$ -component column vector whose elements are all equal to  $-1$ . One then has

$$\mathbf{c}' = \begin{pmatrix} y_1/\sigma_1 \\ y_2/\sigma_2 \\ \vdots \\ y_n/\sigma_n \end{pmatrix} \quad , \quad A' = \begin{pmatrix} -1/\sigma_1 \\ -1/\sigma_2 \\ \vdots \\ -1/\sigma_n \end{pmatrix} \quad , \quad A'^T A = \sum_{j=1}^n \frac{1}{\sigma_j^2} \quad ,$$

and (9.2.25) becomes

$$\tilde{\mathbf{x}} = \left( \sum_{j=1}^n \frac{1}{\sigma_j^2} \right)^{-1} \sum_{j=1}^n \frac{y_j}{\sigma_j^2} \quad ,$$

which is identical to (9.1.9).

The solution of (9.2.26) represents a linear relation between the solution vector  $\tilde{\mathbf{x}}$  and the vector of measurements  $\mathbf{y}$ , since  $\mathbf{c} = \mathbf{y} + \mathbf{a}_0$ . We can thus apply the error propagation techniques of Sect. 3.8. Using (3.8.2) and (3.8.4) one immediately obtains

$$C_x = G_x^{-1} = [(A^T G_y A)^{-1} A^T G_y] G_y^{-1} [(A^T G_y A)^{-1} A^T G_y]^T \quad .$$

The matrices  $G_y$ ,  $G_y^{-1}$ , and  $(A^T G_y A)$  are symmetric, i.e., they are identical to their transposed matrices. Using the rule (A.1.8), this expression simplifies to

$$\begin{aligned} G_{\tilde{x}}^{-1} &= (A^T G_y A)^{-1} A^T G_y G_y^{-1} G_y A (A^T G_y A)^{-1} \\ &= (A^T G_y A)^{-1} (A^T G_y A) (A^T G_y A)^{-1} \quad , \\ G_{\tilde{x}}^{-1} &= (A^T G_y A)^{-1} = (A'^T A')^{-1} \quad . \end{aligned} \quad (9.2.27)$$

We have thus obtained a simple expression for the covariance matrix of the estimators  $\tilde{\mathbf{x}}$  for the unknowns  $\mathbf{x}$ . The square roots of the diagonal elements of this matrix can be viewed as “measurement errors”, although the quantities  $\mathbf{x}$  were not directly measured.

We can also use the result (9.2.26) to improve the original measurements  $\mathbf{y}$ . By substituting (9.2.26) into (9.2.11) one obtains a vector of estimators of the measurement errors  $\boldsymbol{\varepsilon}$ ,

$$\tilde{\boldsymbol{\varepsilon}} = A\tilde{\mathbf{x}} + \mathbf{c} = -A(A^T G_y A)^{-1} A^T G_y \mathbf{c} + \mathbf{c} \quad . \quad (9.2.28)$$

These measurement errors can now be used to compute improved measured values,

$$\begin{aligned} \tilde{\boldsymbol{\eta}} = \mathbf{y} - \tilde{\boldsymbol{\varepsilon}} &= \mathbf{y} + A(A^T G_y A)^{-1} A^T G_y \mathbf{c} - \mathbf{c} \quad , \\ \tilde{\boldsymbol{\eta}} &= A(A^T G_y A)^{-1} A^T G_y \mathbf{c} - \mathbf{a}_0 \quad . \end{aligned} \quad (9.2.29)$$

The  $\tilde{\boldsymbol{\eta}}$  are again linear in  $\mathbf{y}$ . We can again use error propagation to determine the covariance matrix of the improved measurements,

$$\begin{aligned} G_{\tilde{\boldsymbol{\eta}}}^{-1} &= [A(A^T G_y A)^{-1} A^T G_y] G_y^{-1} [A(A^T G_y A)^{-1} A^T G_y]^T \quad , \\ G_{\tilde{\boldsymbol{\eta}}}^{-1} &= A(A^T G_y A)^{-1} A^T = A G_{\tilde{\mathbf{x}}}^{-1} A^T \quad . \end{aligned} \quad (9.2.30)$$

The improved measurements  $\tilde{\boldsymbol{\eta}}$  satisfy (9.2.1) if the unknowns are replaced by their estimators  $\tilde{\mathbf{x}}$ .

### 9.3 Fitting a Straight Line

We will examine in detail the simple but in practice frequently occurring task of fitting a straight line to a set of measurements  $y_j$  at various values  $t_j$  of a so-called *controlled variable*  $t$ . The values of these variables will be assumed to be known exactly, i.e., without error. The variable  $t_i$  could be, for example, the time at which an observation  $y_i$  is made, or a temperature or voltage that is set in an experiment. (If  $t$  also has an error, then fitting a line in the  $(t, y)$  plane becomes a nonlinear problem. It will be treated in Sect. 9.10, Example 9.11.)

In the present case the relation (9.2.1) has the simple form

$$\eta_j = y_j - \varepsilon_j = x_1 + x_2 t_j$$

or using vector notation,

$$\boldsymbol{\eta} - \mathbf{x}_1 - \mathbf{x}_2 \mathbf{t} = \mathbf{0} \quad .$$

We will attempt to determine the unknown parameters

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

from the measurements in Table 9.2.

A comparison of our problem to Eqs. (9.2.2) through (9.2.6) gives  $\mathbf{a}_0 = 0$ ,

$$A = - \begin{pmatrix} 1 & t_1 \\ 1 & t_2 \\ 1 & t_3 \\ 1 & t_4 \end{pmatrix} = - \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{pmatrix} , \quad \mathbf{y} = \mathbf{c} = \begin{pmatrix} 1.4 \\ 1.5 \\ 3.7 \\ 4.1 \end{pmatrix} .$$

The matrices  $G_y$  and  $H$  are found by substitution of the last line of Table 9.2 into (9.2.9) and (9.2.21),

$$G_y = \begin{pmatrix} 4 & & & \\ & 25 & & \\ & & 1 & \\ & & & 4 \end{pmatrix} , \quad H = \begin{pmatrix} 2 & & & \\ & 5 & & \\ & & 1 & \\ & & & 2 \end{pmatrix} .$$

One thus has

$$A' = - \begin{pmatrix} 2 & 0 \\ 5 & 5 \\ 1 & 2 \\ 2 & 6 \end{pmatrix} , \quad \mathbf{c}' = \begin{pmatrix} 2.8 \\ 7.5 \\ 3.7 \\ 8.2 \end{pmatrix} , \quad A'^T \mathbf{c}' = - \begin{pmatrix} 63.2 \\ 94.1 \end{pmatrix} ,$$

$$(A'^T A')^{-1} = \begin{pmatrix} 34 & 39 \\ 39 & 65 \end{pmatrix}^{-1} \\ = \frac{1}{689} \begin{pmatrix} 65 & -39 \\ -39 & 34 \end{pmatrix} = \begin{pmatrix} 0.0943 & -0.0556 \\ -0.0566 & 0.0493 \end{pmatrix} .$$

To invert the  $2 \times 2$  matrix we use (A.6.8). The solution (9.2.25) is then

$$\tilde{\mathbf{x}} = \begin{pmatrix} 0.0943 & -0.0556 \\ -0.0566 & 0.0493 \end{pmatrix} \begin{pmatrix} 63.2 \\ 94.1 \end{pmatrix} = \begin{pmatrix} 0.636 \\ 1.066 \end{pmatrix} .$$

The covariance matrix of for  $\tilde{\mathbf{x}}$  is

$$C_{\tilde{\mathbf{x}}} = G_{\tilde{\mathbf{x}}}^{-1} = (A'^T A')^{-1} = \begin{pmatrix} 0.0943 & -0.0556 \\ -0.0566 & 0.0493 \end{pmatrix} .$$

Its diagonal elements are the variances of  $\tilde{x}_1$  and  $\tilde{x}_2$ , and their square roots are the errors,

$$\Delta \tilde{x}_1 = 0.307 , \quad \Delta \tilde{x}_2 = 0.222 .$$

**Table 9.2:** Data for fitting a straight line.

| $j$        | 1   | 2   | 3   | 4   |
|------------|-----|-----|-----|-----|
| $t_j$      | 0.0 | 1.0 | 2.0 | 3.0 |
| $y_j$      | 1.4 | 1.5 | 3.7 | 4.1 |
| $\sigma_j$ | 0.5 | 0.2 | 1.0 | 0.5 |

For the correlation coefficient between  $\tilde{x}_1$  and  $\tilde{x}_2$  one finds

$$\rho = \frac{-0.0566}{0.307 \cdot 0.222} = -0.830 \quad .$$

The improved measurements are

$$\tilde{\eta} = -A\tilde{x} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{pmatrix} \begin{pmatrix} 0.636 \\ 1.066 \end{pmatrix} = \begin{pmatrix} 0.636 \\ 1.702 \\ 2.768 \\ 3.834 \end{pmatrix} \quad .$$

They lie on a line given by  $\tilde{\eta} = -A\tilde{x}$ , which of course is different in general from the “true” solution. The “residual errors” of  $\tilde{\eta}$  can be obtained with (9.2.30),

$$\begin{aligned} G_{\tilde{\eta}}^{-1} = C_{\tilde{\eta}} &= \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{pmatrix} \begin{pmatrix} 0.0943 & -0.0566 \\ -0.0566 & 0.0493 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 \end{pmatrix} \\ &= \begin{pmatrix} 0.0943 & 0.0377 & -0.0189 & -0.0755 \\ 0.0377 & 0.0305 & 0.0232 & 0.0160 \\ -0.0189 & 0.0232 & 0.0653 & 0.1074 \\ -0.0755 & 0.0160 & 0.1074 & 0.1988 \end{pmatrix} \quad . \end{aligned}$$

The square roots of the diagonal elements are

$$\Delta\tilde{\eta}_1 = 0.31 \quad , \quad \Delta\tilde{\eta}_2 = 0.17 \quad , \quad \Delta\tilde{\eta}_3 = 0.26 \quad , \quad \Delta\tilde{\eta}_4 = 0.45 \quad .$$

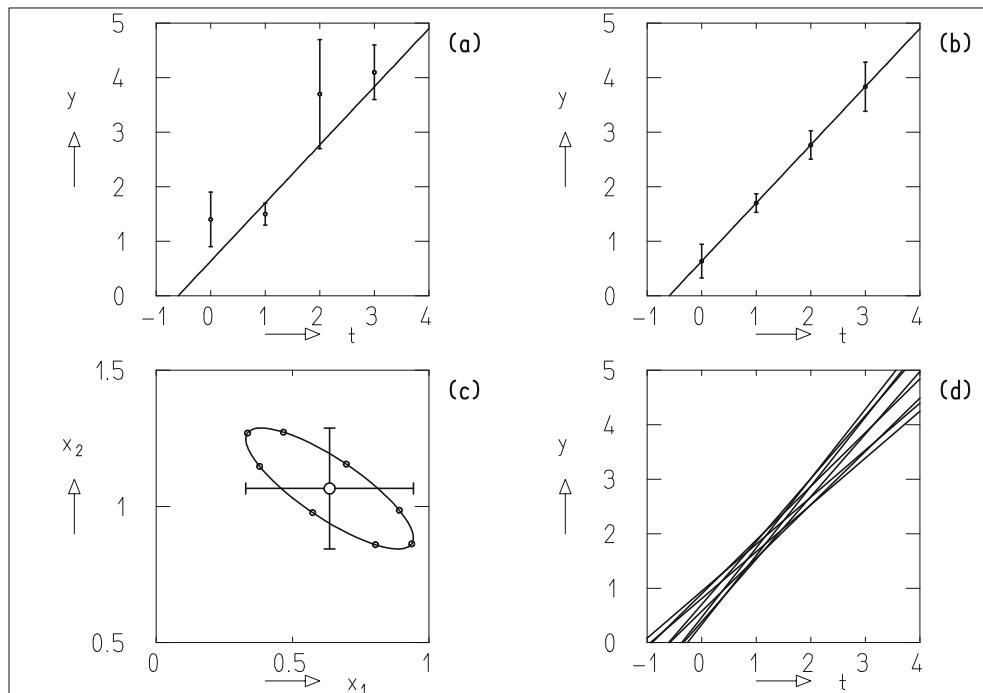
The fit procedure, in which more measurements (four) were used than were necessary to determine the two unknowns, has noticeably reduced the individual errors of the measurements in comparison to the original values  $\sigma_j$ .

Finally we will compute the value of the minimum function (9.2.16),

$$M = \tilde{\epsilon}^T G_y \tilde{\epsilon} = (\mathbf{y} - \tilde{\eta})^T G_y (\mathbf{y} - \tilde{\eta}) = \left( \sum_{j=1}^n \frac{y_j - \tilde{\eta}_j}{\sigma_j} \right)^2 = 4.507 \quad .$$

With this result we can carry out a  $\chi^2$ -test on the goodness-of-fit of a straight line to our data. Since we started with  $n = 4$  measured values and have determined  $r = 2$  unknown parameters, we still have  $n - r = 2$  degrees of freedom available (cf. Sect. 9.7). If we choose a significance level of 5 %, we find from Table I.7 for two degrees of freedom  $\chi^2_{0.95} = 5.99$ . There is thus no reason to reject the hypothesis of a straight line.

The results of the fit are shown in Fig. 9.2. The measurements  $y_j$  are shown as functions of the variables  $t$ . The vertical bars give the measurement errors. They cover the range  $y_j \pm \sigma_j$ . The plotted line corresponds to the result  $\tilde{x}_1, \tilde{x}_2$ . The improved measurements lie on this line. They are shown in Fig. 9.2b together with the residual errors  $\Delta\tilde{\eta}_j$ . In order to illustrate the accuracy of the estimates  $\tilde{x}_1, \tilde{x}_2$ , we consider the covariance matrix  $C_{\tilde{x}}$ . It determines a covariance ellipse (Sect. 5.10) in a plane spanned by the variables  $x_1, x_2$ . This ellipse is shown in Fig. 9.2c. Points on the ellipse correspond to fits of equal probability. Each of these points determines a line in the  $(t, y)$  plane. Some of the points are indicated in Fig. 9.2c and the corresponding lines are plotted in Fig. 9.2d. The points on the covariance ellipse thus correspond to a bundle of lines. The line determined by the “true”



**Fig. 9.2:** Fit of a *straight line* to data from Table 9.2. (a) Original measured values and errors; (b) improved measurement values and residual errors; (c) covariance ellipse for the fitted quantities  $x_1, x_2$ ; (d) various lines corresponding to individual points on the covariance ellipse.

values of the unknowns lies in this bundle with the probability  $1 - e^{-1/2}$ ; cf. (5.10.18).

## 9.4 Algorithms for Fitting Linear Functions of the Unknowns

The starting point of the ideas in Sect. 9.2 was the assumption (9.2.6) of a linear relationship between the “true” values  $\eta$  of the measured quantities  $\mathbf{y}$  and the unknowns  $\mathbf{x}$ . We will write this relation in the form

$$\eta = \mathbf{h}(\mathbf{x}) = -\mathbf{a}_0 - A\mathbf{x} \quad , \quad (9.4.1)$$

or in terms of components,

$$\eta_j = h_j(\mathbf{x}) = -a_{0j} - A_{j1}x_1 - A_{j2}x_2 - \cdots - A_{jr}x_r \quad . \quad (9.4.2)$$

Often it is useful to consider the index  $j$  as specifying that the measurement  $y_j$  corresponds to a value  $t_j$  of a controlled variable, which is taken to be known without error. Then (9.4.2) can be written in the form

$$\eta_j = h(\mathbf{x}, t_j) \quad . \quad (9.4.3)$$

This relation describes a curve in the  $(t, \eta)$  plane. It is characterized by the parameters  $\mathbf{x}$ . The determination of the parameters is thus equivalent to fitting a curve to the measurements  $y_j = y(t_j)$ . Usually the individual measurements  $y_j$  are uncorrelated. The weight matrix  $G_y$  is diagonal; it has the Cholesky decomposition (9.2.20). One then has simply

$$A'_{jk} = A_{jk}/\sigma_j \quad , \quad c'_j = c/\sigma_j \quad ,$$

see (9.2.22).

### 9.4.1 Fitting a Polynomial

As a particularly simple but useful example of a function for (9.4.3), let us consider the relation

$$\eta_j = h_j = x_1 + x_2 t_j + x_3 t_j^2 + \cdots + x_r t_j^{r-1} \quad . \quad (9.4.4)$$

This is a polynomial in  $t_j$ , but is linear in the unknowns  $\mathbf{x}$ . The special case  $r = 2$  has been treated in detail in Sect. 9.3.

A comparison of (9.4.4) with (9.4.2) yields directly

$$a_{0j} = 0 \quad , \quad A_{j\ell} = -t_j^{\ell-1}$$

or more completely,

$$\mathbf{a}_0 = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \quad , \quad A = - \begin{pmatrix} 1 & t_1 & t_1^2 & \dots & t_1^{r-1} \\ 1 & t_2 & t_2^2 & \dots & t_2^{r-1} \\ \vdots & & & & \\ 1 & t_n & t_n^2 & \dots & t_n^{r-1} \end{pmatrix} .$$

The class `LsqPol` performs the fit of a polynomial to data.

### Example 9.2: Fitting of various polynomials

It is often interesting to fit polynomials of various orders to measured data, as long as the number of degrees of freedom of the fit is at least one, i.e.,  $n > \ell + 1$ . As a numerical example we will use measurements from an elementary particle physics experiment. Consider an investigation of elastic scattering of negative K mesons on protons with a fixed K meson energy. The distribution of the cosine of the scattering angle  $\Theta$  in the center-of-mass system of the collision is characteristic of the angular momentum of possible intermediate states in the collision process. If, in particular, the distribution is considered as a polynomial in  $\cos \Theta$ , the order of the polynomial can be used to determine the spin quantum numbers of such intermediate states.

The measured values  $y_j$  ( $j = 1, 2, \dots, 10$ ) are simply the numbers of collisions for which  $\cos \Theta$  was observed in a small interval around  $t_j = \cos \Theta_j$ . As measurement errors the statistical errors were used, i.e., the square roots of the number of observations. The data are given in Table 9.3. The results of the fit of polynomials of various orders are summarized in Table 9.4 and Fig. 9.3. With the  $\chi^2$ -test we can check successively whether a polynomial of order zero, one, ... gives a good fit to the data.

We see that the first two hypotheses (a constant and a straight line) are not in agreement with the experimental data. This can be seen in Fig. 9.3 and is also reflected in the values of the minimum function. The hypothesis  $r = 3$ , a second-order polynomial, gives qualitative agreement. Most of the measurements, however, do not fall on the fitted parabola within the error bars. The  $\chi^2$ -test fails with a significance level of 0.0001. For the hypotheses  $r = 4, 5$ , and 6, however, the agreement is very good. The fitted curves go through the error bars and are almost identical. The  $\chi^2$ -test does not call for a rejection of the hypothesis even at  $\alpha = 0.5$ . We can therefore conclude that a third order polynomial is sufficient to describe the data. An even more careful investigation of the question as to what order a polynomial must be used to describe the data is possible with orthogonal polynomials; cf. Sect. 12.1. ■

**Table 9.3:** Data for Example 9.2. One has  $\sigma_j = \sqrt{y_j}$ .

| $j$ | $t_j = \cos \Theta_j$ | $y_j$ |
|-----|-----------------------|-------|
| 1   | -0.9                  | 81    |
| 2   | -0.7                  | 50    |
| 3   | -0.5                  | 35    |
| 4   | -0.3                  | 27    |
| 5   | -0.1                  | 26    |
| 6   | 0.1                   | 60    |
| 7   | 0.3                   | 106   |
| 8   | 0.5                   | 189   |
| 9   | 0.7                   | 318   |
| 10  | 0.9                   | 520   |

**Table 9.4:** Summary of results from Example 9.2 ( $n = 10$  measured points,  $r$  parameters,  $f = n - r$  degrees of freedom).

| $r$ | $\tilde{x}_1$ | $\tilde{x}_2$ | $\tilde{x}_3$ | $\tilde{x}_4$ | $\tilde{x}_5$ | $\tilde{x}_6$ | $f$ | $M$    |
|-----|---------------|---------------|---------------|---------------|---------------|---------------|-----|--------|
| 1   | 57.85         |               |               |               |               |               | 9   | 833.55 |
| 2   | 82.66         | 99.10         |               |               |               |               | 8   | 585.45 |
| 3   | 47.27         | 185.96        | 273.61        |               |               |               | 7   | 36.41  |
| 4   | 37.94         | 126.55        | 312.02        | 137.59        |               |               | 6   | 2.85   |
| 5   | 39.62         | 119.10        | 276.49        | 151.91        | 52.60         |               | 5   | 1.68   |
| 6   | 39.88         | 121.39        | 273.19        | 136.58        | 56.90         | 16.72         | 4   | 1.66   |

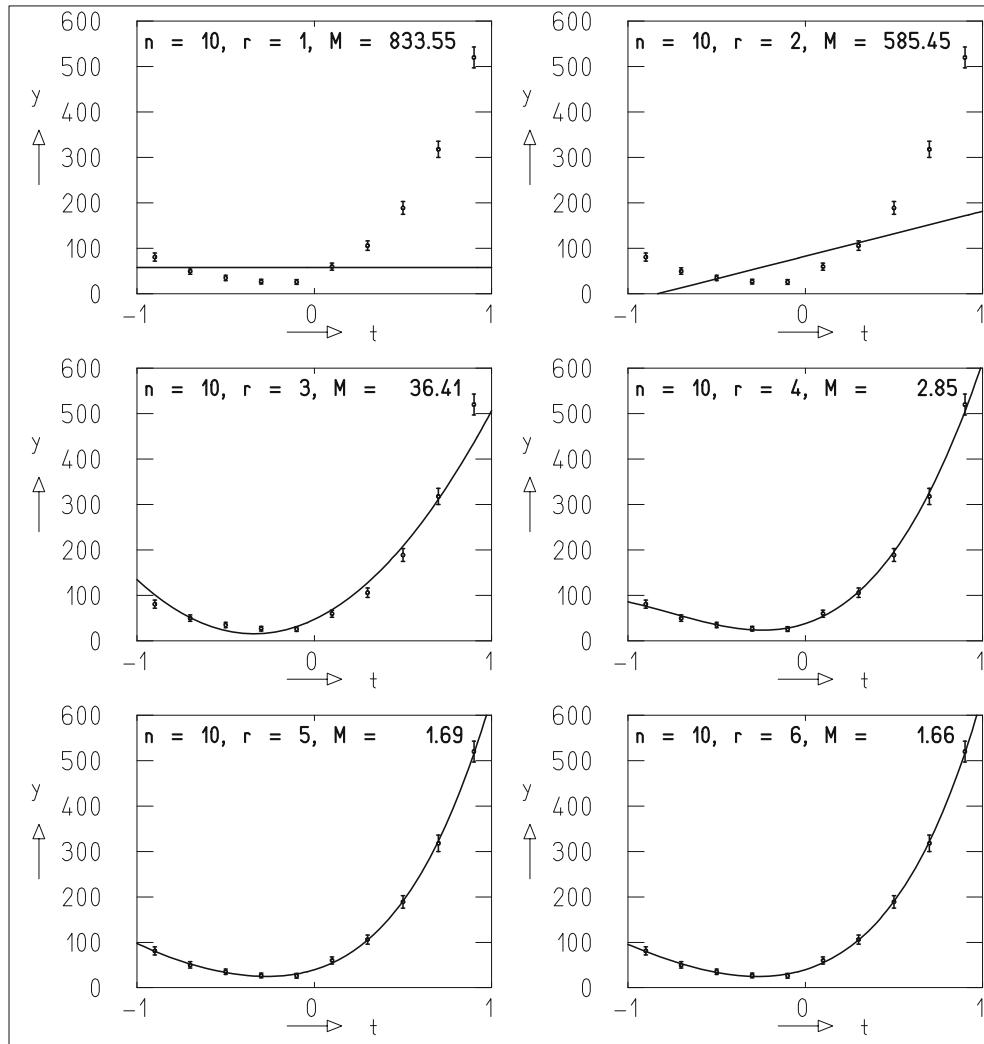
#### 9.4.2 Fit of an Arbitrary Linear Function

The matrix  $A$  and the vector  $\mathbf{c}$  enter into the solution of the problem of Sect. 9.2. They depend on the form of the function to be fitted and must therefore be provided by the user. (In Sect. 9.4.1 the function was known, so the user did not have to worry about computing  $A$  and  $\mathbf{c}$ .) The class `LsqLin` performs the fit of an arbitrary linear function to data.

#### Example 9.3: Fitting a proportional relation

Suppose from the construction of an experiment it is known that the true value  $\eta_j$  of the measurement  $y_j$  is directly proportional to the value of the controlled variable  $t_j$ :

$$\eta_j = x_1 t_j \quad .$$



**Fig. 9.3:** Fit of polynomials of various orders ( $0, 1, \dots, 5$ ) to the data from Example 9.2.

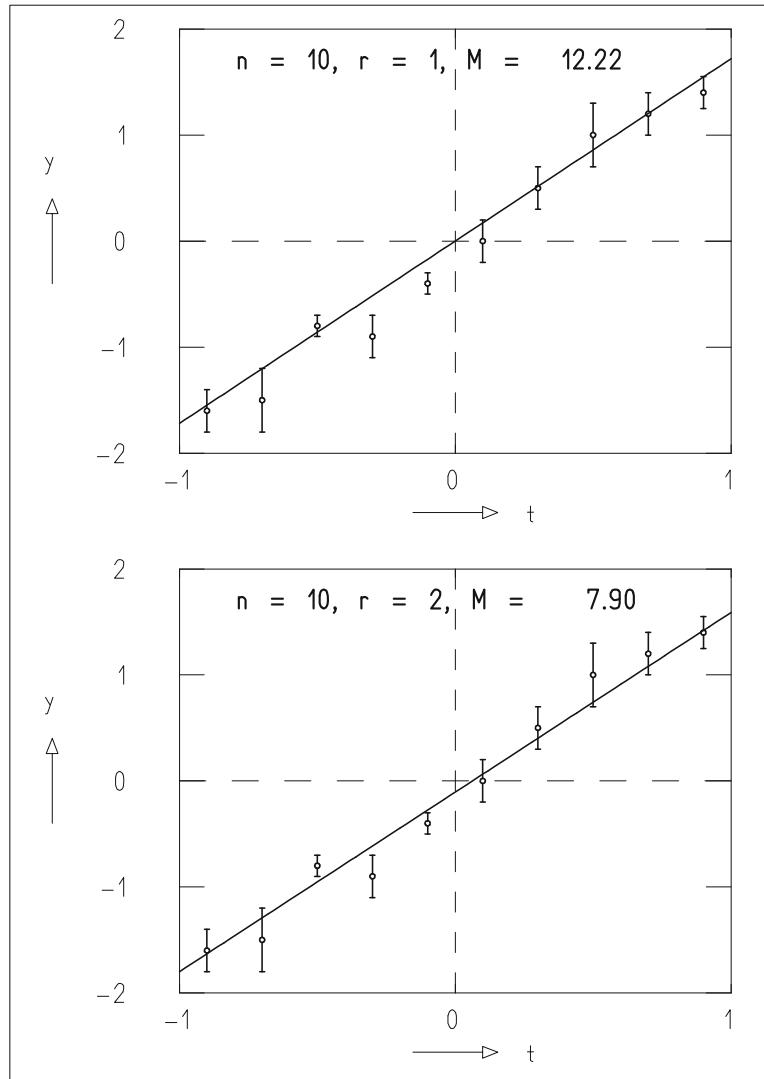
The constant of proportionality  $x_1$  is to be determined from the measurements. This relation is simpler than a first-order polynomial, which contains two constants. A comparison with (9.4.2) gives

$$a_{0j} = 0 \quad , \quad A_{j1} = -t_j$$

and thus

$$\mathbf{c} = - \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \quad , \quad A = - \begin{pmatrix} t_1 \\ \vdots \\ t_n \end{pmatrix} \quad .$$

Shown in Fig. 9.4 are the results of the fit of a line through the origin, i.e., a proportionality, and the fit of a first-order polynomial. The value of the



**Fig. 9.4:** Fit of a proportional relation (above) and a first-order polynomial (below) to data.

minimum function is clearly smaller in the case of the general line, and the fit is visibly better. The number of degrees of freedom, however, is less, and the desired constant of proportionality is not determined. ■

## 9.5 Indirect Measurements: Nonlinear Case

If the relation between the  $n$ -vector  $\eta$  of true values of the measured quantities  $y$  and the  $r$ -vector of unknowns is given by a function,

$$\eta = \mathbf{h}(\mathbf{x}) \quad ,$$

that is not – as (9.4.1) – linear in  $\mathbf{x}$ , then our previous procedure fails in the determination of the unknowns.

Instead of (9.2.2) one has

$$f_j(\mathbf{x}, \boldsymbol{\eta}) = \eta_j - h_j(\mathbf{x}) = 0 \quad , \quad (9.5.1)$$

or in vector notation,

$$\mathbf{f}(\mathbf{x}, \boldsymbol{\eta}) = 0 \quad . \quad (9.5.2)$$

We can, however, relate this situation to the linear case if we expand the  $f_j$  in a Taylor series and keep only the first term. We carry out the expansion about the point  $\mathbf{x}_0 = (x_{10}, x_{20}, \dots, x_{r0})$ , which is a first approximation for the unknowns, which has been obtained in some way,

$$f_j(\mathbf{x}, \boldsymbol{\eta}) = f_j(\mathbf{x}_0, \boldsymbol{\eta}) + \left( \frac{\partial f_j}{\partial x_1} \right)_{\mathbf{x}_0} (x_1 - x_{10}) + \cdots + \left( \frac{\partial f_j}{\partial x_r} \right)_{\mathbf{x}_0} (x_r - x_{r0}) \quad . \quad (9.5.3)$$

If we now define

$$\boldsymbol{\xi} = \mathbf{x} - \mathbf{x}_0 = \begin{pmatrix} x_1 - x_{10} \\ x_2 - x_{20} \\ \vdots \\ x_r - x_{r0} \end{pmatrix} \quad , \quad (9.5.4)$$

$$a_{j\ell} = \left( \frac{\partial f_j}{\partial x_\ell} \right)_{\mathbf{x}_0} = - \left( \frac{\partial h_j}{\partial x_\ell} \right)_{\mathbf{x}_0} \quad , \quad A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1r} \\ a_{21} & a_{22} & \cdots & a_{2r} \\ \vdots & & & \\ a_{n1} & a_{n2} & \cdots & a_{nr} \end{pmatrix} \quad , \quad (9.5.5)$$

$$c_j = f_j(\mathbf{x}_0, \mathbf{y}) = y_j - h_j(\mathbf{x}_0) \quad , \quad \mathbf{c} = \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{pmatrix} \quad , \quad (9.5.6)$$

and use the relation (9.2.10), then we obtain

$$f_j(\mathbf{x}_0, \boldsymbol{\eta}) = f_j(\mathbf{x}_0, \mathbf{y} - \boldsymbol{\varepsilon}) = f_j(\mathbf{x}_0, \mathbf{y}) - \boldsymbol{\varepsilon} \quad . \quad (9.5.7)$$

Thus we can now write the system of equations (9.5.2) in the form

$$\mathbf{f} = A\boldsymbol{\xi} + \mathbf{c} - \boldsymbol{\varepsilon} = 0 \quad , \quad \boldsymbol{\varepsilon} = A\boldsymbol{\xi} + \mathbf{c} \quad . \quad (9.5.8)$$

The least-squares condition (9.2.16) is then

$$M = (\mathbf{c} + A\boldsymbol{\xi})^T G_y (\mathbf{c} + A\boldsymbol{\xi}) = \min \quad (9.5.9)$$

in complete analogy to (9.2.19). Using the quantities defined in (9.2.20) through (9.2.22) we can obtain the solution directly from (9.2.24),

$$\tilde{\xi} = -A'^+ \mathbf{c}' . \quad (9.5.10)$$

Corresponding to (9.5.4) one can use  $\tilde{\xi}$  to find a better approximation

$$\mathbf{x}_1 = \mathbf{x}_0 + \tilde{\xi} \quad (9.5.11)$$

and compute new values of  $A$  and  $\mathbf{c}$  at the point  $\mathbf{x}_1$ . A solution  $\tilde{\xi}$  (9.5.10) for (9.5.9) with these values gives  $\mathbf{x}_2 = \mathbf{x}_1 + \tilde{\xi}$ , etc. This iterative procedure can be terminated when the minimum function in the last step has not significantly decreased in comparison to the result of the previous step.

There is no guaranty, however, for the convergence of this procedure. Heuristically it is clear, however, that the expectation for convergence is greater when the Taylor series, truncated after the first term, is a good approximation in the region over which  $\mathbf{x}$  is varied in the procedure. This is at least the region between  $\mathbf{x}_0$  and the solution  $\tilde{\mathbf{x}}$ . (Intermediate steps can also lie outside of this region.) It is therefore important, particularly in highly nonlinear problems, to start from a good first approximation.

If the solution  $\tilde{\mathbf{x}} = \mathbf{x}_n = \mathbf{x}_{n-1} + \tilde{\xi}$  is reached in  $n$  steps, then it can be expressed as a linear function of  $\tilde{\xi}$ . Using error propagation one finds that the covariance matrices of  $\tilde{\mathbf{x}}$  and  $\tilde{\xi}$  are then identical and one finds that

$$C_{\tilde{\mathbf{x}}} = G_{\tilde{\mathbf{x}}}^{-1} = (A^T G_y A)^{-1} = (A'^T A')^{-1} . \quad (9.5.12)$$

The covariance matrix loses its validity, however, if the linear approximation (9.5.3) is not a good description in the region  $\tilde{x}_i \pm \Delta x_i$ ,  $i = 1, \dots, r$ . (Here  $\Delta x_i = \sqrt{C_{ii}}$ .)

## 9.6 Algorithms for Fitting Nonlinear Functions

It is sometimes useful to set one or several of the  $r$  unknown parameters equal to given values, i.e., to treat them as constants and not as adjustable parameters. This can clearly be done by means of a corresponding definition of  $\mathbf{f}$  in (9.5.1). For the user, however, it is more convenient to write only one subprogram that computes the function  $\mathbf{f}$  for the  $r$  original parameters and when needed, to communicate to the program that the number of adjustable parameters is to be reduced from  $r$  to  $r'$ . Of course a list with  $r$  elements  $\ell_i$  must also be given, in which one specifies which parameters  $x_i$  are to be held constant ( $\ell_i = 0$ ) and which should remain adjustable ( $\ell_i = 1$ ).

Two more difficulties come up when implementing the considerations of the previous section in a program:

On the one hand, the elements of the matrix  $A$  must be found by constructing the function to be fitted and differentiating it with respect to the parameters. Of course it is particularly convenient for the user if he does not have to program these derivatives himself, but rather can turn this task over to a routine for numerical differentiation. In Sect. E.1 we provide such a subroutine, which we will call from the program discussed below. Numerical differentiation implies, however, a loss of precision and an increase in computing time. In addition, the method can fail. Our programs communicate such an occurrence by means of an output parameter. Thus in some cases the user will be forced to program the derivatives by hand.

The second difficulty is related to the fact that the minimum function

$$M = (\mathbf{y} - \mathbf{h}(\mathbf{x}))^T G_y (\mathbf{y} - \mathbf{h}(\mathbf{x})) \quad (9.6.1)$$

is no longer a simple quadratic form of the unknowns like (9.2.17) and (9.2.23). One consequence of this is that the position of the minimum  $\tilde{\mathbf{x}}$  cannot be reached in a single step. In addition, the convergence of the iterative procedure strongly depends on whether the first approximation  $\mathbf{x}_0$  is in a region where the minimum function is sufficiently similar to a quadratic form. The determination of a good first approximation must be handled according to the given problem. Some examples are given below. If one constructs the iterative procedure as indicated in the previous section, then it can easily happen that the minimum function does not decrease with every step. In order to ensure convergence despite this, two methods can be applied, which are described in Sects. 9.6.1 and 9.6.2. The first (reduction of step size) is simpler and faster. The second (the Marquardt procedure) has, however, a larger region of convergence. We give programs for both methods, but recommend applying Marquardt procedure in cases of doubt.

### 9.6.1 Iteration with Step-Size Reduction

As mentioned, the inequality

$$M(\mathbf{x}_i) = M(\mathbf{x}_{i-1} + \tilde{\xi}) < M(\mathbf{x}_{i-1}) \quad (9.6.2)$$

does not hold in every case for the result  $\mathbf{x}_i$  of step  $i$ . The following consideration helps us in handling such steps. Let us consider the expression  $M(\mathbf{x}_{i-1} + s\tilde{\xi})$  as a function of the quantity  $s$  with  $0 \leq s \leq 1$ . If we replace  $\tilde{\xi}$  by  $s\tilde{\xi}$  in (9.5.9), then we obtain

$$M = (\mathbf{c} + s A \tilde{\xi})^T G_y (\mathbf{c} + s A \tilde{\xi}) = (\mathbf{c}' + s A' \tilde{\xi})^2 \quad .$$

Differentiating with respect to  $s$  gives

$$M' = 2(\mathbf{c}' + sA'\tilde{\xi})^T A'\tilde{\xi}$$

or with  $\mathbf{c}' = -A'\tilde{\xi}$ , cf. (9.5.10),

$$M' = 2(s-1)\tilde{\xi}^T A'^T A'\tilde{\xi} \quad ,$$

and thus

$$M'(s=0) < 0 \quad ,$$

if  $A'^T A' = A^T G_y A$  is positive definite. (This is always the case near the minimum.) The matrix  $A'^T A'$  gives the curvature of the function  $M$  in the space spanned by the unknowns  $x_1, \dots, x_r$ . For only one such unknown, the region of positive curvature is the region between the points of inflection around the minimum (cf. Fig. 10.1). Since the function  $M$  is continuous in  $s$ , there exists a value  $\lambda > 0$  such that

$$M'(s) < 0 \quad , \quad 0 \leq s \leq \lambda \quad .$$

After an iteration  $i+1$ , for which (9.6.2) does not hold, one multiplies  $\tilde{\xi}$  by a number  $s$ , e.g.,  $s = 1/2$ , and checks whether

$$M(x_{i-1} + s\tilde{\xi}) < M(x_{i-1}) \quad .$$

If this is the case, then one sets  $\mathbf{x}_i = \mathbf{x}_{i-1} + s\tilde{\xi}$ . If it is not the case, then one multiplies again with  $s$ , and so forth.

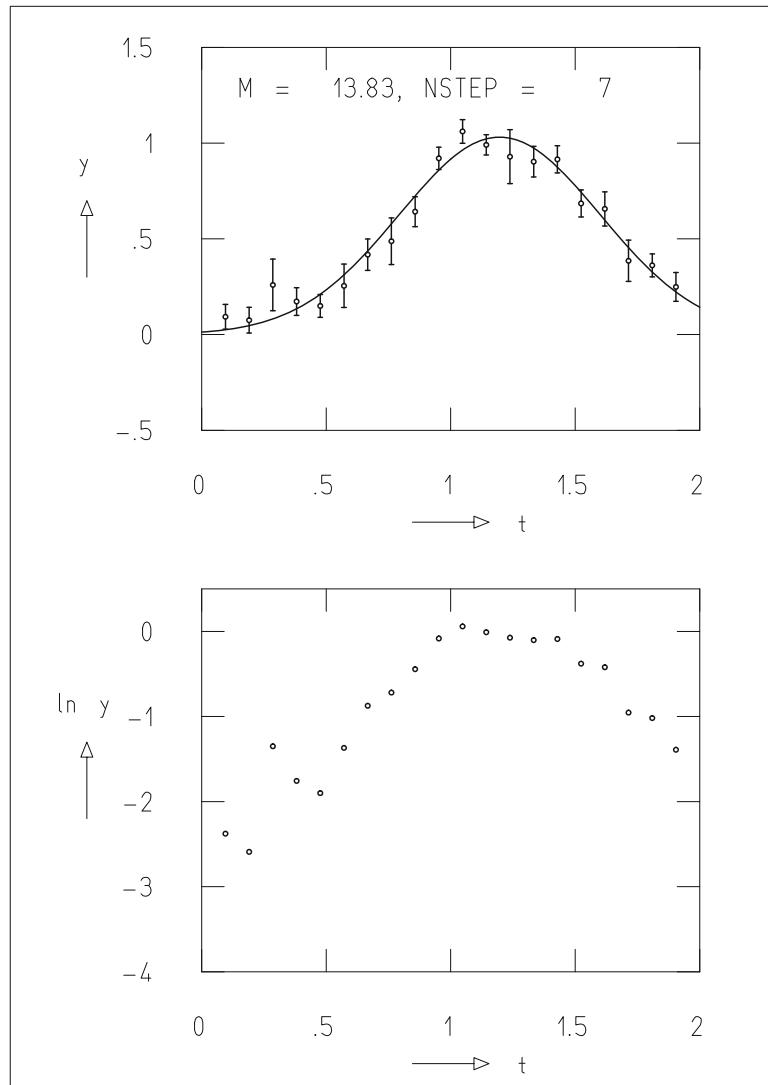
The class LsqNon operates according to this procedure of iteration with step-size reduction. As in the linear case we consider the measurements as being dependent on a controlled variable  $t$ , i.e.,  $y_j = y_j(t_j)$ . For the true values  $\eta_j$  corresponding to the measurements one has

$$\eta_j = h(\mathbf{x}, t_j)$$

or [cf. (9.6.1)]

$$\eta = h(\mathbf{x}, t) \quad .$$

This function has to be programmed by the user. That is done within an extension of the abstract class DataUserFunction, see the example programs in Sect. 9.14. The matrix  $A$  is computed by numerical differentiation with the class AuxDri. Should the accuracy of that method not suffice, the user has to provide a method with the same name and the same method declarations, computing the matrix  $A$  by analytic differentiation.



**Fig. 9.5:** Measured values and fitted Gaussian curve (above); logarithm of the measured values (below).

### Example 9.4: Fitting a Gaussian curve

In many experiments one has signals  $y(t)$  that have the form of a Gaussian curve,

$$y(t) = x_1 \exp(-(t - x_2)^2 / 2x_3^2) \quad . \quad (9.6.3)$$

One wishes to determine the parameters  $x_1$ ,  $x_2$ ,  $x_3$  that give the amplitude, position of the maximum, and the width of the signal.

Figure 9.5 shows the result of the fit to data. The values  $x_1 = 0.5$ ,  $x_2 = 1.5$ ,  $x_3 = 0.2$  were used as a first approximation. In fact we could have estimated significantly better initial values directly from the plot of the measurements,

i.e.,  $x_1 \approx 1$  for the amplitude,  $x_2 \approx 1.25$  for the position of the maximum, and  $x_3 \approx 0.4$  for the width.

Here we have just mentioned a particularly useful aid in determining the initial approximations: the analysis of a plot of the data by eye. One can also proceed more formally, however, and consider the logarithm of (9.6.3),

$$\begin{aligned}\ln y(t) &= \ln x_1 - \frac{(t - x_2)^2}{2x_3^2} \\ &= \left( \ln x_1 - \frac{x_2^2}{2x_3^2} \right) + t \left( \frac{x_2}{x_3^2} \right) - t^2 \left( \frac{1}{2x_3^2} \right) .\end{aligned}$$

This is a polynomial in  $t$ , which is linear in the three terms in parentheses,

$$a_1 = \ln x_1 - \frac{x_2^2}{2x_3^2} , \quad a_2 = \frac{x_2}{x_3^2} , \quad a_3 = -\frac{1}{2x_3^2} .$$

By taking the logarithm, however, the distribution for the measurement errors, originally Gaussian, has been changed, so that strictly speaking one cannot fit a polynomial to determine  $a_1$ ,  $a_2$ , and  $a_3$ . For purposes of determining the first approximation we can disregard this difficulty and set the errors of the  $\ln y(t_i)$  equal to one. We then determine the quantities  $a_1$ ,  $a_2$ ,  $a_3$ , and from them the values  $x_1$ ,  $x_2$ ,  $x_3$ , and use these as initial values.

The logarithms of the measured values  $y_i$  are shown in the lower part of Fig. 9.5. One can see that fit range for the parabola must be restricted to points in the bell-shaped region of the curve, since the points in the tails are subject to large fluctuations. ■

### Example 9.5: Fit of an exponential function

In studies of radioactivity, for example, a function of the form

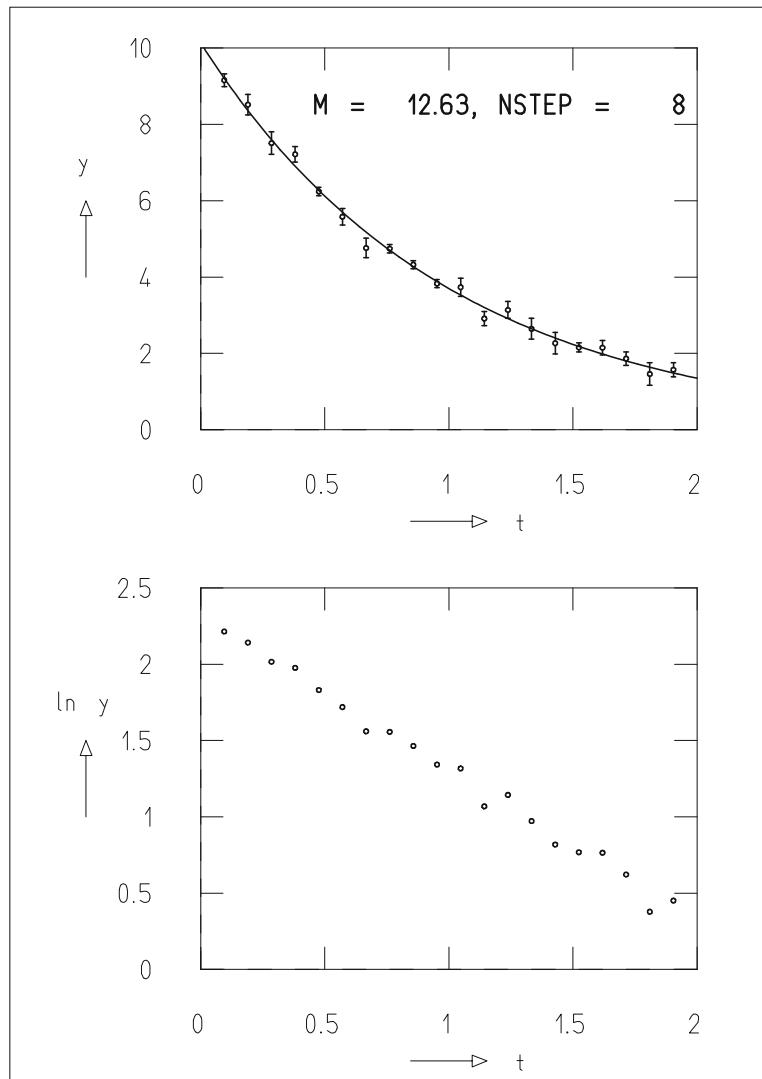
$$y(t) = x_1 \exp(-x_2 t) \tag{9.6.4}$$

must be fitted to measured values  $y_i(t_i)$ . The program to be provided here by the user can have the following form.

In Fig. 9.6 the result of a fit to data is shown. The determination of the first approximation for the unknowns can again be obtained by fitting a straight line (graphically or numerically) to the logarithm of the function  $y(t)$ ,

$$\ln y(t) = \ln x_1 - x_2 t .$$

Here one usually uses only the values at small  $t$ , since these points have smaller fluctuations. ■



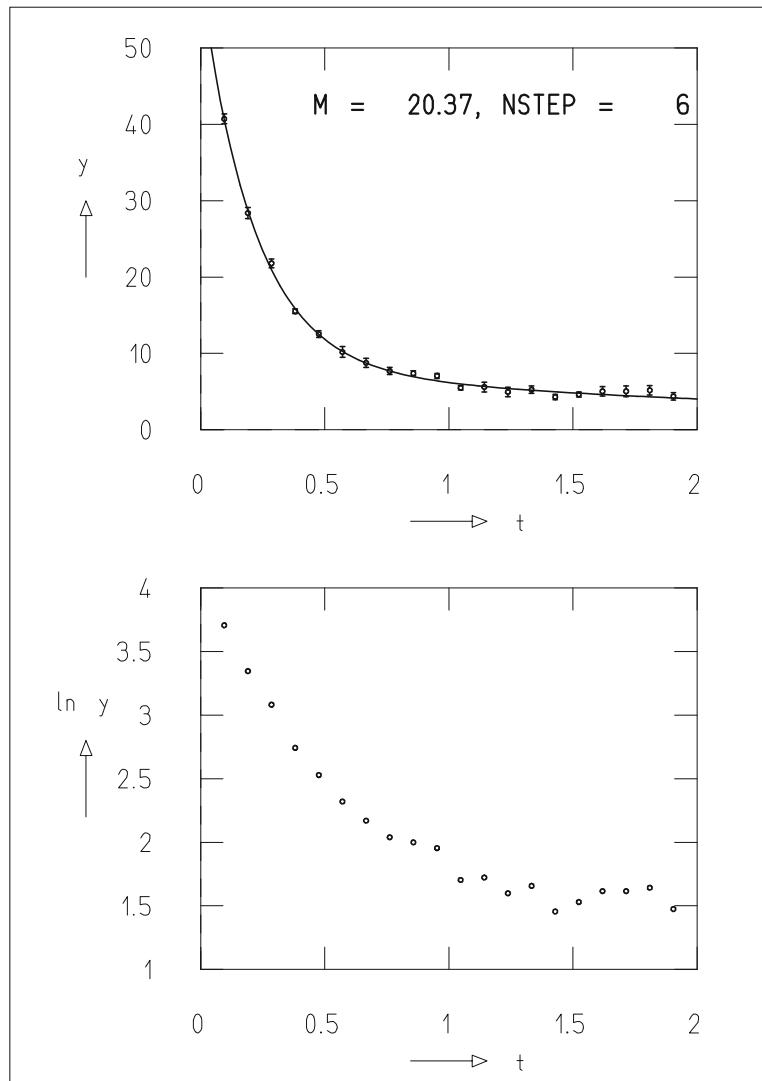
**Fig. 9.6:** Measured values and fitted exponential function (above); logarithms of the measured values (below).

### Example 9.6: Fitting a sum of exponential functions

A radioactive substance often consists of a mixture of components with different decay times. One must therefore fit a sum of several exponential functions. We will consider the case of two functions

$$y(t) = x_1 \exp(-x_2 t) + x_3 \exp(-x_4 t) . \quad (9.6.5)$$

Figure 9.7 shows the result of fitting a sum of two exponential functions to the data. A first approximation can be determined by fitting two different lines to  $\ln y_i(t_i)$  for the regions of smaller and larger  $t_i$ . ■



**Fig. 9.7:** Measured values with the fitted sum of two exponential functions (*above*); logarithm of the measured values (*below*).

### 9.6.2 Marquardt Iteration

The procedure with step-size reduction discussed in Sect. 9.6.1 leads to a minimum when  $\mathbf{x}$  is already in a region where  $A^T A'$  is positive definite, i.e., in a region around the minimum where the function  $M(\mathbf{x})$  has a positive curvature. (In the one-dimensional case of Fig. 10.1, this is the region between the two points of inflection on either side of the minimum.) It is clear, however, that it must be possible to extend the region of convergence to the region between the two maxima surrounding the minimum. This possibility is offered by the Marquardt procedure, which is presented in Sect. 10.15 in a somewhat more

general form. The class LSqMar treats the nonlinear case of least squares; the user's task is much the same as in the case of LsqNon. Readers who want to become familiar with this class should first study Sect. 10.15 and the introductory sections of Chap. 10 and finally Sect. A.17

**Example 9.7:** Fitting a sum of two Gaussian functions and a polynomial

In practice it is usually not so easy to find the amplitude, position, and width of signals as was shown in Example 9.4. One usually has more than one signal, lying on a background that varies slowly with the controlled variable  $t$ . Since in general this background is not well known, it is approximated by a line or a second-order polynomial. We will consider the sum of such a polynomial and two Gaussian distributions, i.e., a function of nine unknown parameters,

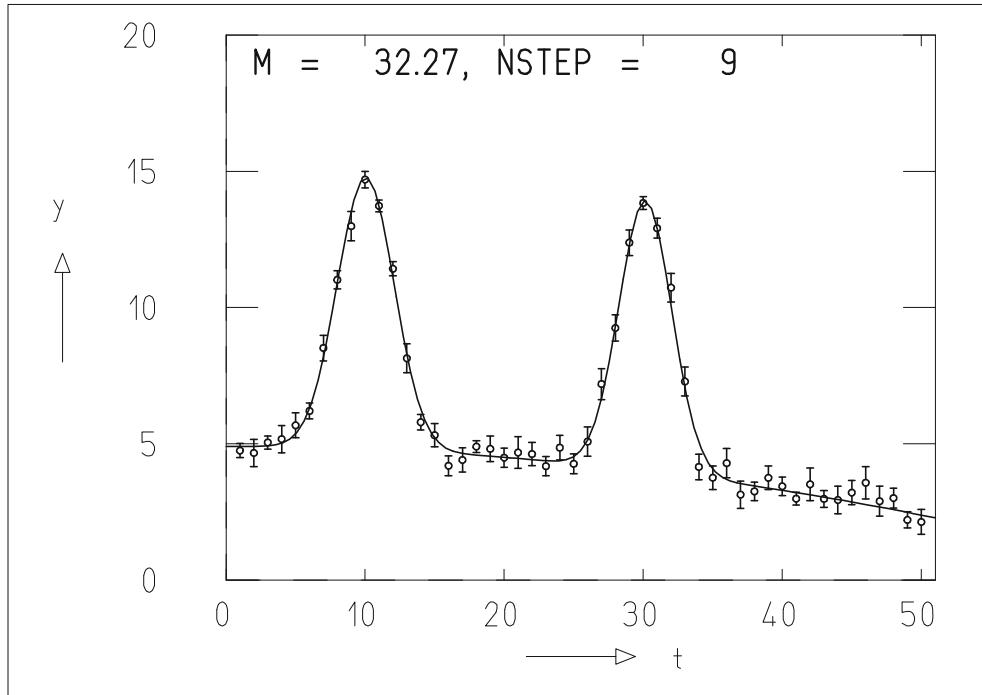
$$h(\mathbf{x}, t) = x_1 + x_2 t + x_3 t^2 + x_4 \exp\{-(x_5 - t)^2 / 2x_6^2\} + x_7 \exp\{-(x_8 - t)^2 / 2x_9^2\} . \quad (9.6.6)$$

The derivatives (9.5.5) are

$$\begin{aligned} -a_{j1} &= \frac{\partial h_j}{\partial x_1} = 1 , \\ -a_{j2} &= \frac{\partial h_j}{\partial x_2} = t_j , \\ -a_{j3} &= \frac{\partial h_j}{\partial x_3} = t_j^2 , \\ -a_{j4} &= \frac{\partial h_j}{\partial x_4} = \exp\{-(x_5 - t_j)^2 / 2x_6^2\} , \\ -a_{j5} &= \frac{\partial h_j}{\partial x_5} = 2x_4 \exp\{-(x_5 - t_j)^2 / 2x_6^2\} \frac{t_j - x_5}{2x_6^2} , \\ -a_{j6} &= \frac{\partial h_j}{\partial x_6} = x_4 \exp\{-(x_5 - t_j)^2 / 2x_6^2\} \frac{(t_j - x_5)^2}{x_6^3} , \\ &\vdots . \end{aligned}$$

If the numerical differentiation fails, the derivatives must be computed with a specially written version of the routineAuxDri. For the numerical example shown in Fig. 9.8, however, this is not necessary. The user must supply, of course, an extension of DataUserFunction to compute (9.6.6).

Figure 9.8 shows the result of fitting to a total of 50 measurements. It is perhaps interesting to look at some of the intermediate steps in Fig. 9.9. As a first approximation the parameters of the polynomial were set to zero ( $x_1 = x_2 = x_3 = 0$ ). For both of the clearly visible signals, rough estimates of the amplitude, position, and width were used for the initial values. The function (9.6.6) with the parameters of this first approximation is shown in the first

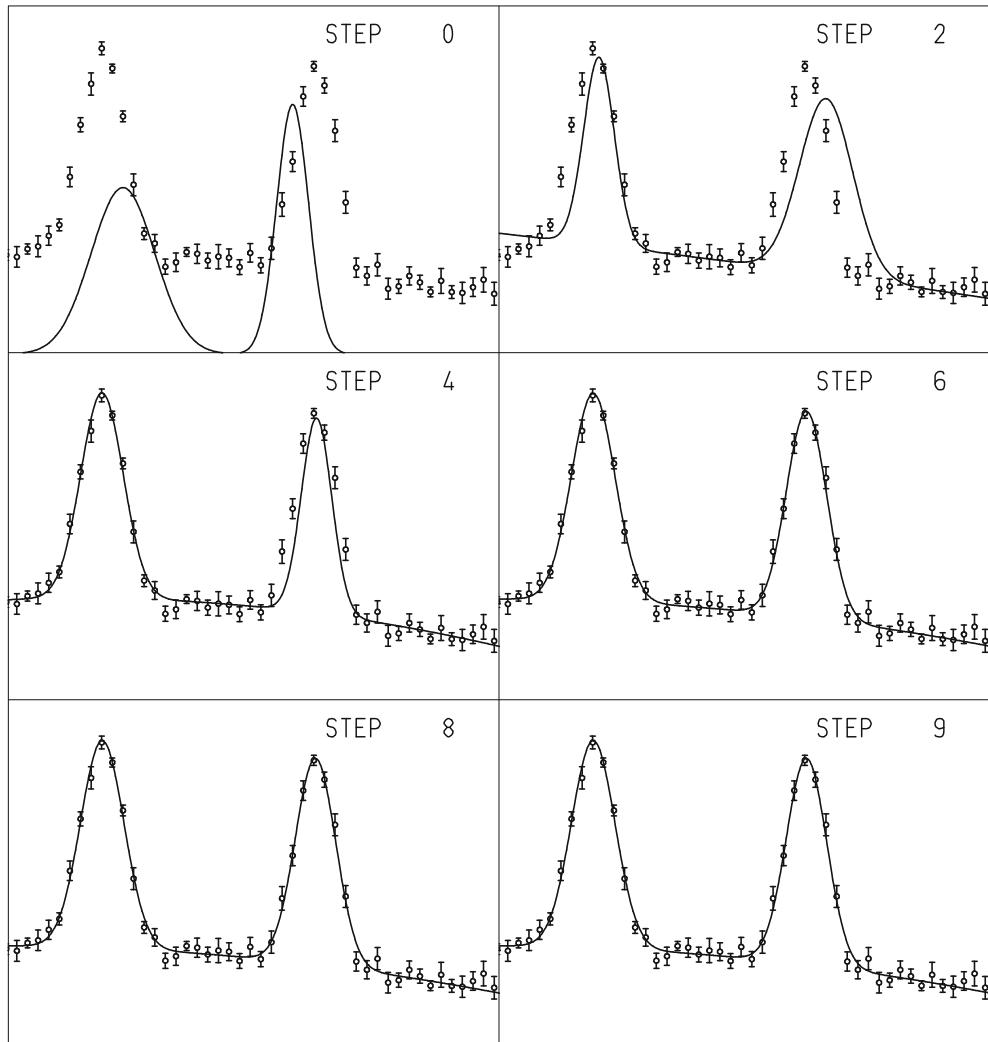


**Fig. 9.8:** Measured values and fitted sum of a second-order polynomial and two Gaussian functions.

frame(STEP 0). In the following frames one sees the change of the function after 2, 4, 6, and 8 steps, and finally at convergence of the class LsqMar after a total of 9 steps. ■

## 9.7 Properties of the Least-Squares Solution: $\chi^2$ -Test

Up to now the method of least squares has merely been an application of the maximum-likelihood method to a linear problem. The prescription of least squares (9.2.15) was obtained directly from the maximization of the likelihood function (9.2.14). In order to be able to specify this likelihood function in the first place, it was necessary to know the distribution of the measurement errors. We assumed a normal distribution. But also when there is no exact knowledge of the error distribution, one can still apply the relation (9.2.15) and with it the remaining formulas of the last sections. Such a procedure seems to lack a theoretical justification. The Gauss–Markov theorem, however, states that in this case as well, the method of least squares provides results with desirable properties. Before entering into this, let us list once more the properties of the maximum-likelihood solution.



**Fig. 9.9:** Successive approximations of the fit function to the measurements.

(a) The solution  $\tilde{\mathbf{x}}$  is asymptotically unbiased, i.e., for very large samples,

$$E(\tilde{x}_i) = x_i \quad , \quad i = 1, 2, \dots, r \quad .$$

(b) It is a minimum variance estimator, i.e.,

$$\sigma^2(\tilde{x}_i) = E\{(\tilde{x}_i - x_i)^2\} = \min \quad .$$

(c) The quantity (9.2.16)

$$M = \boldsymbol{\varepsilon}^T G_y \boldsymbol{\varepsilon}$$

follows a  $\chi^2$ -distribution with  $n - r$  degrees of freedom.

The properties (a) and (b) are familiar from Chap. 7. We will demonstrate the validity of (c) for the simple case of direct measurements ( $r = 1$ ), for which the matrix  $G_y$  is diagonal:

$$G_y = \begin{pmatrix} 1/\sigma_1^2 & & & 0 \\ & 1/\sigma_2^2 & & \\ & & \ddots & \\ 0 & & & 1/\sigma_n^2 \end{pmatrix} .$$

The quantity  $M$  then simply becomes a sum of squares,

$$M = \sum_{j=1}^n \varepsilon_j^2 / \sigma_j^2 . \quad (9.7.1)$$

Since each  $\varepsilon_j$  comes from a normally distributed population with mean of zero and variance  $\sigma_j^2$ , the quantities  $\varepsilon_j^2 / \sigma_j^2$  are described by a standard Gaussian distribution. Thus the sum of squares follows a  $\chi^2$ -distribution with  $n - 1$  degrees of freedom.

If the distribution of the errors  $\varepsilon_j$  is not known, then the least-squares solution has the following properties:

- (a) The solution is unbiased.
- (b) Of all solutions  $\mathbf{x}^*$  that are unbiased estimators for  $\mathbf{x}$  and are linear combinations of the measurements  $\mathbf{y}$ , the least-squares solution has the smallest variance. (This is the GAUSS–MARKOV theorem.)
- (c) The expectation value of

$$M = \mathbf{\epsilon}^T G_y \mathbf{\epsilon}$$

is

$$E(M) = n - r .$$

(This is exactly the expectation value of a  $\chi^2$  variable for  $n - r$  degrees of freedom.)

The quantity  $M$  is often simply called  $\chi^2$ , although it does not necessarily follow a  $\chi^2$ -distribution. Together with the matrices  $C_{\tilde{x}}$  and  $C_{\tilde{\eta}}$  it provides a convenient measure of the quality of a fit with the least-squares method. If the value of  $M$  obtained from the measurements comes out much larger than  $n - r$ , then the assumptions of the calculation must be carefully checked. The result should not simply be accepted without critical examination.

The number  $f = n - r$  is called the *number of degrees of freedom of a fit* or the *number of equations of constraint of the fit*. It is clear from the

beginning (see Appendix A), that the problem of least squares can only be solved for  $f \geq 0$ . Only for  $f > 0$ , however, is the quantity  $M$  meaningfully defined and usable for testing the quality of a fit.

If it is known that the errors are normally distributed, then a  $\chi^2$ -test can be done in conjunction with the fit. One rejects the result of the fit if

$$M = \boldsymbol{\varepsilon}^T G_y \boldsymbol{\varepsilon} > \chi_{1-\alpha}^2(n-r) , \quad (9.7.2)$$

i.e., if the quantity  $M$  exceeds the quantile of the  $\chi^2$ -distribution with  $n - r$  degrees of freedom at a significance level of  $\alpha$ . A larger value of  $M$  can be caused by the following reasons (and also by an error of the first kind):

- (a) The assumed functional dependence  $\mathbf{f}(\mathbf{x}, \boldsymbol{\eta}) = 0$  between the measured values  $\boldsymbol{\eta}$  and the unknown parameters  $\mathbf{x}$  is false. Either the function  $\mathbf{f}(\mathbf{x}, \boldsymbol{\eta})$  is completely wrong, or some of the parameters taken to be known are not correct.
- (b) The function  $\mathbf{f}(\mathbf{x}, \boldsymbol{\eta})$  is correct, but the series expansion with only one term is not a sufficiently good approximation in the region of parameter space covered in the computation.
- (c) The initial approximation  $\mathbf{x}_0$  is too far from the true value  $\mathbf{x}$ . Better values for  $\mathbf{x}_0$  could lead to more acceptable values of  $M$ . This point is clearly related to (b).
- (d) The covariance matrix of the measurements  $C_y$ , which is often only based on rough estimates or assumptions, is not correct.

These four points must be carefully taken into consideration if the method of least squares is to be applied successfully. In many cases the least-squares computation is repeated many times for different data sets. One can then look at the empirical frequency distribution of the quantity  $M$  and compare it to a  $\chi^2$ -distribution with the corresponding number of degrees of freedom. Such a comparison is particularly useful for a good estimate of  $C_y$ . At the start of a new experiment, the apparatus that provides the measured values  $\mathbf{y}$  is usually checked by measuring known quantities. In this way the parameters  $\mathbf{x}$  for several data sets are determined. The covariance matrix of the measurements can then be determined such that the distribution (i.e., the histogram) of  $M$  agrees as well as possible with a  $\chi^2$ -distribution. This investigation is particularly illustrative when one considers the distribution of  $F(M)$  instead of that of  $M$ , where  $F$  is the distribution function (6.6.11) of the  $\chi^2$ -distribution. If  $M$  follows a  $\chi^2$ -distribution, then  $F(M)$  follows a uniform distribution.

## 9.8 Confidence Regions and Asymmetric Errors in the Nonlinear Case

We begin this section with a brief review of the meaning of the covariance matrix  $C_{\tilde{x}}$  of the unknown parameters for the linear least-squares case. The probability that the true value of the unknowns is given by  $\mathbf{x}$  is given by a normal distribution of the form (5.10.1),

$$\phi(\mathbf{x}) = k \exp\left\{-\frac{1}{2}(\mathbf{x} - \tilde{\mathbf{x}})^T B(\mathbf{x} - \tilde{\mathbf{x}})\right\} , \quad (9.8.1)$$

where  $B = C_{\tilde{x}}^{-1}$ . The exponent (multiplied by  $-2$ ),

$$g(\mathbf{x}) = (\mathbf{x} - \tilde{\mathbf{x}})^T B(\mathbf{x} - \tilde{\mathbf{x}}) , \quad (9.8.2)$$

for  $g = 1 = \text{const}$  describes the covariance ellipsoid (cf. Sect. 5.10). For other values of  $g(\mathbf{x}) = \text{const}$  one obtains confidence ellipsoids for a given probability  $W$  [see (5.10.19)]. In this way confidence ellipsoids can easily be constructed that have the point  $\mathbf{x} = \tilde{\mathbf{x}}$  at the center and which contain the true value  $\mathbf{x}$  of the unknown with probability  $W = 0.95$ .

We can now find a relation between the confidence ellipsoid and the minimum function. For this we use the expression (9.2.19), which is exactly true in the linear case, and we compute the difference between the minimum function  $M(\mathbf{x})$  at a point  $\mathbf{x}$  and the minimum function  $M(\tilde{\mathbf{x}})$  at the point  $\tilde{\mathbf{x}}$  where it is a minimum,

$$M(\mathbf{x}) - M(\tilde{\mathbf{x}}) = (\mathbf{x} - \tilde{\mathbf{x}})^T A^T G_y A(\mathbf{x} - \tilde{\mathbf{x}}) . \quad (9.8.3)$$

According to (9.2.27), one has

$$B = C_{\tilde{x}}^{-1} = G_{\tilde{x}} = A^T G_y A . \quad (9.8.4)$$

So the difference (9.8.3) is exactly the function introduced above,  $g(\mathbf{x})$ .

Thus the covariance ellipsoid is in fact the hypersurface in the space of the  $r$  variables  $x_1, \dots, x_r$ , where the function  $M(\mathbf{x})$  has the value  $M(\mathbf{x}) = M(\tilde{\mathbf{x}}) + 1$ . Correspondingly, the confidence ellipsoid with probability  $W$  is the hypersurface for which

$$M(\mathbf{x}) = M(\tilde{\mathbf{x}}) + g , \quad (9.8.5)$$

where the constant  $g$  is the quantile of the  $\chi^2$ -distribution with  $f = n - r$  degrees of freedom according to (5.10.21),

$$g = \chi_W^2(f) . \quad (9.8.6)$$

In the nonlinear case of least squares, our considerations remain approximately valid. The approximation becomes better when the nonlinear deviations of the expression (9.5.3) are less in the region where the unknown

parameters are varied. The deviations are not only small when the derivatives are almost constant, i.e., the function is almost linear, but also when the variation of the unknown parameters is small. A measure of the variation of an unknown is, however, its error. Because of error propagation, the errors of the unknowns are small when the original errors of the measurements are small. In the nonlinear case, therefore, the covariance matrix retains the same interpretation as in the linear case as long as the measurement errors are small.

If the measurement errors are large we retain the interpretation (9.8.5), i.e., we determine the hypersurface with (9.8.5) and state that the true value  $\mathbf{x}$  lies with probability  $W$  within the *confidence region* around the point. This region is, however, no longer an ellipsoid.

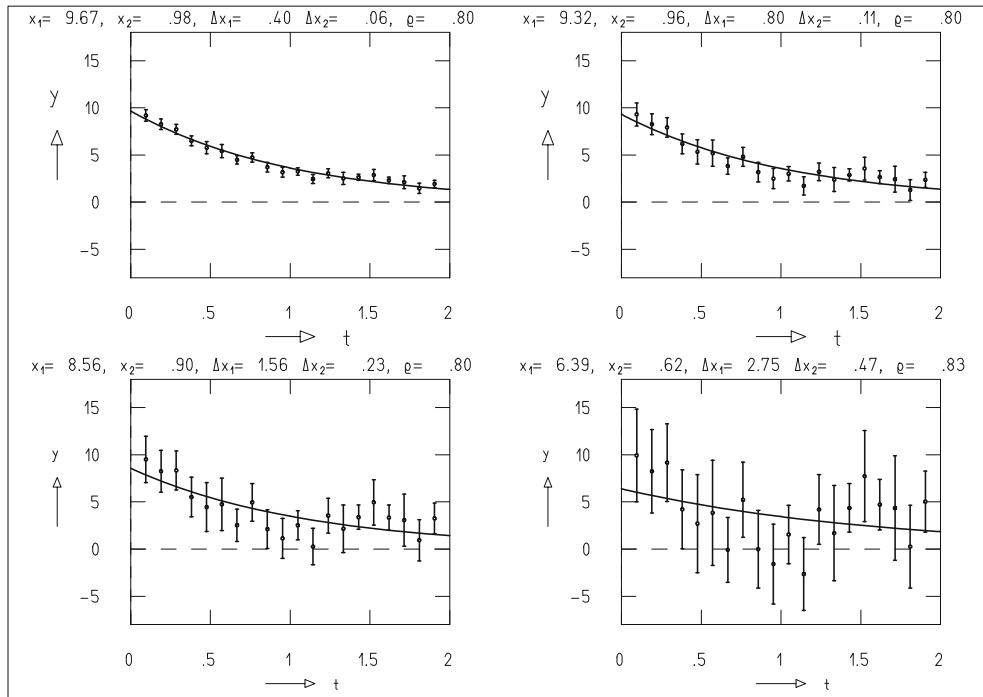
For only one parameter  $x$ , the curve  $M = M(x)$  can easily be plotted. The confidence region is simply a segment of the  $x$  axis. For two parameters  $x_1, x_2$  the boundary of the confidence region is a curve in the  $(x_1, x_2)$  plane. It is the contour line (9.8.5) of the function  $M = M(\mathbf{x})$ . The graphical representation of these contours is performed by the method `DatanGraphics.drawContour`, cf. Appendix F.5. For more than two parameters, a graphical representation is only possible in the form of slices of the confidence region in the  $(x_i, x_j)$  plane in the  $(x_1, x_2, \dots, x_r)$  space through the point  $\tilde{\mathbf{x}} = (\tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_r)$ , where  $x_i, x_j$  can be any possible pair of variables.

**Example 9.8:** The influence of large measurement errors on the confidence region of the parameters for fitting an exponential function

The results of fitting an exponential function as in Example 9.5 to data points with errors of different sizes are shown in Fig. 9.10. The two fitted parameters  $x_1$  and  $x_2$  and their errors and correlation coefficient are also shown in the figure. The quantities  $\Delta x_1$ ,  $\Delta x_2$ , and  $\rho$  were computed directly from the elements of the covariance matrix  $C_{\tilde{x}}$ . As expected, these errors increase for increasing measurement errors.

The covariance ellipses for this fit are shown as thin lines in the various frames of Fig. 9.11. The small circle in the center indicates the solution  $\tilde{\mathbf{x}}$ . In addition the vertical bars through the point  $\tilde{\mathbf{x}}$  mark the regions  $\tilde{x}_1 \pm \Delta x_1$  and  $\tilde{x}_2 \pm \Delta x_2$ . As a thicker line a contour is indicated which surrounds the confidence region  $M(\mathbf{x}) = M(\tilde{\mathbf{x}}) + 1$ . We see that for small measurement errors the confidence region and covariance ellipse are practically identical, while for large measurement errors they clearly differ from each other. ■

Computing and plotting the confidence regions requires, to the extent that they differ from the confidence ellipsoids, considerable effort. It is therefore important to be able to decide whether a clear difference between the two exists. For clarity we will stay with the case of two variables and consider again Example 9.8 and in particular Fig. 9.11. The errors  $\Delta x_i = \sigma_i =$



**Fig. 9.10:** Fit of an exponential function  $\eta = x_1 \exp(-x_2 t)$  to measurements with errors of different sizes.

$\sqrt{C_{\tilde{x}}(i, i)}$  obtained from the covariance matrix  $C_{\tilde{x}}$  have the following property. The lines  $x_i = \tilde{x}_i \pm \Delta x_i$  are tangent to the covariance ellipse. If the covariance ellipse must be replaced by a confidence region of less regular form, then we can nevertheless find the horizontal and vertical tangents, and at those places in Fig. 9.11 one has

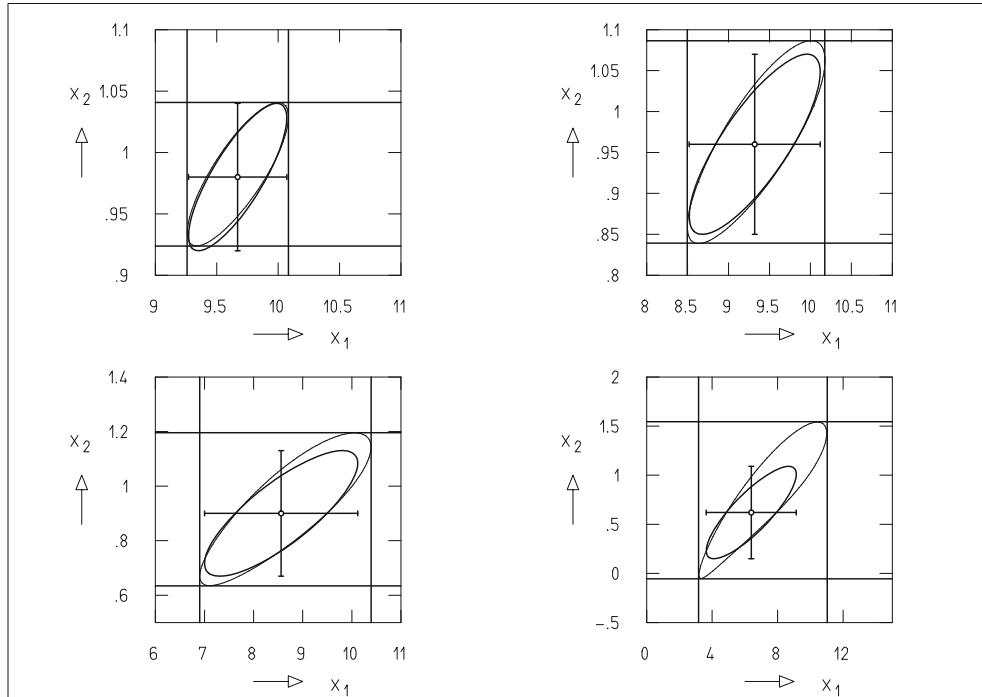
$$x_i = \tilde{x}_i + \Delta x_{i+} \quad , \quad x_i = \tilde{x}_i - \Delta x_{i-} \quad . \quad (9.8.7)$$

Because of the loss of symmetry, the errors  $\Delta x_+$  and  $\Delta x_-$  are in general different. One speaks of *asymmetric errors*. For  $r$  variables one has tangent hypersurfaces of dimension  $r - 1$  instead of tangent lines.

We now give a procedure to compute the asymmetric errors  $\Delta x_{i\pm}$ . One is only interested in asymmetric confidence regions when the asymmetric errors are significantly different from the symmetric ones. The values  $x_{i\pm} = \tilde{x}_i \pm \Delta x_{i\pm}$  have the property that the minimum of the function  $M$  for a fixed value of  $x_i = x_{i\pm}$  has the value

$$\min\{M(\mathbf{x}; x_i = x_{i\pm})\} = M(\tilde{\mathbf{x}}) + g \quad (9.8.8)$$

with  $g = 1$ . For other values of  $g$  one obtains corresponding *asymmetric confidence boundaries*. If we bring all the terms in (9.8.8) to the left-hand side,



**Fig. 9.11:** Results of the fit from Fig. 9.10 in the parameter space  $x_1, x_2$ . Solution  $\tilde{\mathbf{x}}$  (small circle), symmetric errors (error bars), covariance ellipse, asymmetric error limits (horizontal and vertical lines), confidence region corresponding to the probability of the covariance ellipse (dark contour).

$$\min\{M(\mathbf{x}; x_i = x_{i\pm})\} - M(\tilde{\mathbf{x}}) - g = 0 \quad , \quad (9.8.9)$$

then we see the problem to be solved is a combination of minimization and zero-finding. We find the zeros using an iterative procedure as in Sect. E.2. The zero, i.e., the point  $x_i$  fulfilling (9.8.9), is first bracketed by the values  $x_{\text{small}}$  and  $x_{\text{big}}$ , corresponding to negative and positive values of (9.8.9). Next, this interval is reduced by successively dividing it in half until the expression (9.8.9) differs from zero by no more than  $g/100$ . The minimum in (9.8.9) is found with LsqNon or LsqMar. The classes, by which the asymmetric errors are determined, are LsqAsn, if LsqNon is used, and lsqAsm for the use of LsqMar. The asymmetric errors for Example 9.8 are shown in Fig. 9.11.

## 9.9 Constrained Measurements

We now return to the case of Sect. 9.1, where the quantities of interest were directly measured. The  $N$  measurements are, however, no longer completely

independent, but rather are related to each other by  $q$  equations of constraint. One could measure, for example, the three angles of a triangle. The equation of constraint says that their sum is equal to  $180^\circ$ . We again ask for the best estimates  $\tilde{\eta}_j$  for the quantities  $\eta_j$ . The measurements give instead the values

$$y_j = \eta_j + \varepsilon_j \quad , \quad j = 1, 2, \dots, n \quad . \quad (9.9.1)$$

As above let us assume a normal distribution about zero for the measurement errors  $\varepsilon_j$ :

$$E(\varepsilon_j) = 0 \quad , \quad E(\varepsilon_j^2) = \sigma_j^2 \quad .$$

The  $q$  equations of constraint have the form

$$f_k(\boldsymbol{\eta}) = 0 \quad , \quad k = 1, 2, \dots, q \quad . \quad (9.9.2)$$

Let us first consider the simple case of linear equations of constraint. The Eqs. (9.9.2) are then of the form

$$\begin{aligned} b_{10} + b_{11}\eta_1 + b_{12}\eta_2 + \dots + b_{1n}\eta_n &= 0 \quad , \\ b_{20} + b_{21}\eta_1 + b_{22}\eta_2 + \dots + b_{2n}\eta_n &= 0 \quad , \\ &\vdots \\ b_{q0} + b_{q1}\eta_1 + b_{q2}\eta_2 + \dots + b_{qn}\eta_n &= 0 \quad , \end{aligned} \quad (9.9.3)$$

or in matrix notation,

$$B\boldsymbol{\eta} + \mathbf{b}_0 = 0 \quad . \quad (9.9.4)$$

### 9.9.1 The Method of Elements

Although not well suited for automatic processing of data, an illustrative procedure is the *method of elements*. We can use the  $q$  equations (9.9.3) to eliminate  $q$  of the  $n$  quantities  $\eta$ . The remaining  $n - q$  quantities  $\alpha_i$  ( $i = 1, 2, \dots, n - q$ ) are called *elements*. They can be chosen arbitrarily from the original  $\eta$  or they can be a linear combination of them. We can then express the full vector  $\eta$  as a set of linear combinations of these elements,

$$\eta_j = f_{j0} + f_{j1}\alpha_1 + f_{j2}\alpha_2 + \dots + f_{j,n-q}\alpha_{n-q} \quad , \quad j = 1, 2, \dots, n \quad , \quad (9.9.5)$$

or

$$\boldsymbol{\eta} = F\boldsymbol{\alpha} + \mathbf{f}_0 \quad . \quad (9.9.6)$$

Equation (9.9.6) is of the same type as (9.2.2). The solution must thus be of the form of (9.2.26), i.e.,

$$\tilde{\boldsymbol{\alpha}} = (F^T G_y F)^{-1} F^T G_y (\mathbf{y} - \mathbf{f}_0) \quad (9.9.7)$$

describes the estimation of the elements  $\alpha$  according to the method of least squares. The corresponding covariance matrix is

$$G_{\tilde{\alpha}}^{-1} = (F^T G_y F)^{-1} , \quad (9.9.8)$$

cf. (9.2.27). The improved measurements are obtained by substituting (9.9.7) into (9.9.6):

$$\tilde{\eta} = F\tilde{\alpha} + \mathbf{f}_0 = F(F^T G_y F)^{-1} F^T G_y (\mathbf{y} - \mathbf{f}_0) + \mathbf{f}_0 . \quad (9.9.9)$$

By error propagation the covariance matrix is found to be

$$G_{\tilde{\eta}}^{-1} = F(F^T G_y F)^{-1} F^T = FG_{\tilde{\alpha}}^{-1} F^T . \quad (9.9.10)$$

### Example 9.9: Constraint between the angles of a triangle

Suppose measurements of the angles of a triangle have yielded the values  $y_1 = 89^\circ$ ,  $y_2 = 31^\circ$ ,  $y_3 = 61^\circ$ , i.e.,

$$\mathbf{y} = \begin{pmatrix} 89 \\ 31 \\ 62 \end{pmatrix} .$$

The linear equation of constraint is

$$\eta_1 + \eta_2 + \eta_3 = 180 .$$

It can be written as

$$B\eta + \mathbf{b}_0 = 0$$

with

$$B = (1, 1, 1) , \quad \mathbf{b}_0 = b_0 = -180 .$$

As elements we choose  $\eta_1$  and  $\eta_2$ . The system (9.9.5) then becomes

$$\eta_1 = \alpha_1 , \quad \eta_2 = \alpha_2 , \quad \eta_3 = 180 - \alpha_1 - \alpha_2$$

or

$$\eta = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ -1 & -1 \end{pmatrix} \alpha + \begin{pmatrix} 0 \\ 0 \\ 180 \end{pmatrix} ,$$

i.e.,

$$F = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ -1 & -1 \end{pmatrix} , \quad \mathbf{f}_0 = \begin{pmatrix} 0 \\ 0 \\ 180 \end{pmatrix} .$$

We assume a measurement error for the angle of  $1^\circ$ , i.e.,

$$C_y = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = I \quad , \quad G_y = C_y^{-1} = I \quad .$$

Using these in (9.9.7) gives

$$\begin{aligned} \tilde{\alpha} &= \left[ \left( \begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \end{pmatrix} I \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ -1 & -1 \end{pmatrix} \right)^{-1} \left( \begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \end{pmatrix} I \begin{pmatrix} 89 \\ 31 \\ -118 \end{pmatrix} \right. \right. \\ &= \left( \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}^{-1} \begin{pmatrix} 207 \\ 149 \end{pmatrix} \right. = \frac{1}{3} \left( \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix} \begin{pmatrix} 207 \\ 149 \end{pmatrix} \right. \\ &= \left( \begin{pmatrix} 88\frac{1}{3} \\ 30\frac{1}{3} \end{pmatrix} \quad . \right. \end{aligned}$$

Using (9.9.9) one finally obtains

$$\tilde{\eta} = F\tilde{\alpha} + \mathbf{f}_0 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ -1 & -1 \end{pmatrix} \begin{pmatrix} 88\frac{1}{3} \\ 30\frac{1}{3} \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ 180 \end{pmatrix} = \begin{pmatrix} 88\frac{1}{3} \\ 30\frac{1}{3} \\ 61\frac{1}{3} \end{pmatrix} \quad .$$

This result was clearly to be expected. The “excess” in the measured sum of angles of  $2^\circ$  is subtracted equally from the three measurements. This would not have been the case, however, if the individual measurements had errors of different magnitudes. The reader can easily repeat the computation for such a case. The residual errors of the improved measurements can be determined by application of (9.9.10),

$$\begin{aligned} &G_{\tilde{\eta}}^{-1} \\ &= \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ -1 & -1 \end{pmatrix} \left[ \left( \begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \end{pmatrix} I \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ -1 & -1 \end{pmatrix} \right)^{-1} \left( \begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \end{pmatrix} \right. \right. \\ &= \frac{1}{3} \left( \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ -1 & -1 \end{pmatrix} \left( \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix} \begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \end{pmatrix} \right) \right. \\ &= \frac{1}{3} \left( \begin{pmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{pmatrix} \quad . \right. \end{aligned}$$

The residual error of each angle is thus equal to  $\sqrt{2/3} \approx 0.82$ . ■

At this point we want to make a general statement on measurements related by equations of constraint. Although in the statistical methods used so far we have found no way of dealing with systematic errors, equations of constraint offer such a possibility in many cases. If, for example, the sum of angles in many measurements is observed to be greater than  $180^\circ$  more frequently than less, then one can conclude that the measurement apparatus has a systematic error.

### 9.9.2 The Method of Lagrange Multipliers

Instead of elements, one more frequently uses the method of *Lagrange multipliers*. Although both methods clearly give the same results, the latter has the advantage that all unknowns are treated in the same way and thus the user is spared having to choose elements. The method of Lagrange multipliers is a well-known procedure in differential calculus for determination of extrema in problems with additional constraints.

We begin again from the linear system of equations of constraint (9.9.4)

$$B\eta + \mathbf{b}_0 = 0$$

and we recall that the measured quantities are the sum of the true value  $\eta$  and the measurement error  $\varepsilon$ ,

$$\mathbf{y} = \eta + \varepsilon .$$

Thus one has

$$B\mathbf{y} - B\varepsilon + \mathbf{b}_0 = 0 . \quad (9.9.11)$$

Since  $\mathbf{y}$  is known from the measurement, and  $\mathbf{b}_0$  and  $B$  are also constructed from known quantities, we can construct a column vector with  $q$  elements,

$$\mathbf{c} = B\mathbf{y} + \mathbf{b}_0 , \quad (9.9.12)$$

which does not contain any unknowns. Thus (9.9.11) can be written in the form

$$\mathbf{c} - B\varepsilon = 0 . \quad (9.9.13)$$

We now introduce an additional column vector with  $q$  elements, whose elements, not yet known, are the *Lagrange multipliers*,

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_q \end{pmatrix} . \quad (9.9.14)$$

Using this we extend the original minimum function (9.2.16)

$$M = \boldsymbol{\varepsilon}^T G_y \boldsymbol{\varepsilon}$$

to

$$L = \boldsymbol{\varepsilon}^T G_y \boldsymbol{\varepsilon} + 2\mu^T (\mathbf{c} - B\boldsymbol{\varepsilon}) . \quad (9.9.15)$$

The function  $L$  is called the *Lagrange function*. The requirement

$$M = \min$$

with the constraint

$$\mathbf{c} - B\boldsymbol{\varepsilon} = 0$$

is then fulfilled when the total differential of the Lagrange function vanishes, i.e., when

$$dL = 2\boldsymbol{\varepsilon}^T G_y d\boldsymbol{\varepsilon} - 2\boldsymbol{\mu}^T B d\boldsymbol{\varepsilon} = 0 .$$

This is equivalent to

$$\boldsymbol{\varepsilon}^T G_y - \boldsymbol{\mu}^T B = 0 . \quad (9.9.16)$$

The system (9.9.16) consists of  $n$  equations containing a total of  $n + q$  unknowns,  $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$  and  $\mu_1, \mu_2, \dots, \mu_q$ . In addition we have the  $q$  equations of constraint (9.9.13). We transpose (9.9.16) and obtain

$$G_y \boldsymbol{\varepsilon} = B^T \boldsymbol{\mu} , \quad \boldsymbol{\varepsilon} = G_y^{-1} B^T \boldsymbol{\mu} . \quad (9.9.17)$$

By substitution into (9.9.13) we obtain

$$\mathbf{c} - BG_y^{-1}B^T \boldsymbol{\mu} = 0 ,$$

which can easily be solved for  $\boldsymbol{\mu}$ :

$$\tilde{\boldsymbol{\mu}} = (BG_y^{-1}B^T)^{-1} \mathbf{c} . \quad (9.9.18)$$

With (9.9.17) we thus have the least-squares estimators of the measurement errors,

$$\tilde{\boldsymbol{\varepsilon}} = G_y^{-1} B^T (BG_y^{-1}B^T)^{-1} \mathbf{c} . \quad (9.9.19)$$

The estimators of the unknowns are then given by (9.9.1),

$$\tilde{\boldsymbol{\eta}} = \mathbf{y} - \tilde{\boldsymbol{\varepsilon}} = \mathbf{y} - G_y^{-1} B^T (BG_y^{-1}B^T)^{-1} \mathbf{c} . \quad (9.9.20)$$

With the abbreviation

$$G_B = (BG_y^{-1}B^T)^{-1}$$

this becomes

$$\tilde{\boldsymbol{\eta}} = \mathbf{y} - G_y^{-1} B^T G_B \mathbf{c} . \quad (9.9.21)$$

The covariance matrices of  $\tilde{\boldsymbol{\mu}}$  and  $\tilde{\boldsymbol{\eta}}$  are easily obtained by applying error propagation to the linear system of equations (9.9.18) and (9.9.19),

$$G_{\tilde{\boldsymbol{\mu}}}^{-1} = (BG_y^{-1}B^T)^{-1} = G_B , \quad (9.9.22)$$

$$G_{\tilde{\boldsymbol{\eta}}}^{-1} = G_y^{-1} - G_y^{-1} B^T G_B B G_y^{-1} . \quad (9.9.23)$$

**Example 9.10:** Application of the method of Lagrange multipliers to Example 9.9

We apply the method of Lagrange multipliers to the problem of Example 9.9. We then have

$$\mathbf{c} = B\mathbf{y} + \mathbf{b}_0 = (1, 1, 1) \begin{pmatrix} 89 \\ 31 \\ 62 \end{pmatrix} - 180 = 182 - 180 = 2 \quad ,$$

and in addition,

$$G_B = (BG_y B^T)^{-1} = \left[ (1, 1, 1)I \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \right]^{-1} = 3^{-1} = \frac{1}{3}$$

and

$$G_y B^T = I \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \quad .$$

We can now compute (9.9.21),

$$\tilde{\eta} = \begin{pmatrix} 89 \\ 31 \\ 62 \end{pmatrix} - \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \frac{2}{3} = \begin{pmatrix} 88\frac{1}{3} \\ 30\frac{1}{3} \\ 61\frac{1}{3} \end{pmatrix} \quad .$$

The covariance matrices are then

$$\begin{aligned} G_{\tilde{\mu}}^{-1} &= \frac{1}{3} \quad , \\ G_{\tilde{\eta}}^{-1} &= I - I \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \frac{1}{3} (1, 1, 1) I \\ &= I - \frac{1}{3} \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix} = \frac{1}{3} \begin{pmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{pmatrix} \quad . \blacksquare \end{aligned}$$

We now generalize the method of Lagrange multipliers for the case of nonlinear equations of constraint having the general form (9.9.2), i.e.,

$$f_k(\boldsymbol{\eta}) = 0 \quad , \quad k = 1, 2, \dots, q \quad .$$

These equations can be expanded in a series about  $\boldsymbol{\eta}_0$ ,

$$f_k(\boldsymbol{\eta}) = f_k(0) + \left( \frac{\partial f_k}{\partial \eta_1} \right)_{\boldsymbol{\eta}_0} (\eta_1 - \eta_{10}) + \dots + \left( \frac{\partial f_k}{\partial \eta_n} \right)_{\boldsymbol{\eta}_0} (\eta_n - \eta_{n0}) \quad . \quad (9.9.24)$$

Here the  $\eta_0$  are the first approximations for the true values  $\eta$ . Using the definitions

$$\begin{aligned} b_{kl} &= \left( \frac{\partial f_k}{\partial \eta_l} \right)_{\eta_0}, & B = \begin{pmatrix} b_{11} & b_{12} & \cdots & b_{1n} \\ b_{21} & b_{22} & \cdots & b_{2n} \\ \vdots & & & \\ b_{q1} & b_{q2} & \cdots & b_{qn} \end{pmatrix}, \\ c_k &= f_k(\eta_0), & \mathbf{c} = \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_q \end{pmatrix}, \\ \delta_k &= (\eta_k - \eta_{k0}), & \boldsymbol{\delta} = \begin{pmatrix} \delta_1 \\ \delta_2 \\ \vdots \\ \delta_n \end{pmatrix}, \end{aligned}$$

we can write (9.9.24) in the form

$$B\boldsymbol{\delta} + \mathbf{c} = 0. \quad (9.9.25)$$

Except for a sign this relation corresponds to (9.9.13). The solution  $\tilde{\boldsymbol{\delta}}$  therefore can be read off (9.9.19),

$$\tilde{\boldsymbol{\delta}} = -G_y^{-1}B^T(BG_y^{-1}B^T)^{-1}\mathbf{c}. \quad (9.9.26)$$

As first approximations  $\eta_0$  we use the measured values  $\mathbf{y}$ ,

$$\eta_0 = \mathbf{y} \quad (9.9.27)$$

and obtain

$$\tilde{\boldsymbol{\eta}} = \eta_0 + \tilde{\boldsymbol{\delta}}. \quad (9.9.28)$$

For linear constraint equations this already is the solution.

If the equations are nonlinear an iteration is performed. The prescription for step  $i$  of the iteration is described for the general case at the end of the next section. If, in the formulas given there, all terms containing the matrix  $A$  are set to zero, one obtains the iteration procedure for the case of constraint measurements.

For each step  $i$  one computes

$$M_i = \boldsymbol{\varepsilon}_i^T G_y \boldsymbol{\varepsilon}_i, \quad \boldsymbol{\varepsilon}_i = \mathbf{y} - \boldsymbol{\eta}_i. \quad (9.9.29)$$

The procedure is terminated as convergent, if a further step leads to no appreciable reduction of  $M_i$ . We call the result  $\tilde{\boldsymbol{\eta}}$ .

The covariance matrix is still given by (9.9.23), i.e.,

$$G_{\tilde{\eta}}^{-1} = G_y^{-1} - G_y^{-1} B^T G_B B G_y^{-1} \quad , \quad (9.9.30)$$

if the elements of  $B$  are computed using at  $\tilde{\eta}$ . The best estimates of the measurement errors are computed using

$$\tilde{\varepsilon} = \mathbf{y} - \tilde{\eta} \quad . \quad (9.9.31)$$

With them the minimum function is found to be

$$\tilde{M} = \tilde{\varepsilon}^T G_y \tilde{\varepsilon} \quad . \quad (9.9.32)$$

This quantity again can be used for a  $\chi^2$  test with  $q$  degrees of freedom.

Although the method of Lagrange multipliers is mathematically elegant, in programs provided here we use the method of orthogonal transformations (see Sect. 9.12).

## 9.10 The General Case of Least-Squares Fitting

After the preparation of the previous sections we can now take up the general case of fitting with the method of least squares.

We first recall the notation. The  $r$  unknown parameters are placed in a vector  $\mathbf{x}$ . The quantities to be measured form an  $n$ -vector  $\boldsymbol{\eta}$ . The values  $\mathbf{y}$  actually measured differ from  $\boldsymbol{\eta}$  by the errors  $\boldsymbol{\varepsilon}$ . We will assume a normal distribution for the individual errors  $\varepsilon_j$  ( $j = 1, 2, \dots, n$ ), i.e., a normal distribution for the  $n$  variables  $\varepsilon_j$  with the null vector as the vector of expectation values, and a covariance matrix  $C_y = G_y^{-1}$ . The vectors  $\mathbf{x}$  and  $\boldsymbol{\eta}$  are related by  $m$  functions

$$f_k(\mathbf{x}, \boldsymbol{\eta}) = f_k(\mathbf{x}, \mathbf{y} - \boldsymbol{\varepsilon}) = 0 \quad , \quad k = 1, 2, \dots, m \quad . \quad (9.10.1)$$

We will further assume that we have already obtained in some way a first approximation for the unknowns  $\mathbf{x}_0$ . As a first approximation for  $\boldsymbol{\eta}$  we use  $\boldsymbol{\eta}_0 = \mathbf{y}$  as in Sect. 9.9. Finally we require that the functions  $f_k$  can be approximated by linear functions in the range of variability of our problem, i.e., in the region around  $(\mathbf{x}_0, \boldsymbol{\eta}_0)$ , which is given by the differences  $\mathbf{x} - \mathbf{x}_0$  and  $\boldsymbol{\eta} - \boldsymbol{\eta}_0$ . We can then write

$$\begin{aligned} f_k(\mathbf{x}, \boldsymbol{\eta}) &= f_k(\mathbf{x}_0, \boldsymbol{\eta}_0) \\ &+ \left( \frac{\partial f_k}{\partial x_1} \right)_{\mathbf{x}_0, \boldsymbol{\eta}_0} (x_1 - x_{10}) + \cdots + \left( \frac{\partial f_k}{\partial x_r} \right)_{\mathbf{x}_0, \boldsymbol{\eta}_0} (x_r - x_{r0}) \\ &+ \left( \frac{\partial f_k}{\partial \eta_1} \right)_{\mathbf{x}_0, \boldsymbol{\eta}_0} (\eta_1 - \eta_{10}) + \cdots + \left( \frac{\partial f_k}{\partial \eta_n} \right)_{\mathbf{x}_0, \boldsymbol{\eta}_0} (\eta_n - \eta_{n0}) \quad . \end{aligned} \quad (9.10.2)$$

With the abbreviations

$$a_{k\ell} = \left( \frac{\partial f_k}{\partial x_\ell} \right)_{\mathbf{x}_0, \mathbf{\eta}_0}, \quad A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1r} \\ a_{21} & a_{22} & \cdots & a_{2r} \\ \vdots & & & \\ a_{m1} & a_{m2} & \cdots & a_{mr} \end{pmatrix}, \quad (9.10.3)$$

$$b_{k\ell} = \left( \frac{\partial f_k}{\partial \eta_\ell} \right)_{\mathbf{x}_0, \mathbf{\eta}_0}, \quad B = \begin{pmatrix} b_{11} & b_{12} & \cdots & b_{1n} \\ b_{21} & b_{22} & \cdots & b_{2n} \\ \vdots & & & \\ b_{m1} & b_{m2} & \cdots & b_{mn} \end{pmatrix}, \quad (9.10.4)$$

$$c_k = f_k(\mathbf{x}_0, \mathbf{\eta}_0), \quad \mathbf{c} = \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_m \end{pmatrix}, \quad (9.10.5)$$

$$\xi = \mathbf{x} - \mathbf{x}_0, \quad \delta = \mathbf{\eta} - \mathbf{\eta}_0 \quad (9.10.6)$$

the system of equations (9.10.2) can be written as follows:

$$A\xi + B\delta + \mathbf{c} = 0. \quad (9.10.7)$$

The Lagrange function is

$$L = \delta^T G_y \delta + 2\mu^T (A\xi + B\delta + \mathbf{c}) . \quad (9.10.8)$$

Here  $\mu$  is an  $m$ -vector of the Lagrange multipliers. We require that the total derivative of (9.10.8) with respect to  $\delta$  vanishes. This is equivalent to requiring

$$G_y \delta + B^T \mu = 0$$

or

$$\delta = -G_y^{-1} B^T \mu . \quad (9.10.9)$$

Substitution into (9.10.7) gives

$$A\xi - BG_y^{-1} B^T \mu + \mathbf{c} = 0 \quad (9.10.10)$$

or

$$\mu = G_B (A\xi + \mathbf{c}) , \quad (9.10.11)$$

where

$$G_B = (BG_y^{-1}B^T)^{-1} . \quad (9.10.12)$$

With (9.10.9) we can now write

$$\delta = -G_y^{-1} B^T G_B (A\xi + \mathbf{c}) . \quad (9.10.13)$$

Since the Lagrange function  $L$  is a minimum also with respect to  $\xi$ , the total derivative of (9.10.8) with respect to  $\xi$  must also vanish, i.e.,

$$2\mu^T A = 0 \quad .$$

By transposing and substituting (9.10.11) one obtains

$$2A^T G_B (A\xi + \mathbf{c}) = 0$$

or

$$\tilde{\xi} = -(A^T G_B A)^{-1} A^T G_B \mathbf{c} \quad . \quad (9.10.14)$$

Substituting (9.10.14) into (9.10.13) and (9.10.11) immediately gives the estimates of the deviations  $\delta$  and the Lagrange multipliers  $\mu$ ,

$$\tilde{\delta} = -G_y^{-1} B^T G_B (\mathbf{c} - A(A^T G_B A)^{-1} A^T G_B \mathbf{c}) \quad , \quad (9.10.15)$$

$$\tilde{\mu} = G_B (\mathbf{c} - A(A^T G_B A)^{-1} A^T G_B \mathbf{c}) \quad . \quad (9.10.16)$$

The estimates for the parameters  $\mathbf{x}$  and for the improved measurements  $\eta$  are

$$\tilde{\mathbf{x}} = \mathbf{x}_0 + \tilde{\xi} \quad , \quad (9.10.17)$$

$$\tilde{\eta} = \eta_0 + \tilde{\delta} \quad . \quad (9.10.18)$$

From (9.10.14), (9.10.4), and (9.10.5) we obtain for the matrix of derivatives of the elements of  $\xi$  with respect to the elements of  $\mathbf{y}$

$$\frac{\partial \tilde{\xi}}{\partial \mathbf{y}} = -(A^T G_B A)^{-1} A^T G_B \frac{\partial \mathbf{c}}{\partial \mathbf{y}} = -(A^T G_B A)^{-1} A^T G_B B \quad .$$

Using error propagation one obtains the covariance matrix

$$G_{\tilde{x}}^{-1} = G_{\tilde{\xi}}^{-1} = (A^T G_B A)^{-1} \quad . \quad (9.10.19)$$

Correspondingly one finds

$$G_{\tilde{\eta}}^{-1} = G_y^{-1} - G_y^{-1} B^T G_B B G_y^{-1} + G_y^{-1} B^T G_B A (A^T G_B A)^{-1} A^T G_B B G_y^{-1} \quad . \quad (9.10.20)$$

One can show that under the assumed conditions, i.e., sufficient linearity of (9.10.2) and normally distributed measurement errors, the minimum function  $M$ , which can also be written in the form

$$M = (B\tilde{\epsilon})^T G_B (B\tilde{\epsilon}) \quad , \quad \tilde{\epsilon} = \mathbf{y} - \tilde{\eta} \quad , \quad (9.10.21)$$

follows a  $\chi^2$ -distribution with  $m - r$  degrees of freedom.

If the Eqs. (9.10.1) are linear, then the relations (9.10.17) to (9.10.20) already are the solutions. In nonlinear cases one can perform an iterative procedure, which we now discuss in detail for step  $i$  with  $i = 1, 2, \dots$ . For the functions  $f_k$  the following holds:

$$\begin{aligned} f_k^{(i)}(\mathbf{x}, \boldsymbol{\eta}) &= f_k(\mathbf{x}_{i-1}, \boldsymbol{\eta}_{i-1}) \\ &\quad + \left( \frac{\partial f_k}{\partial x_1} \right)_{\mathbf{x}_{i-1}, \boldsymbol{\eta}_{i-1}} (x_1 - x_{1,i-1}) + \dots \\ &\quad + \left( \frac{\partial f_k}{\partial x_r} \right)_{\mathbf{x}_{i-1}, \boldsymbol{\eta}_{i-1}} (x_r - x_{r,i-1}) \\ &\quad + \left( \frac{\partial f_k}{\partial \eta_1} \right)_{\mathbf{x}_{i-1}, \boldsymbol{\eta}_{i-1}} (\eta_1 - \eta_{1,i-1}) + \dots \\ &\quad + \left( \frac{\partial f_k}{\partial \eta_n} \right)_{\mathbf{x}_{i-1}, \boldsymbol{\eta}_{i-1}} (\eta_n - \eta_{n,i-1}) \quad . \end{aligned}$$

With  $A^{(i)}, B^{(i)}, \mathbf{c}^{(i)}$  we denote the quantities  $A, B, \mathbf{c}$ , evaluated at  $\mathbf{x}_{i-1}, \boldsymbol{\eta}_{i-1}$ . Furthermore let

$$\boldsymbol{\xi}^{(i)} = \mathbf{x}_i - \mathbf{x}_{i-1} \quad , \quad \boldsymbol{\delta}^{(i)} = \boldsymbol{\eta}_i - \boldsymbol{\eta}_{i-1} \quad .$$

Then

$$A^{(i)}\boldsymbol{\xi}^{(i)} + B^{(i)}\boldsymbol{\delta}^{(i)} + \mathbf{c}^{(i)} = 0 \quad .$$

We now denote with

$$\mathbf{s}^{(i)} = \sum_{\ell=1}^{i-1} \boldsymbol{\delta}^{(\ell)}$$

the sum of the contributions of all previous steps to improve the measurements and find for the difference between the measurements  $\mathbf{y}$  and the approximation

$\boldsymbol{\eta}_i$

$$\mathbf{y} - \boldsymbol{\eta}_i = \mathbf{y} - (\boldsymbol{\eta}_0 + \mathbf{s}^{(i)} + \boldsymbol{\delta}^{(i)}) = -(\mathbf{s}^{(i)} + \boldsymbol{\delta}^{(i)}) \quad ,$$

since  $\boldsymbol{\eta}_0 = \mathbf{y}$ . The first term of the Lagrangian function is

$$(\mathbf{y} - \boldsymbol{\eta}_i)^T G_y (\mathbf{y} - \boldsymbol{\eta}_i) = (\mathbf{s}^{(i)} + \boldsymbol{\delta}^{(i)})^T G_y (\mathbf{s}^{(i)} + \boldsymbol{\delta}^{(i)})$$

and the full Lagrangian is

$$L = (\mathbf{s}^{(i)} + \boldsymbol{\delta}^{(i)})^T G_y (\mathbf{s}^{(i)} + \boldsymbol{\delta}^{(i)}) + 2\boldsymbol{\mu}^{(i)T} (A^{(i)}\boldsymbol{\xi}^{(i)} + B^{(i)}\boldsymbol{\delta}^{(i)} + \mathbf{c}^{(i)}) \quad .$$

We can now proceed as above and get, with  $G_B^{(i)} = (B^{(i)}G_y^{-1}B^{(i)T})^{-1}$ ,

$$\boldsymbol{\xi}^{(i)} = -(A^{(i)T}G_B^{(i)}A^{(i)})^{-1} A^{(i)T}G_B^{(i)} (\mathbf{c}^{(i)} - B^{(i)}\mathbf{s}^{(i)}) \quad .$$

and

$$\begin{aligned}\boldsymbol{\delta}^{(i)} = & -G_y^{-1} B^{(i)\top} G_B^{(i)} (\mathbf{c}^{(i)} - B^{(i)} \mathbf{s}^{(i)} - A^{(i)} (A^{(i)\top} G_B^{(i)} A^{(i)})^{-1} A^{(i)\top} \\ & \times G_B^{(i)} (\mathbf{c}^{(i)} - B^{(i)} \mathbf{s}^{(i)})) .\end{aligned}$$

For every step  $i$  we compute

$$M_i = \boldsymbol{\varepsilon}_i^\top G_y \boldsymbol{\varepsilon}_i , \quad \boldsymbol{\varepsilon}_i = \mathbf{y} - \boldsymbol{\eta}_i . \quad (9.10.22)$$

The procedure is terminated as convergent, if a new step does not yield an appreciable further reduction of  $M_i$ . The results of the iteration are denoted by  $\tilde{\mathbf{x}}, \tilde{\boldsymbol{\eta}}$ . The corresponding covariance matrices are given by (9.10.19) and (9.10.20), if the matrices  $A$  and  $B$  are evaluated at  $\tilde{\mathbf{x}}, \tilde{\boldsymbol{\eta}}$ . It is, of course, possible for the iteration process to diverge. In this case points (a)–(d), raised at the end of Sect. 9.7, should be considered.

In the following section we describe a different way to determine the solutions  $\tilde{\mathbf{x}}, \tilde{\boldsymbol{\eta}}$ . Also for that procedure the formulas (9.10.19), (9.10.20), and (9.10.21) for the computation of the covariance matrices  $G_{\tilde{\mathbf{x}}}^{-1}, G_{\tilde{\boldsymbol{\eta}}}^{-1}$  and the minimum function  $M$  remain valid, if the matrices  $A$  and  $B$  are evaluated at the position of the solution.

## 9.11 Algorithm for the General Case of Least Squares

In the Java class LSqGen, treating the general case of least squares, we do not use the method of Lagrange multipliers but rather the procedure of Sect. A.18, which is based on orthogonal transformations.

At every step of the iteration we must determine the  $r$ -vector  $\boldsymbol{\xi}$  and the  $n$ -vector  $\boldsymbol{\delta}$ . We combine both into an  $(r+n)$ -vector  $\mathbf{u}$ ,

$$\mathbf{u} = \begin{pmatrix} \boldsymbol{\xi} \\ \boldsymbol{\delta} \end{pmatrix} . \quad (9.11.1)$$

The  $m \times r$  matrix  $A$  and the  $m \times n$  matrix  $B$  are also combined into an  $m \times (r+n)$  matrix  $E$ ,

$$E = (A, B) . \quad (9.11.2)$$

The vector containing the solutions  $\mathbf{u}$  must satisfy the constraint (9.10.7), i.e.,

$$E\mathbf{u} = \mathbf{d} , \quad \mathbf{d} = -\mathbf{c} . \quad (9.11.3)$$

We now consider the minimum function in the  $i$ th iterative step. It depends on

$$\boldsymbol{\eta} = \mathbf{y} + \sum_{\ell=1}^i \boldsymbol{\delta}_\ell = \mathbf{y} + \sum_{\ell=1}^{i-1} \boldsymbol{\delta}_\ell + \boldsymbol{\delta}_i = \mathbf{y} + \mathbf{s} + \boldsymbol{\delta} , \quad \boldsymbol{\delta} = \boldsymbol{\delta}_i , \quad (9.11.4)$$

where

$$\mathbf{s} = \sum_{\ell=1}^{i-1} \boldsymbol{\delta}_\ell \quad (9.11.5)$$

is the result of all of the previous steps that changed  $\mathbf{y}$ . One then has

$$M = (\boldsymbol{\eta} - \mathbf{y})^T G_y (\boldsymbol{\eta} - \mathbf{y}) = (\boldsymbol{\delta} + \mathbf{s})^T G_y (\boldsymbol{\delta} + \mathbf{s}) = \min \quad . \quad (9.11.6)$$

We now extend the  $n \times n$  matrix  $G_y$  to an  $(r+n) \times (r+n)$  matrix

$$G = \begin{pmatrix} 0 & 0 \\ 0 & G_y \end{pmatrix} \begin{matrix} \}r \\ \}n \end{matrix}, \quad (9.11.7)$$

for which we find a Cholesky decomposition according to Sect. A.9,

$$G = F^T F \quad . \quad (9.11.8)$$

Then (9.11.6) becomes

$$\begin{aligned} M &= (\mathbf{u} + \mathbf{t})^T G (\mathbf{u} + \mathbf{t}) \\ &= (F\mathbf{u} + F\mathbf{t})^2 = \min \end{aligned}$$

or

$$(F\mathbf{u} - \mathbf{b})^2 = \min \quad (9.11.9)$$

with

$$\mathbf{t} = \begin{pmatrix} \mathbf{0} \\ \mathbf{s} \end{pmatrix} \begin{matrix} \}r \\ \}n \end{matrix}, \quad \mathbf{b} = -F\mathbf{t} \quad . \quad (9.11.10)$$

Now one must merely solve the problem (9.11.9) with the constraint (9.11.3), e.g., with the procedure of Sect. A.18. With the solution

$$\tilde{\mathbf{u}} = \begin{pmatrix} \tilde{\boldsymbol{\xi}} \\ \tilde{\boldsymbol{\delta}} \end{pmatrix}$$

or rather with the vectors  $\tilde{\boldsymbol{\xi}}$ ,  $\tilde{\boldsymbol{\delta}}$  one finds improved values for  $\mathbf{x}$  [cf. (9.10.17)],  $\boldsymbol{\eta}$  [cf. (9.10.18)], and for  $\mathbf{s}$ , as well as for  $\mathbf{t}$  [cf. (9.11.5) and (9.11.10)], with which an additional iterative step can be carried out.

The procedure can be regarded as having converged and terminated when the minimum function (9.11.9) in two successive steps only changes by an insignificant amount, or it can be terminated without success if after a given number of steps convergence has not been reached. In the case of convergence, the covariance of the unknowns can be computed according to (9.10.19). The calculation of the covariance matrix of the “improved” measurements  $\tilde{\boldsymbol{\eta}}$  according to (9.10.20) is possible as well. It is, however, rarely

of interest. Finally, the value of  $M$  obtained in the last step can be used for a  $\chi^2$ -test of the goodness-of-fit with  $m - r$  degrees of freedom.

All these operations are performed by the class LsqGen. This includes the numerical computation of the derivatives for the matrix  $E$ . The user only has to program the relation (9.10.1), which depends on the problem at hand. That is done by an extension of the abstract class DatanUserFunction. There exist example programs (Sect. 9.14) for the following examples in this chapter, including realizations of such classes.

**Example 9.11:** Fitting a line to points with measurement errors in both the abscissa and ordinate

Suppose a number of measured points  $(t_i, s_i)$  in the  $(t, s)$  plane are given. Each point has measurement errors  $\Delta t_i$ ,  $\Delta s_i$ , which can, in general, be correlated. The covariance between the measurement errors  $\Delta t_i$  and  $\Delta s_i$  is  $c_i$ . We identify  $t_i$  and  $s_i$  with elements of the  $n$ -vector  $\mathbf{y}$  of measured quantities

$$y_1 = t_1 \quad , \quad y_2 = s_1 \quad , \quad \dots \quad , \quad y_{n-1} = t_{n/2} \quad , \quad y_n = s_{n/2} \quad .$$

The covariance matrix is

$$C_y = \begin{pmatrix} (\Delta t_1)^2 & c_1 & 0 & 0 \\ c_1 & (\Delta s_1)^2 & 0 & 0 \\ 0 & 0 & (\Delta t_2)^2 & c_2 \\ 0 & 0 & c_2 & (\Delta s_2)^2 \\ & & & \ddots \end{pmatrix} \quad .$$

A straight line in the  $(t, s)$  plane is described by the equation  $s = x_1 + x_2 t$ . For the assumption of such a line through the measured points, the equations of constraint (9.10.1) take on the form

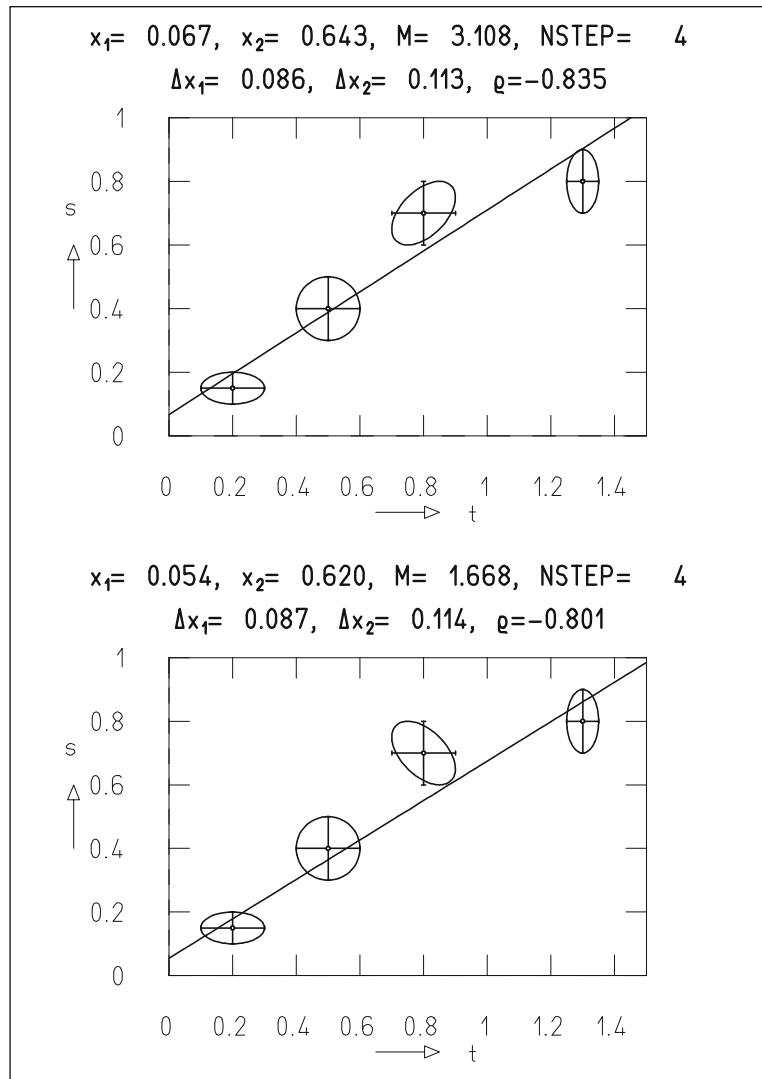
$$f_k(\mathbf{x}, \boldsymbol{\eta}) = \eta_{2k} - x_1 - x_2 \eta_{2k-1} = 0 \quad , \quad k = 1, 2, \dots, n/2 \quad . \quad (9.11.11)$$

Because of the term  $x_2 \eta_{2k-1}$ , the equations are not linear. The derivatives with respect to  $\eta_{2k-1}$  depend on  $x_2$  and those with respect to  $x_2$  depend on  $\eta_{2k-1}$ .

The results of fitting to four measured points are shown in Fig. 9.12. The examples in the two individual plots differ only in the correlation coefficient of the third measurement,  $\rho_3 = 0.5$  and  $\rho_3 = -0.5$ . One can see a noticeable effect of the sign of  $\rho_3$  on the fit result and in particular on the value of the minimum function. ■

**Example 9.12:** Fixing parameters

In Fig. 9.13 the results of fitting a line to measured points with errors in the abscissa and ordinate are shown, where in each plot one parameter of the line was held fixed. In the upper plot the intercept of the vertical axis  $x_1$  was fixed, and in the lower plot the slope  $x_2$  was fixed. ■

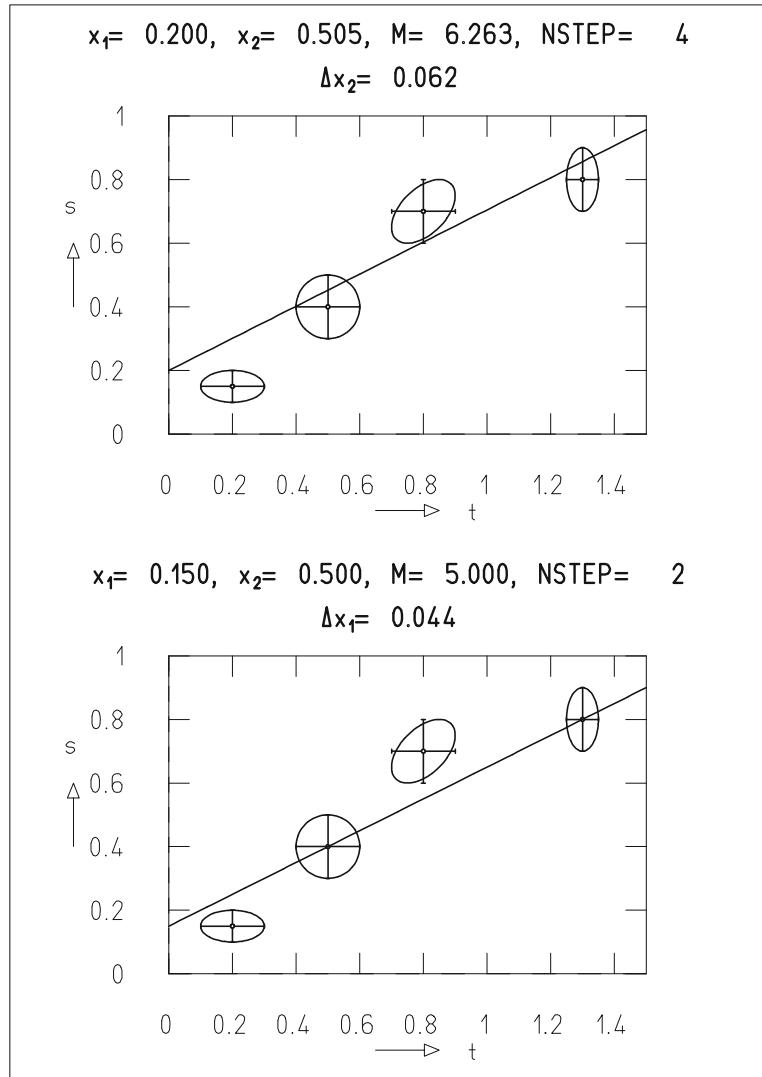


**Fig. 9.12:** Fitting a line to four measured points in the  $(t, s)$  plane. The points are shown with measurement errors (in  $t$  and  $s$ ) and covariance ellipses. The individual plots show the results of the fit, the errors, and correlation coefficients.

## 9.12 Applying the Algorithm for the General Case to Constrained Measurements

If all of the variables  $x_1, \dots, x_r$  are fixed, then there are no more unknowns in the least-squares problem.

In the equations of constraint (9.10.2) only the components of  $\eta$  are variable. Thus all terms containing the matrix  $A$  vanish from the formulas of the previous sections. As previously, however, the improved measurements  $\tilde{\eta}$  can be computed. The quantity  $M$  can in addition be used for a  $\chi^2$ -test with  $m$



**Fig. 9.13:** Fitting a line to the same points as in the first plot of Fig. 9.12. The intercept with the vertical axis  $x_1$  was held constant in the *upper plot*, and below the slope  $x_2$  was held constant.

degrees of freedom for how well the equations of constraint are satisfied by the measurements. Finally, the covariance matrix of the improved measurements  $C_{\tilde{\eta}}$  can be determined.

Mathematically – as well as when using the program LsqGen – it does not matter whether all variables are fixed ( $r \neq 0, r' = 0$ ) or whether from the start the equations of constraint do not depend on the variables  $\mathbf{x}$  ( $r = 0$ ). In both cases LsqGen gives the same solution.

**Example 9.13:**  $\chi^2$ -test of the description of measured points with errors in abscissa and ordinate by a given line

We use the same measurements as used already in Examples 9.11 and 9.12. The results of the analysis of these measurements with LsqGen with fixed parameters are shown in Fig. 9.14. For the upper plot,  $x_1$  and  $x_2$  were fixed to the values obtained from fitting with two adjustable parameters in Example 9.11, Fig. 9.12. Clearly we also obtain the same value of  $M$  as previously. For the lower plot, arbitrarily crude estimates ( $x_1 = 0$ ,  $x_2 = 0.5$ ) were used. They give a significantly higher value of  $M$ . This value would lead to rejection of the hypothesis with a confidence level of 99 % that the data points are described by a linear relation with these parameter values ( $\chi^2_{0.99} = 13.28$  for four degrees of freedom).

It is also interesting to consider the improved measurements  $\tilde{\eta}$  and their errors, which are shown in Fig. 9.14. The improved measurements naturally lie on the line. The measurement errors are the square roots of the diagonal elements of  $C_{\tilde{\eta}}$ . The correlations between the errors of a measured point in  $s$  and  $t$  are obtained from the corresponding off-diagonal elements  $C_{\tilde{\eta}}$ . They are exactly equal to one. The covariance ellipses of the individual improved measurements collapse to line segments which lie on the line given by  $x_1$ ,  $x_2$ .

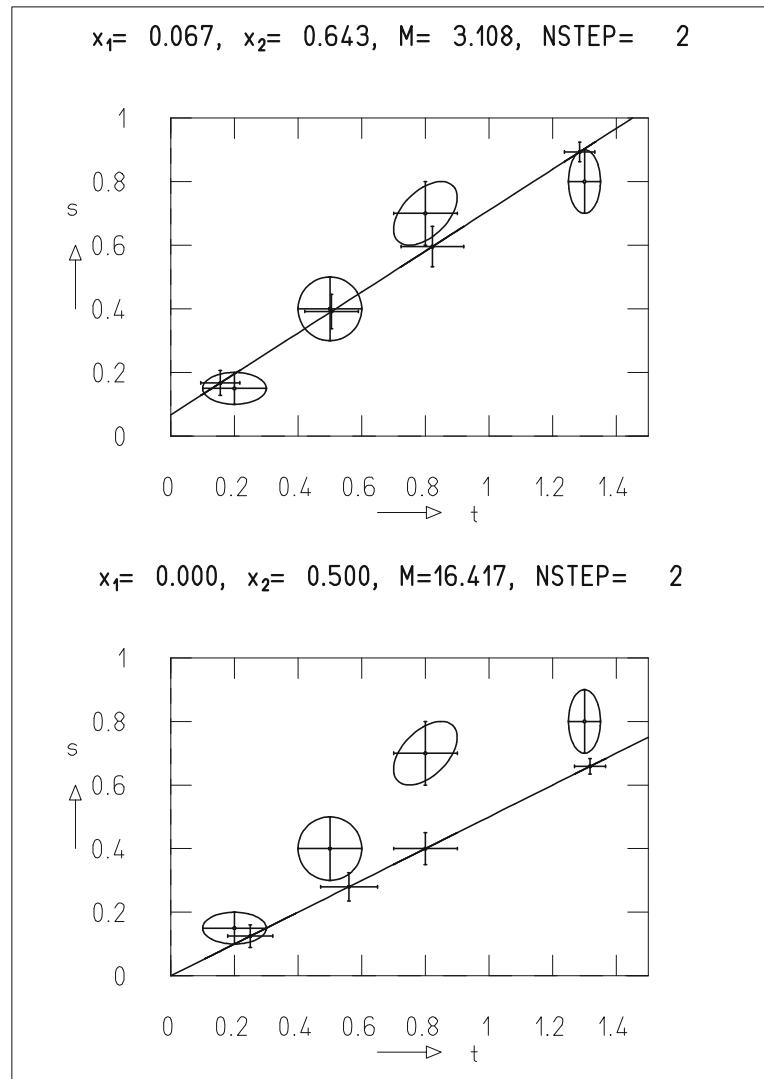
■

## 9.13 Confidence Region and Asymmetric Errors in the General Case

The results obtained in Sect. 9.8 on confidence regions and asymmetric errors are also valid for the general case. We therefore limit ourselves to stating that asymmetric errors are computed by the class LsqAsg and to presenting an example.

**Example 9.14:** Asymmetric errors and confidence region for fitting a straight line to measured points with errors in the abscissa and ordinate

Figure 9.15 shows the result of fitting to four points with large measurement errors. From Sect. 9.8 we already know that large measurement errors can lead to asymmetric errors in the fitted parameters. In fact, we see highly asymmetric errors and large differences between the covariance ellipse and the corresponding confidence region. ■



**Fig. 9.14:** Constrained measurements. The hypothesis was tested whether the true values of the measured points, indicated by their covariance ellipses, lie on the line  $s = x_1 + x_2t$ . With the numerical values of  $M$ , a  $\chi^2$ -test with four degrees of freedom can be carried out. Shown as well are the improved measurements, which lie on the line, and their errors.

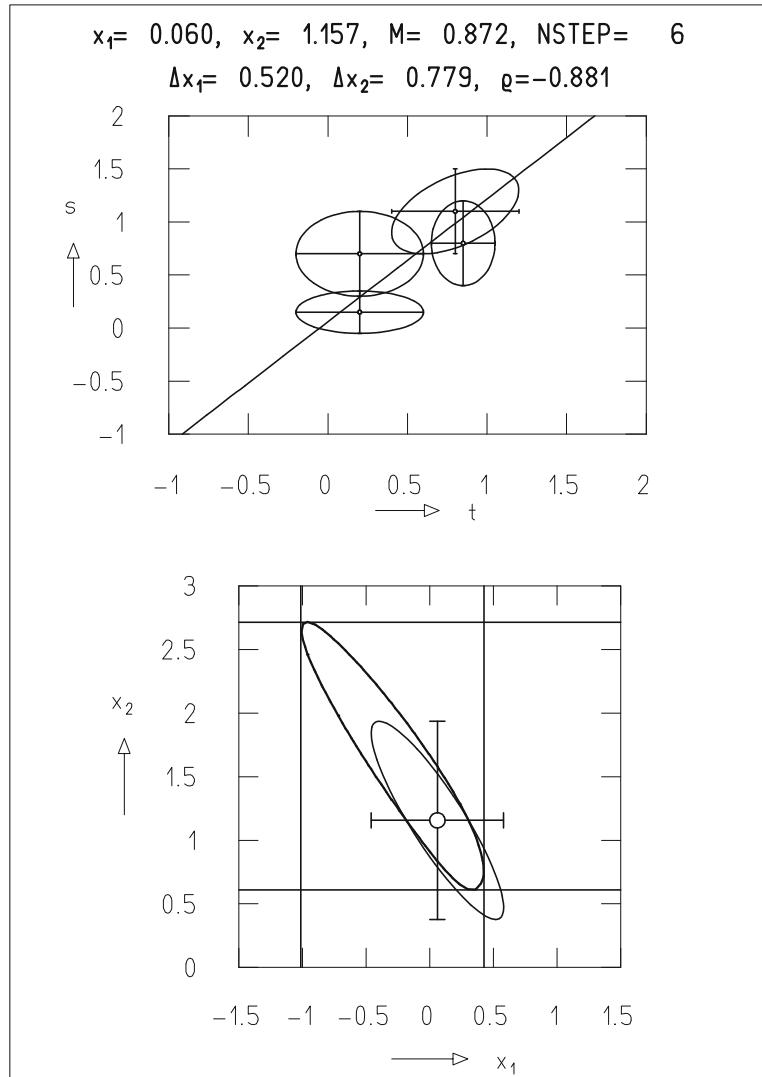
## 9.14 Java Classes and Example Programs

### Java Classes for Least-Squares Problems

LsqPol handles the fitting of a polynomial (Sect. 9.4.1).

LsqLin handles the linear case of indirect measurements (Sect. 9.4.2).

LsqNon handles the nonlinear case of indirect measurements (Sect. 9.6.1).



**Fig. 9.15:** (Above) Measured points with covariance ellipses and fitted line. (Below) Result of the fit, given in the plane spanned by the parameters  $x_1, x_2$ . Shown are the fitted parameter values (circle), symmetric errors (crossed bars), covariance ellipse, asymmetric errors (horizontal and vertical lines), and the confidence region (dark contour).

LsqMar handles the nonlinear case using Marquardt's method (Sect. 9.6.2).

LsqAsn yields asymmetric errors or confidence limits in the nonlinear case (Sect. 9.8).

LsqAsm yields asymmetric errors or confidence limits using Marquardt's method.

LsqGen handles the general case of least squares (Sect. 9.11).

LsqAsg yields asymmetric errors or confidence limits in the general case (Sect. 9.13).

**Example Program 9.1:** The class E1Lsq demonstrates the use of LsqPol

The short program uses the data of Table 9.3 and computes vectors  $\tilde{\mathbf{x}}$  of coefficients and their covariance matrix  $C_x$  for  $r = 1, 2, 3, 4$ . Here  $r - 1$  is the degree of the polynomial, i.e.,  $r$  is the number of elements in  $\mathbf{x}$ . The results are presented numerically

**Suggestions:** (a) Modify the program (by modifying a single statement) so that the cases  $r = 1, 2, \dots, 10$  are treated. Which peculiarity do you expect for  $r = 10$ ? (b) Instead of the data of Table 9.3 use different data determined without error by a polynomial, e.g.,  $y = t^2$ , and let the program determine the parameters of the polynomial from the data. Use different (although obviously incorrect) sets of values for the errors  $\Delta y_i$ , e.g.,  $\Delta y_i = \sqrt{y_i}$  for one run through the program and  $\Delta y_i = 1$  for another run. What is the influence of the choice of  $\Delta y_i$  on the coefficients  $\tilde{\mathbf{x}}$ , on the minimum function, and on the covariance matrix?

**Example Program 9.2:** The class E2Lsq demonstrates the use of LsqLin

The program uses the data of Fig. 9.4 and sets up the matrix  $A$  and the vector  $\mathbf{c}$ ; these are needed for the fit of a proportional relation  $y = x_1 t$  to the data. Next the fit is performed by calling LsqLin. The results are displayed numerically.

**Suggestions:** (a) Modify the program so that in addition a first degree polynomial is fitted. Set up the matrix  $A$  yourself, i.e., do not use LsqPol. Compare your result with Fig. 9.4. (b) Display the results graphically as in Fig. 9.4.

**Example Program 9.3:** The class E3Lsq demonstrates the use of LsqNon

The program solves the following problem. First 20 pairs of values  $(t_i, y_i)$  are generated. The values  $t_i$  of the controlled variables are  $1/21, 2/21, \dots, 20/21$ . The values  $y_i$  are given by

$$y_i = x_1 \exp(-(t - x_2)^2 / 2x_3^2) + \varepsilon_i \quad .$$

Here  $\varepsilon_i$  is an error taken from a normal distribution with expectation value zero and width  $\sigma_i$ . The width  $\sigma_i$  is different from point to point. It is taken from a uniform distribution with the limits  $\sigma/2$  and  $3\sigma/2$ . Thus, the  $y_i$  are points scattered within their errors around a Gaussian curve defined by the parameters  $\mathbf{x} = (x_1, x_2, x_3)$ . The widths of the error distributions are known, i.e.,  $\Delta y_i = \sigma_i$ . The data points are generated for the parameter values  $x_1 = 1, x_2 = 1.2, x_3 = 0.4$ . (They are identical with those shown in Fig. 9.5.) The program now successively performs four different fits of a Gaussian curve to the data. In a first step all three parameters in the fit procedure are considered variable. Then successively one, two, and finally all parameters are fixed. Before each fit, first approximations for the variables are set which, of course, are not modified during the fitting procedure for the fixed parameters. The results are presented numerically.

**Suggestions:** (a) Choose different first approximations for the non-fixed parameters and observe the influence of the choice on the results. (b) Obtain first approximations by computation, e.g., by the procedure described in Example 9.4, in which a parabola is fitted to the logarithms of the data. (c) Modify the program by adding graphical output corresponding to Fig. 9.5.

**Example Program 9.4:** The class E4Lsq demonstrates the use of LsqMar

The program solves the problem of Example 9.7. First a set of 50 data points  $(t_i, y_i)$  is generated. They are scattered according to their errors around a curve corresponding to the sum of a second degree polynomial and two Gaussians, cf. (9.6.6). The nine parameters of this function are combined to form the vector  $\mathbf{x}$ . The measurement errors  $\Delta y_i$  are generated by the procedure described in Sect. 9.3. The data points are generated for predefined values of  $\mathbf{x}$ . Next, by calling LSQMAR (with significantly different values  $\mathbf{x}$  as first approximations) solutions  $\tilde{\mathbf{x}}$  are obtained by a fit to the data points. The results are presented numerically and also in graphical form.

**Suggestions:** (a) Fix all parameters except for  $x_5$  and  $x_8$ , which determine the mean values of the two Gaussians used for generating the data. (In the program, as customary in Java, where indexing begins with 0, they are denoted by  $\mathbf{x}[4]$  and  $\mathbf{x}[7]$ .) Allow for interactive input of  $x_5$  and  $x_8$ . For different input values of  $(x_5, x_8)$ , e.g.,  $(10, 30)$ ,  $(19, 20)$ ,  $(19, 15)$ ,  $(10, 11)$ , try to determine whether you can still separate the two Gaussians by the fit, i.e., whether you obtain significantly different values for  $\tilde{x}_5$  and  $\tilde{x}_8$ , considering their errors  $\Delta \tilde{x}_5$ ,  $\Delta \tilde{x}_8$ . (b) Repeat (a), but for smaller measurement errors. Choose, e.g.,  $\sigma = 0.1$  or  $0.01$  instead of  $\sigma = 0.4$ .

**Example Program 9.5:** The class E5Lsq demonstrates the use of LsqAsn

The program solves the problem of Example 9.8. First, pairs  $(t_i, y_i)$  of data are produced. The  $y_i$  are scattered according to their measurement errors around a curve given by the function  $y_i(t_i) = x_1 \exp(-x_2 t_i)$ , cf. (9.6.4). The measurement errors  $\Delta y_i$  are generated by the procedure already used in Sect. 9.3. Starting from a given first approximation for the parameters  $\mathbf{x}_1$ , values  $\tilde{\mathbf{x}}$  of these parameters are fitted to the data using LsqNon. Finally the asymmetric errors are found using LsqAsn. Two plots are produced. One shows the measurements and the fitted curve. The second displays, in the  $(x_1, x_2)$  plane, the fitted parameters with symmetric and asymmetric errors, covariance ellipse, and confidence region.

**Example Program 9.6:** The class E6Lsq demonstrates the use of LsqGen

The program fits a straight line to points with measurement errors in the abscissa and ordinate, i.e., it solves the problem of Example 9.11. From the measured values, their errors and covariances the vector  $\mathbf{y}$  and the covariance matrix  $C_y$  are set up. Determination of the two parameters  $x_1$  (ordinate intercept) and  $x_2$  (slope) is done with LsqGen. For this the function (9.11.11) is needed. It is implemented in the method `getValue` of the subclass `StraightLine` of `E6Lsq`, which itself is an extension of the abstract class `DatanUserFunction`. The first approximations of  $x_1$  and  $x_2$  are obtained by constructing a straight line through the outermost two points. A loop extends over the two cases of Fig. 9.12. The results are shown numerically and graphically.

**Example Program 9.7:** The class E7Lsq demonstrates the use of LsqGen  
with some variables fixed

The program also treats Example 9.11. Again the first approximations of  $x_1$  and  $x_2$  are obtained by constructing a straight line through the outer two measured points.

A loop extends over two cases. In the first one  $x_1$  is fixed at  $x_1 = 0.2$ , in the second  $x_2$  is fixed at  $x_2 = 0.5$ . The results correspond to Fig. 9.13.

**Example Program 9.8:** The class E8Lsq demonstrates the use of LsqGen with all variables are fixed and produces a graphical representation of improved measurements

The problem of Example 9.13 is solved. The results are those of Fig. 9.14.

**Example Program 9.9:** The class E9Lsq demonstrates the use of LsqAsg and draws the confidence-region contour for the fitted variables

Here, the problem of Example 9.14 is solved. The results are those of Fig. 9.15.

# 10. Function Minimization

Locating extreme values (maxima and minima) is particularly important in data analysis. This task occurs in solving the least-squares problem in the form  $M(\mathbf{x}, \mathbf{y}) = \min$  and in the maximum likelihood problem as  $L = \max$ . By means of a simple change of sign, the latter problem can also be treated as locating a minimum. We always speak therefore of *minimization*.

## 10.1 Overview: Numerical Accuracy

We consider first the simple *quadratic form*

$$M(x) = c - bx + \frac{1}{2}Ax^2 . \quad (10.1.1)$$

It has an extremum where the first derivative vanishes,

$$\frac{dM}{dx} = 0 = -b + Ax , \quad (10.1.2)$$

that is, at the value

$$x_m = \frac{b}{A} . \quad (10.1.3)$$

With  $M(x_m) = M_m$ , Eq. (10.1.1) can easily be put into the form

$$M(x) - M_m = \frac{1}{2}A(x - x_m)^2 . \quad (10.1.4)$$

Although the function whose minimum we want to find does not usually have the simple form of (10.1.1), it can nevertheless be approximated by a quadratic form in the region of the minimum, where one has the Taylor expansion around the point  $x_0$ ,

$$M(x) = M(x_0) - b(x - x_0) + \frac{1}{2}A(x - x_0)^2 + \dots , \quad (10.1.5)$$

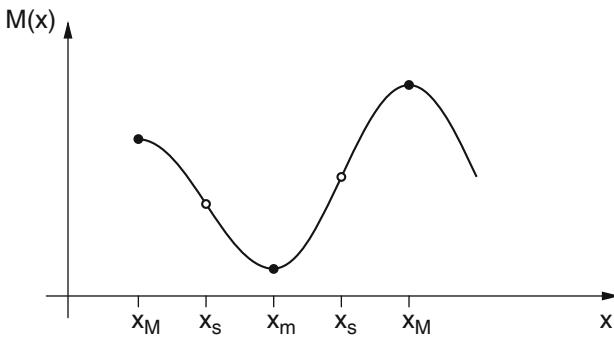
where

$$b = -M'(x_0) , \quad A = M''(x_0) . \quad (10.1.6)$$

In this approximation the minimum is given by the point where the derivative  $M'(x)$  is zero, i.e.,

$$x_{\text{mp}} = x_0 + \frac{b}{A} . \quad (10.1.7)$$

This holds only if  $x_0$  is sufficiently close to the minimum so that terms of order higher than quadratic in (10.1.5) can be neglected. The situation is depicted in Fig. 10.1. The function  $M(x)$  has a minimum at  $x_m$ , maxima at  $x_M$ , and points of inflection at  $x_s$ . For the second derivative in the region  $x > x_m$  one has  $M''(x) > 0$  for  $x < x_s$  and  $M''(x) < 0$  for  $x > x_s$ . If we now choose  $x_0$  to be in the region  $x_m < x_0 < x_M$ , then the first derivative  $M'(x)$  there is always positive. Therefore  $x_{\text{mp}}$  lies closer to  $x_m$  only if  $x_0 < x_s$ . Clearly the point  $x_0$  is not in general chosen arbitrarily, but rather as close as possible to where the minimum is expected to be. We can call this estimated value the *zeroth approximation* of  $x_m$ . Various strategies are available to obtain successively better approximations:



**Fig. 10.1:** The function  $M(x)$  has a minimum at  $x_m$ , maxima at the points  $x_M$ , and points of inflection at  $x_s$ .

(i) *Use of the function and its first and second derivatives at  $x_0$ .*

One computes  $x_{\text{mp}}$  according to (10.1.7), takes  $x_{\text{mp}}$  as a first approximation, i.e.,  $x_0$  is replaced by  $x_{\text{mp}}$ , and obtains by repeated application of (10.1.7) a second approximation. The procedure is repeated until two successive approximations differ by less than a given value  $\varepsilon$ . From the discussion above it follows that this procedure does not converge if the zeroth approximation lies outside the points of inflection in Fig. 10.1.

(ii) *Use of the function and its first derivative at  $x_0$ .*

The sign of the derivative  $M'(x_0) = -b$  at the point  $x_0$  determines the direction in which the function increases. It is assumed that one

should search for the minimum in the direction in which the function decreases. The quantity

$$x_1 = x_0 + b \quad (10.1.8)$$

is computed, i.e., one replaces the second derivative by the value unity.

Instead of  $x_0$  one now uses  $x_1$  as the approximation, and so forth. Alternatively one can also use instead of (10.1.8) the rule

$$x_1 = x_0 + cb \quad , \quad (10.1.9)$$

where  $c$  is an arbitrary positive constant. Both rules ensure that the step from  $x_0$  to  $x_1$  proceeds in the direction of the minimum. If in addition one chooses  $c$  to be small, then the step is small, so that it does not go beyond (or not far beyond) the minimum.

(iii) *Use of the function at various points.*

The procedure (i) can be carried out without knowing the derivative of the function if the function itself is known at three points. One can then uniquely fit a parabola through these three points and take the extreme value as an approximation of the minimum of the function. One speaks of locating the minimum by *quadratic interpolation*. It is, however, by no means certain that the extreme value of the parabola is a minimum and not a maximum. As in procedure (i) it is therefore important that the three chosen points are already in the region of the minimum of the function.

(iv) *Successive reduction of an interval containing the minimum.*

In none of the procedures discussed up to this point were we able to guarantee that the minimum of the function would actually be found. The minimum can be found with certainty provided one knows an interval  $x_a < x < x_b$  containing the minimum. If such an interval is known, one can locate the minimum with arbitrary accuracy by successively subdividing and checking in which subinterval the minimum is located.

In Sects. 10.2–10.7 we shall examine the minimization of a function of only one variable. In Sect. 10.2 the formula for a parabola determined by three points will be given. In Sect. 10.3 it is shown that the minimization of a function of one variable is equivalent to the minimization of a function of  $n$  variables on a line in an  $n$ -dimensional space. Section 10.4 describes a procedure for locating an interval containing the minimum. In Sect. 10.5 a minimum search by means of interval division is described. This is combined in Sect. 10.5 with the procedure of quadratic interpolation in a way that the

interpolation is only used when it leads quickly to the minimum of the function. If this is not the case, one continues with interval division. In this way we possess a procedure with the certainty of interval division combined with – as much as possible – the speed of quadratic interpolation. The same procedure can also be used for a function of  $n$  variables if the search for the minimum is restricted to a line in the  $n$ -dimensional space. This problem is addressed in Sect. 10.7.

Next we turn to the task of searching for the minimum of a function of  $n$  variables. We begin in Sect. 10.8 with the particularly elegant simplex method. This is followed by the discussion of various procedures of successive minimization along fixed directions in the  $n$ -dimensional space. These directions can simply be the directions of the coordinates (Sect. 10.9), or for a function that depends only quadratically on the variables they can be chosen such that the minimum is reached in at most  $n$  steps (Sects. 10.10 and 10.11).

Finally we discuss a procedure of  $n$ -dimensional minimization which employs aspects of methods (i) and (ii) of the one-dimensional case. If  $\mathbf{x}$  is the  $n$ -dimensional vector of the variables, then the general quadratic form, i.e., the generalization of (10.1.1) to  $n$  variables, is

$$\begin{aligned} M(\mathbf{x}) &= c - \mathbf{b} \cdot \mathbf{x} + \frac{1}{2} \mathbf{x}^T \mathbf{A} \mathbf{x} \\ &= c - \sum_k b_k x_k + \frac{1}{2} \sum_{k,\ell} x_k A_{k\ell} x_\ell . \end{aligned} \quad (10.1.10)$$

Here  $\mathbf{A}$  is a symmetric matrix,  $A_{k\ell} = A_{\ell k}$ . The partial derivative with respect to  $x_i$  is

$$\begin{aligned} \frac{\partial M}{\partial x_i} &= -b_i + \frac{1}{2} \left( \sum_\ell A_{i\ell} x_\ell + \sum_k x_k A_{ki} \right) \\ &= -b_i + \sum_\ell A_{i\ell} x_\ell . \end{aligned} \quad (10.1.11)$$

Expressing all of the partial derivatives as a vector  $\nabla M$  gives

$$\nabla M = -\mathbf{b} + \mathbf{A} \mathbf{x} . \quad (10.1.12)$$

At the minimum point the vector of derivatives vanishes. The minimum is therefore located at

$$\mathbf{x}_m = \mathbf{A}^{-1} \mathbf{b} , \quad (10.1.13)$$

in analogy to (10.1.3).

Clearly the function  $M(\mathbf{x})$  does not in general have the simple form of (10.1.10). We can, however, expand it in a series around the point  $\mathbf{x}_0$ ,

$$M(\mathbf{x}) = M(\mathbf{x}_0) - \mathbf{b}(\mathbf{x} - \mathbf{x}_0) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_0)^T A(\mathbf{x} - \mathbf{x}_0) + \dots , \quad (10.1.14)$$

with the negative *gradient*

$$\mathbf{b} = -\nabla M(\mathbf{x}_0) , \quad \text{i.e.,} \quad b_i = -\left. \frac{\partial M}{\partial x_i} \right|_{\mathbf{x}=\mathbf{x}_0} , \quad (10.1.15)$$

and the *Hessian matrix* of second derivatives

$$A_{ik} = \left. \frac{\partial^2 M}{\partial x_i \partial x_k} \right|_{\mathbf{x}=\mathbf{x}_0} . \quad (10.1.16)$$

The series (10.1.14) is the starting point for various minimization procedures:

(i) *Minimization in the direction of the gradient.*

Starting from the point  $\mathbf{x}_0$  one searches for the minimum along the direction given by the gradient  $\nabla M(\mathbf{x}_0)$  and calls the point where it is found  $\mathbf{x}_1$ . Starting from  $\mathbf{x}_1$  one looks for the minimum along the direction  $\nabla M(\mathbf{x}_1)$ , and so forth. We will discuss this procedure, called *minimization in the direction of steepest descent*, in Sect. 10.12.

(ii) *Step of given size in the gradient direction.*

One computes in analogy to (10.1.9)

$$\mathbf{x}_1 = \mathbf{x}_0 + c\mathbf{b} , \quad \mathbf{b} = -\nabla M(\mathbf{x}_0) , \quad (10.1.17)$$

with a given positive  $c$ . That is, one takes a *step in the direction of steepest descent* of the function, without, however, searching exactly for the minimum in this direction. Next one computes the gradient at  $\mathbf{x}_1$ , steps from  $\mathbf{x}_1$  in the direction of this gradient, etc. In Sect. 10.13 we will combine this method with the following one.

(iii) *Use of the gradient and the Hessian matrix at  $\mathbf{x}_0$ .*

If we truncate (10.1.14) after the quadratic term, we obtain a function whose minimum is, according to (10.1.13), given by

$$\mathbf{x}_{\text{mp}} = \mathbf{x}_0 + A^{-1}\mathbf{b} . \quad (10.1.18)$$

We take  $\mathbf{x}_1 = \mathbf{x}_{\text{mp}}$  as the first approximation, compute for this point the gradient and Hessian matrix, obtain by corresponding use of (10.1.18) the second approximation, and so forth. This procedure is discussed

in Sect. 10.14. It converges quickly if the zeroth approximation  $\mathbf{x}_0$  is sufficiently close to the minimum. If that is not the case, however, then it gives – as for the corresponding one dimensional procedure – no reasonable solution. We will combine it, therefore, in Sect. 10.15 with method (ii), in order to obtain, when possible, the speed of (iii), but when necessary, the certainty of (ii).

In Sects. 10.8 through 10.15 very different methods for solving the same problem, the minimization of a function of  $n$  variables, will be discussed. In Sect. 10.16 we give information on how to choose one of the methods appropriate for the problem in question. Section 10.17 is dedicated to considerations of errors. In Sect. 10.18 several examples are discussed in detail.

Before we find the minimum  $x_m$  of a function, we would like to inquire briefly about the numerical accuracy we expect for  $x_m$ . The minimum is after all almost always determined by a comparison of values of the function at points close in  $x$ . If we solve (10.1.4) for  $(x - x_m)$ , we obtain

$$(x - x_m) = \sqrt{\frac{2[M(x) - M(x_m)]}{A}} \quad .$$

We assume that  $A$ , i.e., the second derivative of the function  $M$ , is of order of magnitude unity close to the minimum. (This need only be true approximately. In fact, in numerical calculations one always scales all of the quantities such that they are of order unity, i.e., not something like  $10^6$  or  $10^{-6}$ .) If we compute the function  $M$  with the precision  $\delta$  then the difference  $M(x) - M(x_m)$  is also known at best with a precision of  $\delta$ , i.e., two function values can not be considered as being significantly different if they differ by only  $\delta$ . For the corresponding  $x$  values one then has

$$(x - x_m) \approx \sqrt{\frac{2\delta}{A}} \approx \sqrt{\delta} \quad . \quad (10.1.19)$$

If the computer has  $n$  binary places available for representing the mantissa, then a value  $x$  can be represented with the precision (4.2.7)

$$\frac{\Delta x}{x} = 2^{-n} \quad .$$

For computing a value  $x$  one chooses therefore a relative precision

$$\varepsilon \geq 2^{-n} \quad ,$$

since it is clearly pointless to try to compute a value with a higher precision than that with which it can be represented. If  $x$  is computed iteratively, i.e., one

computes a series  $x_0, x_1, \dots$  of approximations for  $x$ , then one can truncate this series as soon as

$$\frac{|x_k - x_{k-1}|}{|x_k|} < \varepsilon$$

or

$$|x_k - x_{k-1}| < \varepsilon |x_k| \quad (10.1.20)$$

for a given  $\varepsilon$ . With this prescription we will have difficulties, however, if  $x_k = 0$ . We introduce, therefore, in addition to  $\varepsilon$  a constant  $t \neq 0$  and extend (10.1.20) to

$$|x_k - x_{k-1}| < \varepsilon |x_k| + t \quad . \quad (10.1.21)$$

The last task remaining is to choose the numerical values for  $\varepsilon$  and  $t$ . If  $x$  is the position of the minimum, then by (10.1.19) a value for  $\varepsilon$  must be chosen that is greater than or equal to the square root of the relative precision for the representation of a floating point number. With computations using “double precision” in Java there are  $n = 53$  binary places available for the representation of the mantissa. Then only the values

$$\varepsilon > 2^{-n/2} \approx 2 \cdot 10^{-8}$$

are reasonable. The quantity  $t$  corresponds to an absolute precision. Therefore it can be chosen to be considerably smaller.

## 10.2 Parabola Through Three Points

If three points  $(x_a, y_a), (x_b, y_b), (x_c, y_c)$  of a function are known, then we can determine the parabola

$$y = a_0 + a_1 x + a_2 x^2 \quad (10.2.1)$$

that passes through these points. Instead of (10.2.1) we can also represent the parabola by

$$y = c_0 + c_1(x - x_b) + c_2(x - x_b)^2 \quad . \quad (10.2.2)$$

This relationship is naturally valid for three given points, i.e.,

$$y_b = c_0$$

and

$$(y_a - y_b) = c_1(x_a - x_b) + c_2(x_a - x_b)^2 \quad ,$$

$$(y_c - y_b) = c_1(x_c - x_b) + c_2(x_c - x_b)^2 \quad .$$

From this we obtain

$$c_1 = C[(x_c - x_b)^2(y_a - y_b) - (x_a - x_b)^2(y_c - y_b)] \quad , \quad (10.2.3)$$

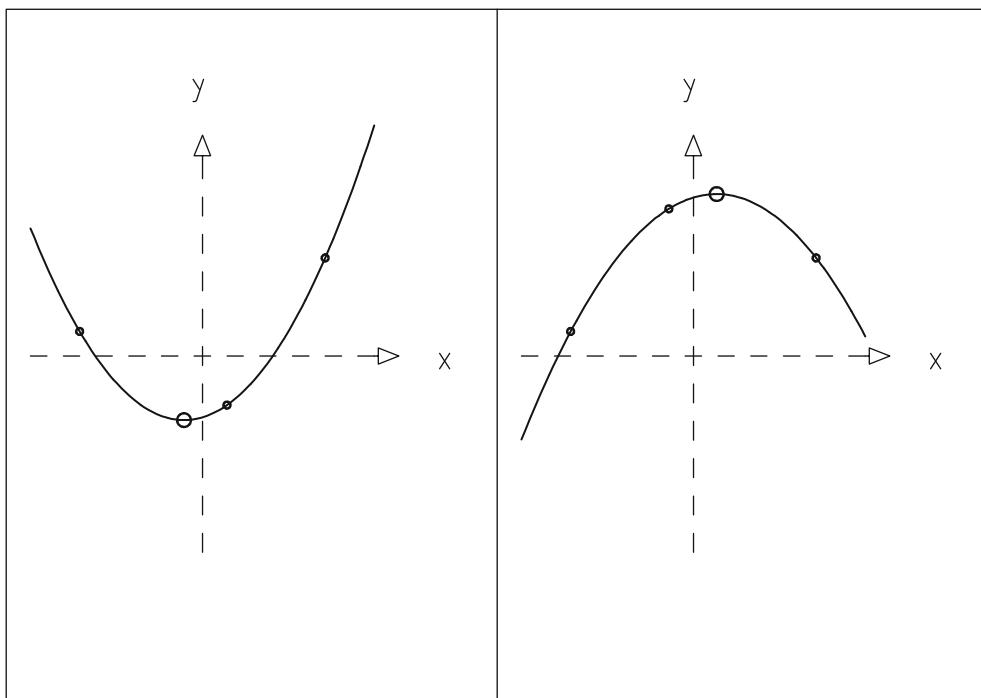
$$c_2 = C[-(x_c - x_b)(y_a - y_b) + (x_a - x_b)(y_c - y_b)] \quad (10.2.4)$$

with

$$C = \frac{1}{(x_a - x_b)(x_c - x_b)^2 - (x_c - x_b)(x_a - x_b)^2}$$

and for the extremum of the parabola

$$x_{\text{mp}} = x_b - \frac{c_1}{2c_2} \quad . \quad (10.2.5)$$



**Fig. 10.2:** Parabola through three points (*small circles*) and its extremum (*large circle*). In the *left figure* there is a minimum, and on the *right* a maximum.

The class `MinParab` performs this simple calculation. We must still determine whether the extremum of the parabola is a minimum or a maximum (cf. Fig. 10.2). One has a minimum if the second derivative of (10.2.2) with respect to  $x - x_b$  is positive, i.e., if  $c_2 > 0$ . We now order the three given points such that

$$x_a < x_b < x_c$$

and find that then

$$1/C = (x_c - x_b)(x_a - x_b)(x_c - x_a) = -(x_c - x_b)(x_b - x_a)(x_c - x_a) < 0 \quad .$$

With this one has for the sign of  $c_2$

$$\operatorname{sign} c_2 = \operatorname{sign}[(x_c - x_b)(y_a - y_b) + (x_b - x_a)(y_c - y_b)] .$$

Both expressions  $(x_c - x_b)$  and  $(x_b - x_a)$  are positive. Therefore for the extremum to be a minimum it is sufficient that

$$y_a > y_b , \quad y_c > y_b . \quad (10.2.6)$$

The condition is not necessary, but has the advantage of greater clarity. In the interval  $x_a < x < x_c$  one clearly has a minimum if there is a point  $(x_b, y_b)$  in the interval where the function has a value smaller than at the two end points. Clearly this statement is also valid if the function is not a parabola. We will make use of this fact in the next section.

### 10.3 Function of $n$ Variables on a Line in an $n$ -Dimensional Space

Locating the minimum of the function  $M(x)$  of a single variable  $x$  in an interval of the  $x$  axis is equivalent to locating the minimum of a function  $M(\mathbf{x})$  of an  $n$ -dimensional vector of variables  $\mathbf{x} = (x_1, x_2, \dots, x_n)$  with respect to a given line in the  $n$ -dimensional space. If  $\mathbf{x}_0$  is a fixed point and  $\mathbf{d}$  is a fixed vector, then

$$\mathbf{x}_0 + a\mathbf{d} , \quad -\infty < a < \infty , \quad (10.3.1)$$

describes a fixed line (see Fig. 10.3), and

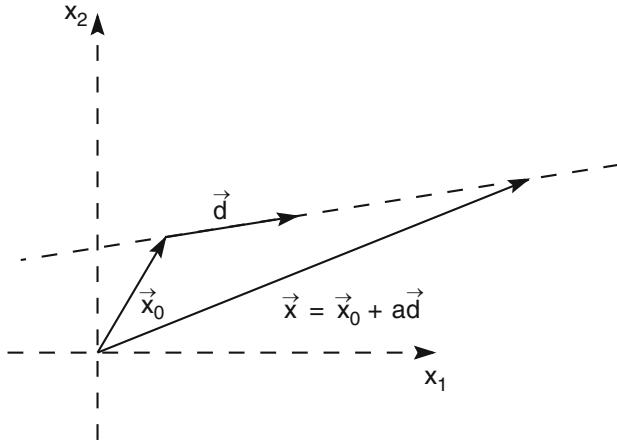
$$f(a) = M(\mathbf{x}_0 + a\mathbf{d}) \quad (10.3.2)$$

is the value of the function at the point  $a$  on this line. For  $n = 1$ ,  $\mathbf{x}_0 = 0$ ,  $\mathbf{d} = 1$  and with the change of notation  $a = x$ , that is,  $f(x) = M(x)$ , one recovers the original problem.

The class `FunctionOnline` computes the value (10.3.2); it makes use of an extension of the abstract class `DatanUserFunction`, to be provided by the user, which defines the function  $M(\mathbf{x})$ . In Sects. 10.4 through 10.6 we consider the minimum of a function of a single variable. The programs also treat, however, the case of a minimum of a function of  $n$  variables on a line in the  $n$ -dimensional space.

### 10.4 Bracketing the Minimum

For many minimization procedures it is important to know ahead of time that the minimum  $x_m$  is located in a specific interval,



**Fig. 10.3:** The line given by (10.3.1) in two dimensions.

$$x_a < x_m < x_c \quad . \quad (10.4.1)$$

By systematically reducing the interval, the position of the minimum can then be further constrained until finally for a given precision  $\varepsilon$  one has

$$|x_a - x_c| < \varepsilon \quad . \quad (10.4.2)$$

There is, in fact, a minimum in the interval (10.4.1) if there is an  $x$  value  $x_b$  such that

$$M(x_b) < M(x_a) \quad , \quad M(x_b) < M(x_c) \quad , \quad x_a < x_b < x_c \quad . \quad (10.4.3)$$

The class `MinEnclose` attempts to bracket the minimum of a function by giving values  $x_a, x_b, x_c$  with the property (10.4.3). The program is based on a similar subroutine by PRESS et al. [12]. Starting from the input values  $x_a, x_b$ , which, if necessary, are relabeled so that  $y_b \leq y_a$ , a value  $x_c = x_b + p(x_b - x_a)$  is computed that presumably lies in the direction of decreasing function value, i.e., closer to the minimum. The factor  $p$  in our program is set to  $p = 1.618034$ . In this way the original interval  $(x_a, x_b)$  is enlarged by the ratio of the golden section (cf. Sect. 10.5). The goal is reached if  $y_c > y_b$ . If this is not the case, a parabola is constructed through the three points  $(x_a, y_a), (x_b, y_b), (x_c, y_c)$ , whose minimum is at  $x_m$ .

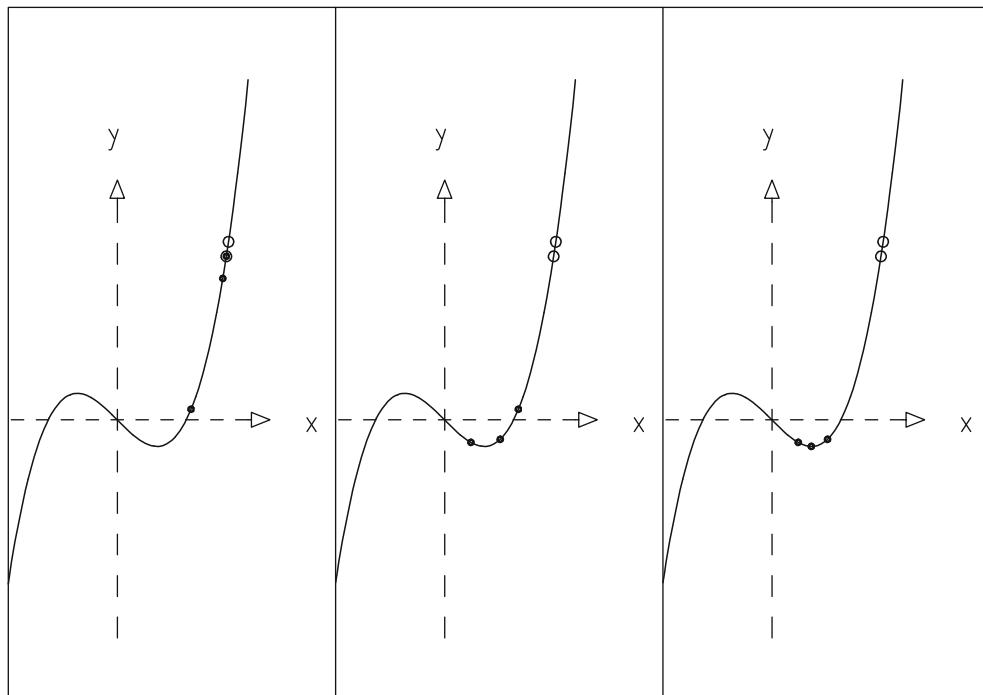
We now examine the point  $(x_m, y_m)$ . Here one must distinguish between various cases:

(a)  $x_b < x_m < x_c$ :

- (a1)  $y_m < y_c$ :  $(x_b, x_m, x_c)$  is the desired interval.
- (a2)  $y_b < y_m$ :  $(x_a, x_b, x_m)$  is the desired interval.
- (a3)  $y_m > y_c$  and  $y_m < y_b$ : There is no minimum. The interval will be extended further to the right.

- (b)  $x_c < x_m < x_{\text{end}}$  and  $x_{\text{end}} = x_b + f(x_c - x_b)$  and  $f = 10$  in our program.
- (b1)  $y_m > y_c$ :  $(x_b, x_c, x_m)$  is the desired interval.
- (b2)  $y_m < y_c$ : There is no minimum. The interval will be extended further to the right.
- (c)  $x_{\text{end}} < x_m$ : As a new interval  $(x_b, x_c, x_{\text{end}})$  is used.
- (d)  $x_m < x_b$ : This result is actually impossible. It can, however, be caused by a rounding error. The interval will be extended further to the right.

If the goal is not reached in the current step, a further step is taken with the new interval. Figure 10.4 shows an example of the individual steps carried out until the bracketing of the minimum is reached.



**Fig. 10.4:** Bracketing of a minimum with three points  $a, b, c$  according to (10.4.3). The initial values are shown as *larger circles*, and the results of the individual steps are shown as *small circles*.

## 10.5 Minimum Search with the Golden Section

As soon as the minimum has been enclosed by giving three points  $x_a, x_b, x_c$  with the property (10.4.3), the bracketing can easily be tightened further. One chooses a point  $x$  inside the larger of the two subintervals  $(x_a, x_b)$  and  $(x_b, x_c)$ .

If the value of the function at  $x$  is smaller than at  $x_b$ , then the subinterval containing  $x$  is taken as the new interval. If the value of the function is greater, then  $x$  is taken as the endpoint of the new interval.

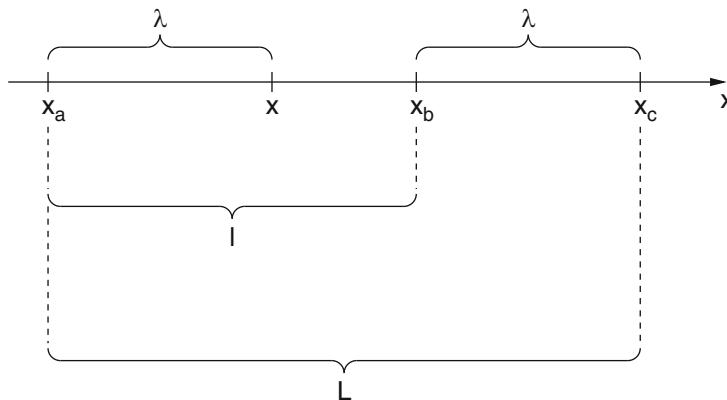
A particularly clever division of the intervals is possible with the *golden section*. Let us assume (see Fig. 10.5) that

$$g = \frac{\ell}{L} \quad , \quad g > \frac{1}{2} \quad , \quad (10.5.1)$$

is the length of the subinterval  $(x_a, x_b)$  (to be determined later) measured in units of the length of the full interval  $(x_a, x_c)$ . We now want to be able to divide the subinterval  $(x_a, x_b)$  again with a point  $x$  corresponding to a fraction  $g$ ,

$$g = \frac{\lambda}{\ell} \quad . \quad (10.5.2)$$

In addition, the points  $x$  and  $x_b$  should be situated symmetrically with respect to each other in the interval  $(x_a, x_c)$ , i.e.,



**Fig. 10.5:** The golden section.

$$\lambda = L - \ell \quad . \quad (10.5.3)$$

It follows that

$$\frac{\ell}{\lambda + \ell} = \frac{\lambda}{\ell} \quad ,$$

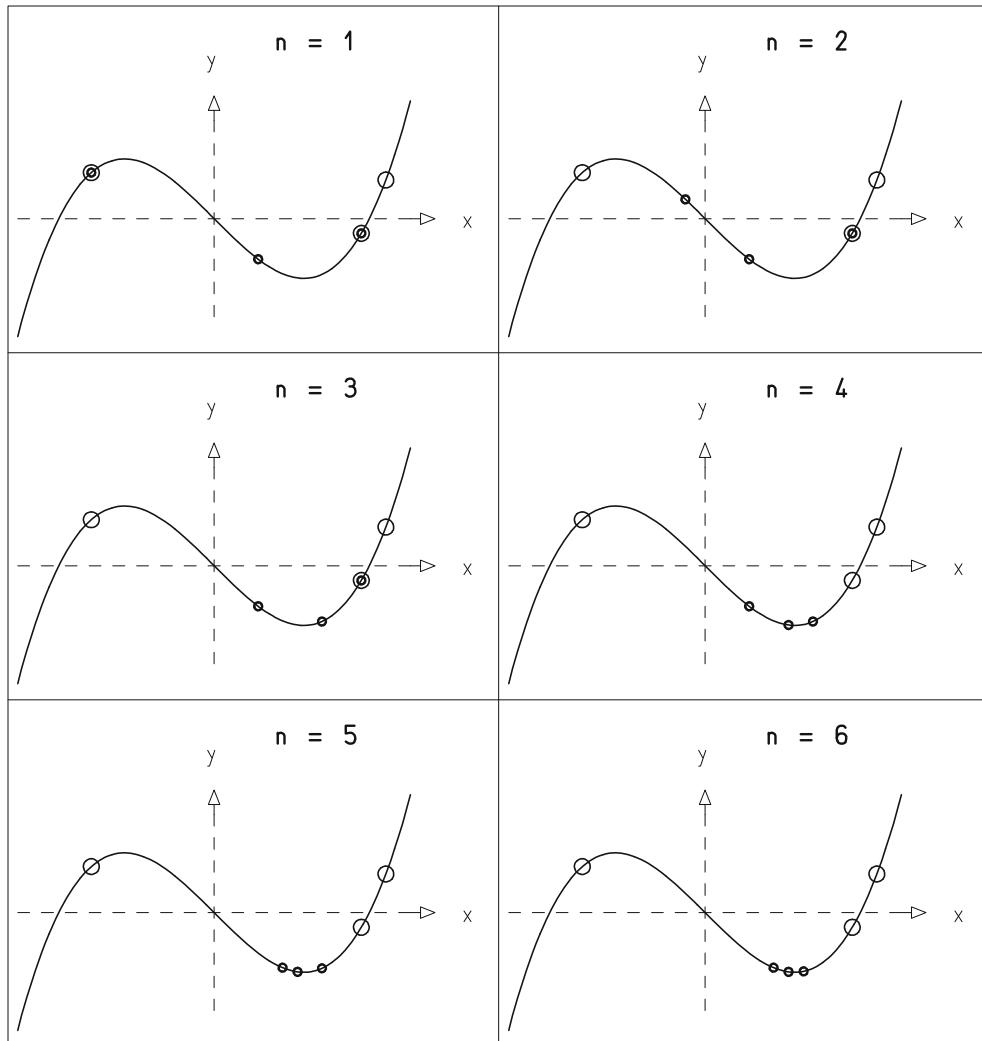
that is,

$$\lambda = \frac{\sqrt{5} - 1}{2} \ell$$

and

$$g = \frac{\sqrt{5} - 1}{2} \approx 0.618034 \quad . \quad (10.5.4)$$

As shown at the beginning of this section (for the case shown in Fig. 10.5  $x_b - x_a > x_c - x_b$ ), the minimum, which was originally only constrained to be in the interval  $(x_a, x_c)$ , now lies either in the interval  $(x_a, x_b)$  or in the interval  $(x, x_c)$ . By subdividing by the golden section one obtains intervals of equal size.



**Fig. 10.6:** Stepwise bracketing of a minimum by subdividing according to the golden section. The original interval (*larger circles*) is reduced with every step (*small circles*).

Figure 10.6 shows the first six steps of an example of minimization with the golden section.

## 10.6 Minimum Search with Quadratic Interpolation

From the example shown in Fig. 10.6 one sees that the procedure of interval subdivision is quite certain, but works slowly. We now combine it, therefore, with quadratic interpolation.

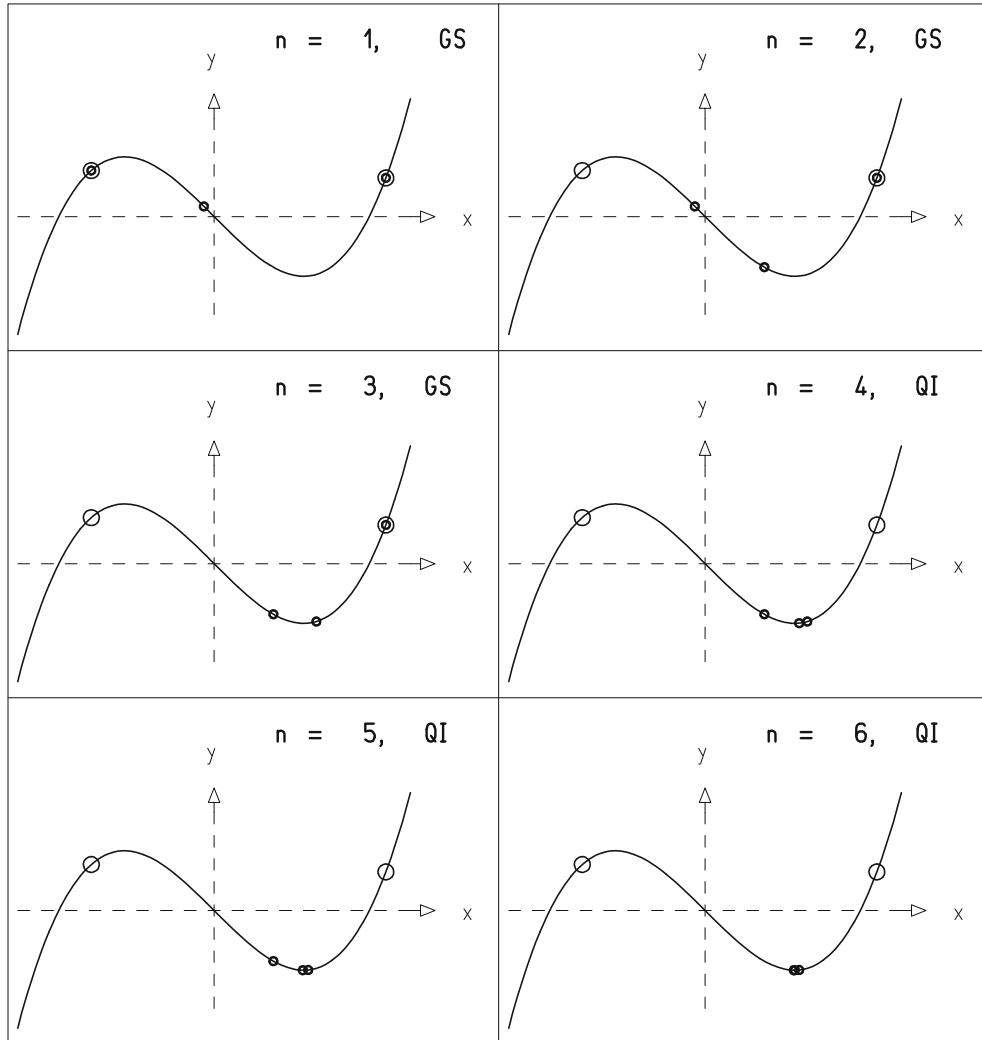
The class `MinCombined` is based on a program developed by BRENT [13], who first combined the two methods. The meanings of the most important variable names in the program are as follows. `a` and `b` denote the  $x$  values  $x_a$  and  $x_b$  that contain the minimum. `xm` is their mean value. `m` is the point at which the function has its lowest value up to this step, `w` is the point with the second lowest, and `v` is the point with the third lowest value of the function. `u` is the point where the function was last computed. One begins with the two initial values,  $x_a$  and  $x_b$ , that contain the minimum and adds a point  $x$ , that divides the interval  $(x_a, x_b)$  according to the golden section. Then in each subsequent iteration parabolic interpolation is attempted as in Sect. 10.2. The result is accepted if it lies in the interval defined by the last step *and* if in this step the change in the minimum is less than half as much as in the previous one. By this condition it is ensured that the procedure converges, i.e., that the steps become smaller on the average, although a temporary increase in step size is tolerated. If both conditions are not fulfilled, a reduction of the interval is carried out according to the golden section.

Numerical questions are handled in a particularly careful way of Brent. Starting from the two parameters  $\varepsilon$  and  $t$ , which define the relative precision, and the current value  $x$  for the position of the minimum, an absolute precision  $\Delta x = \varepsilon x + t = \text{tol}$  is computed according to (10.1.21). The iterations are continued until the half of the interval width falls below the value `tol` (i.e., until the distance from `x` to `xm` is not greater than `tol`) or until the maximum number of steps is reached. In addition the function is not computed at points that are separated by a distance less than `tol`, since such function values would not differ significantly.

The first six steps of an example of minimization according to Brent are shown in Fig. 10.7. Steps according to the golden section are marked by `GS`, and those from quadratic interpolation with `QI`. The comparison with Fig. 10.6 shows the considerably faster convergence achieved by quadratic interpolation.

## 10.7 Minimization Along a Direction in $n$ Dimensions

The class `MinDir` computes the minimum of a function of  $n$  variables along the line defined in Sect. 10.3 by `x0` and `d`. It first uses `MinEnclose` to bracket the minimum and then `MinCombined` to locate it exactly.



**Fig. 10.7:** Stepwise bracketing of a minimum with the combined method of Brent. The initial interval (*large circles*) is reduced with each step (*small circles*). Steps are carried out according to the golden section (GS) or by quadratic interpolation (QI).

The class `MinDir` is the essential tool used to realize a number of different strategies to find a minimum in  $n$  dimensional space, which we discuss in Sects. 10.9–10.15. A different type of strategy is the basis of the simplex method presented in the next section.

## 10.8 Simplex Minimization in $n$ Dimensions

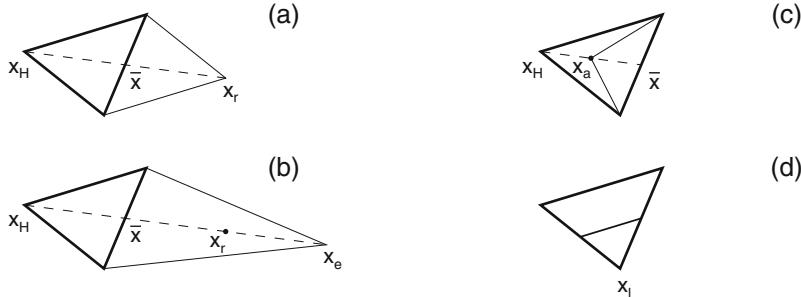
A simple and very elegant (although relatively slow) procedure for determining the minimum of a function of several variables is the simplex method by

NELDER and MEAD [14]. The variables  $x_1, x_2, \dots, x_n$  define an  $n$ -dimensional space. A *simplex* is defined in this space by  $n + 1$  points  $\mathbf{x}_i$ ,

$$\mathbf{x}_i = (x_{1i}, x_{2i}, \dots, x_{ni}) \quad . \quad (10.8.1)$$

A simplex in two dimensions is a triangle with the corner points  $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$ .

We use  $y_i$  to designate the value of the function at  $\mathbf{x}_i$  and use particular indices for labeling special points  $\mathbf{x}_i$ . At the point  $\mathbf{x}_H$  the value of the function is highest, i.e.,  $y_H > y_i, i \neq H$ . At the point  $\mathbf{x}_h$  it is higher than at all other points except  $\mathbf{x}_H$  ( $y_h > y_i, i \neq H, i \neq h$ ) and at the point  $\mathbf{x}_e$  it is lowest ( $y_e < y_i, i \neq e$ ). The simplex is now changed step by step. In each substep one of four operations takes place, these being *reflection*, *stretching*, *flattening*, or *contraction* of the simplex (cf. Fig. 10.8).



**Fig. 10.8:** Transformation of a simplex (*triangle with thick border*) into a new form (*triangle with thin border*) by means of a reflection (a), stretching (b), flattening (c), and contraction (d).

Using  $\bar{\mathbf{x}}$  to designate the center-of-gravity of the (hyper)surface of the simplex opposite to  $\mathbf{x}_H$ ,

$$\bar{\mathbf{x}} = \frac{1}{N-1} \sum_{i \neq H} \mathbf{x}_i \quad , \quad (10.8.2)$$

the *reflection* of  $\mathbf{x}_H$  is

$$\mathbf{x}_r = (1 + \alpha)\bar{\mathbf{x}} - \alpha\mathbf{x}_H \quad (10.8.3)$$

with  $\alpha > 0$  as the *coefficient of reflection*. The reflected simplex differs from the original one only in that  $\mathbf{x}_H$  is replaced by  $\mathbf{x}_r$ .

A *stretching* of the simplex consists of replacing  $\mathbf{x}_H$  by

$$\mathbf{x}_e = \gamma\mathbf{x}_r + (1 - \gamma)\bar{\mathbf{x}} \quad (10.8.4)$$

with the *coefficient of stretching*  $\gamma$ . One therefore chooses (for  $\gamma > 1$ ) a point along the line joining  $\mathbf{x}_H$  and  $\mathbf{x}_r$ , which is still further than the point  $\mathbf{x}_r$ .

In a *flattening*,  $\mathbf{x}_H$  is replaced by a point lying on the line joining  $\mathbf{x}_H$  and  $\bar{\mathbf{x}}$ . This point is located between the two points,

$$\mathbf{x}_a = \beta \mathbf{x}_H + (1 - \beta) \bar{\mathbf{x}} \quad . \quad (10.8.5)$$

The *coefficient of flattening*  $\beta$  is in the range  $0 < \beta < 1$ .

In the three operations discussed up to now only one point of the simplex is changed, that being the point  $\mathbf{x}_H$ , which corresponds to the highest value of the function. The point is displaced along the line by  $\mathbf{x}_H$  and  $\bar{\mathbf{x}}$ . After the displacement it either lies on the same side of  $\bar{\mathbf{x}}$  (flattening) or on the other side of  $\bar{\mathbf{x}}$  (reflection) or even far beyond  $\bar{\mathbf{x}}$  (stretching). In contrast to these operations, in a *contraction*, all points but one are replaced. The point  $\mathbf{x}_\ell$  where the function has its lowest value is retained. All remaining points are moved to the midpoints of the line segments joining them with  $\mathbf{x}_\ell$

$$\mathbf{x}_{ci} = (\mathbf{x}_i + \mathbf{x}_\ell)/2 \quad , \quad i \neq \ell \quad . \quad (10.8.6)$$

For the original simplex and for each one created by an operation, the points  $\bar{\mathbf{x}}$  and  $\mathbf{x}_r$  and the corresponding function values  $\bar{y}$  and  $y_r$  are computed. The next operation is determined as follows:

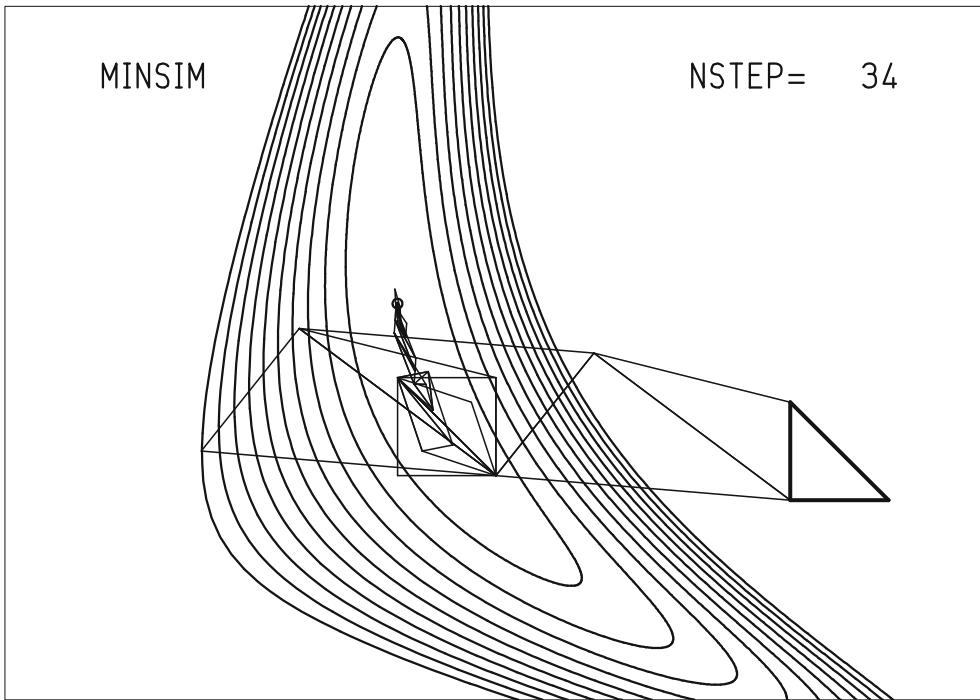
- (a) If  $y_r < y_\ell$ , a stretching is attempted. If this gives  $y_e < y_\ell$ , then the stretching is carried out. Otherwise one performs a *reflection*.
- (b) For  $y_r > y_H$  a reflection is carried out if  $y_r < y_H$ . Otherwise the simplex is unchanged. In each case this is followed by a *flattening*. If one obtains as a result of the flattening a point  $\mathbf{x}_a$  for which the function value is not less than both  $y_H$  and  $\bar{y}$ , the flattening is rejected, and instead a contraction is carried out.

After every step we examine the quantity

$$r = \frac{|y_H - y_\ell|}{|y_H| + |y_\ell|} \quad . \quad (10.8.7)$$

If this falls below a given value, the procedure is terminated and we regard  $\mathbf{x}_\ell$  as the point where the function has its minimum.

The class MinSim determines the minimum of a function of  $n$  variables by the simplex method. The program is illustrated in the example of Fig. 10.9. The triangle with the dark border is the initial simplex. The sequence of triangles with thin borders starting from it correspond to the individual transformations. One clearly recognizes as the first steps: stretching, stretching, reflection, reflection, flattening .... The simplex first finds its way into the



**Fig. 10.9:** Determining the minimum of a function of two variables with the simplex method. The function is shown by *contour lines* on which the function is constant. The function is highest on the outermost contour. The minimum is at the position of the *small circle* within the innermost contour. Each simplex is a *triangle*. The initial simplex is marked by *thicker lines*.

“valley” of the function and then runs along the bottom of the valley towards the minimum. It is reshaped in the process so that it is longest in the direction in which it is progressing, and in this way it can also pass through narrow valleys. It is worth remarking that this method, which is obtained by considering the problem in two and three dimensions, also works in  $n$  dimensions and also in this case possesses a certain descriptiveness.

## 10.9 Minimization Along the Coordinate Directions

Some of the methods for searching for a minimum in an  $n$ -dimensional space are based on the following principle. Starting from a point  $\mathbf{x}_0$  one searches for the minimum along a given direction in the space. Next one minimizes from there along another direction and finds a new minimum, etc. Within this general framework various *strategies* can be developed to choose the individual directions.

The simplest strategy consists in the choice of the coordinate directions in the space of the  $n$  variables  $x_i$ . We can label the corresponding basis vectors with  $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n$ , and they are then chosen in order as directions. After  $\mathbf{e}_n$  one begins again with  $\mathbf{e}_1, \mathbf{e}_2, \dots$ . A partial sequence of minimizations along all coordinate directions starting from  $\mathbf{x}_0$  gives the point  $\mathbf{x}_1$ . After a new partial sequence one obtains  $\mathbf{x}_2$ , etc.

The procedure is ended successfully when for the values of the function  $M(\mathbf{x}_n)$  and  $M(\mathbf{x}_{n-1})$  for two successive steps one has

$$M(\mathbf{x}_{n-1}) - M(\mathbf{x}_n) < \varepsilon |M(\mathbf{x}_n)| + t \quad , \quad (10.9.1)$$

i.e., a condition corresponding to (10.1.21) for given values of  $\varepsilon$  and  $t$ . We compare here, however, the value of the function  $M$  and not the independent variables  $\mathbf{x}$ . Otherwise we would have to compute the distance between two points in an  $n$ -dimensional space.

Figure 10.10 shows the minimization of the same function as in Fig. 10.9 with the coordinate-direction method. After the first comparatively large step, which leads into the “valley” of the function, the following steps are quite small. The individual directions are naturally perpendicular to each other. The “best” point at each step moves along a staircase-like path along the floor of the valley towards the minimum.

## 10.10 Conjugate Directions

The slow convergence seen in Fig. 10.10 clearly stems from the fact that when minimizing along one direction, one loses the result of the minimization with respect to another direction. We will now try to choose the directions in such a way that this does not happen. For this we suppose for simplicity that the function is a quadratic form (10.1.10). Its gradient at the point  $\mathbf{x}$  is then given by (10.1.12),

$$\nabla M = -\mathbf{b} + A\mathbf{x} \quad . \quad (10.10.1)$$

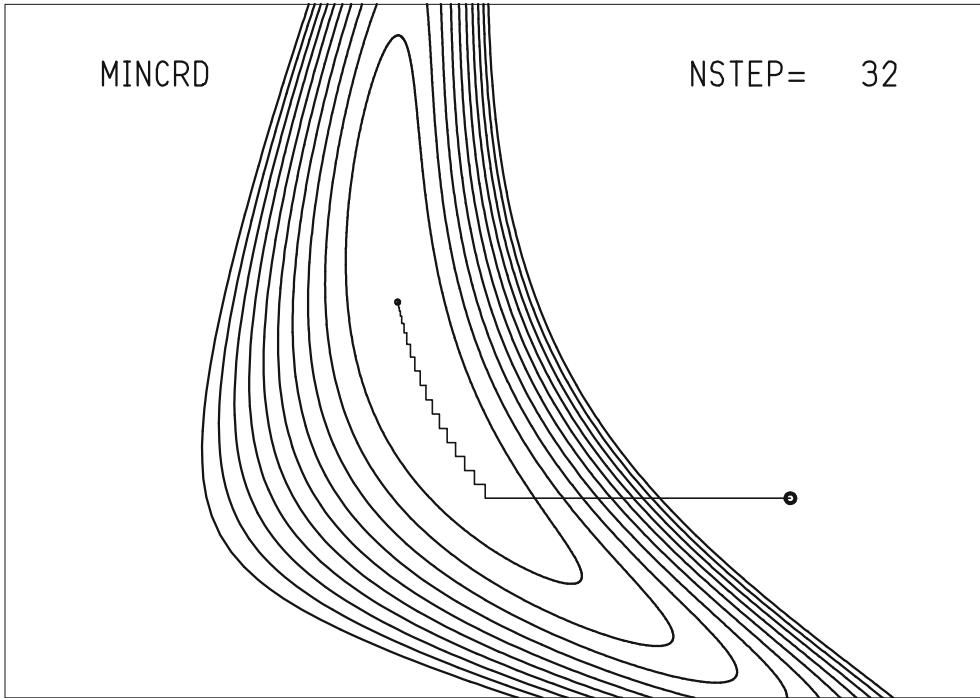
The change of the gradient in moving by  $\Delta\mathbf{x}$  is

$$\Delta(\nabla M) = \nabla M(\mathbf{x} + \Delta\mathbf{x}) - \nabla M(\mathbf{x}) = A\mathbf{x} + A\Delta\mathbf{x} - A\mathbf{x} = A\Delta\mathbf{x} \quad . \quad (10.10.2)$$

This expression is a vector that gives the direction of change of the gradient when the argument is moved in the direction  $\Delta\mathbf{x}$ .

If a minimization has been carried out along a direction  $\mathbf{p}$ , then the reduction in  $M$  with respect to  $\mathbf{p}$  is retained when one moves in a direction  $\mathbf{q}$  such that the gradient only changes perpendicular to  $\mathbf{p}$ , i.e., when one has

$$\mathbf{p} \cdot (A\mathbf{q}) = \mathbf{p}^T A\mathbf{q} = 0 \quad . \quad (10.10.3)$$



**Fig. 10.10:** Minimization along the coordinate directions. The starting point is shown by the *larger circle*, and the end by the *smaller circle*. The *line* shows the results of the individual steps.

The vectors  $\mathbf{p}$  and  $\mathbf{q}$  are said to be *conjugate* to each other with respect to the positive-definite matrix  $A$ . If one has  $n$  variables, i.e., if  $A$  is an  $n \times n$  matrix, then one can find in general  $n$  conjugate linearly independent vectors.

POWELL [15] has given a method to find a set of conjugate directions for a function described by a quadratic form. For this one chooses as a first set of directions  $n$  linearly independent unit vectors  $\mathbf{p}_i$ , e.g., the coordinate directions  $\mathbf{p}_i = \mathbf{e}_i$ . Starting from a point  $\mathbf{x}_0$  one finds successively the minima in the directions  $\mathbf{p}_i$ . The results can be labeled by

$$\begin{aligned}\mathbf{a}_1 &= \alpha_1 \mathbf{p}_1 + \alpha_2 \mathbf{p}_2 + \cdots + \alpha_n \mathbf{p}_n , \\ \mathbf{a}_2 &= \alpha_2 \mathbf{p}_2 + \cdots + \alpha_n \mathbf{p}_n , \\ &\vdots \\ \mathbf{a}_n &= \alpha_n \mathbf{p}_n .\end{aligned}$$

Here  $\mathbf{a}_1$  is the vector representing all  $n$  substeps,  $\mathbf{a}_2$  contains all of the steps except the first one, and  $\mathbf{a}_n$  is just the last substep. The sum of the  $n$  substeps leads then from the point  $\mathbf{x}_0$  to

$$\mathbf{x}_1 = \mathbf{x}_0 + \mathbf{a}_1 .$$

The direction  $\mathbf{a}_1$  describes the average direction of the first  $n$  substeps. Therefore we carry out a step in the direction  $\mathbf{a}_1$ , call the result again  $\mathbf{x}_0$ , determine a new set of directions

$$\begin{aligned}\mathbf{q}_1 &= \mathbf{p}_2 , \\ \mathbf{q}_2 &= \mathbf{p}_3 , \\ &\vdots \\ \mathbf{q}_{n-1} &= \mathbf{p}_n , \\ \mathbf{q}_n &= \mathbf{a}_1/|\mathbf{a}_1| ,\end{aligned}$$

then call these  $\mathbf{q}_i$  again  $\mathbf{p}_i$  and proceed as above. As was shown by POWELL [15], the directions after  $n$  steps, i.e.,  $n(n+1)$  individual minimizations, are individually conjugate to each other if the function is a quadratic form.

## 10.11 Minimization Along Chosen Directions

The procedure given at the end of the last section contains, however, the danger that the directions  $\mathbf{p}_1, \dots, \mathbf{p}_n$  can become almost linearly dependent, because at every step  $\mathbf{p}_1$  is rejected in favor of  $\mathbf{a}_1/|\mathbf{a}_1|$ , and these directions need not be very different from step to step. Powell has therefore suggested not to replace the direction  $\mathbf{p}_1$  by  $\mathbf{a}_1/|\mathbf{a}_1|$ , but rather to replace the direction  $\mathbf{p}_{\max}$ , along which the greatest reduction of the function took place. This sounds at first paradoxical, since what is clearly the best direction is replaced by another. But since out of all the directions  $\mathbf{p}_{\max}$  gave the largest contribution towards reducing the function,  $\mathbf{a}_1/|\mathbf{a}_1|$  has a significant component in the direction  $\mathbf{p}_{\max}$ . By retaining these two similar directions one would increase the danger of a linear dependence.

In some cases, however, after completing a step we will retain the old directions without change. We denote by  $\mathbf{x}_0$  the point before carrying out the step in progress, by  $\mathbf{x}_1$  the point obtained by this step, by

$$\mathbf{x}_e = \mathbf{x}_1 + (\mathbf{x}_1 - \mathbf{x}_0) = 2\mathbf{x}_1 - \mathbf{x}_0 \quad (10.11.1)$$

an extrapolated point that lies in the new direction  $\mathbf{x}_1 - \mathbf{x}_0$  from  $\mathbf{x}_0$  but is still further than  $\mathbf{x}_1$ , and by  $M_0, M_1$ , and  $M_e$  the corresponding function values. If

$$M_e \geq M_0 , \quad (10.11.2)$$

then the function no longer decreases significantly in the direction  $(\mathbf{x}_0 - \mathbf{x}_1)$ . We then remain with the previous directions.

Further we denote by  $\Delta M$  the greatest change in  $M$  along a direction in the step in progress and compute the quantity

$$T = 2(M_0 - 2M_1 + M_e)(M_0 - M_1 - \Delta M)^2 - (M_0 - M_e)^2 \Delta M . \quad (10.11.3)$$

If

$$T \geq 0 ,$$

then we retain the old directions. This requirement is satisfied if either the first or second factor in the first term of (10.11.3) becomes large. The first factor ( $M_0 - 2M_1 + M_e$ ) is proportional to a second derivative of the function  $M$ . If it is large (compared to the first derivative in meaningful units), then we are already close to the minimum. The second factor  $(M_0 - M_1 - \Delta M)^2$  is large when the reduction of the function  $M_0 - M_1$ , with the contribution  $\Delta M$ , does not stem mainly from a single direction.

The class MinPow determines the minimum of a function of  $n$  variables by successive minimization along chosen directions according to Powell. Figure 10.11 shows the advantages of that method. One can see a significantly faster convergence to the minimum compared to that of Fig. 10.10.

## 10.12 Minimization in the Direction of Steepest Descent

In order to get from a point  $\mathbf{x}_0$  to the minimum of a function  $M(\mathbf{x})$ , it is sufficient to always follow the negative gradient  $\mathbf{b}(\mathbf{x}) = -\nabla M(\mathbf{x})$ . It is thus obvious that one should look for the minimum along the direction  $\nabla M(\mathbf{x}_0)$ . One calls this point  $\mathbf{x}_1$ , searches then for the minimum along  $\nabla M(\mathbf{x}_1)$ , and so forth, until the termination requirement (10.9.1) is satisfied.

A comparison of the steps taken with this method, shown in Fig. 10.12, with those from minimization along a coordinate direction (Fig. 10.10) shows, however, a surprising similarity. In both cases, successive steps are perpendicular to each other. Thus the directions cannot be fitted to the function in the course of the procedure and convergence is slow.

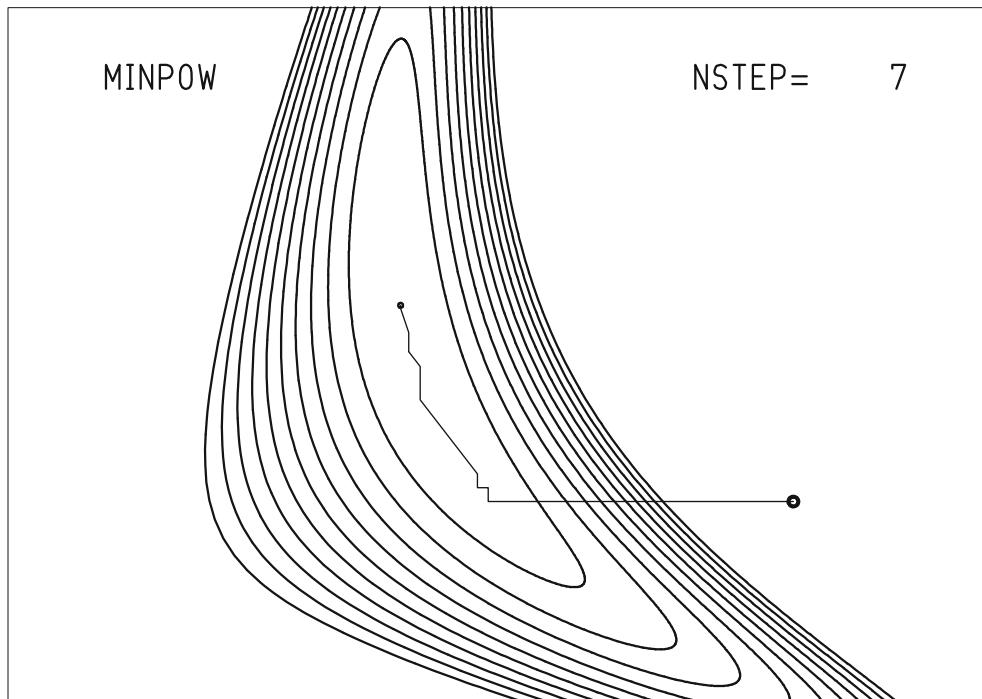
The initially surprising fact that successive gradient directions are perpendicular to each other stems from the construction of the procedure. Searching from  $\mathbf{x}_1$  for the minimum in the direction  $\mathbf{b}_0$  means that the derivative in the direction  $\mathbf{b}_0$  vanishes at the point  $\mathbf{x}_1$ ,

$$\mathbf{b}_0 \cdot \nabla M(\mathbf{x}_1) = 0 ,$$

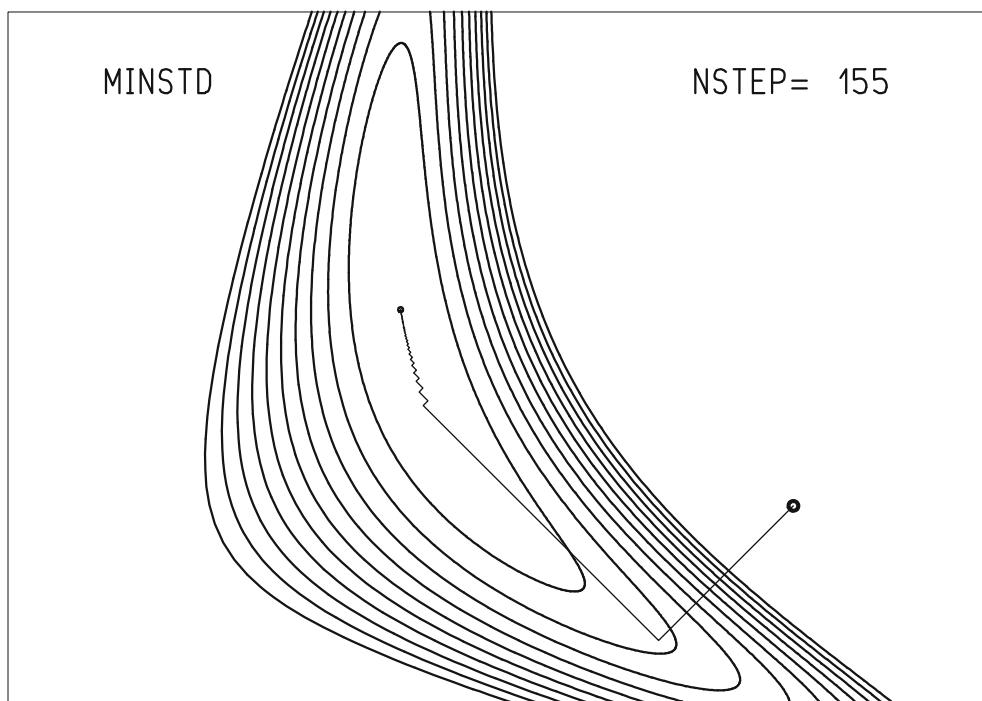
and thus the gradient is perpendicular to  $\mathbf{b}_0$ .

## 10.13 Minimization Along Conjugate Gradient Directions

We take up again the idea of conjugate directions from Sect. 10.10. We construct, starting from an arbitrary vector  $\mathbf{g}_1 = \mathbf{h}_1$ , two sequences of vectors,



**Fig. 10.11:** Minimization along a chosen direction with the method of Powell.



**Fig. 10.12:** Minimization in the direction of steepest descent.

$$\mathbf{g}_{i+1} = \mathbf{g}_i - \lambda_i A \mathbf{h}_i \quad , \quad i = 1, 2, \dots \quad , \quad (10.13.1)$$

$$\mathbf{h}_{i+1} = \mathbf{g}_{i+1} + \gamma_i \mathbf{h}_i \quad , \quad i = 1, 2, \dots \quad , \quad (10.13.2)$$

with

$$\lambda_i = \frac{\mathbf{g}_i^T \mathbf{g}_i}{\mathbf{g}_i^T A \mathbf{h}_i} \quad , \quad \gamma_i = \frac{\mathbf{g}_{i+1}^T A \mathbf{h}_i}{\mathbf{h}_i^T A \mathbf{h}_i} \quad . \quad (10.13.3)$$

Thus one has

$$\mathbf{g}_{i+1}^T \mathbf{g}_i = 0 \quad , \quad (10.13.4)$$

$$\mathbf{h}_{i+1}^T A \mathbf{h}_i = 0 \quad . \quad (10.13.5)$$

This means that successive vectors  $\mathbf{g}$  are orthogonal, and successive vectors  $\mathbf{h}$  are conjugate to each other.

Rewriting the relation (10.13.3) gives

$$\gamma_i = \frac{\mathbf{g}_{i+1}^T \mathbf{g}_{i+1}}{\mathbf{g}_i^T \mathbf{g}_i} = \frac{(\mathbf{g}_{i+1} - \mathbf{g}_i)^T \mathbf{g}_{i+1}}{\mathbf{g}_i^T \mathbf{g}_i} \quad , \quad (10.13.6)$$

$$\lambda_i = \frac{\mathbf{g}_i^T \mathbf{h}_i}{\mathbf{h}_i^T A \mathbf{h}_i} \quad . \quad (10.13.7)$$

We now try to construct the vectors  $\mathbf{g}_i$  and  $\mathbf{h}_i$  without explicit knowledge of the Hessian matrix  $A$ . To do this we again assume that the function to be minimized is a quadratic form (10.1.10). At a point  $\mathbf{x}_i$  we define the vector  $\mathbf{g}_i = -\nabla M(\mathbf{x}_i)$ . If we now search for the minimum starting from  $\mathbf{x}_i$  along the direction  $\mathbf{h}_i$ , we find it at  $\mathbf{x}_{i+1}$  and construct there  $\mathbf{g}_{i+1} = -\nabla M(\mathbf{x}_{i+1})$ . Then  $\mathbf{g}_{i+1}$  and  $\mathbf{g}_i$  are orthogonal, since from (10.1.12) one has

$$\mathbf{g}_i = -\nabla M(\mathbf{x}_i) = \mathbf{b} - A \mathbf{x}_i$$

and

$$\mathbf{g}_{i+1} = -\nabla M(\mathbf{x}_{i+1}) = b - A(\mathbf{x}_i + \lambda_i \mathbf{h}_i) = \mathbf{g}_i - \lambda_i A \mathbf{h}_i \quad . \quad (10.13.8)$$

Here  $\lambda_i$  has been chosen such that  $\mathbf{x}_{i+1}$  is the minimum along the direction  $\mathbf{h}_i$ . This means that there the gradient is perpendicular to  $\mathbf{h}_i$ ,

$$\mathbf{h}_i^T \nabla M(\mathbf{x}_{i+1}) = -\mathbf{h}_i^T \mathbf{g}_{i+1} = 0 \quad . \quad (10.13.9)$$

Substituting into (10.13.8) gives indeed

$$0 = \mathbf{h}_i^T \mathbf{g}_{i+1} = \mathbf{h}_i^T \mathbf{g}_i - \lambda_i \mathbf{h}_i^T A \mathbf{h}_i \quad ,$$

in agreement with (10.13.7).

From these results we can now construct the following algorithm. We begin at  $\mathbf{x}_0$ , construct there the gradient  $\nabla M(\mathbf{x}_0)$ , and set its negative equal to the two vectors

$$\mathbf{g}_1 = -\nabla M(\mathbf{x}_0) \quad , \quad \mathbf{h}_1 = -\nabla M(\mathbf{x}_0) \quad .$$

We minimize along  $\mathbf{h}_1$ . At the point of the minimum  $\mathbf{x}_1$  we construct  $\mathbf{g}_2 = -\nabla M(\mathbf{x}_1)$  and compute from (10.13.6)

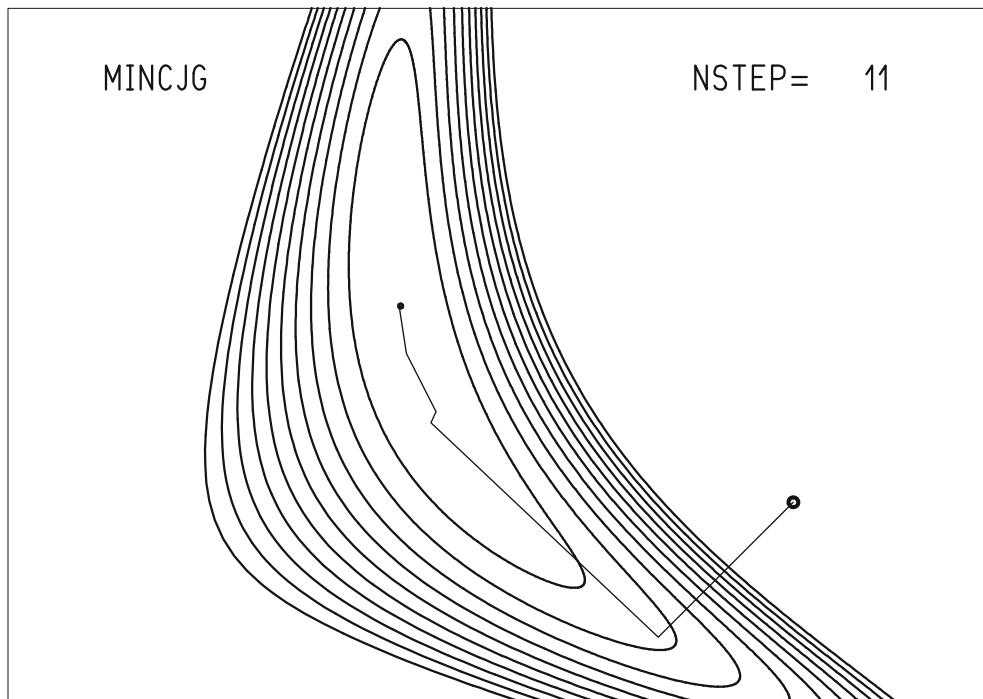
$$\gamma_1 = \frac{(\mathbf{g}_2 - \mathbf{g}_1)^T \mathbf{g}_1}{\mathbf{g}_1^T \mathbf{g}_1}$$

and from (10.13.2)

$$\mathbf{h}_2 = \mathbf{g}_1 + \gamma_1 \mathbf{h}_1 \quad .$$

From  $\mathbf{x}_1$  we then minimize along  $\mathbf{h}_2$ , and so forth.

The class **MinCJg** determines the minimum of a function of  $n$  variables by successive minimization along conjugate gradient directions. A comparison of Fig. 10.13 with Fig. 10.12 shows the superiority of the method of conjugate gradients over determination of direction according to steepest descent, especially near the minimum.



**Fig. 10.13:** Minimization along conjugate gradient directions.

## 10.14 Minimization with the Quadratic Form

If the function to be minimized  $M(\mathbf{x})$  is of the simple form (10.1.10), then the position of the minimum is given directly by (10.1.13). Otherwise one can always expand  $M(\mathbf{x})$  about a point  $\mathbf{x}_0$ ,

$$M(\mathbf{x}) = M(\mathbf{x}_0) - b(\mathbf{x} - \mathbf{x}_0) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_0)^T A(\mathbf{x} - \mathbf{x}_0) + \dots \quad (10.14.1)$$

with

$$\mathbf{b} = -\nabla M(\mathbf{x}_0) \quad , \quad A_{ik} = \frac{\partial^2 M}{\partial x_i \partial x_k} \quad (10.14.2)$$

and obtain as an approximation for the minimum

$$\mathbf{x}_1 = \mathbf{x}_0 + A^{-1}\mathbf{b} \quad . \quad (10.14.3)$$

One can now compute again  $\mathbf{b}$  and  $A$  as derivatives at the point  $\mathbf{x}_1$  and from this obtain a further approximation  $\mathbf{x}_2$  according to (10.14.3), and so forth.

For the case where the approximation (10.14.2) gives a good description of the function  $M(\mathbf{x})$ , the procedure converges quickly, since it tries to jump directly to the minimum. Otherwise it might not converge at all. We have already discussed the difficulties for the corresponding one-dimensional case in Sect. 10.1 with Fig. 10.1.

The class `MinQdr` finds the minimum of a function of  $n$  variables with the quadratic form. Figure 10.14 illustrates the operation of the method. One can observe that the minimum is in fact reached in very few steps.

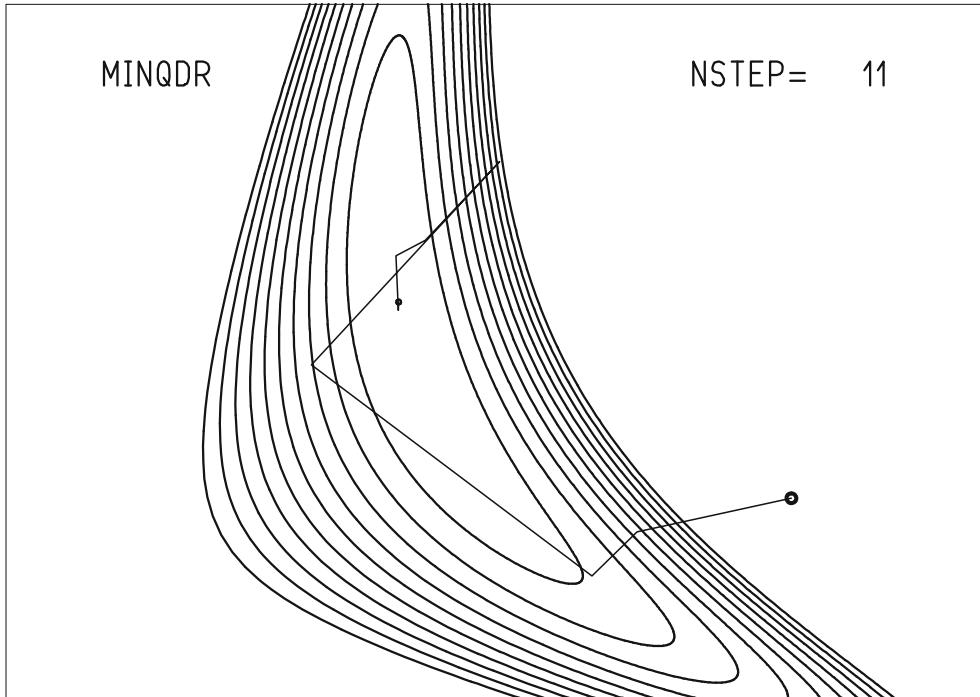
## 10.15 Marquardt Minimization

MARQUARDT [16] has given a procedure that combines the speed of minimization with the quadratic form near the minimum with the robustness of the method of steepest descent, where one finds the minimum even starting from a point far away. It is based on the following simple consideration.

The prescription (10.14.3), written as a computation of the  $i$ th approximation for the position of the minimum,

$$\mathbf{x}_i = \mathbf{x}_{i-1} + A^{-1}\mathbf{b} \quad , \quad (10.15.1)$$

means that one obtains  $\mathbf{x}_i$  from the point  $\mathbf{x}_{i-1}$  by taking a step along the vector  $A^{-1}\mathbf{b}$ . Here  $\mathbf{b} = -\nabla M(\mathbf{x}_{i-1})$  is the negative gradient, i.e., a vector in the direction of steepest descent of the function  $M$  at the point  $\mathbf{x}_{i-1}$ . If in (10.15.1) one had the unit matrix multiplied by a constant instead of the matrix  $A$ , i.e., if instead of (10.15.1) one were to use the prescription



**Fig. 10.14:** Minimization with quadratic form.

$$\mathbf{x}_i = \mathbf{x}_{i-1} + (\lambda I)^{-1} \mathbf{b} \quad , \quad (10.15.2)$$

then one would step by the vector  $\mathbf{b}/\lambda$  from  $\mathbf{x}_{i-1}$ . This is a step in the direction of steepest descent of the function, which is smaller for larger values of the constant  $\lambda$ . A sufficiently small step in the direction of steepest descent is, however, always a step towards the minimum (at least when one is still in the “approach” to the minimum, i.e., in the one-dimensional case of Fig. 10.1, between the two maxima). The Marquardt procedure consists of interpolating between the prescriptions (10.15.1) and (10.15.2) in such a way that the function  $M$  is reduced with every step, and such that the fast convergence of (10.15.1) is exploited as much as possible.

In place of (10.15.1) or (10.15.2) one computes

$$\mathbf{x}_i = \mathbf{x}_{i-1} + (A + \lambda I)^{-1} \mathbf{b} \quad . \quad (10.15.3)$$

Here  $\lambda$  is determined in the following way. One first chooses a fixed number  $v > 1$  and denotes by  $\lambda^{(i-1)}$  the value of  $\lambda$  from the previous step. As an initial value one chooses, e.g.,  $\lambda^{(0)} = 0.01$ . The value obtained from (10.15.3) of  $\mathbf{x}_i$  clearly depends on  $\lambda$ . One computes two points  $\mathbf{x}_i(\lambda^{(i-1)})$  and  $\mathbf{x}_i(\lambda^{(i-1)}/v)$ , where for  $\lambda$  one chooses the values  $\lambda^{(i-1)}$  and  $\lambda^{(i-1)}/v$ , and the corresponding function values  $M_i = M(\mathbf{x}_i(\lambda^{(i-1)}))$  and  $M_i^{(v)} = M(\mathbf{x}_i(\lambda^{(i-1)}/v))$ . These

are compared with the function value  $M_{i-1} = M(\mathbf{x}_{i-1})$ . The result of the comparison determines what happens next. The following cases are possible:

$$(i) \quad M_i^{(v)} \leq M_{i-1}:$$

One sets  $\mathbf{x}_i = \mathbf{x}_i(\lambda^{(i-1)}/v)$  and  $\lambda^{(i)} = \lambda^{(i-1)}/v$ .

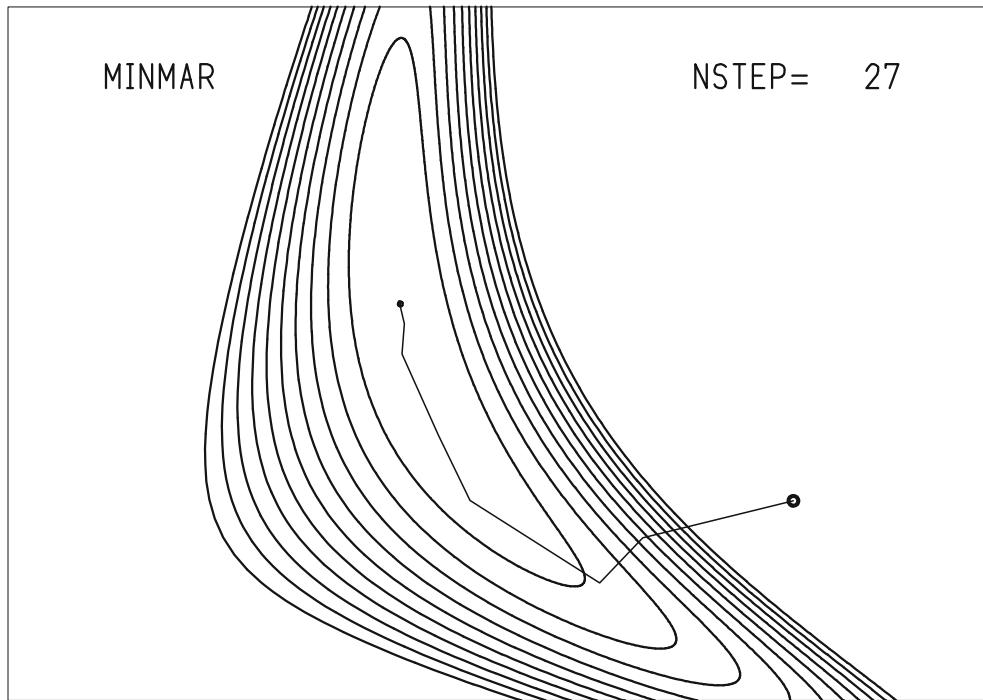
$$(ii) \quad M_i^{(v)} > M_{i-1} \text{ and } M_i \leq M_{i-1}:$$

One sets  $\mathbf{x}_i = \mathbf{x}_i(\lambda^{(i-1)})$  and  $\lambda^{(i)} = \lambda^{(i-1)}$ .

$$(iii) \quad M_i^{(v)} > M_{i-1} \text{ and } M_i > M_{i-1}:$$

One replaces  $\lambda^{(i-1)}$  by  $\lambda^{(i-1)}v$  and repeats the computation of  $\mathbf{x}_i(\lambda^{(i-1)}/v)$  and  $\mathbf{x}_i(\lambda^{(i-1)})$  and the corresponding function values, and repeats the comparisons.

In this way one ensures that the function value in fact decreases with each step and that the value of  $\lambda$  is always as small as possible when adjusted to the local situation. Clearly (10.15.3) becomes (10.15.1) for  $\lambda \rightarrow 0$ , i.e., it describes minimization with the quadratic form. For very large values of  $\lambda$ , Eq. (10.15.3) becomes the relation (10.15.2), which prescribes a small but sure step in the direction of steepest descent.



**Fig. 10.15:** Minimization with the Marquardt procedure.

The class `MinMar` finds the minimum of a function of  $n$  variables by Marquardt minimization. In Fig. 10.15 one sees the operation of the

Marquardt method. It follows a systematic path to the minimum. Comparing with Fig. 10.14 one sees that the method of the quadratic form converged in fewer steps. The Marquardt procedure, however, leads to the minimum in many cases where the method of the quadratic form would fail, e.g., in cases with poorly determined initial values.

## 10.16 On Choosing a Minimization Method

Given the variety of minimization methods, the user is naturally faced with the question of choosing a method appropriate to the task. Before we give recommendations for this, we will first recall the various methods once again.

The *simplex method* (routine MinSim) is particularly robust. Only function values  $M(\mathbf{x})$  are computed. The method is slow, however. Faster, but still quite robust is *minimization along a chosen direction* (routine MinPow). This also requires only function values.

The *method of conjugate gradients* (routine MinCjg) requires, as the name indicates, the computation of not only the function, but also its gradient. The number of iteration steps is, however, approximately equal to that of MinPow.

For *minimization with the quadratic form* (routine MinQdr) and *Marquardt minimization* (routine MinMar) one requires in addition the Hessian matrix of second derivatives. The derivatives are computed numerically with utility routines. The user can replace these utility routines by routines in which the analytic formulas for the derivatives are programmed. If the starting values are sufficiently accurate, MinQdr converges after just a few steps. The convergence is slower for MinMar. The method is, however, more robust; it often converges when starting from values from which MinQdr would fail.

From these characteristics of the methods we arrive at the following recommendations:

1. For problems that need only to be solved once or not many times, that is, in which the computing time does not play an important role, one should choose MinSim or MinPow.
2. For problems occurring repeatedly (with different numerical values) one should use MinMar. If one always has an accurate starting approximation, MinQdr can be used.
3. For repeated problems the derivatives needed for MinMar or MinQdr should be calculated analytically. Although this entails additional programming work, one gains precision compared to numerical derivatives and saves in many cases computing time.

At this point we should make an additional remark about the comparison of the minimization methods of this chapter with the method of least squares from Chap. 9. The method of least squares is a special case of minimization. The function to be minimized is a sum of squares, e.g., (9.1.8), or the generalization of a sum of squares, e.g., (9.5.9). In this generalized sum of squares there appears the matrix of derivatives  $A$ . These are not, however, the derivatives of the function to be minimized, but rather derivatives of a function  $\mathbf{f}$ , which characterizes the problem under consideration; cf. (9.5.2). Second derivatives are never needed. In addition, if one uses the singular value decomposition, as has been done in our programs in Chap. 9 to solve the least-squares problem, one works in numerically critical cases with a considerably higher accuracy than in the computation of sums of squares (cf. Sect. A.13, particularly Example A.4).

Least-squares problems should therefore always be carried out with the routines of Chap. 9. This applies particularly to problems of fitting of functions like those in the examples of Sect. 9.6, when one has many measured points. The matrix  $A$  then contains many rows, but few columns. In computing the function to be minimized the product  $A^T A$  is encountered, and one is threatened with the above mentioned loss of accuracy in comparison to the singular value decomposition.

## 10.17 Consideration of Errors

In data analysis minimization procedures are used for determining best estimates  $\tilde{\mathbf{x}}$  for unknown values  $\mathbf{x}$ . The function to be minimized  $M(\mathbf{x})$  is here usually a sum of squares (Chap. 9) or a log-likelihood function (multiplied by  $-1$ ) as in Chap. 7. In using the equations from Chap. 7 one must note, however, that there the  $n$ -vector that was called  $\lambda$  is now denoted by  $\mathbf{x}$ . The variables in Chap. 7 that were called  $x^{(i)}$  are the measured values (usually called  $\mathbf{y}$  in Chap. 9).

Information on the error of  $\tilde{\mathbf{x}}$  is obtained directly from results of Sects. 9.7, 9.8, and 9.13 by defining

$$H_{ik} = \left( \frac{\partial^2 M}{\partial x_i \partial x_k} \right)_{\mathbf{x}=\tilde{\mathbf{x}}} \quad (10.17.1)$$

as the elements of the symmetric matrix of second derivatives (the *Hessian matrix*) of the minimization function. Here one must note that the factor  $f_{QL}$  takes on the numerical value

$$f_{QL} = 1 \quad (10.17.2)$$

when the function being minimized is a sum of squares. If the function is a log-likelihood function (times  $-1$ ) then this must be set equal to

$$f_{QL} = 1/2 \quad . \quad (10.17.3)$$

1. *Covariance matrix. Symmetric errors.* The covariance matrix of  $\tilde{\mathbf{x}}$  is

$$C_{\tilde{\mathbf{x}}} = 2 f_{QL} H^{-1} \quad . \quad (10.17.4)$$

The square roots of the diagonal elements are the (symmetric) *errors*

$$\Delta \tilde{x}_i = \sqrt{c_{ii}} \quad . \quad (10.17.5)$$

It is only meaningful to give the covariance matrix if the measurement errors are small and/or there are many measurements, i.e., the requirements for (7.5.8) are fulfilled.

2. *Confidence ellipsoid. Symmetric confidence limits.* The covariance matrix defines the covariance ellipsoid; cf. Sects. 5.10 and A.11. Its center is given by  $\mathbf{x} = \tilde{\mathbf{x}}$ . The probability that the true value of  $\mathbf{x}$  is contained inside the ellipsoid is given by (5.10.20). The ellipsoid for which this probability has a given value  $W$ , the *confidence level*, is given by the *confidence matrix*

$$C_{\tilde{\mathbf{x}}}^{(W)} = \chi_W^2(n_f) C_{\tilde{\mathbf{x}}} \quad . \quad (10.17.6)$$

Here  $\chi_W^2(n_f)$  is the quantile of the  $\chi^2$ -distribution for  $n_f$  degrees of freedom and probability  $P = W$ ; cf. (5.10.19) and (C.5.3). The number of degrees of freedom  $n_f$  is equal to the number of measured values minus the number of parameters determined by the minimization. The square roots of the diagonal elements of  $C_{\tilde{\mathbf{x}}}^{(W)}$  are the distances from the *symmetric confidence limits*,

$$x_{i\pm}^{(W)} = \tilde{x}_i \pm \sqrt{c_{ii}^{(W)}} \quad . \quad (10.17.7)$$

The class `MinCov` yields the covariance or confidence matrix, respectively, for parameters determined by minimization.

3. *Confidence region.* If giving a covariance or confidence ellipsoid is not meaningful, it is still possible to give a confidence region with confidence level  $W$ . It is determined by the hypersurface

$$M(\mathbf{x}) = M(\tilde{\mathbf{x}}) + \chi_W^2(n_f) f_{QL} \quad . \quad (10.17.8)$$

With the help of the following routine a contour of a cross section through this hypersurface is drawn in a plane that contains the point  $\tilde{\mathbf{x}}$  and is parallel to  $(x_i, x_j)$ . Here  $x_i$  and  $x_j$  are two components of the vector of parameters  $\mathbf{x}$ . The boundary of the confidence region in a plane spanned by two parameters, which were found by minimization, can be graphically shown with the method `DatanGraphics.drawContour`, cf. Examples 10.1–10.3 and Example Programs 10.2–10.4.

4. *Asymmetric errors and confidence limits.* If the confidence region is not an ellipsoid, the asymmetric *confidence limits* for the variable  $x_i$  can still be determined by

$$\min \left\{ M(\tilde{\mathbf{x}}); x_i = x_{i\pm}^{(W)} \right\} = M(\tilde{\mathbf{x}}) + \chi_W^2(n_f) f_{QL} . \quad (10.17.9)$$

The differences

$$\begin{aligned} \Delta x_{i+}^{(W)} &= x_{i+}^{(W)} - \tilde{x}_i , \\ \Delta x_{i-}^{(W)} &= \tilde{x}_i - x_{i-}^{(W)} \end{aligned} \quad (10.17.10)$$

are the (asymmetric) distances from the confidence limits. If one takes  $\chi_W^2(n_f) = 1$ , the *asymmetric errors*  $\Delta x_{i+}$  and  $\Delta x_{i-}$  are obtained. The class MinAsy yields asymmetric errors or the distances from the confidence limits, respectively, for parameters, determined by minimization.

## 10.18 Examples

**Example 10.1:** Determining the parameters of a distribution from the elements of a sample with the method of maximum likelihood

Suppose one has  $N$  measurements  $y_1, y_2, \dots, y_n$  that can be assumed to come from a normal distribution with expectation value  $a = x_1$  and standard deviation  $\sigma = x_2$ . The likelihood function is

$$L = \prod_{i=1}^N \frac{1}{x_2 \sqrt{2\pi}} \exp \left\{ -\frac{(y_i - x_1)^2}{2x_2^2} \right\} \quad (10.18.1)$$

and its logarithm is

$$\ell = - \sum_{i=1}^N \frac{(y_i - x_1)^2}{2x_2^2} - N \ln\{x_2 \sqrt{2\pi}\} . \quad (10.18.2)$$

The task of determining the maximum-likelihood estimators  $\tilde{x}_1, \tilde{x}_2$  has already been solved in Example 7.8 by setting the analytically calculated derivatives of the function  $\ell(\mathbf{x})$  to zero. Here we will accomplish this by means of numerical minimization of  $\ell(\mathbf{x})$ . For this we must first provide a function that computes the function to be minimized,

$$M(\mathbf{x}) = -\ell(\mathbf{x}) .$$

This simple example is implemented in Example Programs 10.2 and 10.3. The results of the minimization of this user function are shown in Fig. 10.16 for two samples. Confidence regions and covariance ellipses agree reasonably well with each other. The agreement is better for the larger sample. ■

**Example 10.2:** Determination of the parameters of a distribution from the histogram of a sample by maximizing the likelihood

Instead of the original sample  $y_1, y_2, \dots, y_N$  as in Example 10.1, one often uses the corresponding histogram. Denoting by  $n_i$  the number of observations that fall into the interval centered about the point  $t_i$  with width  $\Delta t$ ,

$$t_i - \Delta t/2 \leq y < t_i + \Delta t/2 , \quad (10.18.3)$$

the histogram is given by the pairs of numbers

$$(t_i, n_i) , \quad i = 1, 2, \dots, n . \quad (10.18.4)$$

If the original sample is taken from a normal distribution with expectation value  $x_1 = a$  and standard deviation  $x_2 = \sigma$ , i.e., from the probability density

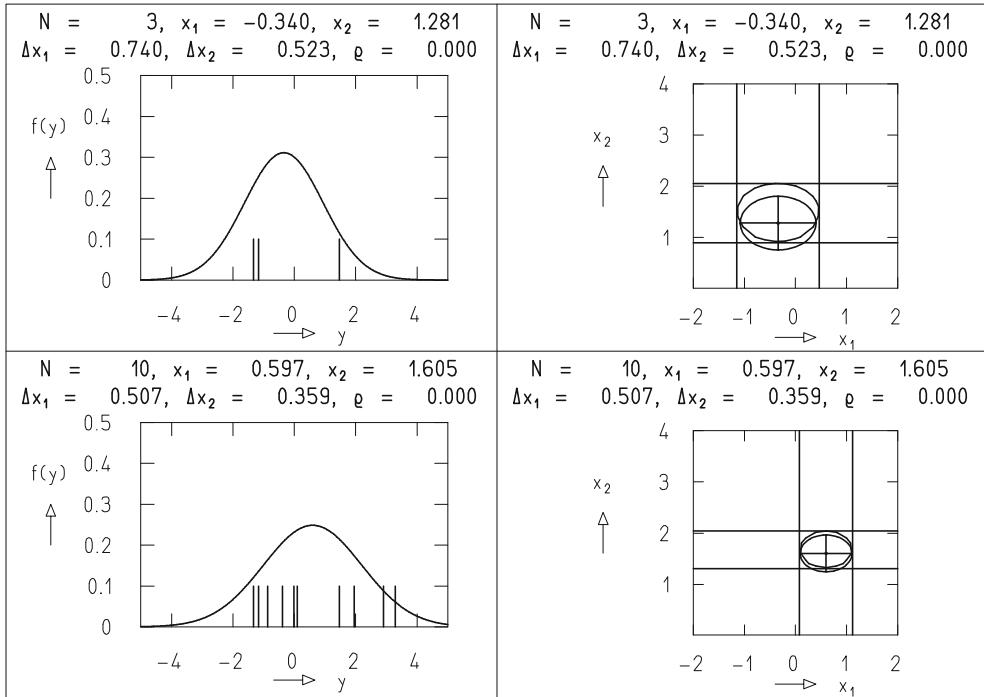
$$f(t; x_1, x_2) = \frac{1}{x_2 \sqrt{2\pi}} \exp\left\{-\frac{(t - x_1)^2}{2x_2^2}\right\} , \quad (10.18.5)$$

then one might expect the  $n_i(t_i)$  (at least in the limit  $N \rightarrow \infty$ ) to be given by

$$g_i = N \Delta t f(t_i; x_1, x_2) . \quad (10.18.6)$$

The quantities  $n_i(t_i)$  are integers, which are clearly not equal in general to  $g_i$ . We can, however, regard each  $n_i(t_i)$  as a sample of size one from a Poisson distribution with the expectation value

$$\lambda_i = g_i . \quad (10.18.7)$$



**Fig. 10.16:** Determination of the parameters  $x_1$  (mean) and  $x_2$  (width) of a Gaussian distribution by maximization of the log-likelihood function of a sample. The plots on the left show two different samples, marked as one-dimensional scatter plots (tick marks) on the  $y$  axis. The curves  $f(y)$  are Gaussian distributions with the fitted parameters. The right-hand plots show in the  $(x_1, x_2)$  plane the values of the parameters obtained with symmetric errors and covariance ellipses, as well as the confidence region for  $\chi^2_W = 1$  and the corresponding confidence boundaries (horizontal and vertical lines).

The a posteriori probability to observe the value  $n_i(t_i)$  is clearly

$$\frac{1}{n_i!} \lambda_i^{n_i} e^{-\lambda_i} . \quad (10.18.8)$$

The likelihood function for the observation of the entire histogram is

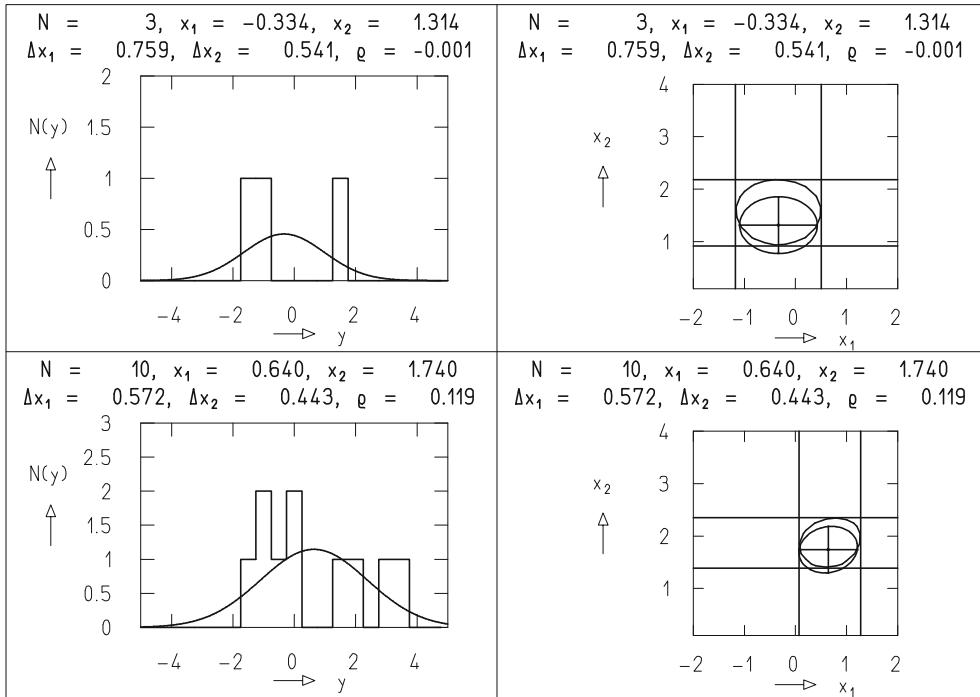
$$L = \prod_{i=1}^n \frac{1}{n_i!} \lambda_i^{n_i} e^{-\lambda_i} , \quad (10.18.9)$$

and its logarithm is

$$\ell = - \sum_{i=1}^n \ln n_i! + \sum_{i=1}^n n_i \ln \lambda_i - \sum_{i=1}^n \lambda_i . \quad (10.18.10)$$

If we use for  $\lambda_i$  the notation of (10.18.7) and find the minimum of  $-\ell$  with respect to  $x_1, x_2$ , then this determines the best estimates of the parameters  $x_1, x_2$ . Of course the same procedure can be applied not only in the

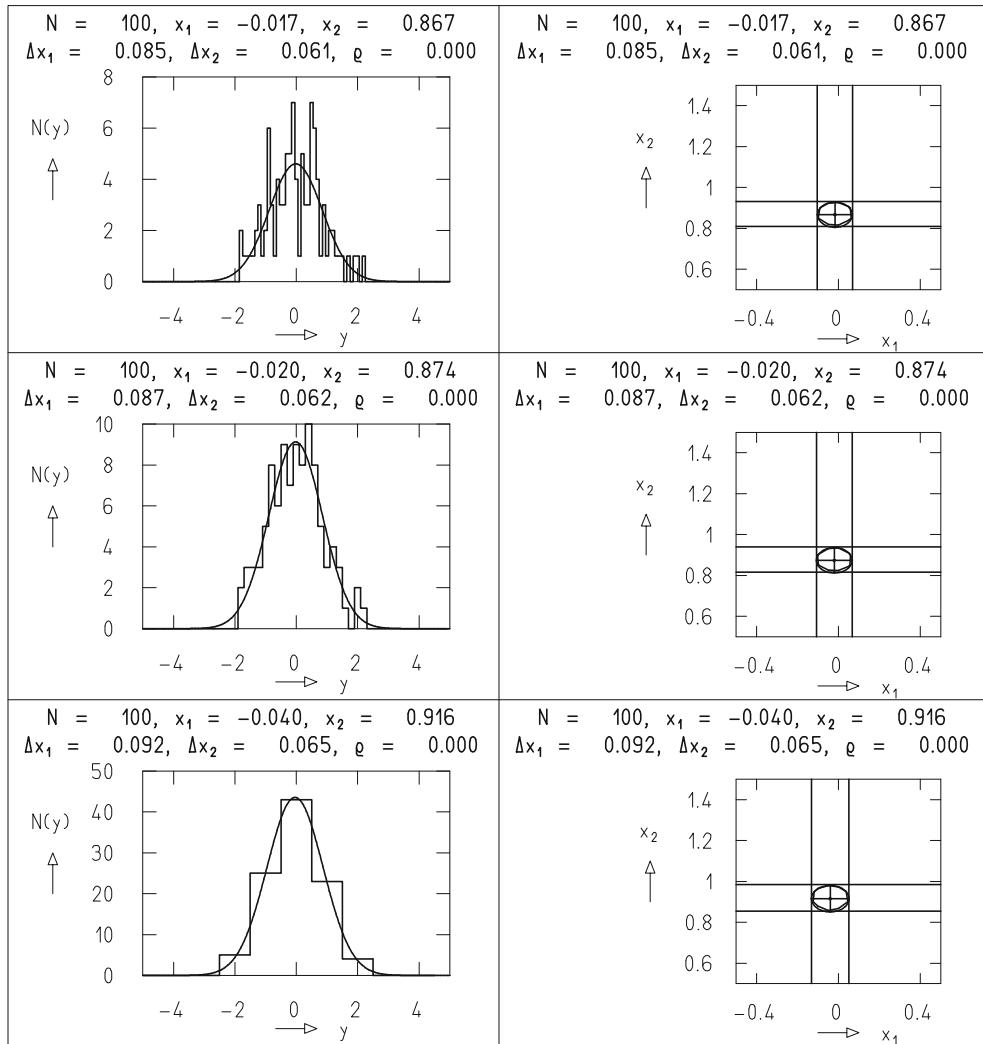
case of a simple Gaussian distribution, but for any distribution that depends on parameters. One must simply use the corresponding probability density in (10.18.5) and (10.18.6). In the user function given below one must change only one instruction in order to replace the Gaussian by another distribution. This example is implemented in Example Program 10.4. There the user function carries the name MinLogLikeHistPoisson.



**Fig. 10.17:** Determination of the parameters  $x_1$  (mean) and  $x_2$  (width) of a Gaussian distribution by maximizing the log-likelihood function of a histogram. The *left-hand plots* show two histograms corresponding to the samples from Fig. 10.16. The curves are the Gaussian distributions normalized to the histograms. The *right-hand plots* show the symmetric errors, covariance ellipse, and confidence region represented as in Fig. 10.16.

The results of the minimization with this user function are shown in Fig. 10.17 for two histograms, which are based on the samples from Fig. 10.16. The results are very similar to those from Example 10.1. The errors of the parameters, however, are somewhat larger. This is to be expected, since some information is necessarily lost when constructing a histogram from the sample.

A histogram can be viewed as a sample in a compressed representation. The compression becomes greater as the bin width of the histogram increases. This is made clear in Fig. 10.18. One sees that for the same sample, the errors of the determined parameters increase for greater bin width. The effect is relatively small, however, for the relatively large sample size in this case. ■



**Fig. 10.18:** As in Fig. 10.17, but for histograms with different interval widths of the same sample.

**Example 10.3:** Determination of the parameters of a distribution from the histogram of a sample by minimization of a sum of squares

If the bin contents  $n_i$  of a histogram (10.18.4) are sufficiently large, then the statistical fluctuations of each  $n_i$  can be approximated by a Gaussian distribution with the standard deviation

$$\Delta n_i = \sqrt{n_i} . \quad (10.18.11)$$

(See Sect. 6.8.) The weighted sum of squares describing the deviation of the histogram contents  $n_i$  from the expected values  $g_i$  from (10.18.6) is then

$$Q = \sum_{i=1}^N \frac{(n_i - g_i)^2}{n_i} . \quad (10.18.12)$$

When carrying out the sum one must take care (in contrast to Example 10.2) that empty bins ( $n_i = 0$ ) are not included. Even better is to not include bins in the sum even with low bin contents, e.g.,  $n_i < 4$ .

Also this example is implemented in Example Program 10.4. Here  $g_i$  is given by (10.18.6) as in the previous example. The sum is carried out over all bins with  $n_i > 0$ . The user function is called MinHistSumOfSquares. The statistical fluctuations of the bin contents are taken as approximately Gaussian.

Figure 10.19 shows the results obtained by minimizing the quadratic sum for the histogram in Fig. 10.18. Since the histograms are based on the same sample, the values  $n_i$  decrease for decreasing bin width, whereby the requirement for using the sum of squares becomes less well fulfilled. Thus we cannot trust as much the results nor the errors for smaller bin width. One sees, however, that the errors given by the procedure in fact increase for decreasing bin width. ■

We emphasize that the determination of parameters from a histogram by quadratic-sum minimization gives less exact results than those obtained by likelihood maximization. This is because the assumption of a normal distribution of the  $n_i$  with the width (10.18.11) is only an approximation, which often requires large bin widths for the histogram and thus implies a loss of information. If, however, enough data, i.e., sufficiently large samples, are available, then the difference between the two procedures is small. One should compare, e.g., Fig. 10.18 (upper right-hand plot) with Fig. 10.19 (lower right-hand plot).

## 10.19 Java Classes and Example Programs

### Java Classes for Minimization Problems

MinParab finds the extremum of a parabola through three given points.

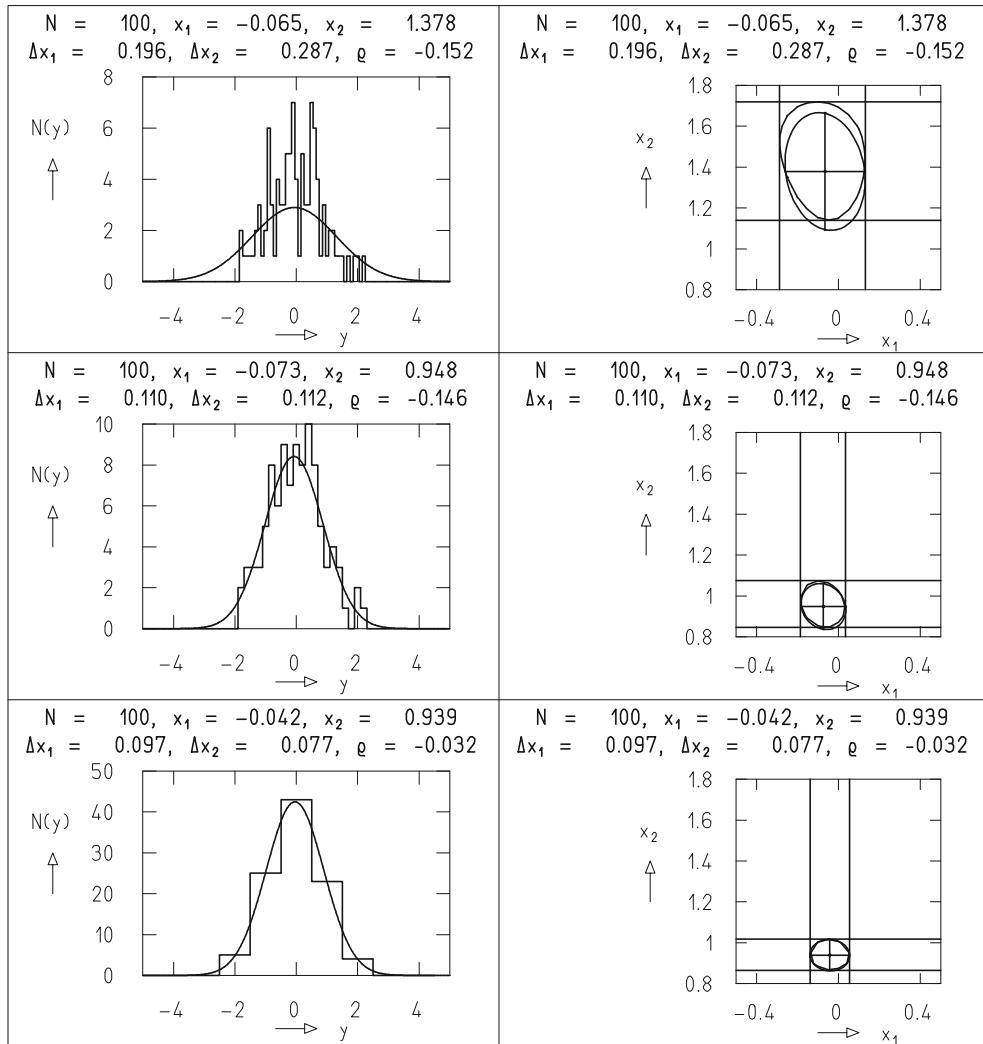
FunctionOnline computes the value of a function on a straight line in  $n$ -dimensional space.

MinEnclose brackets the minimum on a straight line in  $n$ -dimensional space.

MinCombined finds the minimum in a given interval along a straight line with a combined method according the Brent.

MinDir finds the minimum on a straight line in  $n$ -dimensional space.

MinSim finds the minimum in  $n$ -dimensional space using the simplex method.



**Fig. 10.19:** Determination of the parameters  $x_1$  (mean) and  $x_2$  (width) of a Gaussian distribution by minimization of a weighted sum of squares. The histograms on the left are the same as in Fig. 10.18. The fit results and errors, covariance ellipses, and confidence regions are represented as in Figs. 10.17 and 10.18.

MinPow finds the minimum in  $n$ -dimensional space with Powell's method of chosen directions.

MinCjg finds the minimum in  $n$ -dimensional space with the method of conjugate directions.

MinQdr finds the minimum in  $n$ -dimensional space with quadratic form.

MInMar finds the minimum in  $n$ -dimensional space with Marquardt's method.

MinCov finds the covariance matrix of the coordinates of the minimum.

MinAsy finds the asymmetric errors of the coordinates of the minimum.

**Example Program 10.1:** The class E1Min demonstrates the use of  
MinSim, MinPow, MinCjg, MinQdr, and MinMar

The program calls one of the classes, as requested by the user, in order to solve the following problem. One wants to find the minimum of the function  $f = f(\mathbf{x}) = f(x_1, x_2, x_3)$ . The search is started at the point  $\mathbf{x}^{(in)} = (x_1^{(in)}, x_2^{(in)}, x_3^{(in)})$ , which is also input by the user. The program treats consecutively four different cases:

- (i) No variables are fixed,
- (ii)  $x_3$  is fixed,
- (iii)  $x_2$  and  $x_3$  are fixed,
- (iv) All variables are fixed.

The user can choose one of the following functions to be minimized:

$$\begin{aligned}f_1(\mathbf{x}) &= r^2, \quad r = \sqrt{x_1^2 + x_2^2 + x_3^2}, \\f_2(\mathbf{x}) &= r^{10}, \\f_3(\mathbf{x}) &= r, \\f_4(\mathbf{x}) &= -e^{-r^2}, \\f_5(\mathbf{x}) &= r^6 - 2r^4 + r^2, \\f_6(\mathbf{x}) &= r^2 e^{-r^2}, \\f_7(\mathbf{x}) &= -e^{-r^2} - 10e^{-r_a^2}, \quad r_a^2 = (x_1 - 3)^2 + (x_2 - 3)^2 + (x_3 - 3)^2.\end{aligned}$$

**Suggestions:** Discuss the functions  $f_1$  through  $f_7$ . All of them have a minimum at  $x_1 = x_2 = x_3 = 0$ . Some possess additional minima. Study the convergence behavior of the different minimization methods for these functions using different starting points and explain this behavior qualitatively.

**Example Program 10.2:** The class E2Min determines the parameters of a distribution from the elements of a sample and demonstrates the use of MinCov

The program solves the problem in Example 10.1. First a sample is drawn from the standard normal distribution. Next the sample is used to estimate the parameters  $x_1$  (mean) and  $x_2$  (standard deviation) of the population by minimizing the negative of the likelihood function (10.18.2). This is done withMinSim. The covariance matrix of the parameters is determined by callingMinCov. The results are presented numerically. The rest of the program performs the graphical display of the sample as a one-dimensional scatter plot and of the fitted function as in Fig. 10.16 (left-hand plots).

**Suggestion:** Run the program for samples of different size.

**Example Program 10.3:** The class E3Min demonstrates the use of MinAsy and draws the boundary of a confidence region

The class solves the same problem as the previous example. In addition it computes the asymmetric errors of the parameters using MinAsy. Then the solution, the symmetric errors, the covariance ellipse, and the asymmetric errors are displayed graphically in the  $(x_1, x_2)$  plane, and with the help of the method DatanGraphics.drawContour, the contour of the confidence region is shown as well. The plot corresponds to Fig. 10.16 (right-hand plots).

**Example Program 10.4:** The class E4Min determines the parameters of a distribution from the histogram of a sample

The program solves the problem of Examples 10.2 and 10.3. First a sample of size  $n_{\text{ev}}$  is drawn from a standard normal distribution and a histogram with  $n_t$  bins between  $t_0 = -5.25$  and  $t_{\max} = 5.25$  is constructed from the sample. The bin centers are  $t_i$  ( $i = 1, 2, \dots, n_t$ ), and the bin contents are  $n_i$ . As chosen by the user either the likelihood function  $\ell$  given by (10.18.10) is maximized (i.e.,  $-\ell$  is minimized) by MinSim and the user function MinLogLikeHistPoison, or the sum of squares  $Q$  given by (10.18.12) is minimized, again with MinSim, but using MinHistSumOfSquares.

The results are presented in graphical form. One plot contains the histogram and the fitted Gaussian (i.e., a plot corresponding to the plots on the left-hand side of Figs. 10.17 or 10.18), a second one presents the solution in the  $(x_1, x_2)$  plane with symmetric and asymmetric errors, covariance ellipse, and confidence region (corresponding to the plots on the right-hand side of those figures).

**Suggestions:** (a) Choose  $n_{\text{ev}} = 100$ . Show that for likelihood maximization the errors  $\Delta x_1, \Delta x_2$  increase if you decrease the number of bins beginning with  $n_t = 100$ , but that for the minimization of the sum of squares the number of bins has to be small to get meaningful errors. (b) Show that for  $n_{\text{ev}} = 1000$  and  $n_t = 50$  or  $n_t = 20$  there is practically no difference between the results of the methods.

# 11. Analysis of Variance

The analysis of variance (or ANOVA), originally developed by R. A. FISHER, concerns testing the hypothesis of equal means of a number of samples. Such problems occur, for example, in the comparison of a series of measurements carried out under different conditions, or in quality control of samples produced by different machines. One tries to discover what influence the changing of *external variables* (e.g., experimental conditions, the number of a machine) has on a sample. For the simple case of only two samples, this problem can also be solved with Student's difference test (Sect. 8.3).

We speak of *one-way analysis of variance*, or also one-way classification, when only one external variable is changed. The evaluation of a series of measurements of an object micrometer performed with different microscopes can serve as an example. One has a *two- (or more) way analysis of variance* (two-way classification) when several variables are changed simultaneously. In the example above, if different observers carry out the series of measurements with each microscope, then a two-way analysis of variance can investigate influences of both the observer and the instrument on the result.

## 11.1 One-Way Analysis of Variance

Let us consider a sample of size  $n$ , which can be divided into  $t$  groups according to a certain criterion  $A$ . Clearly the criterion must be related to the sampling or measuring process. We say that the groups are constructed according to the *classification A*. We assume that the populations from which the  $t$  subsamples are taken are normally distributed with the same variance  $\sigma^2$ . We now want to test the hypothesis that the mean values of these populations are also equal. If this hypothesis is true, then all of the samples come from the same population. We can then apply the results of Sect. 6.4 (samples from subpopulations). Using the same notation as there, we have  $t$  groups of size  $n_i$  with

$$n = \sum_{i=1}^t n_i$$

and we write the  $j$ th element of the  $i$ th group as  $x_{ij}$ . The sample mean of the  $i$ th group is

$$\bar{x}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij} \quad (11.1.1)$$

and the mean of the entire sample is

$$\bar{x} = \frac{1}{n} \sum_{i=1}^t \sum_{j=1}^{n_i} x_{ij} = \frac{1}{n} \sum_{i=1}^t n_i \bar{x}_i \quad . \quad (11.1.2)$$

We now construct the *sum of squares*

$$\begin{aligned} Q &= \sum_{i=1}^t \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2 = \sum_{i=1}^t \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i + \bar{x}_i - \bar{x})^2 \\ &= \sum_{i=1}^t \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2 + \sum_{i=1}^t \sum_{j=1}^{n_i} (\bar{x}_i - \bar{x})^2 + 2 \sum_{i=1}^t \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)(\bar{x}_i - \bar{x}) \quad . \end{aligned}$$

The last term vanishes because of (11.1.1) and (11.1.2). One therefore has

$$\begin{aligned} Q &= \sum_{i=1}^t \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2 = \sum_{i=1}^t n_i (\bar{x}_i - \bar{x})^2 + \sum_{i=1}^t \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2 \quad , \\ Q &= Q_A + Q_W \quad . \end{aligned} \quad (11.1.3)$$

The first term is the *sum of squares between the groups* obtained with the classification  $A$ . The second term is a sum over the *sums of squares within a group*. The sum of squares  $Q$  is decomposed into a sum of two sums of squares corresponding to different “sources” – the variation of means within the classification  $A$  and the variation of measurements within the groups. If our hypothesis is correct, then  $Q$  is a sum of squares from a normal distribution, i.e.,  $Q/\sigma^2$  follows a  $\chi^2$ -distribution with  $n - 1$  degrees of freedom. Correspondingly, for each group the quantity

$$\frac{Q_i}{\sigma^2} = \frac{1}{\sigma^2} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$$

follows a  $\chi^2$ -distribution with  $n_i - 1$  degrees of freedom. The sum

$$\frac{Q_W}{\sigma^2} = \sum_{i=1}^t \frac{Q_i}{\sigma^2}$$

is then described by a  $\chi^2$ -distribution with  $\sum_i(n_i - 1) = n - t$  degrees of freedom (see Sect. 6.6). Finally,  $Q_A/\sigma^2$  follows a  $\chi^2$ -distribution with  $t - 1$  degrees of freedom.

The expressions

$$\begin{aligned}s^2 &= \frac{Q}{n-1} = \frac{1}{n-1} \sum_i \sum_j (x_{ij} - \bar{x})^2 , \\ s_A^2 &= \frac{Q_A}{t-1} = \frac{1}{t-1} \sum_i n_i (\bar{x}_i - \bar{x})^2 , \\ s_W^2 &= \frac{Q_W}{n-t} = \frac{1}{n-t} \sum_i \sum_j (x_{ij} - \bar{x}_i)^2\end{aligned}\quad (11.1.4)$$

are unbiased estimators of the population variances. (In Sect. 6.5 we called such expressions mean squares.) The ratio

$$F = s_A^2/s_W^2 \quad (11.1.5)$$

can thus be used to carry out an  $F$ -test.

If the hypothesis of equal means is false, then the values  $\bar{x}_i$  of the individual groups will be quite different. Thus  $s_A^2$  will be relatively large, while  $s_W^2$ , which is the mean of the variances of the individual groups, will not change much. This means that the ratio (11.1.5) will be large. Therefore one uses a one-sided  $F$ -test. The hypothesis of equal means is rejected at the significance level  $\alpha$  if

$$F = s_A^2/s_W^2 > F_{1-\alpha}(t-1, n-t) . \quad (11.1.6)$$

The sums of squares can be computed according to two equivalent formulas,

$$\begin{aligned}Q &= \sum_i \sum_j (x_{ij} - \bar{x})^2 = \sum_i \sum_j x_{ij}^2 - n\bar{x}^2 , \\ Q_A &= \sum_i n_i (\bar{x}_i - \bar{x})^2 = \sum_i n_i \bar{x}_i^2 - n\bar{x}^2 , \\ Q_W &= \sum_i \sum_j (x_{ij} - \bar{x}_i)^2 = \sum_i \sum_j x_{ij}^2 - \sum_i n_i \bar{x}_i^2 .\end{aligned}\quad (11.1.7)$$

The expression on the right of each line is usually easier to compute. Since each sum of squares is obtained by computing the difference of two relatively large numbers, one must pay attention to possible problems with errors in rounding. Although one only needs the sums  $Q_A$  and  $Q_W$  in order to compute the ratio  $F$ , it is recommended to compute  $Q$  as well, since one can then perform a check using (11.1.3), i.e.,  $Q = Q_A + Q_W$ . The check is only meaningful when  $Q$  is computed with the left-hand form of (11.1.3). Usually the

results of an analysis of variance are summarized in a so-called *analysis of variance table*, (or ANOVA table) as shown in Table 11.1.

Before carrying out an analysis of variance one must consider whether the requirements are met under which the procedure has been derived. In particular, one must check the assumption of a normal distribution for the measurements within each group. This is by no means certain in every case. If, for example, the measured values are always positive (e.g., the length or weight

**Table 11.1:** ANOVA table for one-way classification.

| Source             | SS<br>(sum of squares) | DF<br>(degrees of freedom) | MS<br>(mean square)         | F                         |
|--------------------|------------------------|----------------------------|-----------------------------|---------------------------|
| Between the groups | $Q_A$                  | $t - 1$                    | $s_A^2 = \frac{Q_A}{t - 1}$ |                           |
| Within the groups  | $Q_W$                  | $n - t$                    | $s_W^2 = \frac{Q_W}{n - t}$ | $F = \frac{s_A^2}{s_W^2}$ |
| Sum                | $Q$                    | $n - 1$                    | $s^2 = \frac{Q}{n - 1}$     |                           |

of an object) and if the standard deviation is of a magnitude comparable to the measured values, then the probability density can be asymmetric and thus not Gaussian. If, however, the original measurements (let us denote them for the moment by  $x'$ ) are transformed using a monotonic transformation such as

$$x = a \log(x' + b) \quad , \quad (11.1.8)$$

where  $a$  and  $b$  are appropriately chosen constants, then a normal distribution can often be sufficiently well approximated. Other transformations sometimes used are  $x = \sqrt{x'}$  or  $x = 1/x'$ .

**Example 11.1:** One-way analysis of variance of the influence of various drugs

The spleens of mice with cancer are often attacked particularly strongly. The weight of the spleen can thus serve as a measure of the reaction to various drugs. The drugs (I–III) were used to treat ten mice each. Table 11.2 contains the measured spleen weights, which have already been transformed according to  $x = \log x'$ , where  $x'$  is the weight in grams. Most of the calculation is presented in Table 11.2. Table 11.3 contains the resulting ANOVA table. Since even at a significance level of 50 % the  $F$ -test gives  $F_{0.5}(2, 24) = 3.4$ , one cannot reject the hypothesis of equal mean values. The experiment thus showed no significant difference in the effectiveness of the three drugs. ■

**Table 11.2:** Data for Example 11.1.

| Experiment<br>number | Group  |        |                                  |
|----------------------|--------|--------|----------------------------------|
|                      | I      | II     | III                              |
| 1                    | 19     | 40     | 32                               |
| 2                    | 45     | 28     | 26                               |
| 3                    | 26     | 26     | 30                               |
| 4                    | 23     | 15     | 17                               |
| 5                    | 36     | 24     | 23                               |
| 6                    | 23     | 26     | 24                               |
| 7                    | 26     | 36     | 29                               |
| 8                    | 33     | 27     | 20                               |
| 9                    | 22     | 28     | —                                |
| 10                   | —      | 19     | —                                |
| $\sum_j x_j$         | 253    | 269    | 201                              |
| $n_i$                | 9      | 10     | 8                                |
|                      |        |        | $\sum_i \sum_j x_{ij}^2 = 20607$ |
|                      |        |        | $\sum_i \sum_j x_{ij} = 723$     |
|                      |        |        | $n\bar{x}^2 = 19360$             |
| $\bar{x}_i$          | 28.11  | 26.90  | 25.13                            |
| $\bar{x}_i^2$        | 790.23 | 723.61 | 631.52                           |
|                      |        |        | $\sum_i n_i \bar{x}_i^2 = 19398$ |

**Table 11.3:** ANOVA table for Example 11.1.

| Source             | SS   | DF | MS   | F     |
|--------------------|------|----|------|-------|
| Between the groups | 38   | 2  | 19.0 | 0.377 |
| Within the groups  | 1209 | 24 | 50.4 |       |
| Sum                | 1247 | 26 | 47.8 |       |

## 11.2 Two-Way Analysis of Variance

Before we turn to analysis of variance with two external variables, we would like to examine more carefully the results obtained for one-way classification. We denoted the  $j$ th measurement of the quantity  $x$  in group  $i$  by  $x_{ij}$ . We now assume for simplicity that each group contains the same number of measurements, i.e.,  $n_i = J$ . In addition we denote the total number of groups by  $I$ . The classification into individual groups was done according to the criterion A, e.g., the production number of a microscope, by which the groups can be distinguished. The labeling according to measurement and group is illustrated in Table 11.4.

We can write the individual means of the groups in the form

$$\begin{aligned}\bar{x}_{..} &= \bar{x} = \frac{1}{IJ} \sum_i \sum_j x_{ij} , \\ \bar{x}_{i..} &= \frac{1}{J} \sum_j x_{ij} , \\ \bar{x}_{.j} &= \frac{1}{I} \sum_i x_{ij} .\end{aligned}\tag{11.2.1}$$

**Table 11.4:** One-way classification.

| Measurement<br>number | Classification A |          |         |          |         |          |
|-----------------------|------------------|----------|---------|----------|---------|----------|
|                       | $A_1$            | $A_2$    | $\dots$ | $A_i$    | $\dots$ | $A_I$    |
| 1                     | $x_{11}$         | $x_{22}$ |         | $x_{i1}$ |         | $x_{I1}$ |
| 2                     | $x_{12}$         | $x_{22}$ |         | $x_{i2}$ |         | $x_{I2}$ |
| $\vdots$              |                  |          |         |          |         |          |
| $j$                   | $x_{1j}$         | $x_{2j}$ |         | $x_{ij}$ |         | $x_{Ij}$ |
| $\vdots$              |                  |          |         |          |         |          |
| $J$                   | $x_{1J}$         | $x_{2J}$ |         | $x_{iJ}$ |         | $x_{IJ}$ |

Here a point denotes summation over the index that it replaces. This notation allows a simple generalization to a larger number of indices. The analysis of variance with one-way classification is based on the assumption that the measurements within a group only differ by the measurement errors, which follow a normal distribution with a mean of zero and variance  $\sigma^2$ . That is, we consider the *model*

$$x_{ij} = \mu_i + \varepsilon_{ij} .\tag{11.2.2}$$

The goal of an analysis of variance was to test the hypothesis

$$H_0(\mu_1 = \mu_2 = \dots = \mu_I = \mu) .\tag{11.2.3}$$

By choosing measurements out of a certain group  $i$  and by applying the maximum-likelihood method to (11.2.2) one obtains the estimator

$$\tilde{\mu}_i = \bar{x}_{i..} = \frac{1}{J} \sum_j x_{ij} .\tag{11.2.4}$$

If  $H_0$  is true, then one has

$$\tilde{\mu} = \bar{x} = \frac{1}{IJ} \sum_i \sum_j x_{ij} = \frac{1}{I} \sum_i \tilde{\mu}_i . \quad (11.2.5)$$

The (composite) alternative hypothesis is that not all of the  $\mu_i$  are equal. We want, however, to retain the concept of the overall mean and we write

$$\mu_i = \mu + a_i .$$

The model (11.2.2) then has the form

$$x_{ij} = \mu + a_i + \varepsilon_{ij} . \quad (11.2.6)$$

Between the quantities  $a_i$ , which represent a measure of the deviation of the mean for the  $i$ th group from the overall mean, one has the relation

$$\sum_i a_i = 0 . \quad (11.2.7)$$

The maximum-likelihood estimators for the  $a_i$  are

$$\tilde{a}_i = \bar{x}_{i\cdot} - \bar{x} . \quad (11.2.8)$$

The one-way analysis of variance of Sect. 11.1 was derived from the identity

$$x_{ij} - \bar{x} = (\bar{x}_{i\cdot} - \bar{x}) + (x_{ij} - \bar{x}_{i\cdot}) , \quad (11.2.9)$$

which describes the deviation of the individual measurements from the overall mean. The sum of squares  $Q$  of these deviations could then be decomposed into the terms  $Q_A$  and  $Q_W$ ; cf. (11.1.3).

After this preparation we now consider a two-way classification, where the measurements are divided into groups according to two criteria,  $A$  and  $B$ . The measurements  $x_{ijk}$  belong to class  $A_i$ , which is given by the classification according to  $A$ , and also to class  $B_j$ . The index  $k$  denotes the measurement number within the group that belongs to both class  $A_i$  and class  $B_j$ .

A two-way classification is said to be *crossed*, when a certain classification  $B_j$  has the same meaning for all classes  $A$ . If, for example, microscopes are classified by  $A$  and observers by  $B$ , and if each observer carries out a measurement with each microscope, then the classifications are crossed. If, however, one compares the microscopes in different laboratories, and if therefore in each laboratory a different group of  $J$  observers makes measurements with a certain microscope  $i$ , then the classification  $B$  is said to be *nested* in  $A$ . The index  $j$  then merely counts the classes  $B$  within a certain class  $A$ .

The simplest case is a crossed classification with only *one observation*. Since then  $k = 1$  for all observations  $x_{ijk}$ , we can drop the index  $k$ . One uses the model

$$x_{ij} = \mu + a_i + b_j + \varepsilon_{ij} \quad , \quad \sum_i a_i = 0 \quad , \quad \sum_j b_j = 0 \quad , \quad (11.2.10)$$

where  $\varepsilon$  is normally distributed with mean zero and variance  $\sigma^2$ . The null hypothesis says that by classification according to  $A$  or  $B$ , no deviation from the overall mean occurs. We write this in the form of two individual hypotheses,

$$H_0^{(A)}(a_1 = a_2 = \dots = a_I = 0) \quad , \quad H_0^{(B)}(b_1 = b_2 = \dots = b_J = 0) \quad . \quad (11.2.11)$$

The least-squares estimators for  $a_i$  and  $b_j$  are

$$\tilde{a}_i = \bar{x}_{i\cdot} - \bar{x} \quad , \quad \tilde{b}_j = \bar{x}_{\cdot j} - \bar{x} \quad .$$

In analogy to Eq. (11.2.9) we can write

$$x_{ij} - \bar{x} = (\bar{x}_{i\cdot} - \bar{x}) + (\bar{x}_{\cdot j} - \bar{x}) + (x_{ij} - \bar{x}_{i\cdot} - \bar{x}_{\cdot j} + \bar{x}) \quad . \quad (11.2.12)$$

In a similar way the sum of squares can be written

$$\sum_i \sum_j (x_{ij} - \bar{x})^2 = Q = Q_A + Q_B + Q_W \quad , \quad (11.2.13)$$

where

$$\begin{aligned} Q_A &= J \sum_i (\bar{x}_{i\cdot} - \bar{x})^2 = J \sum_i \bar{x}_{i\cdot}^2 - IJ\bar{x}^2 \quad , \\ Q_B &= I \sum_j (\bar{x}_{\cdot j} - \bar{x})^2 = I \sum_j \bar{x}_{\cdot j}^2 - IJ\bar{x}^2 \quad , \\ Q_W &= \sum_i \sum_j (x_{ij} - \bar{x}_{i\cdot} - \bar{x}_{\cdot j} + \bar{x})^2 \\ &= \sum_i \sum_j x_{ij}^2 - J \sum_i \bar{x}_{i\cdot}^2 - I \sum_j \bar{x}_{\cdot j}^2 + IJ\bar{x}^2 \quad . \end{aligned} \quad (11.2.14)$$

When divided by the corresponding number of degrees of freedom, these sums are estimators of  $\sigma^2$ , providing that the hypotheses (11.2.11) are correct. The hypotheses  $H_0^{(A)}$  and  $H_0^{(B)}$  can be tested individually by using the ratio

$$F^{(A)} = s_A^2 / s_W^2 \quad , \quad F^{(B)} = s_B^2 / s_W^2 \quad . \quad (11.2.15)$$

Here one uses one-sided  $F$ -tests as in Sect. 11.1. The overall situation can be summarized in an ANOVA table (Table 11.5).

If more than one observation is made in each group, then the crossed classification can be generalized in various ways. The most important generalization involves *interaction* between the classes. One then has the model

$$x_{ijk} = \mu + a_i + b_j + (ab)_{ij} + \varepsilon_{ijk} . \quad (11.2.16)$$

The quantity  $(ab)_{ij}$  is called the *interaction* between the classes  $A_i$  and  $B_j$ . It describes the deviation from the group mean that occurs because of the specific interaction of  $A_i$  and  $B_j$ . The parameters  $a_i, b_j, (ab)_{ij}$  are related by

$$\sum_i a_i = \sum_j b_j = \sum_i \sum_j (ab)_{ij} = 0 . \quad (11.2.17)$$

Their maximum-likelihood estimators are

$$\begin{aligned} \tilde{a}_i &= \bar{x}_{i..} - \bar{x} , & \tilde{b}_j &= \bar{x}_{.j.} - \bar{x} , \\ (\tilde{ab})_{ij} &= \bar{x}_{ij.} + \bar{x} - \bar{x}_{i..} - \bar{x}_{.j.} . \end{aligned} \quad (11.2.18)$$

**Table 11.5:** Analysis of variance table for crossed two-way classification with only one observation.

| Source        | SS    | DF               | MS                                   | F                               |
|---------------|-------|------------------|--------------------------------------|---------------------------------|
| Class. A      | $Q_A$ | $I - 1$          | $s_A^2 = \frac{Q_A}{I - 1}$          | $F^{(A)} = \frac{s_A^2}{s_W^2}$ |
| Class. B      | $Q_B$ | $J - 1$          | $s_B^2 = \frac{Q_B}{J - 1}$          | $F^{(B)} = \frac{s_B^2}{s_W^2}$ |
| Within groups | $Q_W$ | $(I - 1)(J - 1)$ | $s_W^2 = \frac{Q_W}{(I - 1)(J - 1)}$ |                                 |
| Sum           | $Q$   | $IJ - 1$         | $s^2 = \frac{Q}{IJ - 1}$             |                                 |

The null hypothesis can be divided into three individual hypotheses,

$$\begin{aligned} H_0^{(A)}(a_i = 0; i = 1, 2, \dots, I) , & \quad H_0^{(B)}(b_j = 0; j = 1, 2, \dots, J) , \\ H_0^{(AB)}((ab)_{ij} = 0; i = 1, 2, \dots, I; j = 1, 2, \dots, J) , & \end{aligned} \quad (11.2.19)$$

which can then be tested individually. The analysis of variance is based on the identity

$$\begin{aligned} x_{ijk} - \bar{x} &= (\bar{x}_{i..} - \bar{x}) + (\bar{x}_{.j.} - \bar{x}) \\ &\quad + (\bar{x}_{ij.} + \bar{x} - \bar{x}_{i..} - \bar{x}_{.j.}) + (x_{ijk} - \bar{x}_{ij.}) , \end{aligned} \quad (11.2.20)$$

which allows the decomposition of the sum of squares of deviations into four terms,

$$\begin{aligned}
 Q &= \sum_i \sum_j \sum_k (x_{ijk} - \bar{x})^2 \\
 &= Q_A + Q_B + Q_{AB} + Q_W , \\
 Q_A &= JK \sum_i (\bar{x}_{i..} - \bar{x})^2 , \\
 Q_B &= IK \sum_j (\bar{x}_{.j.} - \bar{x})^2 , \\
 Q_{AB} &= K \sum_i \sum_j (\bar{x}_{ij.} + \bar{x} - \bar{x}_{i..} - \bar{x}_{.j.})^2 , \\
 Q_W &= \sum_i \sum_j \sum_k (x_{ijk} - \bar{x}_{ij.})^2 .
 \end{aligned} \tag{11.2.21}$$

The degrees of freedom and mean squares as well as the  $F$ -ratio, which can be used for testing the hypotheses, are given in Table 11.6.

**Table 11.6:** Analysis of variance table for crossed two-way classification.

| Source        | SS       | DF               | MS   | F                                   |
|---------------|----------|------------------|--|-------------------------------------|
| Class. A      | $Q_A$    | $I - 1$          | $s_A^2 = \frac{Q_A}{I - 1}$                | $F^{(A)} = \frac{s_A^2}{s_W^2}$     |
| Class. B      | $Q_B$    | $J - 1$          | $s_B^2 = \frac{Q_B}{J - 1}$                | $F^{(B)} = \frac{s_B^2}{s_W^2}$     |
| Interaction   | $Q_{AB}$ | $(I - 1)(J - 1)$ | $s_{AB}^2 = \frac{Q_{AB}}{(I - 1)(J - 1)}$ | $F^{(AB)} = \frac{s_{AB}^2}{s_W^2}$ |
| Within groups | $Q_W$    | $IJ(K - 1)$      | $s_W^2 = \frac{Q_W}{IJ(K - 1)}$            |                                     |
| Sum           | $Q$      | $IJK - 1$        | $s^2 = \frac{Q}{IJK - 1}$                  |                                     |

Finally, we will give the simplest case of a *nested two-way classification*. Because the classification  $B$  is only defined within the individual classes of  $A$ , the terms  $b_j$  and  $(ab)_{ij}$  from Eq. (11.2.10) are not defined, since they imply a sum over  $i$  for fixed  $j$ . Therefore one uses the model

$$x_{ijk} = \mu + a_i + b_{ij} + \varepsilon_{ijk} \tag{11.2.22}$$

with

$$\begin{aligned}
 \sum_i a_i &= 0 , & \sum_i \sum_j b_{ij} &= 0 , \\
 \tilde{a}_i &= \bar{x}_{i..} - \bar{x} , & \tilde{b}_{ij} &= \bar{x}_{ij.} - \bar{x}_{i..} .
 \end{aligned}$$

The term  $b_{ij}$  is a measure of the deviation of the measurements of class  $B_j$  within class  $A_i$  from the overall mean of class  $A_i$ . The null hypothesis consists of

$$\begin{aligned} H_0^{(A)}(a_i = 0; i = 1, 2, \dots, I) &, \\ H_0^{(B(A))}(b_{ij} = 0; i = 1, 2, \dots, I; j = 1, 2, \dots, J) &. \end{aligned} \quad (11.2.23)$$

An analysis of variance for testing these hypotheses can be carried out with the help of Table 11.7. Here one has

$$\begin{aligned} Q_A &= JK \sum_i (\bar{x}_{i..} - \bar{x})^2 , \\ Q_{B(A)} &= K \sum_i \sum_j (\bar{x}_{ij.} - \bar{x}_{i..})^2 , \\ Q_W &= \sum_i \sum_j \sum_k (x_{ijk} - \bar{x}_{ij.})^2 , \\ Q &= Q_A + Q_{B(A)} + Q_W = \sum_i \sum_j \sum_k (x_{ijk} - \bar{x})^2 . \end{aligned}$$

In a similar way one can construct various models for two-way or multiple classification. For each model the total sum of squares is decomposed into a certain sum of individual sums of squares, which, when divided by the corresponding number of degrees of freedom, can be used to carry out an  $F$ -test. With this, the hypotheses implied by the model can be tested.

Some models are, at least formally, contained within others. For example one finds by comparing Tables 11.6 and 11.7 the relation

$$Q_{B(A)} = Q_B + Q_{AB} . \quad (11.2.24)$$

A similar relation holds for the corresponding number of degrees of freedom,

$$f_{B(A)} = f_B + f_{AB} . \quad (11.2.25)$$

**Table 11.7:** Analysis of variance table for nested two-way classification.

| Source        | SS         | DF          | MS                                       | F                                       |
|---------------|------------|-------------|--|---|
| Class. A      | $Q_A$      | $I - 1$     | $s_A^2 = \frac{Q_A}{I - 1}$              | $F^{(A)} = \frac{s_A^2}{s_W^2}$         |
| Within A      | $Q_{B(A)}$ | $I(J - 1)$  | $s_{B(A)}^2 = \frac{Q_{B(A)}}{I(J - 1)}$ | $F^{(B(A))} = \frac{s_{B(A)}^2}{s_W^2}$ |
| Within groups | $Q_W$      | $IJ(K - 1)$ | $s_W^2 = \frac{Q_W}{IJ(K - 1)}$          |   |
| Sum           | $Q$        | $IJK - 1$   | $s^2 = \frac{Q}{IJK - 1}$                |   |

### **Example 11.2:** Two-way analysis of variance in cancer research

Two groups of rats are injected with the amino acid thymidine containing traces of tritium, a radioactive isotope of hydrogen. In addition, one of the groups receives a certain carcinogen. The incorporation of thymidine into the skin of the rats is investigated as a function of time by measuring the number of tritium decays per unit area of skin. The classifications are crossed since the time dependence is controlled in the same way for both series of test animals. The measurements are compiled in Table 11.8. The numbers are already transformed from the original counting rates  $x'$  according to  $x = 50 \log x' - 100$ . The results, obtained with the class AnalysisOfVariance, are shown in Table 11.9. There is no doubt that the presence or absence of the carcinogen (classification A) has an influence on the result, since the ratio  $F^{(A)}$  is very large. We now want to test the existence of a time dependence (classification B) and of an interaction between A and B at a significance level of  $\alpha = 0.01$ . From Table I.8 we find  $F_{0.99} = 2.72$ . The hypotheses of time independence and vanishing interaction must therefore be rejected. Table 11.9 also contains the values of  $\alpha$  for which the hypothesis would not need to be rejected. They are very small. ■

**Table 11.8:** Data for Example 11.2.

| Obs.<br>no. | Injection  | Time after injection (h) |    |    |    |    |    |    |    |    |    |
|-------------|------------|--------------------------|----|----|----|----|----|----|----|----|----|
|             |            | 4                        | 8  | 12 | 16 | 20 | 24 | 28 | 32 | 36 | 48 |
| 1           |            | 34                       | 54 | 44 | 51 | 62 | 61 | 59 | 66 | 52 | 52 |
| 2           | Thymidine  | 40                       | 57 | 52 | 46 | 61 | 70 | 67 | 59 | 63 | 50 |
| 3           |            | 38                       | 40 | 53 | 51 | 54 | 64 | 58 | 67 | 60 | 44 |
| 4           |            | 36                       | 43 | 51 | 49 | 60 | 68 | 66 | 58 | 59 | 52 |
| 1           |            | 28                       | 23 | 42 | 43 | 31 | 32 | 25 | 24 | 26 | 26 |
| 2           | Thymidine  | 32                       | 23 | 41 | 48 | 45 | 38 | 27 | 26 | 31 | 27 |
| 3           | and        | 34                       | 29 | 34 | 36 | 41 | 32 | 27 | 32 | 25 | 27 |
| 4           | Carcinogen | 27                       | 30 | 39 | 43 | 37 | 34 | 28 | 30 | 26 | 30 |

**Table 11.9:** Printout from Example 11.2.

| Source | Analysis of variance table |                    |             |           |          |
|--------|----------------------------|--------------------|-------------|-----------|----------|
|        | Sum of squares             | Degrees of freedom | Mean square | F Ratio   | Alpha    |
| A      | 9945.80                    | 1                  | 9945.80     | 590.54725 | 0.00E-10 |
| B      | 1917.50                    | 9                  | 213.06      | 12.65050  | 0.54E-10 |
| INT.   | 2234.95                    | 9                  | 248.33      | 14.74485  | 0.03E-10 |
| W      | 1010.50                    | 60                 | 16.84       |           |          |
| TTL.   | 15108.75                   | 79                 | 191.25      |           |          |

## 11.3 Java Class and Example Programs

### Java Class

AnalysisOfVariance performs a crossed as well as a nested two-way analysis of variance.

**Example Program 11.1:** The class E1Anova demonstrates the use of AnalysisOfVariance

The short program analyses the data of Example 11.2. Data and output are presented as in Table 11.9.

**Example Program 11.2:** The class E2Anova simulates data and performs an analysis of variance on them

The program allows interactive input of numerical values for  $\sigma$ , the quantities  $I, J, K$  and three further parameters:  $\Delta_i, \Delta_j, \Delta_k$ . It generates data of the simple form

$$x_{ijk} = i\Delta_i + j\Delta_j + k\Delta_k + \varepsilon_{ijk} .$$

Here the quantities  $\varepsilon_{ijk}$  are taken from a normal distribution with zero mean standard deviation  $\sigma$ . An analysis of variance is performed on the data and the results are presented for crossed and for nested two-way classification.

**Suggestion:** Take a value  $\neq 0$  for only one of the parameters  $\Delta_i, \Delta_j, \Delta_k$  and interpret the resulting analysis of variance table.

# 12. Linear and Polynomial Regression

The fitting of a linear function (or, more generally, of a polynomial) to measured data that depend on a controlled variable is probably the most commonly occurring task in data analysis. This procedure is also referred to as *linear* (or *polynomial*) *regression*. Although we have already treated this problem in Sect. 9.4.1, we take it up again here in greater detail. Here we will use different numerical methods, emphasize the most appropriate choice for the order of the polynomial, treat in detail the question of confidence limits, and also give a procedure for the case where the measurement errors are not known.

## 12.1 Orthogonal Polynomials

In Sect. 9.4.1 we have already fitted a polynomial of order  $r - 1$ ,

$$\eta(t) = x_1 + x_2 t + \cdots + x_r t^{r-1} \quad , \quad (12.1.1)$$

to measured values  $y_i(t_i)$ , which corresponded to a given value  $t_i$  of the controlled variable  $t$ . Here the quantities  $\eta(t)$  were the true values of the measured quantities  $y(t)$ . In Example 9.2 we saw that there can be a maximum reasonable order for the polynomial, beyond which a further increase in the order gives no significant improvement in the fit. We want here to pursue the question of how to find the optimal order of the polynomial. Example 9.2 shows that when increasing the order, all of the coefficients  $x_1, x_2, \dots$  change and that all of the coefficients are in general correlated. Because of this the situation becomes difficult to judge. These difficulties are avoided by using *orthogonal polynomials*.

Instead of (12.1.1) one describes the data by the expression

$$\eta(t) = x_1 f_1(t) + x_2 f_2(t) + \cdots + x_r f_r(t) \quad . \quad (12.1.2)$$

Here the quantities  $f_j$  are polynomials of order  $j - 1$ ,

$$f_j(t) = \sum_{k=1}^j b_{jk} t^{k-1} . \quad (12.1.3)$$

These are chosen to satisfy the *condition of orthogonality* (or more precisely condition of orthonormality) with respect to the values  $t_i$  and the measurement weights  $g_i = 1/\sigma_i^2$ ,

$$\sum_{i=1}^N g_i f_j(t_i) f_k(t_i) = \delta_{jk} . \quad (12.1.4)$$

Defining

$$A_{ij} = f_j(t_i), \quad A = \begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1r} \\ \vdots & & & \\ A_{N1} & A_{N2} & \cdots & A_{Nr} \end{pmatrix} \quad (12.1.5)$$

and using the matrix notation (9.2.9)

$$G_y = \begin{pmatrix} g_1 & & & 0 \\ & g_2 & & \\ & & \ddots & \\ 0 & & & g_N \end{pmatrix} , \quad (12.1.6)$$

Eq. (12.1.4) has the simple form

$$A^T G_y A = I . \quad (12.1.7)$$

The least-squares requirement,

$$\begin{aligned} & \sum_{i=1}^N g_i \left\{ y_i(t_i) - \sum_{j=1}^r x_j f_j(t_i) \right\}^2 \\ &= (\mathbf{y} - A\mathbf{x})^T G_y (\mathbf{y} - A\mathbf{x}) = \min \quad , \end{aligned} \quad (12.1.8)$$

is of the form (9.2.19) and thus has the solution (9.2.26)

$$\tilde{\mathbf{x}} = -A^T G_y \mathbf{y} , \quad (12.1.9)$$

where we have used (9.2.18) and (12.1.7). Because of (9.2.27) and (12.1.7) one has

$$C_{\tilde{x}} = I , \quad (12.1.10)$$

i.e., the covariance matrix of the coefficients  $\tilde{x}_1, \dots, \tilde{x}_r$  is simply the unit matrix. In particular, the coefficients are uncorrelated.

We now discuss the procedure for determining the matrix elements

$$A_{ij} = f_j(t_i) = \sum_{k=1}^j b_{jk} t_i^{k-1} . \quad (12.1.11)$$

For  $j = 1$ , the orthogonality condition gives

$$\sum_{i=1}^N g_i b_{11}^2 = 1, \quad b_{11} = 1 / \sqrt{\sum_i g_i} . \quad (12.1.12)$$

For  $j = 2$  there are two orthogonality conditions. First we obtain

$$\sum_i g_i f_2(t_i) f_1(t_i) = \sum_i g_i (b_{21} + b_{22} t_i) b_{11} = 0$$

and from this

$$b_{21} = -b_{22} \frac{\sum g_i t_i}{\sum g_i} = -b_{22} \bar{t} , \quad (12.1.13)$$

where

$$\bar{t} = \sum g_i t_i / \sum g_i \quad (12.1.14)$$

is the weighted mean of the values of the controlled variable  $t_i$  at which the measurements were made. The second orthogonality condition for  $j = 2$  gives

$$\sum g_i [f_2(t_i)]^2 = \sum g_i (b_{21} + b_{22} t_i)^2 = \sum g_i b_{22}^2 (t_i - \bar{t})^2 = 1$$

or

$$b_{22} = 1 / \sqrt{\sum g_i (t_i - \bar{t})^2} . \quad (12.1.15)$$

Substitution into (12.1.13) gives  $b_{21}$ .

For  $j > 2$  one can obtain the values of  $A_{ij} = f_j(t_i)$  recursively from the quantities for  $j - 1$  and  $j - 2$ . We make the ansatz

$$\gamma f_j(t_i) = (t_i - \alpha) f_{j-1}(t_i) - \beta f_{j-2}(t_i) , \quad (12.1.16)$$

multiply by  $g_i f_{j-1}(t_i)$ , sum over  $i$ , and obtain

$$\begin{aligned} \gamma \sum g_i f_j(t_i) f_{j-1}(t_i) &= 0 \\ &= \sum g_i t_i [f_{j-1}(t_i)]^2 - \alpha \sum g_i [f_{j-1}(t_i)]^2 - \beta \sum g_i f_{j-2}(t_i) f_{j-1}(t_i) . \end{aligned}$$

Because of the orthogonality condition, the second term on the right-hand side is equal to unity, and the third term vanishes. Thus one has

$$\alpha = \sum g_i t_i [f_{j-1}(t_i)]^2 . \quad (12.1.17)$$

By multiplication of (12.1.16) by  $g_i f_{j-2}(t_i)$  and summing one obtains in the same way

$$\beta = \sum g_i t_i f_{j-1}(t_i) f_{j-2}(t_i) . \quad (12.1.18)$$

Finally by computing the expression

$$\sum g_i f_j^2(t_i) = 1$$

and substituting  $f_j(t_i)$  from (12.1.16) one obtains

$$\gamma^2 = \sum g_i [(t_i - \alpha) f_{j-1}(t_i) - \beta f_{j-2}(t_i)]^2 . \quad (12.1.19)$$

Once the quantities  $\alpha$ ,  $\beta$ ,  $\gamma$  have been computed for a given  $j$ , then the coefficients  $b_{jk}$  are determined to be

$$\begin{aligned} b_{j1} &= (-\alpha b_{j-1,1} - \beta b_{j-2,1})/\gamma , \\ b_{jk} &= (b_{j-1,k-1} - \alpha b_{j-1,k} - \beta b_{j-2,k})/\gamma , \quad k = 2, \dots, j-2 , \\ b_{j,j-1} &= (b_{j-1,j-2} - \alpha b_{j-1,j-1})/\gamma , \\ b_{jj} &= b_{j-1,j-1}/\gamma . \end{aligned} \quad (12.1.20)$$

With (12.1.3) and (12.1.5) one thus obtains the quantities  $A_{ij} = f_j(t_i)$ . Since the procedure is recursive, the column  $j$  of the matrix  $A$  is obtained from the elements of the columns to its left. By extending the matrix to the right, the original matrix elements are not changed. This has the consequence that by increasing the number of terms in the polynomial (12.1.1) from  $r$  to  $r'$ , the coefficients  $\tilde{x}_1, \dots, \tilde{x}_r$  are retained, and one merely obtains additional coefficients. All of the  $\tilde{x}_j$  are uncorrelated and have the standard deviation  $\sigma_{\tilde{x}_j} = 1$ . The order of the polynomial sufficient to describe the data can now be determined by the requirement

$$|\tilde{x}_j| < c, \quad j > r .$$

With this the contribution of all higher coefficients  $\tilde{x}_j$  is smaller than  $c\sigma_{\tilde{x}_j}$ . The simplest choice of  $c$  is clearly  $c = 1$ .

For a given  $r$  one can now substitute the values  $b_{jk}$  into (12.1.3) and  $f_j(t_i)$  into (12.1.2) in order to obtain the best estimates  $\tilde{\eta}_i(t_i)$  of the true values  $\eta_i(t_i)$  corresponding to the measurements  $y_i(t_i)$ . One can also compute the quantity

$$M = \sum_{i=1}^N \frac{(\tilde{\eta}_i(t_i) - y_i(t_i))^2}{\sigma_i^2} .$$

If the data are in fact described by the polynomial of the chosen order, then the quantity  $M$  follows a  $\chi^2$ -distribution with  $f = N - r$  degrees of freedom. It can be used for a  $\chi^2$ -test for the goodness-of-fit of the polynomial.

**Example 12.1:** Treatment of Example 9.2 with Orthogonal Polynomials

Application of polynomial regression to the data of Example 9.2 yields the results shown in Table 12.1. The numerical values of the minimization function are, of course, exactly the same as in Example 9.2. The fitted polynomials are also the same. The numerical values  $\tilde{x}_1, \tilde{x}_2, \dots$  are different from those in Example 9.2 since these quantities are now defined differently. We have emphasized that the covariance matrix for  $\tilde{\mathbf{x}}$  is the  $r \times r$  unit matrix. We see that the quantities  $x_6, x_7, \dots, x_{10}$  have magnitudes less than unity and thus are no longer significantly different from zero. It is more difficult to judge the significance of  $x_5 = 1.08$ . In many cases one would not consider this value as being clearly different from zero. This means that a third order polynomial (i.e.,  $r = 4$ ) is sufficient to describe the data. ■

**Table 12.1:** Results of the application of polynomial regression using the data from Example 9.2.

| r  | $\tilde{x}_r$ | M      | Degrees of freedom |
|----|---------------|--------|--------------------|
| 1  | 24.05         | 833.55 | 9                  |
| 2  | 15.75         | 585.45 | 8                  |
| 3  | 23.43         | 36.41  | 7                  |
| 4  | 5.79          | 2.85   | 6                  |
| 5  | 1.08          | 1.69   | 5                  |
| 6  | 0.15          | 1.66   | 4                  |
| 7  | 0.85          | 0.94   | 3                  |
| 8  | -0.41         | 0.77   | 2                  |
| 9  | -0.45         | 0.57   | 1                  |
| 10 | 0.75          | 0.00   | 0                  |

## 12.2 Regression Curve: Confidence Interval

Once the order of the polynomial (12.1.2) in one way or another has been fixed, then every point of the *regression polynomial* or – in reference to the graphical representation – the *regression curve* can be computed. This is done by first substituting the recursively computed values  $b_{jk}$  into (12.1.3) for a given  $t$ , and using the thus computed  $f_j(t)$  and the parameters  $\tilde{\mathbf{x}}$  in (12.1.2):

$$\tilde{\eta}(t) = \sum_{j=1}^r \tilde{x}_j \left( \sum_{k=1}^j b_{jk} t^{k-1} \right) = \mathbf{d}^T(t) \tilde{\mathbf{x}} \quad . \quad (12.2.1)$$

Here  $\mathbf{d}$  is an  $r$ -vector with the elements

$$d_j(t) = \sum_{k=1}^j b_{jk} t^{k-1} . \quad (12.2.2)$$

After error propagation, (3.8.4), the variance of  $\tilde{\eta}(t)$  is

$$\sigma_{\tilde{\eta}(t)}^2 = \mathbf{d}^T(t) \mathbf{d}(t) , \quad (12.2.3)$$

since  $C_{\tilde{x}} = I$ . The reduced variable

$$u = \frac{\tilde{\eta}(t) - \eta(t)}{\sigma_{\tilde{\eta}(t)}} \quad (12.2.4)$$

thus follows a standard Gaussian distribution. This also means that the probability for the position of the true value  $\eta(t)$  is distributed according to a Gaussian of width  $\sigma_{\tilde{\eta}(t)}$  centered about the value  $\tilde{\eta}(t)$ .

We can now easily give a *confidence interval* for  $\eta(t)$ . According to Sect. 5.8 one has with probability  $P$  that

$$|u| \geq \Omega'(P) = \Omega\left(\frac{1}{2}(P+1)\right) , \quad (12.2.5)$$

e.g.,  $\Omega'(0.95) = 1.96$ . Thus the confidence limits at the confidence level  $P$  (e.g.,  $P = 0.95$ ) are

$$\eta(t) = \tilde{\eta}(t) \pm \Omega'(P) \sigma_{\tilde{\eta}(t)} = \tilde{\eta}(t) \pm \delta \tilde{\eta}(t) . \quad (12.2.6)$$

## 12.3 Regression with Unknown Errors

If the measurement errors  $\sigma_i = 1/\sqrt{g_i}$  are not known, but it can be assumed that they are equal,

$$\sigma_i = \sigma, \quad i = 1, \dots, N , \quad (12.3.1)$$

then one can easily obtain an estimate  $s$  for  $\sigma$  directly from the regression itself. The quantity

$$\sum_{i=1}^N \frac{(\tilde{\eta}(t_i) - y_i(t_i))^2}{\sigma^2} = M \quad (12.3.2)$$

follows a  $\chi^2$ -distribution with  $f = N - r$  degrees of freedom. More precisely, if many similar experiments with  $N$  measurements each were to be carried out, then the values of  $M$  computed for each experiment would be distributed like a  $\chi^2$ -variable with  $f = N - r$  degrees of freedom. Its expectation value

is simply  $f = N - r$ . If the value of  $\sigma$  in (12.3.2) is not known, then it can be replaced by an estimate  $s$  such that on the right-hand side one has the expectation value of  $M$ ,

$$s^2 = \sum (\tilde{\eta}(t_i) - y_i(t_i))^2 / (N - r) . \quad (12.3.3)$$

Here all of the steps necessary to compute  $\tilde{\eta}(t_i)$  must be carried out using the formulas of the last section and with the value  $\sigma_i = 1$ . (In fact, for the case (12.3.1) the quantities  $\sigma_i$  and  $g_i$  could be removed completely from the formulas for computing  $\tilde{\mathbf{x}}$  and  $\tilde{\eta}(t)$ .)

If we define

$$\bar{s}_{\tilde{\eta}(t)}^2 = \mathbf{d}^T(t) \mathbf{d}(t)$$

as the value of the variance of  $\tilde{\eta}(t)$  obtained for  $\sigma_i = 1$ , then

$$s_{\tilde{\eta}(t)}^2 = s^2 \bar{s}_{\tilde{\eta}(t)}^2 \quad (12.3.4)$$

is clearly the estimate of the variance of  $\tilde{\eta}(t)$ , which is based on the estimate (12.3.3) for the variance of the measured quantities. Replacing  $\sigma_{\tilde{\eta}(t)}$  by  $s_{\tilde{\eta}(t)}$  in (12.2.4), we obtain the variable

$$v = \frac{\tilde{\eta}(t) - \eta(t)}{s_{\tilde{\eta}(t)}} , \quad (12.3.5)$$

which no longer follows a standard normal distribution but rather Student's  $t$ -distribution for  $f = N - r$  degrees of freedom (cf. Sect. 8.3). For the confidence limits at the confidence level  $P = 1 - \alpha$  one now has

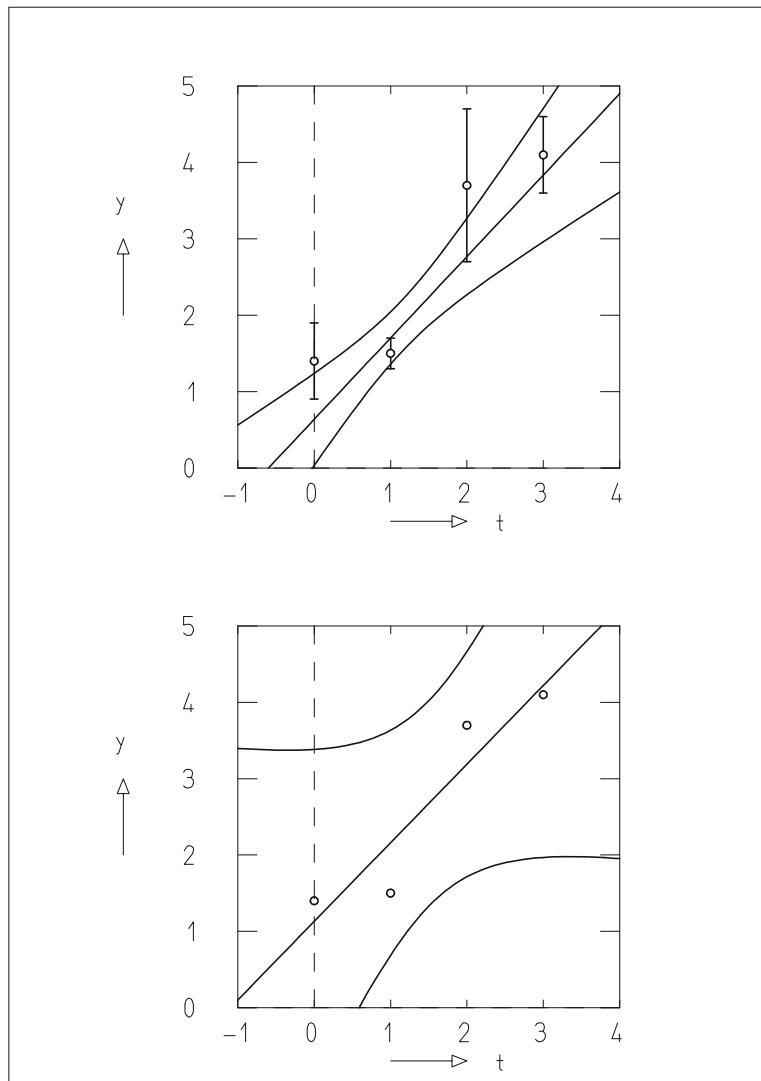
$$\eta(t) = \tilde{\eta}(t) \pm t_{1-\alpha/2} s_{\tilde{\eta}(t)} = \tilde{\eta}(t) \pm \delta \tilde{\eta}(t) , \quad (12.3.6)$$

where  $t_{1-\alpha/2}$  is the quantile of the  $t$ -distribution for  $f = N - r$  degrees of freedom.

It must be emphasized at this point that in the case where the errors are not known, one loses the possibility of applying the  $\chi^2$ -test. The goodness-of-fit for such a polynomial cannot be tested. Thus the procedure described at the end of Sect. 12.1 for determining the order of the polynomial is no longer valid. One must rely on a priori knowledge about the order of the polynomial. Therefore one almost always refrains from treating anything beyond linear dependences between  $\eta$  and  $t$ , i.e., the case  $r = 2$ .

### Example 12.2: Confidence limits for linear regression

In the upper plot of Fig. 12.1, four measured points with their errors, the corresponding regression line, and the limits at a confidence level of 95 % are shown. The measured points are taken from the example of Sect. 9.3. The confidence limits clearly become wider the more one leaves the region of the



**Fig. 12.1:** (Above) Measurements with errors and regression line with 95 % confidence limits. (Below) Measured points with errors assumed to be of unknown but equal magnitude, regression line, and 95 % confidence limits.

measured points. The narrowest region is close to the measured point with the smallest error. The plot is similar to Fig. 9.2d. It is easy to convince oneself that the envelope of the lines of Fig. 9.2d corresponds to the 68.3 % confidence limit. In the lower plot of Fig. 12.1 the same measured points were used, but the errors were treated as equal but of unknown magnitude. The regression line and 95 % confidence limits are shown for this assumption. ■

## 12.4 Java Class and Example Programs

### Java Class

Regression performs a polynomial regression.

#### Example Program 12.1: The class E1Reg demonstrates the use of Regression

The short program contains the data of Examples 9.2 and 12.1. As measurement errors the statistical errors  $\Delta y_i = \sqrt{y_i}$  are taken. A polynomial regression with  $r = 10$  parameters is performed. The results are presented numerically (see also Table 12.1).

#### Example Program 12.2: The class E2Reg demonstrates the use of Regression and presents the results in graphical form

The same problem as in E1Reg is treated. An additional input variable is  $r_{\max}$ , the maximum number of terms in the polynomials to be presented graphically. Graphs of these polynomials are shown together with the data points.

**Suggestion:** Choose consecutively  $r_{\max} = 2, 3, 4, 6, 10$ . Try to explain why the curves for  $r_{\max} = 3$  or 4 seem to be particularly convincing although for  $r_{\max} = 10$  all data points lie exactly on the graph of the polynomial.

#### Example Program 12.3: The class E3Reg demonstrates the use of Regression and graphically represents the regression line and its confidence limits

This program again treats the same problem as E1Reg. Additional input parameters are  $r$ , the order of the polynomial to be fitted, and the probability  $P$ , which determines the confidence limits. Presented in a plot are the data points with their errors, the regression line, and – in a different color – its confidence limits.

#### Example Program 12.4: The class E4Reg demonstrates the linear regression with known and with unknown errors

The program uses the data of Example 12.2. The user can declare the errors to be either known or unknown and chooses a probability  $P$  which determines the confidence limits. The graphics – corresponding to Fig. 12.1 – contains data points, regression line, and confidence limits.

# 13. Time Series Analysis

## 13.1 Time Series: Trend

In the previous chapter we considered the dependence of a random variable  $y$  on a controlled variable  $t$ . As in that case we will assume here that  $y$  consists of two parts, the true value of the measured quantity  $\eta$  and a measurement error  $\varepsilon$ ,

$$y_i = \eta_i + \varepsilon_i \quad , \quad i = 1, 2, \dots, n . \quad (13.1.1)$$

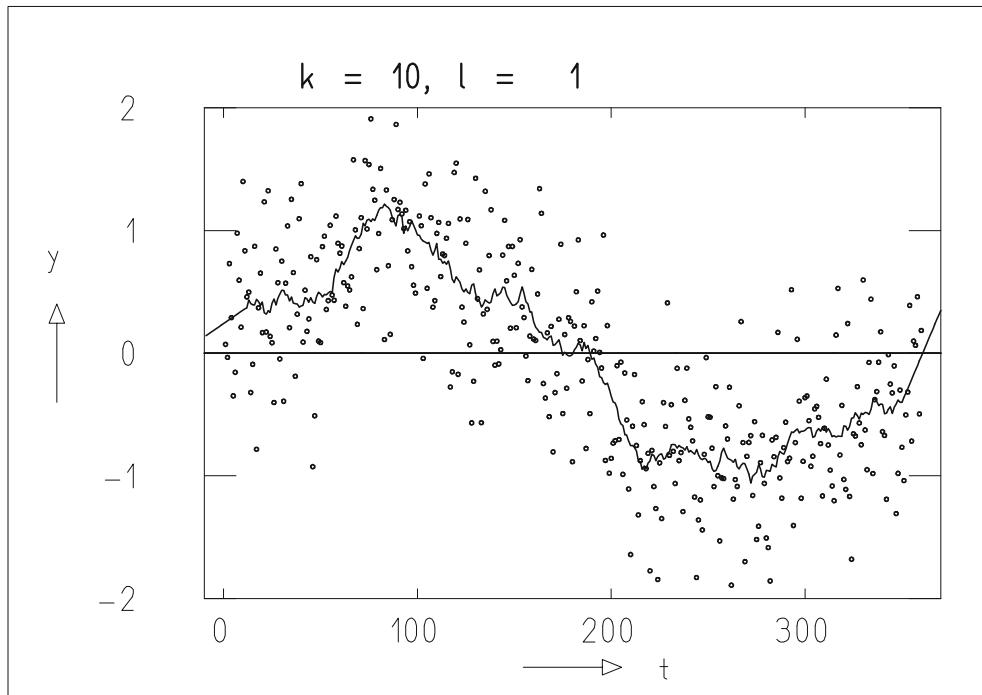
In Chap. 12 we assumed that  $\eta_i$  was a polynomial in  $t$ . The measurement error  $\varepsilon_i$  was considered to be normally distributed about zero.

We now want to make less restrictive assumptions about  $\eta$ . In this chapter we call the controlled variable  $t$  “time”, although in many applications it could be something different. The method we want to discuss is called *time series analysis* and is often applied in economic problems. It can always be used where one has little or no knowledge about the functional relationship between  $\eta$  and  $t$ . In considering time series problems it is common to observe the  $y_i$  at equally spaced points in time,

$$t_i - t_{i-1} = \Delta t = \text{const} \quad , \quad (13.1.2)$$

since this leads to a significant simplification of the formulas.

An example of a time series is shown in Fig. 13.1. If we look first only at the measured points, we notice strong fluctuations from point to point. Nevertheless they clearly follow a certain trend. In the left half of the plot they are mostly positive, and in the right half, mostly negative. One could qualitatively obtain the average time dependence by drawing a smooth curve by hand through the points. Since, however, such curves are not free from personal influences, and are thus not reproducible, we must try to develop an objective method.



**Fig. 13.1:** Data points (*circles*) and moving average (joined by line segments).

We use the notation from (13.1.1) and call  $\eta_i$  the *trend* and  $\varepsilon_i$  the *random component* of the *measurement*  $y_i$ . In order to obtain a smoother function of  $t$ , one can, for example, construct for every value of  $y_i$  the expression

$$u_i = \frac{1}{2k+1} \sum_{j=i-k}^{i+k} y_j , \quad (13.1.3)$$

i.e., the unweighted mean of the measurements for the times

$$t_{i-k}, t_{i-k+1}, \dots, t_{i-1}, t_i, t_{i+1}, \dots, t_{i+k} .$$

The expression (13.1.3) is called a *moving average* of  $y$ .

## 13.2 Moving Averages

Of course the moving average (13.1.3) is a very simple construction. We will show later (in Example 13.1) that the use of a moving average of this form is equivalent to the assumption that  $\eta$  is a linear function of time in the interval considered,

$$\eta_j = \alpha + \beta t_j , \quad j = -k, -k+1, \dots, k . \quad (13.2.1)$$

Here  $\alpha$  and  $\beta$  are constants. They can be estimated from the data by linear regression.

Instead of restricting ourselves to the linear case, we will assume more generally that  $\eta$  can be a polynomial of order  $\ell$ .

In the averaging interval,  $t$  takes on the values

$$t_j = t_i + j \Delta t \quad , \quad j = -k, -k+1, \dots, k \quad . \quad (13.2.2)$$

Because  $\eta$  is a polynomial in  $t$ ,

$$\eta_j = a_1 + a_2 t_j + a_3 t_j^2 + \dots + a_{\ell+1} t_j^\ell \quad , \quad (13.2.3)$$

it is also a polynomial in  $j$ ,

$$\eta_j = x_1 + x_2 j + x_3 j^2 + \dots + x_{\ell+1} j^\ell \quad , \quad (13.2.4)$$

since (13.2.2) describes a linear transformation between  $t_j$  and  $j$ , i.e., it is merely a change of scale. We now want to obtain the coefficients  $x_1, x_2, \dots, x_{\ell+1}$  from the data by fitting with least squares. This task has already been treated in Sect. 9.4.1. We assume (in the absence of any better knowledge) that all of the measurements are of the same accuracy. Thus the matrix  $G_y = aI$  is simply a multiple of the unit matrix  $I$ . According to (9.2.26), the vector of coefficients is thus given by

$$\tilde{\mathbf{x}} = -(A^T A)^{-1} A^T \mathbf{y} \quad , \quad (13.2.5)$$

where  $A$  is a  $(2k+1) \times (\ell+1)$  matrix,

$$A = - \begin{pmatrix} 1 & -k & (-k)^2 & \dots & (-k)^\ell \\ 1 & -k+1 & (-k+1)^2 & \dots & (-k+1)^\ell \\ \vdots & & & & \\ 1 & k & k^2 & \dots & k^\ell \end{pmatrix} \quad . \quad (13.2.6)$$

For the trend  $\tilde{\eta}_0$  at the center of the averaging interval ( $j = 0$ ) we obtain from (13.2.4) the estimate

$$\tilde{\eta}_0 = \tilde{x}_1 \quad . \quad (13.2.7)$$

It is equal to the first coefficient of the polynomial. According to (13.2.5),  $\tilde{x}_1$  is obtained by multiplication of the column vector of measurements  $\mathbf{y}$  on the left with the row vector

$$\mathbf{a} = (-(A^T A)^{-1} A^T)_1 \quad , \quad (13.2.8)$$

i.e., with the first row of the matrix  $-(A^T A)^{-1} A^T$ . We obtain

$$\tilde{\eta}_0 = \mathbf{a} \mathbf{y} = a_{-k} y_{-k} + a_{-k+1} y_{-k+1} + \dots + a_0 y_0 + \dots + a_k y_k \quad . \quad (13.2.9)$$

**Table 13.1:** Components of the vector  $\mathbf{a}$  for computing moving averages.

| $\mathbf{a} = (a_{-k}, a_{-k+1}, \dots, a_k) = \frac{1}{A}(\alpha_{-k}, \alpha_{-k+1}, \dots, \alpha_k)$ |       |               |               |               |               |               |               |               |            |
|--|-------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|------------|
| $\alpha_{-j} = \alpha_j$   |       |               |               |               |               |               |               |               |            |
| <hr/>  |       |               |               |               |               |               |               |               |            |
| $\ell = 2$ and $\ell = 3$  |       |               |               |               |               |               |               |               |            |
| <hr/>  |       |               |               |               |               |               |               |               |            |
| $k$  | $A$   | $\alpha_{-7}$ | $\alpha_{-6}$ | $\alpha_{-5}$ | $\alpha_{-4}$ | $\alpha_{-3}$ | $\alpha_{-2}$ | $\alpha_{-1}$ | $\alpha_0$ |
| 2  | 35    |               |               |               |               |               | -3            | 12            | 17         |
| 3  | 21    |               |               |               |               | -2            | 3             | 6             | 7          |
| 4  | 231   |               |               |               | -21           | 14            | 39            | 54            | 59         |
| 5  | 429   |               |               | -36           | 9             | 44            | 69            | 84            | 89         |
| 6  | 143   |               | -11           | 0             | 9             | 16            | 21            | 24            | 25         |
| 7  | 1105  | -78           | -13           | 42            | 87            | 122           | 147           | 162           | 167        |
| <hr/>  |       |               |               |               |               |               |               |               |            |
| $\ell = 4$ and $\ell = 5$  |       |               |               |               |               |               |               |               |            |
| <hr/>  |       |               |               |               |               |               |               |               |            |
| $k$  | $A$   | $\alpha_{-7}$ | $\alpha_{-6}$ | $\alpha_{-5}$ | $\alpha_{-4}$ | $\alpha_{-3}$ | $\alpha_{-2}$ | $\alpha_{-1}$ | $\alpha_0$ |
| 3  | 231   |               |               |               |               | 5             | -30           | 75            | 131        |
| 4  | 429   |               |               |               | 15            | -55           | 30            | 135           | 179        |
| 5  | 429   |               |               | 18            | -45           | -10           | 60            | 120           | 143        |
| 6  | 2431  |               | 110           | -198          | -135          | 110           | 390           | 600           | 677        |
| 7  | 46189 | 2145          | -2860         | -2937         | -165          | 3755          | 7500          | 10125         | 11063      |

This is a linear function of the measurements within the averaging interval. Here the vector  $\mathbf{a}$  does not depend on the measurements, but rather only on  $\ell$  and  $k$ , i.e., on the order of the polynomial and on the length of the interval. Clearly one must choose

$$\ell < 2k + 1 \quad ,$$

since otherwise there would not remain any degrees of freedom for the least-squares fit. The components of  $\mathbf{a}$  for small values of  $\ell$  and  $k$  can be obtained from Table 13.1.

Equation (13.2.9) describes the moving average corresponding to the assumed polynomial (13.2.4). Once the vector  $\mathbf{a}$  is determined, the moving averages

$$u_i = \tilde{\eta}_0(i) = \mathbf{ay}(i) = a_1 y_{i-k} + a_2 y_{i-k+1} + \cdots + a_{2k+1} y_{i+k} \quad (13.2.10)$$

can easily be computed for each value of  $i$ .

**Example 13.1:** Moving average with linear trend

In the case of a linear trend function,

$$\eta_j = x_1 + x_2 j \quad ,$$

the matrix  $A$  becomes simply

$$A = - \begin{pmatrix} 1 & -k \\ 1 & -k+1 \\ \vdots & \\ 1 & k \end{pmatrix} \quad .$$

One then has

$$\begin{aligned} A^T A &= \begin{pmatrix} 1 & 1 & \dots & 1 \\ -k & -k+1 & \dots & k \end{pmatrix} \begin{pmatrix} 1 & -k \\ 1 & -k+1 \\ \vdots & \vdots \\ 1 & k \end{pmatrix} \\ &= \begin{pmatrix} 2k+1 & 0 \\ 0 & k(k+1)(2k+1)/3 \end{pmatrix} \quad , \\ (A^T A)^{-1} &= \begin{pmatrix} \frac{1}{2k+1} & 0 \\ 0 & \frac{3}{k(k+1)(2k+1)} \end{pmatrix} \quad , \\ \mathbf{a} &= -(A^T A)^{-1} A^T = \frac{1}{2k+1} (1, 1, \dots, 1) \quad . \end{aligned}$$

In this case the moving average is simply the unweighted mean (13.1.3). ■

For more complicated models one can obtain the vectors  $\mathbf{a}$  either by solving (13.2.8) or simply from Table 13.1. Because of the symmetry of  $A$ , one can show that polynomials of odd order (i.e.,  $\ell = 2n$  with  $n$  an integer) have the same values of  $\mathbf{a}$  as those of polynomials of the next lower order  $\ell = 2n - 1$ . One can also easily show that  $\mathbf{a}$  has the symmetry

$$a_j = a_{-j} \quad , \quad j = 1, 2, \dots, k \quad . \quad (13.2.11)$$

### 13.3 Edge Effects

Of course the moving average (13.2.10) can be used to estimate the trend only for points  $i$  that have at least  $k$  neighboring points both to the right and left, since the averaging interval covers  $2k + 1$  points. This means that for the first and last  $k$  points of a time series one must use a different estimator. One obtains the most obvious generalization of the estimator by extrapolating the polynomial (13.2.4) rather than using it only at the center of an interval. One then obtains the estimators

$$\begin{aligned}\tilde{\eta}_i = u_i &= \tilde{x}_1^{(k+1)} + \tilde{x}_2^{(k+1)}(i - k - 1) + \tilde{x}_3^{(k+1)}(i - k - 1)^2 + \dots \\ &\quad + \tilde{x}_{\ell+1}^{(k+1)}(i - k - 1)^\ell, \quad i \leq k, \\ \tilde{\eta}_i = u_i &= \tilde{x}_1^{(n-k)} + \tilde{x}_2^{(n-k)}(i + k - n) + \tilde{x}_3^{(n-k)}(i + k - n)^2 + \dots \\ &\quad + \tilde{x}_{\ell+1}^{(n-k)}(i + k - n), \quad i > n - k.\end{aligned}\quad (13.3.1)$$

Here the notation  $\tilde{x}^{(k+1)}$  and  $\tilde{x}^{(n-k)}$  indicates that the coefficients  $\tilde{x}$  were determined for the first and last intervals of the time series for which the centers are at  $(k + 1)$  and  $(n - k)$ .

The estimators are now defined even for  $i < 1$  and  $i > n$ . They thus offer the possibility to continue the time series (e.g., into the future). Such extrapolations must be treated with great care for two reasons:

- (i) Usually there is no theoretical justification for the assumption that the trend is described by a polynomial. It merely simplifies the computation of the moving average. Without a theoretical understanding for a trend model, the meaning of extrapolations is quite unclear.
- (ii) Even in cases where the trend can rightly be described by a polynomial, the confidence limits quickly diverge from the estimated polynomial in the extrapolated region. The extrapolation becomes very inaccurate.

Whether point (i) is correct must be carefully checked in each individual case. The more general point (ii) is already familiar from the linear regression (cf. Fig. 12.1). We will investigate this in detail in the next section.

### 13.4 Confidence Intervals

We first consider the confidence interval for the moving average  $u_i$  from Eq. (13.2.10). The errors of the measurements  $y_i$  are unknown and must therefore first be estimated. From (12.3.2) one obtains for the sample variance of the  $y_j$  in the interval of length  $2k + 1$ ,

$$s_y^2 = \frac{1}{2k-\ell} \sum_{j=-k}^k (y_j - \tilde{\eta}_j)^2 \quad , \quad (13.4.1)$$

where  $\tilde{\eta}_j$  is given by

$$\tilde{\eta}_j = \tilde{x}_1 + \tilde{x}_2 j + \tilde{x}_3 j^2 + \cdots + \tilde{x}_{\ell+1} j^\ell \quad . \quad (13.4.2)$$

The covariance matrix for the measurements can then be estimated by

$$G_y^{-1} \approx s_y^2 I \quad . \quad (13.4.3)$$

The covariance matrix of the coefficients  $\mathbf{x}$  is then given by (9.2.27),

$$G_{\tilde{x}}^{-1} \approx (A^T G_y A)^{-1} = s_y^2 (A^T A)^{-1} \quad . \quad (13.4.4)$$

Since  $u_i = \tilde{\eta}_0 = \tilde{x}_1$ , we thus have for an estimator of the variance of  $u_i$

$$s_{\tilde{x}_1}^2 = (G_{\tilde{x}}^{-1})_{11} = s_y^2 ((A^T A)^{-1})_{11} = s_y^2 a_0 \quad . \quad (13.4.5)$$

From (13.2.6), (13.2.7), and (13.2.8) one easily obtains that  $(A^T A)_{11}^{-1} = a_0$ , since the middle row of  $A$  is  $-(1, 0, 0, \dots, 0)$ .

Using the same reasoning as in Sect. 12.3 we obtain at a confidence level of  $1 - \alpha$

$$\frac{|\tilde{\eta}_0(i) - \eta_0(i)|}{s_y a_0} \leq t_{1-\frac{1}{2}\alpha} \quad . \quad (13.4.6)$$

For a given  $\alpha$  we can give the confidence limits as

$$\eta_0^\pm(i) = \tilde{\eta}_0(i) \pm a_0 s_y t_{1-\frac{1}{2}\alpha} \quad . \quad (13.4.7)$$

Here  $t_{1-\frac{1}{2}\alpha}$  is a quantile of Student's distribution for  $2k - \ell$  degrees of freedom. The true value of the trend lies within these limits with a confidence level of  $1 - \alpha$ .

Completely analogous, although more difficult computationally, is the determination of confidence limits at the ends of the time series. The moving average is now given by (13.3.1). Labeling the arguments in the expressions (13.3.1)  $j = i - k - 1$  and  $j = i + k - n$ , we obtain

$$\tilde{\eta} = \mathbf{T} \mathbf{x} \quad . \quad (13.4.8)$$

Here  $\mathbf{T}$  is a row vector of length  $\ell + 1$ ,

$$\mathbf{T} = (1, j, j^2, \dots, j^\ell) \quad . \quad (13.4.9)$$

According to the law of error propagation (3.8.4) we obtain

$$G_{\tilde{\eta}}^{-1} = T G_x^{-1} T^T \quad . \quad (13.4.10)$$

With (13.4.4) we finally have

$$G_{\tilde{\eta}}^{-1} \approx s_{\tilde{\eta}}^2 = s_y^2 T (A^T A)^{-1} T^T \quad , \quad (13.4.11)$$

where  $s_y^2$  is again given by (13.4.1).

The quantity  $s_{\tilde{\eta}}^2$  can now be computed for every value of  $j$ , even for values lying outside of the time series itself. Thus we obtain the confidence limits

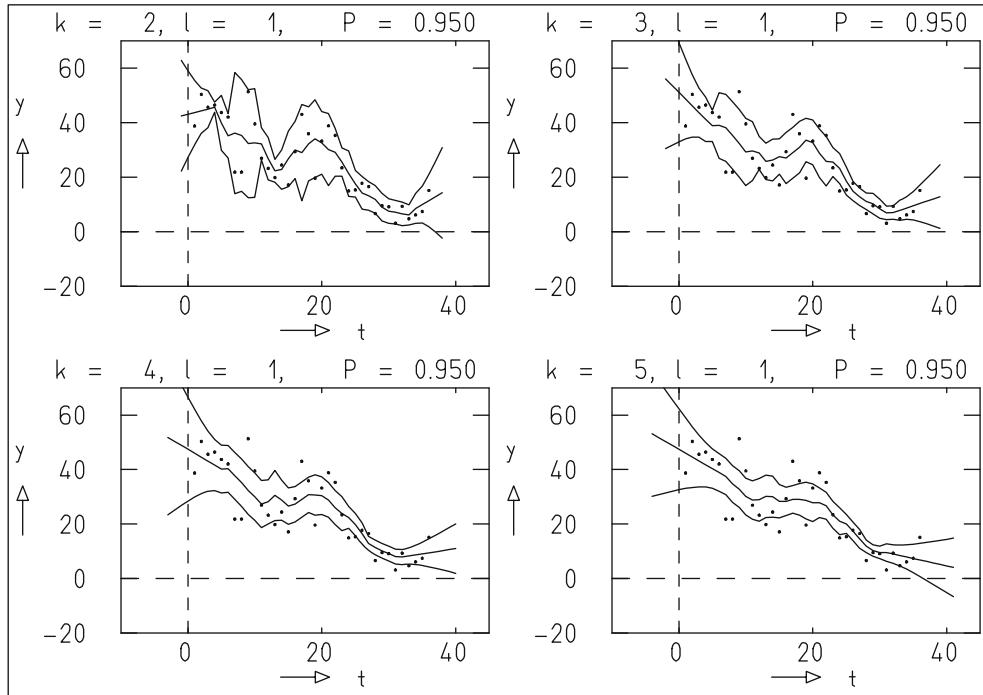
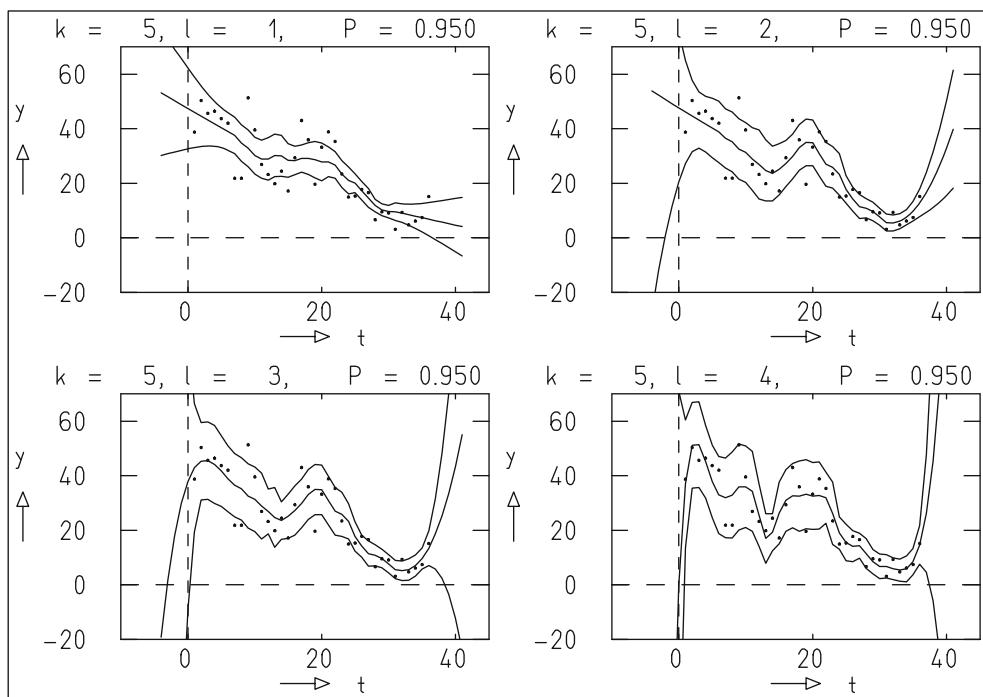
$$\eta^{\pm}(i) = \tilde{\eta}(i) \pm s_{\tilde{\eta}} t_{1-\frac{1}{2}\alpha} \quad . \quad (13.4.12)$$

Caution is always recommended in interpreting the results of a time series analysis. This is particularly true for two reasons:

1. There is usually no a priori justification for the mathematical model on which the time series analysis is based. One has simply chosen a convenient procedure in order to “separate out” statistical fluctuations.
2. The user has considerable freedom in choosing the parameters  $k$  and  $\ell$ , which, however, can have a significant influence on the results. The following example gives an impression of the magnitude of such influences.

**Example 13.2:** Time series analysis of the same set of measurements using different averaging intervals and polynomials of different orders

Figures 13.2 and 13.3 contain time series analyses of the average number of sun spots observed in the 36 months from January 1962 through December 1964. Various values of  $k$  and  $\ell$  were used. The individual plots in Fig. 13.2 show  $\ell = 1$  (linear averaging) but different interval lengths ( $2k + 1 = 5, 7, 9$ , and 11). One can see that the curve of moving averages becomes smoother and the confidence interval becomes narrower when  $k$  increases, but that then the mean deviation of the individual observations from the curve also increases. The extrapolation outside the range of measured points is, of course, a straight line. (For  $\ell = 0$  we would have obtained the same moving averages for the inner points. The outer and extrapolated points would lie, however, on a horizontal line, since a polynomial of order zero is a constant.) The plots in Fig. 13.3 correspond to the interval lengths  $2k + 1 = 7$  and  $\ell = 1, 2, 3, 4$ . The moving averages lie closer to the data points and the confidence interval becomes larger when the value of  $\ell$  increases. ■

**Fig. 13.2:** Time series analyses of the same data with fixed  $\ell$  and various values of  $k$ .**Fig. 13.3:** Time series analyses of the same data with fixed  $k$  and various values of  $\ell$ .

From these observations we can derive the following qualitative rules:

1. The averaging interval should not be chosen larger than the region where one expects that the data can be well described by a polynomial of the given order. That is, for  $\ell = 1$  the interval  $2k + 1$  should be chosen such that the expected nonlinear effects within the interval remain small.
2. On the other hand, the smoothing effect becomes stronger as the length of the averaging interval increases. As a rule of thumb, the smoothing becomes more effective for increasing  $2k + 1 - \ell$ .
3. Caution is required in extrapolation of a time series, especially in non-linear cases.

The art of time series analysis is, of course, much more highly developed than what we have been able to describe in this short chapter. The interested reader is referred to the specialized literature, where, for example, smoothing functions other than polynomials or multidimensional analyses are treated.

## 13.5 Java Class and Example Programs

### Java Class for Time Series Analysis

`TimeSeries` performs a time series analysis.

**Example Program 13.1:** The class `E1TimSer` demonstrates the use of `TimeSeries`

The uses the data of Example 13.2. After setting some parameters it performs a time series analysis by a call of `TimeSeries` with  $k = 2$ ,  $\ell = 2$ . The data, moving averages, and distances to the confidence limits (at a confidence level of 90 %) are output numerically.

**Example Program 13.2:** The class `E2TimSer` performs a time series analysis and yields graphical output

The program starts works on the same data as `E1TimSer`. It allows interactive input for the parameters  $k$  and  $\ell$  and for the confidence level  $P$  and then performs a time series analysis. Subsequently a plot is produced in which the data are displayed as small circles. The floating averages are shown as a polyline. Polylines in a different color indicate the confidence limits.

**Suggestion:** Produce the individual plots of Figs. 13.2 and 13.3.

# Literature

## Literature Cited in the Text

- [1] A. KOLMOGOROV, *Ergebn. Math.* **2** (1933) 3
- [2] D.E. KNUTH, *The Art of Computer Programming*, vol. 2, Addison-Wesley, Reading MA 1981
- [3] P. L'ECUYER, *Comm. ACM* **31** (1988) 742
- [4] B.A. WICHMANN and I.D. HILL, *Appl. Stat.* **31** (1982) 188
- [5] G.E.P. BOX and M.E. MULLER, *Ann. Math. Stat.* **29** (1958) 611
- [6] L. VON BORTKIEWICZ, *Das Gesetz der kleinen Zahlen*, Teubner, Leipzig 1898
- [7] D.J. DE SOLLA PRICE, *Little Science, Big Science*, Columbia University Press, New York 1965
- [8] M.G. KENDALL and A. STUART, *The Advanced Theory of Statistics*, vol. 2, Charles Griffin, London 1968
- [9] S.S. WILKS, *Ann. Math. Stat.*, **9** (1938) 60
- [10] A.H. ROSENFELD, H.A. BARBERO-GALTIERI, W.J. PODOLSKI, L.R. PRICE, P. SÖDING, CH.G. WOHL, M. ROOS, and W.J. WILLIS, *Rev. Mod. Phys.* **33** (1967) 1
- [11] PARTICLE DATA GROUP, *Physics Letters* **B204** (1988) 1
- [12] W.H. PRESS, B.P. FLANNERY, S.A. TEUKOLSKY and W.T. VETTERLING, *Numerical Recipes*, Cambridge University Press, Cambridge 1986

- [13] R.P. BRENT, *Algorithms for Minimization without Derivatives*, Prentice-Hall, Englewood Cliffs NJ 1973
- [14] J.A. NELDER and R. MEAD, *Computer Journal* **7** (1965) 308
- [15] M.J.D. POWELL, *Computer Journal* **7** (1965) 155
- [16] D.W. MARQUARDT, *J. Soc. Ind. Appl. Math.* **11** (1963) 431
- [17] C. LANCZOS, *SIAM J. Numerical Analysis* **1** (1964) 86
- [18] C.L. LAWSON and R.J. HANSON, *Solving Least Squares Problems*, Prentice-Hall, Englewood Cliffs NJ 1974
- [19] G.H. GOLUB and W. KAHN, *SIAM J. Numerical Analysis* **2** (1965) 205
- [20] P.A. BUSINGER and G.H. GOLUB, *Comm. ACM* **12** (1969) 564
- [21] G.H. GOLUB and C. REINSCH, in *Linear Algebra* (J.H. WILKINSON and C. REINSCH, eds.), p. 134, Springer, Berlin 1971
- [22] J.G.F. FRANCIS, *Computer Journal* **4** (1960) 265, 332
- [23] F. JAMES and M. ROOS, *Nuclear Physics* **B172** (1980) 475
- [24] M.L. SWARTZ, *Nuclear Instruments and Methods* **A294** (1966) 278
- [25] V.H. REGENER, *Physical Review* **84** (1951) 161
- [26] G. ZECH, *Nuclear Instruments and Methods* **A277** (1989) 608
- [27] H. RUTISHAUSER, *Numerische Mathematik* **5** (1963) 48
- [28] W. ROMBERG, *Det. Kong. Norske Videnskapers Selskap Forhandlinger* **28** (1955) Nr. 7
- [29] K.S. KOELBIG, in *CERN Computer Centre Program Library*, Program D401, CERN, Geneva 1990

## Bibliography

The following short list contains several books which can serve to extend the breadth and depth of the present material. It is of course in no way complete, but instead merely indicates a number of books of varied levels of difficulty.

## Probability

- L. BREIMAN, *Probability*, Addison-Wesley, Reading MA 1968
- H. CRAMÈR, *The Elements of Probability Theory*, Wiley, New York 1955
- B.V. GNEDENKO, *The Theory of Probability*, 4th ed., Chelsea, New York 1967
- KAI LAI CHUNG, *A Course in Probability Theory*, Harcourt, Brace and World, New York 1968
- K. KRICKEBERG, *Wahrscheinlichkeitsrechnung*, Teubner, Stuttgart 1963

## Mathematical Statistics

- P.R. BEVINGTON, *Data Reduction and Error Analysis for the Physical Sciences*, McGraw-Hill, New York 1969
- B.E. COOPER, *Statistics for Experimentalists*, Pergamon, Oxford 1969
- G. COWAN, *Statistical Data Analysis*, Clarendon Press, Oxford 1998
- H. CRAMÈR, *Mathematical Methods of Statistics*, University Press, Princeton 1946
- W.J. DIXON and F.J. MASSEY, *Introduction to Statistical Analysis*, McGraw-Hill, New York 1969
- D. DUMAS DE RAULY, *L'Estimation Statistique*, Gauthier-Villars, Paris 1968
- W.T. EADIE, D. DRIJARD, F.E. JAMES, M. ROOS and B. SADOULET, *Statistical Methods in Experimental Physics*, North-Holland, Amsterdam 1971
- W. FELLER, *An Introduction to Probability Theory and its Applications*, 2 Vols., Wiley, New York 1968
- M. FISZ, *Probability Theory and Mathematical Statistics*, Wiley, New York 1963
- D.A.S. FRASER, *Statistics; An Introduction*, Wiley, New York 1958
- H. FREEMAN, *Introduction to Statistical Inference*, Addison-Wesley, Reading, MA 1963
- A.G. FRODENSEN and O. SKJEGGESTAD, *Probability and Statistics in Particle Physics*, Universitetsforlaget, Bergen 1979
- P.G. HOEL, *Introduction to Mathematical Statistics*, 4th ed., Wiley, New York 1971

- M.G. KENDALL and A. STUART *The Advanced Theory of Statistics, 4th ed., 3 vols.*, Charles Griffin, London 1977
- L. LYONS, *Statistics for Nuclear and Particle Physicists*, Cambridge University Press, Cambridge 1986
- J. MANDEL, *The Statistical Analysis of Experimental Data*, Interscience, New York 1964
- S.L. MEYER, *Data Analysis for Scientists and Engineers*, Wiley, New York 1975
- L. SACHS, *Statistische Auswertungsmethoden, 5. Aufl.*, Springer, Berlin 1978
- E. SVERDRUP, *Laws and Chance Variations*, 2 Bde., North-Holland, Amsterdam 1967
- B.L. VAN DER WAERDEN, *Mathematische Statistik, 3. Aufl.*, Springer, Berlin 1971
- S.S. WILKS, *Mathematical Statistics*, Wiley, New York 1962
- T. YAMANE, *Elementary Sampling Theory*, Prentice-Hall, Englewood Cliffs, NJ 1967
- T. YAMANE, *Statistics, An Introductory Analysis*, Harper and Row, New York 1967

### Numerical Methods: Matrix Programs

- A. BJÖRK and G. DAHLQUIST, *Numerische Methoden*, R. Oldenbourg Verlag, München 1972
- R.P. BRENT, *Algorithms for Minimization without Derivatives*, Prentice-Hall, Englewood Cliffs, NJ 1973
- C.T. FIKE, *Computer Evaluation of Mathematical Functions*, Prentice-Hall, Englewood Cliffs, NJ 1968
- G.H. GOLUB and C.F. VAN LOAN, *Matrix Computations*, Johns Hopkins University Press, Baltimore 1983
- R.W. HAMMING, *Numerical Methods for Engineers and Scientists, 2nd ed.*, McGraw-Hill, New York 1973
- D.E. KNUTH, *The Art of Computer Programming*, 3 vols., Addison-Wesley, Reading MA 1968
- C.L. LAWSON and R.J. HANSON, *Solving Least Squares Problems*, Prentice-Hall, Englewood Cliffs NJ 1974

- W.H. PRESS, B.P. FLANNERY, S.A. TEUKOLSKY, and W.T. VETTERLING,  
*Numerical Recipes*, Cambridge University Press, Cambridge 1986
- J.R. RICE, *Numerical Methods, Software and Analysis*, McGraw-Hill, New York 1983
- J. STOER and R. BURLISCH, *Einführung in die Numerische Mathematik*, 2 Bde., Springer, Berlin 1983
- J.H. WILKINSON and C. REINSCH, *Linear Algebra*, Springer, Berlin 1971

### Collections of Formulas: Statistical Tables

- M. ABRAMOWITZ and A. STEGUN, *Handbook of Mathematical Functions*, Dover, New York, 1965
- R.A. FISHER and F. YATES, *Statistical Tables for Biological, Agricultural and Medical Research*, Oliver and Boyd, London 1957
- U. GRAF and H.J. HENNING, *Formeln und Tabellen zur Mathematischen Statistik*, Springer, Berlin 1953
- A. HALD, *Statistical Tables and Formulas*, Wiley, New York 1960
- G.A. KORN and T.M. KORN, *Mathematical Handbook for Scientists and Engineers*, 2nd ed., McGraw-Hill, New York 1968
- D.V. LINDLEY and J.C.P. MILLER, *Cambridge Elementary Statistical Tables*, University Press, Cambridge 1961
- D.B. OWEN, *Handbook of Statistical Tables*, Addison-Wesley, Reading MA 1962
- E.S. PEARSON and H.O. HARTLEY, *Biometrika Tables for Statisticians*, University Press, Cambridge 1958
- W. WETZEL, M.D. JÖHNK, and P. NAEVE, *Statistische Tabellen*, Walter de Gruyter, Berlin 1967

# A. Matrix Calculations

The solution of the over-determined linear system of equations  $A\mathbf{x} \approx \mathbf{b}$  is of central significance in data analysis. This can be solved in an optimal way with the singular value decomposition, first developed in the late 1960s. In this appendix the elementary definitions and calculation rules for matrices and vectors are summarized in Sects. A.1 and A.2. In Sect. A.3 orthogonal transformations are introduced, in particular the Givens and Householder transformations, which provide the key to the singular value decomposition.

After a few remarks on determinants (Sect. A.4) there follows in Sect. A.5 a discussion of various cases of matrix equations and a theorem on the orthogonal decomposition of an arbitrary matrix, which is of central importance in this regard. The classical procedure of normal equations, which, however, is inferior to the singular value decomposition, is described here.

Sections A.6–A.8 concern the particularly simple case of exactly determined, non-singular matrix equations. In this case the inverse matrix  $A^{-1}$  exists, and the solution to the problem  $A\mathbf{x} = \mathbf{b}$  is  $\mathbf{x} = A^{-1}\mathbf{b}$ . Methods and programs for finding the solution are given. The important special case of a positive-definite symmetric matrix is treated in Sect. A.9.

In Sect. A.10 we define the pseudo-inverse matrix  $A^+$  of an arbitrary matrix  $A$ . After introducing eigenvectors and eigenvalues in Sect. A.11, the singular value decomposition is presented in Sects. A.12 and A.13. Computer routines are given in Sect. A.14. Modifications of the procedure and the consideration of constraints are the subject of Sects. A.15 through A.18.

It has been attempted to make the presentation illustrative rather than mathematically rigorous. Proofs in the text are only indicated in a general way, or are omitted entirely. As mentioned, the singular value decomposition is not yet widespread. An important goal of this appendix is to make its use possible. For readers with a basic knowledge of matrix calculations, the material covered in Sects. A.3, A.12, A.13, A.14.1, and A.18 is sufficient for this task. Sections A.14.2 through A.14.5 contain technical details on carrying out the singular value decomposition and can be omitted by hurried users.

All procedures described in this Appendix are implemented as methods of the classes `DatanVector` or `DatanMatrix`, respectively. Only in a few cases will we refer to these methods explicitly, in order to establish the connection with some of the more complicated algorithms in the text.

## A.1 Definitions: Simple Operations

By a *vector* in  $m$  dimensions (an  $m$ -vector)  $\mathbf{a}$  we mean an  $m$ -tuple of real numbers, which are the *components* of  $\mathbf{a}$ . The arrangement of the components in the form

$$\mathbf{a} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{pmatrix} \quad (\text{A.1.1})$$

is called a *column vector*.

A  $m \times n$  matrix is a rectangular arrangement of  $m \times n$  numbers in  $m$  rows and  $n$  columns,

$$A = \begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{22} & \cdots & A_{2n} \\ \vdots & \vdots & & \vdots \\ A_{m1} & A_{m2} & \cdots & A_{mn} \end{pmatrix} . \quad (\text{A.1.2})$$

It can be viewed to be composed of  $n$  column vectors. By *transposition* of the  $m \times n$  matrix  $A$  one obtains an  $n \times m$  matrix  $A^T$  with elements

$$A_{ik}^T = A_{ki} \quad . \quad (\text{A.1.3})$$

Under transposition a column vector becomes a *row vector*,

$$\mathbf{a}^T = (a_1, a_2, \dots, a_m) \quad . \quad (\text{A.1.4})$$

A column vector is an  $m \times 1$  matrix; a row vector is a  $1 \times m$  matrix.

For matrices one has the following elementary rules for *addition*, *subtraction* and *multiplication by a constant*

$$A \pm B = C \quad , \quad C_{ik} = A_{ik} \pm B_{ik} \quad , \quad (\text{A.1.5})$$

$$\alpha A = B \quad , \quad B_{ik} = \alpha A_{ik} \quad . \quad (\text{A.1.6})$$

The *product*  $AB$  of two matrices is only defined if the number of columns of the first matrix is equal to the number of rows of the second, e.g.,  $A = A_{m \times \ell}$  and  $B = B_{\ell \times m}$ . One has then

$$AB = C \quad , \quad C_{ik} = \sum_{j=1}^{\ell} A_{ij} B_{jk} \quad . \quad (\text{A.1.7})$$

Since

$$C_{ik}^T = C_{ki} = \sum_{j=1}^{\ell} A_{kj} B_{ji} = \sum_{j=1}^{\ell} A_{jk}^T B_{ij}^T = \sum_{j=1}^{\ell} B_{ij}^T A_{jk}^T \quad ,$$

one has

$$C^T = (AB)^T = B^T A^T \quad . \quad (\text{A.1.8})$$

With (A.1.7) one can also define the product of a row vector  $\mathbf{a}^T$  with a column vector  $\mathbf{b}$ , if both have the same number of elements  $m$ ,

$$\mathbf{a} \cdot \mathbf{b} = \mathbf{a}^T \mathbf{b} = c \quad , \quad c = \sum_{j=1}^m a_j b_j \quad . \quad (\text{A.1.9})$$

The result is then a number, i.e., a *scalar*. The product (A.1.9) is called the *scalar product*. It is usually written without indicating the transposition simply as  $\mathbf{a} \cdot \mathbf{b}$ . The vectors  $\mathbf{a}$  and  $\mathbf{b}$  are *orthogonal* to each other if their scalar product vanishes. Starting from (A.1.9) one obtains the following useful property of the matrix product (A.1.7). The element  $C_{ik}$ , which is located at the intersection of the  $i$ th row and the  $k$ th column of the product matrix  $C$ , is equal to the scalar product of the  $i$ th row of the first matrix  $A$  with the  $k$ th column of the second matrix  $B$ .

The *diagonal elements* of a matrix (A.1.2) are the elements  $A_{ii}$ . They form the *main diagonal* of the matrix  $A$ . If all of the non-diagonal elements vanish,  $A_{ij} = 0, i \neq j$ , then  $A$  is a *diagonal matrix*. An  $n \times n$  diagonal matrix all of whose diagonal elements are unity is the  $n$ -dimensional *unit matrix*  $I_n = I$ ,

$$\begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & & & \\ 0 & 0 & \dots & 1 \end{pmatrix} = \begin{pmatrix} 1 & & & 0 \\ & 1 & & \\ & & \ddots & \\ 0 & & & 1 \end{pmatrix} = I \quad . \quad (\text{A.1.10})$$

The *null matrix* has only zeros as elements:

$$0 = \begin{pmatrix} 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \vdots & & & \\ 0 & 0 & \dots & 0 \end{pmatrix} \quad . \quad (\text{A.1.11})$$

A null matrix with only one column is the *null vector*  $\mathbf{0}$ .

We will now mention several more special types of square matrices.

A square matrix is *symmetric* if

$$A_{ik} = A_{ki} \quad . \quad (\text{A.1.12})$$

If

$$A_{ik} = -A_{ki} \quad , \quad (\text{A.1.13})$$

then the matrix is *antisymmetric*.

A *bidiagonal matrix*  $B$  possesses non-vanishing elements only on the main diagonal ( $b_{ii}$ ) and on the parallel diagonal directly above it ( $b_{i,i+1}$ ).

A *tridiagonal matrix* possesses in addition non-vanishing elements directly below the main diagonal. A *lower triangular matrix* has non-vanishing elements only on and below the main diagonal, an *upper triangular matrix* only on and above the main diagonal.

The *Euclidian norm* or the *absolute value* of a vector is

$$|\mathbf{a}| = \|\mathbf{a}\|_2 = a = \sqrt{\mathbf{a}^T \mathbf{a}} = \sqrt{\sum_j a_j^2} \quad . \quad (\text{A.1.14})$$

A vector with unit norm is called a *unit vector*. We write this in the form

$$\hat{\mathbf{a}} = \mathbf{a}/a \quad .$$

More general *vector norms* are

$$\begin{aligned} \|\mathbf{a}\|_p &= \left( \sum_j |a_j|^p \right)^{1/p} \quad , \quad 1 \leq p < \infty \quad , \\ \|\mathbf{a}\|_\infty &= \max_j |a_j| \quad . \end{aligned} \quad (\text{A.1.15})$$

For every vector norm  $\|\mathbf{x}\|$  one defines a *matrix norm*  $\|A\|$  as

$$\|A\| = \max_{\mathbf{x} \neq 0} \|A\mathbf{x}\|/\|\mathbf{x}\| \quad . \quad (\text{A.1.16})$$

Matrix norms have the following properties:

$$\|A\| > 0 \quad , \quad A \neq 0 \quad ; \quad \|A\| = 0 \quad , \quad A = 0 \quad , \quad (\text{A.1.17})$$

$$\|\alpha A\| = |\alpha| \|A\| \quad , \quad \alpha \text{ real} \quad , \quad (\text{A.1.18})$$

$$\|A + B\| \leq \|A\| + \|B\| \quad , \quad (\text{A.1.19})$$

$$\|AB\| \leq \|A\| \|B\| \quad . \quad (\text{A.1.20})$$

## A.2 Vector Space, Subspace, Rank of a Matrix

An *n-dimensional vector space* is the set of all *n*-dimensional vectors. If  $\mathbf{u}$  and  $\mathbf{v}$  are vectors in this space, then  $\alpha\mathbf{u}$  and  $\mathbf{u} + \mathbf{v}$  are also in the space, i.e., the vector space is *closed* under vector addition and under multiplication with a scalar  $\alpha$ . The vectors  $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_k$  are *linearly independent* if

$$\sum_{j=1}^k \alpha_j \mathbf{a}_j \neq 0 \quad (\text{A.2.1})$$

for all  $\alpha_j$  except for  $\alpha_1 = \alpha_2 = \dots = \alpha_k = 0$ . Otherwise they are *linearly dependent*. The maximum number  $k_{\max}$  of vectors that can be linearly independent is equal to the *dimension* of the vector space  $n$ . An arbitrary set of  $n$  linearly independent vectors  $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n$  forms a *basis* of the vector space. Any vector  $\mathbf{a}$  can be expressed as a *linear combination* of the basis vectors,

$$\mathbf{a} = \sum_{j=1}^n \alpha_j \mathbf{a}_j \quad . \quad (\text{A.2.2})$$

A special basis is

$$\mathbf{e}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad \mathbf{e}_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad \dots, \quad \mathbf{e}_n = \begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix}. \quad (\text{A.2.3})$$

These basis vectors are *orthonormal*, i.e.,

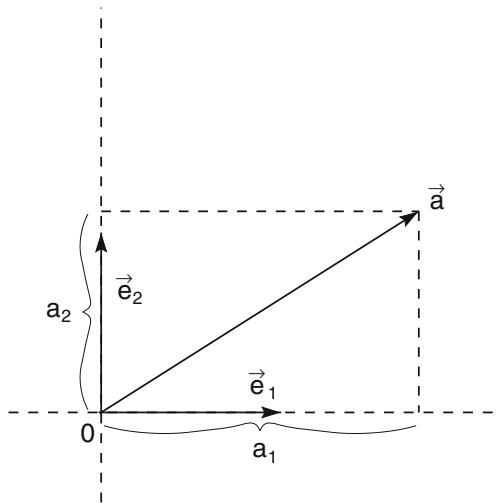
$$\mathbf{e}_i \cdot \mathbf{e}_j = \delta_{ij} = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases} . \quad (\text{A.2.4})$$

The component  $a_j$  of the vector  $\mathbf{a}$  is the scalar product of  $\mathbf{a}$  with the basis vector  $\mathbf{e}_j$ ,

$$\mathbf{a} \cdot \mathbf{e}_j = a_j , \quad (\text{A.2.5})$$

cf. (A.1.1) and (A.1.9).

For  $n \leq 3$  vectors can be visualized geometrically. A vector  $\mathbf{a}$  can be represented as an arrow of length  $a$ . The basis vectors (A.2.3) are perpendicular to each other and are of unit length. The perpendicular projections of  $\mathbf{a}$  onto the directions of the basis vectors are the components (A.2.5), as shown in Fig. A.1.



**Fig.A.1:** The vector  $\mathbf{a}$  in the system of orthonormal basis vectors  $\mathbf{e}_1, \mathbf{e}_2$ .

A subset  $T$  of a vector space  $S$  is called a *subspace* if it is itself closed under vector addition and multiplication with a scalar. The greatest possible number of linearly independent vectors in  $T$  is the dimension of  $T$ . The product of an  $m \times n$  matrix  $A$  with an  $n$ -vector  $\mathbf{a}$  is an  $m$ -vector  $\mathbf{b}$ ,

$$\mathbf{b} = A\mathbf{a} \quad . \quad (\text{A.2.6})$$

The relation (A.2.6) can be regarded as a *mapping* or *transformation* of the vector  $\mathbf{a}$  onto the vector  $\mathbf{b}$ .

The *span* of a set of vectors  $\mathbf{a}_1, \dots, \mathbf{a}_k$  is the vector space defined by the set of all linear combinations  $\mathbf{u}$  of these vectors,

$$\mathbf{u} = \sum_{j=1}^k \alpha_j \mathbf{a}_j \quad . \quad (\text{A.2.7})$$

It has a dimension  $m \leq k$ . The *column space* of an  $m \times n$  matrix  $A$  is the span of the  $n$  column vectors of  $A$ ; in this case the  $m$ -vectors  $\mathbf{u}$  have the form  $\mathbf{u} = \mathbf{Ax}$  with arbitrary  $n$ -vectors  $\mathbf{x}$ . Clearly the dimension of the column space is  $\leq \min(m, n)$ . Similarly, the *row space* of  $A$  is the span of the  $m$  row vectors.

The *null space* or *kernel* of  $A$  consists of the set of vectors  $\mathbf{x}$  for which

$$\mathbf{Ax} = 0 \quad . \quad (\text{A.2.8})$$

Column and row spaces of an  $m \times n$  matrix have the same dimension. This is called the *rank* of the matrix. An  $m \times n$  has *full rank* if

$$\text{Rang}(A) = \min(m, n) \quad . \quad (\text{A.2.9})$$

Otherwise it has *reduced rank*. An  $n \times n$  matrix with  $\text{Rang}(A) < n$  is said to be *singular*.

A vector  $\mathbf{a}$  is orthogonal to a subspace  $T$  if it is orthogonal to every vector  $t \in T$ . (A trivial example is  $\mathbf{t} = t_1 \mathbf{e}_1 + t_2 \mathbf{e}_2$ ,  $\mathbf{a} = a \mathbf{e}_3$ ,  $\mathbf{a} \cdot \mathbf{t} = 0$ .) A subspace  $U$  is orthogonal to the subspace  $T$  if for every pair of vectors  $\mathbf{u} \in U$ ,  $\mathbf{t} \in T$  one has  $\mathbf{u} \cdot \mathbf{t} = 0$ . The set of all vectors  $\mathbf{u} + \mathbf{t}$  forms a vector space  $V$ , called the *direct sum* of  $T$  and  $U$ ,

$$V = T \oplus U \quad . \quad (\text{A.2.10})$$

Its dimension is

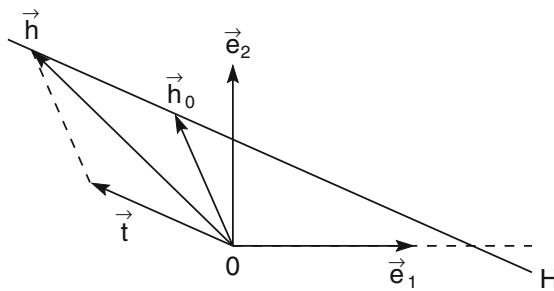
$$\dim(V) = \dim(T) + \dim(U) \quad . \quad (\text{A.2.11})$$

If (A.2.10) holds, then  $T$  and  $U$  are subspaces of  $V$ . They are called *orthogonal complements*,  $T = U^\perp$ ,  $U = T^\perp$ . If  $T$  is a subspace of  $S$ , then there always exists an orthogonal complement  $T^\perp$ , such that  $S = T \oplus T^\perp$ . Every vector  $\mathbf{a} \in S$  can then be uniquely decomposed into the form  $\mathbf{a} = \mathbf{t} + \mathbf{u}$  with  $\mathbf{t} \in T$  and  $\mathbf{u} \in T^\perp$ . For the norms of the vectors the relation  $a^2 = t^2 + u^2$  holds.

If  $T$  is an  $(n-1)$ -dimensional subspace of an  $n$ -dimensional vector space  $S$  and if  $\mathbf{s}$  is a fixed vector in  $S$ , then the set of all vectors  $\mathbf{h} = \mathbf{s} + \mathbf{t}$  with  $\mathbf{t} \in T$  forms an  $(n-1)$ -dimensional *hyperplane*  $H$  in  $S$ . If  $H$  is given and  $\mathbf{h}_0$  is an arbitrary fixed vector in  $H$ , then  $T$  is the set of all vectors  $\mathbf{t} = \mathbf{h} - \mathbf{h}_0$ ,  $\mathbf{h} \in H$ , as shown in Fig. A.2. If  $\hat{\mathbf{u}}$  is a unit vector in the one-dimensional subspace  $T = H^\perp$  orthogonal to  $H$ , then the scalar product

$$\hat{\mathbf{u}} \cdot \mathbf{h} = d \quad (\text{A.2.12})$$

has the same value for all  $\mathbf{h} \in H$ , where  $d$  is the distance of the hyperplane from the origin (see Fig. A.3).

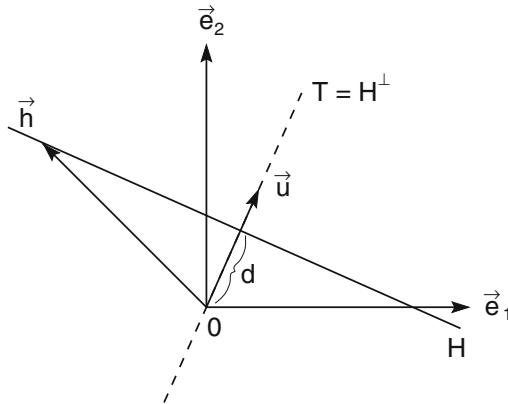


**Fig. A.2:** Hyperplane  $H$  in a two-dimensional vector space.

For a given  $\hat{\mathbf{u}}$  and  $d$ , Eq. (A.2.12) defines a hyperplane  $H$ . It divides the  $n$ -dimensional vector space into two *half spaces*, which consist of the set of vectors  $\mathbf{x}$  for which  $\hat{\mathbf{u}} \cdot \mathbf{x} < 0$  and  $\hat{\mathbf{u}} \cdot \mathbf{x} > 0$ .

### A.3 Orthogonal Transformations

According to (A.2.6), the mapping of an  $n$ -vector  $\mathbf{a}$  onto an  $n$ -vector  $\mathbf{b}$  is performed by multiplication by a square  $n \times n$  matrix,  $\mathbf{b} = Q\mathbf{a}$ . If the length of



**Fig.A.3:** Hyperplane  $H$  and complementary one-dimensional vector space  $T$ .

the vector (A.1.14) remains unchanged, one speaks of an *orthogonal transformation*. For such a case one has  $b = a$  or  $b^2 = a^2$ , i.e.,

$$\mathbf{b}^T \mathbf{b} = \mathbf{a}^T Q^T Q \mathbf{a} = \mathbf{a}^T \mathbf{a} \quad ,$$

and thus

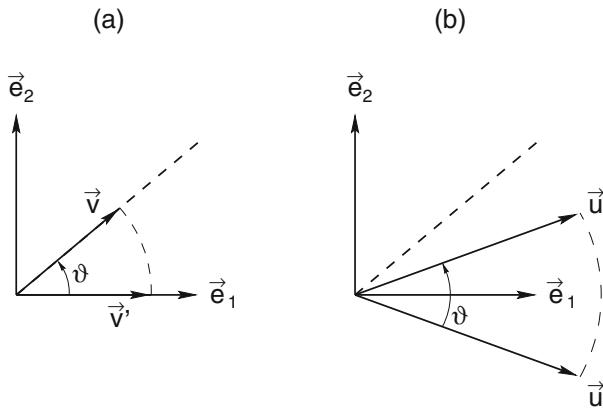
$$Q^T Q = I \quad . \quad (\text{A.3.1})$$

A square matrix  $Q$  that fulfills (A.3.1) is said to be *orthogonal*.

It is clear that transformations are orthogonal when the transformed vector  $\mathbf{b}$  is obtained from  $\mathbf{a}$  by means of a spatial rotation and/or reflection. We will examine some orthogonal transformations that are important for the applications of this appendix.

### A.3.1 Givens Transformation

The *Givens rotation* is a transformation that affects only components in a plane spanned by two orthogonal basis vectors. For simplicity we will first consider only two-dimensional vectors.



**Fig.A.4:** Application of the Givens transformation (a) to the vector  $v$  that defines the transformation and (b) to an arbitrary vector  $\mathbf{u}$ .

A vector  $\mathbf{v}$  can be represented as

$$\mathbf{v} = \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = v \begin{pmatrix} c \\ s \end{pmatrix}, \quad c = \cos \vartheta = v_1/v, \quad s = \sin \vartheta = v_2/v. \quad (\text{A.3.2})$$

A rotation by an angle  $-\vartheta$  transforms the vector  $\mathbf{v}$  into the vector  $\mathbf{v}' = G\mathbf{v}$ , whose second component vanishes, as in Fig. A.4a. Clearly one has

$$G = \begin{pmatrix} c & s \\ -s & c \end{pmatrix}, \quad \mathbf{v}' = G\mathbf{v} = \begin{pmatrix} v \\ 0 \end{pmatrix}. \quad (\text{A.3.3})$$

Of course the transformation thus defined with the vector  $\mathbf{v}$  can also be applied to any other vector  $\mathbf{u}$ , as shown in Fig. A.4b. In  $n$  dimensions the Givens rotation leads to the transformation

$$\mathbf{v} \rightarrow \mathbf{v}' = G\mathbf{v},$$

such that  $v'_k = 0$ . The components  $v'_\ell = v_\ell$ ,  $\ell \neq k$ ,  $\ell \neq i$ , remain unchanged and  $v'_i$  is determined such that the norm of the vector remains unchanged,  $v' = v$ . This is clearly given by

$$G = \begin{pmatrix} 1 & & & & & & & & \\ & \ddots & & & & & & & \\ & & 1 & & & & & & \\ & & & c & & s & & & \\ & & & & 1 & & & & \\ & & & & & \ddots & & & \\ & & & & & & 1 & & \\ & & & & & & & c & \\ & & & & & & & & 1 \\ & & & & & & & & & \ddots \\ & & & & & & & & & & 1 \\ & & & & & & & & & & & \ddots \\ & & & & & & & & & & & & 1 \end{pmatrix} \quad \begin{matrix} \leftarrow & 1 \\ & \vdots \\ \leftarrow & i \\ & \vdots \\ \leftarrow & k \\ & \vdots \\ \leftarrow & n \end{matrix}. \quad (\text{A.3.4})$$

In practical applications, however, the full matrix is not needed. The method `DatanMatrix.defineGivensTransformation` defines a Givens transformation by the input of the two components  $v_1, v_2$ , `DatanMatrix._ApplyGivensTransformation` applies that transformation to two components of another vector. The method `DatanMatrix.defineANDApplyGivensTransformation` defines a transformation and directly applies it to the defining vector.

### A.3.2 Householder Transformation

The Givens rotation is used to transform a vector in such a way that a given vector component vanishes. A more general transformation is the *Householder transformation*. If the original vector is

$$\mathbf{v} = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix}, \quad (\text{A.3.5})$$

then for the transformed vector we want to have

$$\mathbf{v}' = H\mathbf{v} = \begin{pmatrix} v_1 \\ \vdots \\ v_{p-1} \\ v'_p \\ v_{p+1} \\ \vdots \\ v_{\ell-1} \\ 0 \\ \vdots \\ 0 \end{pmatrix}. \quad (\text{A.3.6})$$

That is, the components  $v'_\ell, v'_{\ell+1}, \dots, v'_n$  should vanish. The remaining components should (with the exception of  $v'_p$ ) remain unchanged. The component  $v'_p$  must be changed in such a way that one has  $\mathbf{v} = \mathbf{v}'$ . From the Pythagorean theorem in  $n - \ell + 1$  dimensions one has

$$v'^2_p = v_H^2 = v_p^2 + \sum_{i=\ell}^n v_i^2$$

or

$$v'_p = -\sigma v_H = -\sigma \sqrt{v_p^2 + \sum_{i=\ell}^n v_i^2} \quad (\text{A.3.7})$$

with  $\sigma = \pm 1$ . We choose

$$\sigma = \text{sign}(v_p) \quad . \quad (\text{A.3.8})$$

We now construct the matrix  $H$  of (A.3.6). To do this we decompose the vector  $\mathbf{v}$  into a sum,

$$\mathbf{v} = \mathbf{v}_H + \mathbf{v}_{H^\perp} \quad , \quad \mathbf{v}_H = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ v_p \\ 0 \\ \vdots \\ 0 \\ v_\ell \\ \vdots \\ v_n \end{pmatrix} \quad , \quad \mathbf{v}_{H^\perp} = \begin{pmatrix} v_1 \\ \vdots \\ v_{p-1} \\ 0 \\ v_{p+1} \\ \vdots \\ v_{\ell-1} \\ 0 \\ \vdots \\ 0 \end{pmatrix} \quad . \quad (\text{A.3.9})$$

Here the vector  $\mathbf{v}_H$  is in the subspace spanned by the basis vectors  $\mathbf{e}_p, \mathbf{e}_\ell, \mathbf{e}_{\ell+1}, \dots, \mathbf{e}_n$ , and  $\mathbf{v}_{H^\perp}$  is in the subspace orthogonal to it. We now construct

$$\mathbf{u} = \mathbf{v}_H + \sigma v_H \mathbf{e}_p \quad (\text{A.3.10})$$

and

$$H = I_n - \frac{2\mathbf{u}\mathbf{u}^T}{u^2} \quad . \quad (\text{A.3.11})$$

If we now decompose an arbitrary vector  $\mathbf{a}$  into a sum of vectors parallel and perpendicular to  $\mathbf{u}$ ,

$$\mathbf{a} = \mathbf{a}_\parallel + \mathbf{a}_\perp \quad ,$$

with

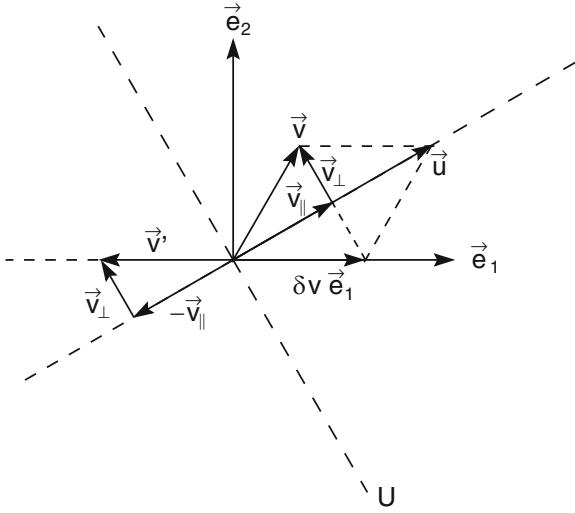
$$\mathbf{a}_\parallel = \frac{\mathbf{u}\mathbf{u}^T}{u^2} \mathbf{a} = \frac{\mathbf{u}}{u^2} (\mathbf{u} \cdot \mathbf{a}) = \hat{\mathbf{u}} (\hat{\mathbf{u}} \cdot \mathbf{a}) \quad , \quad \mathbf{a}_\perp = \mathbf{a} - \mathbf{a}_\parallel \quad ,$$

then one has

$$\mathbf{a}'_\parallel = H \mathbf{a}_\parallel = -\mathbf{a}_\parallel \quad , \quad \mathbf{a}'_\perp = H \mathbf{a}_\perp = \mathbf{a}_\perp \quad .$$

Thus we see that the transformation is a reflection in the subspace that is orthogonal to the vector  $\mathbf{u}$ , as in Fig. A.5. One can easily verify that  $H$  in fact yields the transformation (A.3.6).

The transformation is uniquely determined by the vector  $\mathbf{u}$ . According to (A.3.10) one has for the components of this vector  $u_p = v_p + \sigma v_H, u_\ell = v_\ell$ ,



**Fig.A.5:** The vector  $\mathbf{v}$  is mapped onto  $\mathbf{v}'$  according to a Householder transformation such that  $v'_2 = 0$ . The mapping corresponds to a reflection in the subspace  $U$  that is orthogonal to the auxiliary vector  $\mathbf{u}$ .

$u_{\ell+1} = v_{\ell+1}, \dots, u_n = v_n$ , and  $u_i = 0$  for all other  $i$ . If the vector  $\mathbf{v}$  and the indices  $p$  and  $\ell$  are given, then only  $u_p$  must be computed. The quantity  $u^2$  appearing in (A.3.11) is then

$$\begin{aligned} u^2 &= u_p^2 + \sum_{i=\ell}^n u_i^2 = (v_p + \sigma v_H)^2 + \sum_{i=\ell}^n v_i^2 \\ &= v_p^2 + \sum_{i=\ell}^n v_i^2 + v_H^2 + 2\sigma v_H v_p \\ &= 2v_H^2 + 2\sigma v_H v_p = 2v_H(v_H + \sigma v_p) = 2v_H u_p \quad . \end{aligned}$$

We can thus write (A.3.11) in the form

$$H = I_n - b\mathbf{u}\mathbf{u}^T \quad , \quad b = (v_H u_p)^{-1} \quad . \quad (\text{A.3.12})$$

The matrix  $H$ , however, is not needed explicitly to compute a transformed vector,

$$\mathbf{c}' = H\mathbf{c} \quad .$$

It is sufficient to know the vector  $\mathbf{u}$  and the constant  $b$ . Since, however,  $\mathbf{u}$  only differs from  $\mathbf{v}$  in the element  $u_p$  (and in the vanishing elements), it is sufficient, starting from  $\mathbf{v}$ , to first compute the quantities  $u_p$  and  $b$  and when applying the transformation to use in addition the elements  $v_\ell, v_{\ell+1}, \dots, v_n$ . These are at the same time the corresponding elements of  $\mathbf{u}$ .

By the method `DatanMatrix.defineHouseholderTransformation` a transformation is defined; with `DatanMatrix.applyHouseholderTransformation` it is applied to a vector.

### A.3.3 Sign Inversion

If the diagonal element  $I_{ii}$  of the unit matrix is replaced by  $-1$ , then one obtains a symmetric orthogonal matrix,

$$R^{(i)} = \begin{pmatrix} 1 & & & \\ & \ddots & & \\ & & -1 & \\ & & & \ddots \\ & & & & 1 \end{pmatrix} .$$

Applying this to the vector  $\mathbf{a}$ ,

$$\mathbf{a}' = R^{(i)}\mathbf{a} ,$$

changes the sign of the element  $a_i$  and leaves all the other elements unchanged. Clearly  $R^{(i)}$  is a Householder matrix which produces a reflection in the subspace orthogonal to the basis vector  $\mathbf{e}_i$ . This can be seen immediately by substituting  $\mathbf{u} = \mathbf{e}_i$  in (A.3.11).

### A.3.4 Permutation Transformation

The  $n \times n$  unit matrix (A.1.10) can easily be written as an arrangement of the basis vectors (A.2.3) in a square matrix,

$$I_n = (\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n) = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & & & \\ 0 & 0 & \dots & 1 \end{pmatrix} .$$

It is clearly an orthogonal matrix. The orthogonal transformation  $I\mathbf{a} = \mathbf{a}$  leaves the vector  $\mathbf{a}$  unchanged. If we now exchange two of the basis vectors  $\mathbf{e}_i$  and  $\mathbf{e}_k$ , then we obtain the symmetric orthogonal matrix  $P^{ik}$ . As an example we show this for  $n = 4$ ,  $i = 2$ ,  $k = 4$ ,

$$P^{(ik)} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} .$$

The transformation  $\mathbf{a}' = P^{(ik)}\mathbf{a}$  leads to an exchange of the elements  $a_i$  and  $a_k$ . All other elements remain unchanged:

$$\mathbf{a} = \begin{pmatrix} a_1 \\ \vdots \\ a_i \\ \vdots \\ a_k \\ \vdots \\ a_n \end{pmatrix}, \quad \mathbf{a}' = P^{(ik)} \mathbf{a} = \begin{pmatrix} a_1 \\ \vdots \\ a_k \\ \vdots \\ a_i \\ \vdots \\ a_n \end{pmatrix}.$$

Multiplication of an  $n \times m$  matrix  $A$  on the left with  $P^{(ik)}$  permutes the lines  $i$  and  $k$  of  $A$ . Multiplication of an  $m \times n$  matrix  $A$  from the right exchanges the columns  $i$  and  $k$ . If  $D$  is an  $n \times n$  diagonal matrix, then the elements  $D_{ii}$  and  $D_{kk}$  are exchanged by the operation

$$D' = P^{(ik)} D P^{(ik)}.$$

## A.4 Determinants

To every  $n \times n$  matrix  $A$  one can associate a number, its *determinant*,  $\det A$ . A determinant, like the corresponding matrix, is written as a square arrangement of the matrix elements, but is enclosed, however, by vertical lines. The determinants of orders two and three are defined by

$$\det A = \begin{vmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{vmatrix} = A_{11}A_{22} - A_{12}A_{21} \quad (\text{A.4.1})$$

and

$$\begin{aligned} \det A &= \begin{vmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{vmatrix} \\ &= + A_{11}A_{22}A_{33} - A_{11}A_{23}A_{32} \\ &\quad + A_{12}A_{23}A_{31} - A_{12}A_{21}A_{33} \\ &\quad + A_{13}A_{21}A_{32} - A_{13}A_{22}A_{31} \end{aligned} \quad (\text{A.4.2})$$

or, written in another way,

$$\begin{aligned} \det A &= - A_{11}(A_{22}A_{33} - A_{23}A_{32}) \\ &\quad + A_{12}(A_{21}A_{33} - A_{23}A_{31}) \\ &\quad + A_{13}(A_{21}A_{32} - A_{22}A_{31}). \end{aligned} \quad (\text{A.4.3})$$

A general determinant of order  $n$  is written in the form

$$\det A = \begin{vmatrix} A_{11} & A_{12} & \dots & A_{1n} \\ A_{21} & A_{22} & \dots & A_{2n} \\ \vdots & & & \\ A_{n1} & A_{n2} & \dots & A_{nn} \end{vmatrix}. \quad (\text{A.4.4})$$

The *cofactor*  $A_{ij}^\dagger$  of the element  $A_{ij}$  of a matrix is a determinant of order  $(n - 1)$ , calculated from the matrix obtained by deleting the  $i$ th row and  $j$ th column of the original matrix, and multiplying by  $(-1)^{i+j}$ ,

$$A_{ij}^\dagger = (-1)^{i+j} \begin{vmatrix} A_{11} & A_{12} & \dots & A_{1,j-1} & A_{1,j+1} & \dots & A_{1n} \\ A_{21} & A_{22} & \dots & A_{2,j-1} & A_{2,j+1} & \dots & A_{2n} \\ \vdots & & & & & & \\ A_{i-1,1} & A_{i-1,2} & \dots & A_{i-1,j-1} & A_{i-1,j+1} & \dots & A_{i-1,n} \\ A_{i+1,1} & A_{i+1,2} & \dots & A_{i+1,j-1} & A_{i+1,j+1} & \dots & A_{i+1,n} \\ \vdots & & & & & & \\ A_{n1} & A_{n2} & \dots & A_{n,j-1} & A_{n,j+1} & \dots & A_{nn} \end{vmatrix}. \quad (\text{A.4.5})$$

Determinants of higher order can be written as the sum of all elements of any row or column multiplied with the corresponding cofactors,

$$\det A = \sum_{k=1}^n A_{ik} A_{ik}^\dagger = \sum_{k=1}^n A_{kj} A_{kj}^\dagger. \quad (\text{A.4.6})$$

One can easily show that the result is independent of the choice of row  $i$  or column  $j$ . Equation (A.4.3) already shows that (A.4.6) is correct for  $n = 3$ . Determinants of arbitrary order can be computed by decomposing them according to their cofactors until one reaches, for example, the order two. A singular matrix, i.e., a square matrix whose rows or columns are not linearly independent, has *determinant zero*.

From  $A$  we can construct a further matrix by replacing each element  $ij$  by the cofactor of the element  $ji$ . In this way we obtain the *adjoint matrix* of  $A$ ,

$$A^\dagger = \begin{pmatrix} A_{11}^\dagger & A_{21}^\dagger & \dots & A_{n1}^\dagger \\ A_{12}^\dagger & A_{22}^\dagger & \dots & A_{n2}^\dagger \\ \vdots & & & \\ A_{1n}^\dagger & A_{2n}^\dagger & \dots & A_{nn}^\dagger \end{pmatrix}. \quad (\text{A.4.7})$$

For determinants the following rules hold:

$$\det A = \det A^T, \quad (\text{A.4.8})$$

$$\det AB = \det A \det B. \quad (\text{A.4.9})$$

For an orthogonal matrix  $Q$  one has  $QQ^T = I$ , i.e.,

$$\det I = 1 = \det Q \det Q$$

and thus

$$\det Q = \pm 1 \quad . \quad (\text{A.4.10})$$

## A.5 Matrix Equations: Least Squares

A system of  $m$  linear equations with  $n$  unknowns  $x_1, x_2, \dots, x_n$  has the general form

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n - b_1 &= 0 \quad , \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n - b_2 &= 0 \quad , \\ \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n - b_m &= 0 \end{aligned} \quad (\text{A.5.1})$$

or in matrix notation

$$Ax - \mathbf{b} = 0 \quad . \quad (\text{A.5.2})$$

In finding the solution to this equation  $\mathbf{x}$  we must distinguish between various cases, which can be characterized by the values of  $m$ ,  $n$ , and

$$k = \text{Rang}(A) \quad .$$

The vector  $A\mathbf{x}$  is in the column space of  $A$ , which is of dimension  $k$ . Since  $\mathbf{b}$  is an  $m$ -vector, the equation can in general only be fulfilled if  $k = m$ , i.e., for  $k = n = m$  and for  $k = m < n$ , since  $k \leq \min(m, n)$ . For  $k = n = m$  one has  $n$  independent equations (A.5.1) with  $n$  unknowns, which have a *unique* solution. If  $k = m < n$ , then there are arbitrarily many  $n$ -vectors  $\mathbf{x}$  that can be mapped on to the  $m$ -vector  $A\mathbf{x} = \mathbf{b}$  such that (A.5.2) is fulfilled. The system of equations is *underdetermined*. The solution is not unique.

For  $k = \text{Rang}(A) \neq m$  there is, in general, no solution of (A.5.2). In this case we look for a vector  $\tilde{\mathbf{x}}$ , for which the left-hand side of (A.5.2) is a vector of minimum Euclidian norm. That is, we replace Eq. (A.5.2) by

$$r^2 = (Ax - \mathbf{b})^2 = \min \quad , \quad (\text{A.5.3})$$

i.e., we look for a vector  $\tilde{\mathbf{x}}$  for which the mapping  $A\tilde{\mathbf{x}}$  differs as little as possible from  $\mathbf{b}$ .

Given a  $m$ -vector  $\mathbf{c}$ , only for  $k = \text{Rang}(A) = n$  does there exist only one  $n$ -vector  $\mathbf{x}$ , such that  $A\mathbf{x} = \mathbf{c}$ . Therefore, only for the case  $k = n$  is there a unique solution  $\tilde{\mathbf{x}}$ . Thus there exists for  $\text{Rang}(A) = n$  and  $m \geq n$  a *unique solution*  $\tilde{\mathbf{x}}$  of (A.5.3). For  $n = m$  one has  $r = 0$ .

The relation (A.5.3) is also often written simply in the form

$$A\mathbf{x} - \mathbf{b} \approx 0 \quad (\text{A.5.4})$$

or even, not entirely correctly, in the form (A.5.2). One calls the solution vector  $\tilde{\mathbf{x}}$  the *least-squares solution of (A.5.4)*. In Table A.1 we list again the various cases that result from different relationships between  $m$ ,  $n$ , and  $k$ .

**Table A.1:** Behavior of the solutions of (A.5.3) for various cases.  $m$  is the row number,  $n$  is the column number and  $k$  is the rank of the matrix  $A$ .

| Case |         | Rang( $A$ ) | Residual   | Solution unique |
|------|---------|-------------|------------|-----------------|
| 1a   | $m = n$ | $k = n$     | $r = 0$    | Yes             |
| 1b   |         | $k < n$     | $r \geq 0$ | No              |
| 2a   | $m > n$ | $k = n$     | $r \geq 0$ | Yes             |
| 2b   |         | $k < n$     | $r \geq 0$ | No              |
| 3a   | $m < n$ | $k = m$     | $r = 0$    | No              |
| 3b   |         | $k < m$     | $r \geq 0$ | No              |

We now want to state more formally what we have determined in this section with respect to the solution of (A.5.4).

### Theorem on the orthogonal decomposition of a matrix:

Every  $m \times n$  matrix  $A$  of rank  $k$  can be written in the form

$$A = H R K^T \quad , \quad (\text{A.5.5})$$

where  $H$  is an  $m \times m$  orthogonal matrix,  $K$  is an  $n \times n$  orthogonal matrix, and  $R$  is an  $m \times n$  matrix of the form

$$R = \begin{pmatrix} R_{11} & 0 \\ 0 & 0 \end{pmatrix} \quad (\text{A.5.6})$$

and where  $R_{11}$  is a  $k \times k$  matrix of rank  $k$ .

Substituting (A.5.5) into (A.5.4) and multiplying from the left by  $H^T$  leads to

$$R K^T \mathbf{x} \approx H^T \mathbf{b} \quad . \quad (\text{A.5.7})$$

We define

$$H^T \mathbf{b} = \mathbf{g} = \begin{pmatrix} \mathbf{g}_1 \\ \mathbf{g}_2 \end{pmatrix} \quad \begin{matrix} \{k \\ \}_{m-k} \end{matrix} \quad (\text{A.5.8})$$

and

$$K^T \mathbf{x} = \mathbf{p} = \begin{pmatrix} \mathbf{p}_1 \\ \mathbf{p}_2 \end{pmatrix} \quad \begin{matrix} \{k \\ \}_{n-k} \end{matrix} \quad , \quad (\text{A.5.9})$$

so that (A.5.7) takes on the form

$$R\mathbf{p} \approx \mathbf{g} . \quad (\text{A.5.10})$$

Because of (A.5.6), this breaks into two independent relations

$$R_{11}\mathbf{p}_1 = \mathbf{g}_1 \quad (\text{A.5.11})$$

and

$$0 \cdot \mathbf{p}_2 \approx \mathbf{g}_2 . \quad (\text{A.5.12})$$

If  $m = k$  and/or  $n = k$ , then the corresponding lower partial vectors are absent in (A.5.8) and/or (A.5.9) and the corresponding lower matrices are absent in (A.5.6). Since in (A.5.11)  $R_{11}$  is a  $k \times k$  matrix of rank  $k$ , and  $\mathbf{p}_1$  and  $\mathbf{g}_1$  are  $k$ -vectors, there exists a solution vector  $\tilde{\mathbf{p}}_1$  for which equality holds. Because of the null matrix on the left-hand side of (A.5.12), we cannot derive any information about  $\mathbf{p}_2$  from this relation.

**Theorem on the solutions of  $\mathbf{Ax} \approx \mathbf{b}$ :** If  $\tilde{\mathbf{p}}_1$  is the unique solution vector of (A.5.11), then the following statements hold:

- (i) All solutions of (A.5.3) have the form

$$\hat{\mathbf{x}} = K \begin{pmatrix} \tilde{\mathbf{p}}_1 \\ \tilde{\mathbf{p}}_2 \end{pmatrix} . \quad (\text{A.5.13})$$

Here  $\tilde{\mathbf{p}}_2$  is an arbitrary  $(n - k)$ -vector, i.e., the solution is unique for  $k = n$ . There is always, however, a *unique solution of minimum absolute value*:

$$\tilde{\mathbf{x}} = K \begin{pmatrix} \tilde{\mathbf{p}}_1 \\ 0 \end{pmatrix} . \quad (\text{A.5.14})$$

- (ii) All solutions  $\hat{\mathbf{x}}$  have the same residual vector

$$\mathbf{r} = \mathbf{b} - A\hat{\mathbf{x}} = H \begin{pmatrix} 0 \\ \mathbf{g}_2 \end{pmatrix} \quad (\text{A.5.15})$$

with the absolute value  $r = g_2 = |\mathbf{g}_2|$ . The residual vanishes for  $k = m$ .

The problem  $\mathbf{Ax} \approx \mathbf{b}$  should always be handled with orthogonal decompositions, preferably with the singular value decomposition and singular value analysis described in Sects. A.12 and A.13. Numerically the results are at least as accurate as with other methods, and are often more accurate (cf. Sect. A.13, Example A.4).

Nevertheless we will briefly present as well the method of *normal equations*. The method is very transparent compared to the orthogonal decomposition and is therefore always described in textbooks.

We consider the square (A.5.3) of the residual vector

$$\begin{aligned} r^2 &= (\mathbf{Ax} - \mathbf{b})^T(\mathbf{Ax} - \mathbf{b}) = \mathbf{x}^T A^T A \mathbf{x} - 2\mathbf{b}^T A \mathbf{x} + \mathbf{b}^T \mathbf{b} \\ &= \sum_{i=1}^m \sum_{j=1}^n \sum_{\ell=1}^n A_{ij} A_{i\ell} x_j x_\ell - 2 \sum_{i=1}^m \sum_{j=1}^n b_i A_{ij} x_j + \sum_{i=1}^m b_i^2 . \end{aligned}$$

The requirement  $r^2 = \min$  leads to

$$\frac{\partial r^2}{\partial x_k} = 2 \sum_{i=1}^m \sum_{\ell=1}^n A_{ik} A_{i\ell} x_\ell - 2 \sum_{i=1}^m b_i A_{ik} = 0 , \quad k = 1, \dots, n .$$

These  $n$  linear equations are called normal equations. They can be arranged in a matrix,

$$A^T A \mathbf{x} = A^T \mathbf{b} . \quad (\text{A.5.16})$$

This is a system of  $n$  equations with  $n$  unknowns. If the equations are linearly independent, then the  $n \times n$  matrix  $(A^T A)$  is of full rank  $n$ ; it is not singular. According to Sect. A.6 there then exists an inverse  $(A^T A)^{-1}$ , such that  $(A^T A)^{-1}(A^T A) = I$ . Thus the desired solution to (A.5.3) is

$$\tilde{\mathbf{x}} = (A^T A)^{-1} A^T \mathbf{b} . \quad (\text{A.5.17})$$

This simple prescription is, however, useless if  $A^T A$  is singular or nearly singular (cf. Example A.4).

## A.6 Inverse Matrix

For every non-singular  $n \times n$  matrix  $A$ , the *inverse matrix*  $A^{-1}$  is defined by

$$A A^{-1} = I_n = A^{-1} A . \quad (\text{A.6.1})$$

It is also an  $n \times n$  matrix.

If  $A^{-1}$  is known, then the solution of the matrix equation

$$A \mathbf{x} = \mathbf{b} \quad (\text{A.6.2})$$

is given simply by

$$\mathbf{x} = A^{-1} \mathbf{b} . \quad (\text{A.6.3})$$

As we will show,  $A^{-1}$  only exists for non-singular square matrices, so that (A.6.3) only gives the solution for the case 1a of Table A.1.

In order to determine  $A^{-1}$ , we set  $A^{-1} = X$  and express the column vectors of  $X$  as  $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ . Equation (A.6.1) is then decomposed into  $n$  equations:

$$A\mathbf{x}_i = \mathbf{e}_i \quad , \quad i = 1, 2, \dots, n \quad . \quad (\text{A.6.4})$$

The right-hand sides are the basis vectors (A.2.3). For the case  $n = 2$  we can write the system (A.6.4) in various equivalent ways, e.g.,

$$\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \begin{pmatrix} X_{11} \\ X_{21} \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} , \quad \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \begin{pmatrix} X_{12} \\ X_{22} \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix} , \quad (\text{A.6.5})$$

or alternatively

$$\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \begin{pmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} ,$$

or as a system of four equations with four unknowns,

$$\begin{aligned} A_{11}X_{11} + A_{12}X_{21} &= 1 & , \\ A_{21}X_{11} + A_{22}X_{21} &= 0 & , \\ A_{11}X_{12} + A_{12}X_{22} &= 0 & , \\ A_{21}X_{12} + A_{22}X_{22} &= 1 & . \end{aligned} \quad (\text{A.6.6})$$

By elimination and substitution one easily finds

$$\begin{aligned} X_{11} &= \frac{A_{22}}{A_{11}A_{22} - A_{12}A_{21}} & , \\ X_{12} &= \frac{-A_{12}}{A_{11}A_{22} - A_{12}A_{21}} & , \\ X_{21} &= \frac{-A_{21}}{A_{11}A_{22} - A_{12}A_{21}} & , \\ X_{22} &= \frac{A_{11}}{A_{11}A_{22} - A_{12}A_{21}} & \end{aligned} \quad (\text{A.6.7})$$

or in matrix notation

$$X = A^{-1} = \frac{1}{\det A} \begin{pmatrix} A_{22} & -A_{12} \\ -A_{21} & A_{11} \end{pmatrix} . \quad (\text{A.6.8})$$

The matrix on the right-hand side is the adjoint of the original matrix, i.e.,

$$A^{-1} = \frac{A^\dagger}{\det A} . \quad (\text{A.6.9})$$

One can show that this relation holds for square matrices of arbitrary order. From (A.6.9) it is clear that the inverse of a singular matrix, i.e., of a matrix with vanishing determinant, is not determined.

In practice the inverse matrix is not computed using (A.6.9) but rather as in our example of the  $2 \times 2$  matrix by elimination and substitution from the system (A.6.4). This system consists of  $n$  sets of equations of the form

$$Ax = b \quad , \quad (\text{A.6.10})$$

each consisting of  $n$  equations with  $n$  unknowns. Here  $A$  is a non-singular  $n \times n$  matrix. We will first give the solution algorithm for square non-singular matrices  $A$ , but we will later remove these restrictions.

## A.7 Gaussian Elimination

We will write Eq. (A.6.10) for an  $n \times n$  matrix  $A$  in components,

$$\begin{aligned} A_{11}x_1 + A_{12}x_2 + \cdots + A_{1n}x_n &= b_1 \quad , \\ A_{21}x_1 + A_{22}x_2 + \cdots + A_{2n}x_n &= b_2 \quad , \\ \vdots \\ A_{n1}x_1 + A_{n2}x_2 + \cdots + A_{nn}x_n &= b_n \quad , \end{aligned} \quad (\text{A.7.1})$$

and we will solve the system by *Gaussian elimination*. For this we define  $n - 1$  multipliers

$$m_{i1} = \frac{A_{i1}}{A_{11}} \quad , \quad i = 2, 3, \dots, n \quad , \quad (\text{A.7.2})$$

multiply the first equation by  $m_{21}$ , and subtract it from the second. We then multiply the first equation by  $m_{31}$  and subtract it from the third, and so forth. We obtain the system

$$\begin{aligned} A_{11}^{(1)}x_1 + A_{12}^{(1)}x_2 + \cdots + A_{1n}^{(1)}x_n &= b_1^{(1)} \quad , \\ A_{22}^{(2)}x_2 + \cdots + A_{2n}^{(2)}x_n &= b_2^{(2)} \quad , \\ \vdots \\ A_{n2}^{(2)}x_2 + \cdots + A_{nn}^{(2)}x_n &= b_n^{(2)} \quad , \end{aligned} \quad (\text{A.7.3})$$

where the unknown  $x_1$  has disappeared from all the equations except the first. The coefficients  $A_{ij}^{(2)}, b_i^{(2)}$  are given by the equations

$$\begin{aligned} A_{ij}^{(2)} &= A_{ij}^{(1)} - m_{i1}A_{1j}^{(1)} \quad , \\ b_i^{(2)} &= b_i^{(1)} - m_{i1}b_1^{(1)} \quad . \end{aligned}$$

The procedure is then repeated with the last  $n - 1$  equations by defining

$$m_{i2} = \frac{A_{i2}^{(2)}}{A_{22}^{(2)}} \quad , \quad i = 3, 4, \dots, n \quad ,$$

multiplying the second equation of the system (A.7.3) with the corresponding  $m_{i2}$ , and then subtracting it from the third, fourth,  $\dots$ ,  $n$ th equation. In the  $k$ th step of the procedure, the multipliers

$$m_{ik} = \frac{A_{ik}^{(k)}}{A_{kk}^{(k)}} \quad , \quad i = k+1, k+2, \dots, n \quad , \quad (\text{A.7.4})$$

are used and the new coefficients

$$\begin{aligned} A_{ij}^{(k+1)} &= A_{ij}^{(k)} - m_{ik} A_{kj}^{(k)} \quad , \\ b_i^{(k+1)} &= b_i^{(k)} - m_{ik} b_k^{(k)} \end{aligned} \quad (\text{A.7.5})$$

are computed. After  $n - 1$  steps we have produced the following *triangular system of equations*:

$$\begin{array}{lcl} A_{11}^{(1)}x_1 + A_{12}^{(1)}x_2 + \cdots + A_{1n}^{(1)}x_n & = & b_1^{(1)} \quad , \\ A_{22}^{(2)}x_2 + \cdots + A_{2n}^{(2)}x_n & = & b_2^{(2)} \quad , \\ \vdots & & \vdots \\ A_{nn}^{(n)}x_n & = & b_n^{(n)} \quad . \end{array} \quad (\text{A.7.6})$$

The last equation contains only  $x_n$ . By substituting into the next higher equation one obtains  $x_{n-1}$ , i.e., in general

$$x_i = \frac{1}{A_{ii}^{(i)}} \left\{ b_i^{(i)} - \sum_{\ell=i+1}^n A_{i\ell}^{(i)} x_\ell \right\} \quad , \quad i = n, n-1, \dots, 1 \quad . \quad (\text{A.7.7})$$

One should note that the change of the elements of the matrix  $A$  does not depend on the right-hand side  $\mathbf{b}$ . Therefore one can reduce several systems of equations with different right-hand sides simultaneously. That is, instead of  $\mathbf{Ax} = \mathbf{b}$ , one can solve the more general system

$$\mathbf{AX} = \mathbf{B} \quad (\text{A.7.8})$$

at the same time, where  $B$  and  $X$  are  $n \times m$  matrices, the matrix  $X$  being unknown.

**Example A.1:** Inversion of a  $3 \times 3$  matrix

As a numerical example let us consider the inversion of a  $3 \times 3$  matrix, i.e.,  $B = I$ . The individual computations for the example

$$\begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & -2 \\ 1 & 1 & 2 \end{pmatrix} X = I$$

are shown in Table A.2. The result is

$$X = \frac{1}{5} \begin{pmatrix} -4 & 1 & 7 \\ 6 & 1 & -8 \\ -1 & -1 & 3 \end{pmatrix} . \blacksquare$$

Following the individual steps of the calculation one sees that division is carried out in two places, namely in Eqs. (A.7.4) and (A.7.7). The denominator is in both cases a coefficient

$$A_{ii}^{(i)} , \quad i = 1, 2, \dots, n-1 ,$$

i.e., the upper left-hand coefficient of the system, the so-called *pivot* for the step  $i-1$  of the reduction process. Our procedure fails if this coefficient is equal to zero. In such a case one can simply exchange the  $i$ th line of the system with some other lower line whose first coefficient is not zero. The system of equations itself is clearly not changed by exchanging two equations. The procedure still fails if all of the coefficients of a column vanish. In this case the matrix  $A$  is singular, and there is no solution. In practice (at least when using computers where the extra work is negligible) it is advantageous to always carry out a reshuffling so that the pivot is the coefficient with the largest absolute value among those in the first column of the reduced system. One then always has the largest possible denominator. In this way rounding errors are kept as small as possible. The procedure just described is called *Gaussian elimination with pivoting*.

Following it, the method `DatanMatrix.matrixEquation` solves Eq. (A.7.8). In the same way the method `DatanMatrix.inverse` determines the inverse of a square nonsingular matrix.

## A.8 LR-Decomposition

For the elements  $A_{ij}^{(k)}$  transformed by Gaussian elimination [cf. (A.7.6)], one has on and above the main diagonal

$$A_{ij}^{(n)} = A_{ij}^{(n-1)} = \dots = A_{ij}^{(i)} , \quad i \leq j , \quad (\text{A.8.1})$$

**Table A.2:** Application of Gaussian elimination to Example A.1.  
Reduction

|        | Matrix A |    |               | Matrix B       |                |   | Multiplier    |
|--------|----------|----|---------------|----------------|----------------|---|---------------|
| Step 0 | 1        | 2  | 3             | 1              | 0              | 0 | —             |
|        | 2        | 1  | -2            | 0              | 1              | 0 | 2             |
|        | 1        | 1  | 2             | 0              | 0              | 1 | 1             |
| Step 1 |          | -3 | -8            | -2             | 1              | 0 | —             |
|        |          | -1 | -1            | -1             | 0              | 1 | $\frac{1}{3}$ |
| Step 2 |          |    | $\frac{5}{3}$ | $-\frac{1}{3}$ | $-\frac{1}{3}$ | 1 | —             |

### Substitution

|          | $j = 1$  |
|----------|--|
| $x_{3j}$ | $-\frac{1}{5}$   |
| $x_{2j}$ | $-\frac{1}{3}(-2 - 8 \times \frac{1}{5}) = \frac{6}{5}$          |
| $x_{1j}$ | $1 - 2 \times \frac{6}{5} + 3 \times \frac{1}{5} = -\frac{4}{5}$ |
|          | $j = 2$  |
| $x_{3j}$ | $-\frac{1}{5}$   |
| $x_{2j}$ | $-\frac{1}{3}(1 - \frac{8}{5}) = \frac{1}{5}$                    |
| $x_{1j}$ | $-2 \times \frac{1}{5} + 3 \times \frac{1}{5} = \frac{1}{5}$     |
|          | $j = 3$  |
| $x_{3j}$ | $\frac{3}{5}$  |
| $x_{2j}$ | $-\frac{1}{3}(0 + 8 \times \frac{3}{5}) = -\frac{8}{5}$          |
| $x_{1j}$ | $2 \times \frac{8}{5} - 3 \times \frac{3}{5} = \frac{7}{5}$      |

and below the main diagonal

$$A_{ij}^{(n)} = A_{ij}^{(n-1)} = \cdots = A_{ij}^{(j+1)} = 0 \quad , \quad i > j \quad . \quad (\text{A.8.2})$$

The transformation (A.7.5) is thus only carried out for  $k = 1, 2, \dots, r$  with  $r = \min(i - 1, j)$ ,

$$A_{ij}^{(k+1)} = A_{ij}^{(k)} - m_{ik} A_{kj}^{(k)} \quad . \quad (\text{A.8.3})$$

Summation over  $k$  gives

$$\sum_{k=1}^r A_{ij}^{(k+1)} - \sum_{k=1}^r A_{ij}^{(k)} = A_{ij}^{(r+1)} - A_{ij} = - \sum_{k=1}^r m_{ik} A_{kj}^{(k)} \quad .$$

Thus the elements  $A_{ij}^{(1)} = A_{ij}$  of the original matrix  $A$  can be written in the form

$$\begin{aligned} A_{ij} &= A_{ij}^{(i)} + \sum_{k=1}^{i-1} m_{ik} A_{kj}^{(k)} \quad , \quad i \leq j \quad , \\ A_{ij} &= 0 + \sum_{k=1}^j m_{ik} A_{kj}^{(k)} \quad , \quad i > j \quad . \end{aligned} \quad (\text{A.8.4})$$

Noting that the multipliers  $m_{ik}$  have only been defined up to now for  $i > k$ , we can in addition set

$$m_{ii} = 1 \quad , \quad i = 1, 2, \dots, n \quad ,$$

and reduce the two relations (A.8.4) to one,

$$A_{ij} = \sum_{k=1}^p m_{ik} A_{kj}^{(k)} \quad , \quad 1 \leq i, j \leq n \quad , \quad p = \min(i, j) \quad .$$

The equation shows directly that the matrix  $A$  can be represented as the product of two matrices  $L$  and  $R$ . Here  $L$  is a lower and  $R$  an upper triangular matrix,

$$\begin{aligned} A &= L R \quad , \quad (\text{A.8.5}) \\ L &= \begin{pmatrix} m_{11} & & & \\ m_{21} & m_{22} & & \\ \vdots & & & \\ m_{n1} & m_{n2} & \dots & m_{nn} \end{pmatrix} \quad , \quad m_{ii} = 1 \quad , \\ R &= \begin{pmatrix} r_{11} & r_{12} & \dots & r_{1n} \\ r_{22} & \dots & r_{2n} \\ \vdots & & & \\ r_{nn} & & & \end{pmatrix} = \begin{pmatrix} A_{11}^{(1)} & A_{12}^{(1)} & \dots & A_{1n}^{(1)} \\ A_{22}^{(2)} & \dots & A_{2n}^{(2)} \\ \vdots & & & \\ A_{nn}^{(n)} & & & \end{pmatrix} \quad . \end{aligned}$$

The original system of equations (A.6.10)

$$A\mathbf{x} = L R \mathbf{x} = \mathbf{b}$$

is thus equivalent to two triangular systems of equations,

$$L\mathbf{y} = \mathbf{b} \quad , \quad R\mathbf{x} = \mathbf{y} \quad .$$

Instead of resorting to the formulas from Sect. A.7, we can compute the elements of  $L$  and  $R$  directly from (A.8.5). We obtain

$$\left. \begin{aligned} r_{kj} &= A_{kj} - \sum_{\ell=1}^{k-1} m_{k\ell} r_{\ell j}, \quad j = k, k+1, \dots, n \\ m_{ik} &= \left( A_{ik} - \sum_{\ell=1}^{k-1} m_{i\ell} r_{\ell k} \right) / r_{kk}, \quad i = k+1, \dots, n \end{aligned} \right\} k = 1, \dots, n \quad , \quad (\text{A.8.6})$$

i.e., for  $k = 1$  one computes the first row of  $R$  (which is equal to the first row of  $A$ ) and the first column of  $L$ ; for  $k = 2$  one computes the second row of  $R$  and the second column of  $L$ . In a computer program the elements of  $R$  and  $L$  can overwrite the original matrix  $A$  with the same indices, since when computing  $r_{kj}$  one only needs the element  $A_{kj}$  and other elements of  $R$  and  $L$  that have already been computed according to (A.8.6). The corresponding consideration holds for the computation of  $m_{ik}$ .

The algorithm (A.8.6) is called the Doolittle *LR-decomposition*. By including pivoting this is clearly equivalent to Gaussian elimination.

## A.9 Cholesky Decomposition

If  $A$  is a real symmetric positive-definite matrix (cf. Sect. A.11), then it can be uniquely expressed as

$$A = U^T U \quad . \quad (\text{A.9.1})$$

Here  $U$  is a real upper triangular matrix with positive diagonal elements. The  $n \times n$  matrix  $A$  has the required property for the validity of (A.9.1), in particular in those cases where it is equal to the product of an arbitrary real  $n \times m$  matrix  $B$  (with  $m \geq n$  and full rank) with its transpose  $B^T$ ,  $A = B B^T$ .

To determine  $U$  we first carry out the Doolittle LR-decomposition, define a diagonal matrix  $D$  whose diagonal elements are equal to those of  $R$ ,

$$D = \text{diag}(r_{11}, r_{22}, \dots, r_{nn}) \quad , \quad (\text{A.9.2})$$

and introduce the matrices

$$D^{-1} = \text{diag}(r_{11}^{-1}, r_{22}^{-1}, \dots, r_{nn}^{-1}) \quad (\text{A.9.3})$$

and

$$D^{-1/2} = \text{diag}(r_{11}^{-1/2}, r_{22}^{-1/2}, \dots, r_{nn}^{-1/2}) \quad . \quad (\text{A.9.4})$$

We can then write

$$A = L R = L D D^{-1} R = L D R' \quad , \quad R' = D^{-1} R \quad . \quad (\text{A.9.5})$$

Because of the assumed symmetry one has

$$A = A^T = (R')^T D L^T \quad . \quad (\text{A.9.6})$$

Comparing with (A.9.5) gives  $L^T = R'$ , i.e.,

$$L = R^T D^{-1} \quad .$$

With

$$U = D^{-1/2} R = D^{1/2} L^T \quad (\text{A.9.7})$$

Eq. (A.9.1) is indeed fulfilled. The relation (A.9.7) means for the elements  $u_{ij}$  of  $U$

$$u_{ij} = r_{ii}^{-1/2} r_{ij} = r_{ii}^{1/2} m_{ji} \quad .$$

Thus the values  $r$  and  $m$  can be eliminated in favor of  $u$  from (A.8.6), and we obtain with

$$\left. \begin{aligned} u_{kk} &= \left( A_{kk} - \sum_{\ell=1}^{k-1} u_{\ell k}^2 \right)^{1/2} \\ u_{kj} &= \left( A_{kj} - \sum_{\ell=1}^{k-1} u_{\ell k} u_{\ell j} \right) / u_{kk}, \quad j = k+1, \dots, n \end{aligned} \right\} k = 1, \dots, n$$

the algorithm for the *Cholesky decomposition* of a real positive-definite symmetric matrix  $A$ . Since for such matrices all of the diagonal elements are different from zero, one does not need in principle any pivoting. One can also show that it would bring no advantage in numerical accuracy. The method `DatanMatrix.choleskyDecomposition` performs the Cholesky decomposition of a symmetric positive definite square matrix.

Positive-definite symmetric matrices play an important role as weight and covariance matrices. In such cases one often requires the Cholesky decomposition  $A = U^T U$  as well as multiplication of  $U$  by a matrix. It is easier computationally if this multiplication is done by the method `DatanMatrix.choleskyMultiply`, which takes the triangular form of  $U$  into account.

Of particular interest is the inversion of a symmetric positive-definite  $n \times n$  matrix  $A$ . For this we will solve the  $n$  matrix equations (A.6.4) with the previously described Cholesky decomposition of  $A$ ,

$$A \mathbf{x}_i = U^T U \mathbf{x}_i = U^T \mathbf{y}_i = \mathbf{e}_i \quad , \quad i = 1, \dots, n \quad . \quad (\text{A.9.8})$$

We denote the  $\ell$ th component of the three vectors  $\mathbf{x}_i$ ,  $\mathbf{y}_i$ , and  $\mathbf{e}_i$  by  $x_{i\ell}$ ,  $y_{i\ell}$ , and  $e_{i\ell}$ . Clearly one has  $x_{i\ell} = (A^{-1})_{i\ell}$  and  $e_{i\ell} = \delta_{i\ell}$ , since  $e_{i\ell}$  is the element  $(i, \ell)$  of the unit matrix.

We now determine  $y_{i\ell}$  by means of *forward substitution*. From (A.9.8) it follows directly that

$$\sum_{k=1}^{\ell} U_{k\ell} y_{ik} = e_{i\ell} = \delta_{i\ell} \quad ,$$

i.e.,

$$\begin{aligned} U_{11}y_{i1} &= \delta_{i1} \quad , \\ U_{12}y_{i1} + U_{22}y_{i2} &= \delta_{i2} \quad , \\ &\vdots \end{aligned}$$

and thus

$$\begin{aligned} y_{i1} &= \delta_{i1}/U_{11} \quad , \\ &\vdots \\ y_{i\ell} &= \frac{1}{U_{\ell\ell}} \left( \delta_{i\ell} - \sum_{k=1}^{\ell-1} U_{k\ell} y_{ik} \right) \quad . \end{aligned}$$

Since, however,  $\delta_{i\ell} = 0$  for  $i \neq \ell$ , this expression simplifies to

$$\begin{aligned} y_{i\ell} &= 0 \quad , \quad \ell < i \quad , \\ y_{i\ell} &= \frac{1}{U_{\ell\ell}} \left( \delta_{i\ell} - \sum_{k=i}^{\ell-1} U_{k\ell} y_{ik} \right) \quad , \quad \ell \geq i \quad . \end{aligned}$$

We can now obtain  $x_{i\ell}$  by means of *backward substitution* of  $y_{i\ell}$  into

$$U \mathbf{x}_i = \mathbf{y}_i$$

or, written in components,

$$\sum_{k=\ell}^n U_{\ell k} x_{ik} = y_{i\ell}$$

or

$$\begin{aligned} U_{11}x_{i1} + U_{12}x_{i2} + \cdots + U_{1n}x_{in} &= y_{i1} \quad , \\ U_{22}x_{i2} + \cdots + U_{2n}x_{in} &= y_{i2} \quad , \\ &\vdots \\ U_{nn}x_{in} &= y_{in} \quad . \end{aligned}$$

We obtain

$$\begin{aligned} x_{in} &= y_{in}/U_{nn} \quad , \\ &\vdots \\ x_{i\ell} &= \frac{1}{U_{\ell\ell}} \left( y_{i\ell} - \sum_{k=\ell+1}^n U_{\ell k} x_{ik} \right) \quad . \end{aligned}$$

If we compute  $x_{i\ell}$  only for  $\ell \geq i$ , then by backward substitution one only encounters the elements  $y_{i\ell}$  for  $\ell \geq i$ , i.e., only non-vanishing  $y_{i\ell}$ . The vanishing elements  $y_{i\ell}$  thus do not need to be stored. The elements  $x_{i\ell}$  for  $\ell < i$  follow simply from the symmetry of the original matrix,  $x_{i\ell} = x_{\ell i}$ .

The method `DatanMatrix.choleskyInversion` first performs the Cholesky decomposition of  $A$ . Then a loop is carried out over the various right-hand sides  $e_i$ ,  $i = 1, \dots, n$ , of (A.9.8). As the result of running through the loop once one obtains the elements  $x_{in}$ ,  $x_{i,n-1}$ , ...,  $x_{ii}$  of row  $i$ . They are stored as the corresponding elements of the output matrix  $A$ . The elements of the vector  $\mathbf{y}$ , which is only used for intermediate results, can be stored in the last row of  $A$ . Finally the elements below the main diagonal are filled by copies of their mirror images.

## A.10 Pseudo-inverse Matrix

We now return to the problem of Sect. A.5, that of solving the Eq. (A.5.4)

$$A\mathbf{x} \approx \mathbf{b} \tag{A.10.1}$$

for an arbitrary  $m \times n$  matrix  $A$  of rank  $k$ . According to (A.5.14), the unique solution of minimum norm is

$$\tilde{\mathbf{x}} = K \begin{pmatrix} \mathbf{p}_1 \\ 0 \end{pmatrix} \quad .$$

The vector  $\mathbf{p}_1$  is the solution of Eq. (A.5.11) and therefore

$$\tilde{\mathbf{p}}_1 = R_{11}^{-1} \mathbf{g}_1 \quad ,$$

since  $R_{11}$  is non-singular. Because of (A.5.8) one has finally

$$\tilde{\mathbf{x}} = K \begin{pmatrix} R_{11}^{-1} & 0 \\ 0 & 0 \end{pmatrix} H^T \mathbf{b} \quad . \tag{A.10.2}$$

In analogy to (A.6.3) we write

$$\tilde{\mathbf{x}} = A^+ \mathbf{b} \tag{A.10.3}$$

and call the  $n \times m$  matrix

$$A^+ = K \begin{pmatrix} R_{11}^{-1} & 0 \\ 0 & 0 \end{pmatrix} H^T \quad (\text{A.10.4})$$

the *pseudo-inverse* of the  $m \times n$  matrix  $A$ .

The matrix  $A^+$  is uniquely determined by  $A$  and does not depend on the particular orthogonal decomposition (A.5.5). This can easily be seen if one denotes the  $j$ th column vector of  $A^+$  by  $\mathbf{a}_j^+ = A^+ \mathbf{e}_j$ , with  $\mathbf{e}_j$  the  $j$ th column vector of the  $m$ -dimensional unit matrix. According to (A.10.3) the vector  $\mathbf{a}_j^+$  is the minimum-length solution of the equation  $A \mathbf{a}_j^+ = \mathbf{e}_j$ , and is therefore unique.

## A.11 Eigenvalues and Eigenvectors

We now consider the *eigenvalue equation*

$$G\mathbf{x} = \lambda \mathbf{x} \quad (\text{A.11.1})$$

for the  $n \times n$  matrix  $G$ . If this is fulfilled, then the scalar  $\lambda$  is called the *eigenvalue* and the  $n$ -vector  $\mathbf{x}$  the *eigenvector* of  $G$ . Clearly the eigenvector  $\mathbf{x}$  is only determined up to an arbitrary factor. One can choose this factor such that  $|\mathbf{x}| = 1$ .

We consider first the particularly simple case where  $G$  is a diagonal matrix with non-negative diagonal elements,

$$G = S^T S = S^2 = \begin{pmatrix} s_1^2 & & & \\ & s_2^2 & & \\ & & \ddots & \\ & & & s_n^2 \end{pmatrix}, \quad (\text{A.11.2})$$

which can be expressed as the product of an arbitrary diagonal matrix  $S$  with itself,

$$S = \begin{pmatrix} s_1 & & & \\ & s_2 & & \\ & & \ddots & \\ & & & s_n \end{pmatrix}. \quad (\text{A.11.3})$$

The eigenvalue equation  $S^2 \mathbf{x} = \lambda \mathbf{x}$  then has the eigenvalues  $s_i^2 = \lambda_i$  and the normalized eigenvectors are the basis vectors  $\mathbf{x}_i = \mathbf{e}_i$ .

In place of  $S$  we now set

$$A = U S V^T \quad (\text{A.11.4})$$

with orthogonal matrices  $U$  and  $V$  and

$$G = A^T A = V S^T S V^T \quad . \quad (\text{A.11.5})$$

We can write the eigenvalue equation of  $G$  in the form

$$G\mathbf{x} = \lambda\mathbf{x} \quad (\text{A.11.6})$$

or

$$V S^T S V^T \mathbf{x} = \lambda \mathbf{x} \quad .$$

Multiplying on the left with  $V^T$ ,

$$S^T S V^T \mathbf{x} = \lambda V^T \mathbf{x} \quad ,$$

and comparing with (A.11.1) and (A.11.2) shows that  $G$  has the same eigenvalues  $\lambda_i = s_i^2$  as  $S^2$ , but has the orthogonally transformed eigenvectors

$$\mathbf{x}_i = \mathbf{e}'_i = V \mathbf{e}_i \quad . \quad (\text{A.11.7})$$

One can clearly find the eigenvalues and eigenvectors of  $G$  if one knows the orthogonal matrix  $V$  that transforms  $G$  into a diagonal matrix,

$$V^T G V = S^T S = S^2 \quad . \quad (\text{A.11.8})$$

The transformation (A.11.8) is called a *principal axis transformation*. The name becomes clear by considering the equation

$$\mathbf{r}^T G \mathbf{r} = 1 \quad . \quad (\text{A.11.9})$$

We are interested in the geometrical position of all points  $\mathbf{r}$  that fulfill (A.11.9).

For the vector  $\mathbf{r}$  the following two representations are completely equivalent:

$$\mathbf{r} = \sum_i r_i \mathbf{e}_i \quad (\text{A.11.10})$$

and

$$\mathbf{r} = \sum_i r'_i \mathbf{e}'_i \quad , \quad (\text{A.11.11})$$

with the components  $r_i$  and  $r'_i$  taken with respect to the basis vectors  $\mathbf{e}_i$  and  $\mathbf{e}'_i$ , respectively. With the representation (A.11.11) and by using (A.11.5) and (A.11.7) we obtain

$$\begin{aligned} 1 &= \mathbf{r}^T G \mathbf{r} = \mathbf{r}^T V S^2 V^T \mathbf{r} \\ &= \sum_i (r'_i \mathbf{e}'_i)^T V S^2 V^T \sum_j (r'_j \mathbf{e}'_j) \\ &= \sum_i (r'_i \mathbf{e}_i^T V^T) V S^2 V^T \sum_j (r'_j V \mathbf{e}_j) \\ &= \sum_i (r'_i \mathbf{e}_i^T) S^2 \sum_j (r'_j \mathbf{e}_j) \end{aligned}$$

and finally

$$\sum_{i=1}^n r_i'^2 s_i^2 = \sum_{i=1}^n r_i'^2 / a_i^2 = 1 \quad . \quad (\text{A.11.12})$$

This is clearly the equation of an ellipsoid in  $n$  dimensions with half-diameters in the directions  $\mathbf{e}'_i$  and having the lengths

$$a_i = \frac{1}{s_i} \quad . \quad (\text{A.11.13})$$

The vectors

$$\mathbf{a}_i = \mathbf{e}'_i / s_i = V \mathbf{e}_i / s_i \quad (\text{A.11.14})$$

are the *principal axes* of the ellipsoid. They have the directions of the eigenvectors of  $G$ . Their lengths  $a_i = 1/\sqrt{s_i^2}$  are determined by the eigenvalues  $s_i^2$ .

The matrix

$$C = G^{-1} = (A^T A)^{-1} = V (S^2)^{-1} V^T$$

clearly has the same eigenvectors as  $G$ , but has the eigenvalues  $1/s_i^2$ . The lengths of the half-diameters of the ellipsoid described above are then directly equal to the square roots of the eigenvalues of  $C$ .

The matrix  $C$  is called the *unweighted covariance matrix* of  $A$ . The ellipsoid is called the *unweighted covariance ellipsoid*.

Please note that all considerations were done for matrices of the type (A.11.5) which, by construction, are symmetric and *non-negative definite*, i.e., they have real, non-negative eigenvalues. Ellipses with finite semiaxes are obtained only with *positive-definite* matrices. If the same eigenvalue occurs several times then the ellipsoid has several principal axes of the same length. In this case there is a certain ambiguity in the determination of the principal axes which, however, can always be chosen orthogonal to each other.

Up to now we have not given any prescription for finding the eigenvalues. The eigenvalue equation is  $G\mathbf{x} = \lambda\mathbf{x}$  or

$$(G - \lambda I)\mathbf{x} = 0 \quad . \quad (\text{A.11.15})$$

Written in this way it can be considered as a linear system of equations for determining  $\mathbf{x}$ , for which the right-hand side is the null vector. Corresponding to (A.6.5) and (A.6.9) there is a non-trivial solution only for

$$\det(G - \lambda I) = 0 \quad . \quad (\text{A.11.16})$$

This is the *characteristic equation* for determining the eigenvalues of  $A$ . For  $n = 2$  this is

$$\begin{vmatrix} g_{11} - \lambda & g_{12} \\ g_{21} & g_{22} - \lambda \end{vmatrix} = (g_{11} - \lambda)(g_{22} - \lambda) - g_{12}g_{21} = 0$$

and has the solutions

$$\lambda_{1,2} = \frac{g_{11} + g_{22}}{2} \pm \sqrt{g_{12}g_{21} + \frac{(g_{11} - g_{22})^2}{4}} .$$

If  $G$  is symmetric, as previously assumed, i.e.,  $g_{12} = g_{21}$ , then the eigenvalues are real. If  $G$  is positive definite, they are positive.

The characteristic equation of an  $n \times n$  matrix has  $n$  solutions. In practice, however, for  $n > 2$  one does not find the eigenvalues by using the characteristic equation, but rather by means of an iterative procedure such as the singular value decomposition (see Sect. A.12).

## A.12 Singular Value Decomposition

We now consider a particular orthogonal decomposition of the  $m \times n$  matrix  $A$ ,

$$A = U S V^T . \quad (\text{A.12.1})$$

Here and in Sects. A.13 and A.14 we assume that  $m \geq n$ . If this is not the case, then one can simply extend the matrix  $A$  with further rows whose elements are all zero until it has  $n$  rows, so that  $m = n$ . The decomposition (A.12.1) is a special case of the decomposition (A.5.5). The  $m \times n$  matrix  $S$ , which takes the place of  $R$ , has the special form

$$S = \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix} , \quad (\text{A.12.2})$$

and  $D$  is a  $k \times k$  diagonal matrix with  $k = \text{Rang}(A)$ . The diagonal elements of  $S$  are called the *singular values* of  $A$ . If  $A$  is of full rank, then  $k = n$  and all  $s_i \neq 0$ . For reduced rank  $k < n$  one has  $s_i = 0$  for  $i > k$ . We will see below that  $U$  and  $V$  can be determined such that all  $s_i$  are positive and ordered to be non-increasing,

$$s_1 \geq s_2 \geq \dots \geq s_k . \quad (\text{A.12.3})$$

The singular values of  $A$  have a very simple meaning. From Sect. A.11 one has directly that the  $s_i$  are the square roots of the eigenvalues of  $G = A^T A$ . Thus the half-diameters of the covariance ellipsoid of  $G$  are  $a_i = 1/s_i$ . If  $G$  is singular, then  $A$  has the reduced rank  $k < n$ , and the  $n - k$  singular values  $s_{k+1}, \dots, s_n$  vanish. In this case the determinant

$$\det G = \det U \det S^2 \det V = \det S^2 = s_1^2 s_2^2 \dots s_n^2 \quad (\text{A.12.4})$$

also vanishes.

With the substitutions  $H \rightarrow U$ ,  $K \rightarrow V$ ,  $R \rightarrow S$ ,  $R_{11} \rightarrow D$  we obtain from Sect. A.5

$$Ax = USV^T \mathbf{x} \approx \mathbf{b} \quad , \quad (\text{A.12.5})$$

$$SV^T \mathbf{x} \approx U^T \mathbf{b} \quad . \quad (\text{A.12.6})$$

With

$$V^T \mathbf{x} = \mathbf{p} = \begin{pmatrix} \mathbf{p}_1 \\ \mathbf{p}_2 \end{pmatrix} \quad \left\{ \begin{array}{l} k \\ n-k \end{array} \right. , \quad U^T \mathbf{b} = \mathbf{g} = \begin{pmatrix} \mathbf{g}_1 \\ \mathbf{g}_2 \end{pmatrix} \quad \left\{ \begin{array}{l} k \\ m-k \end{array} \right. \quad (\text{A.12.7})$$

one has

$$S\mathbf{p} = \mathbf{g} \quad , \quad (\text{A.12.8})$$

i.e.,

$$D\mathbf{p}_1 = \mathbf{g}_1 \quad (\text{A.12.9})$$

and

$$0 \cdot \mathbf{p}_2 = \mathbf{g}_2 \quad . \quad (\text{A.12.10})$$

These have the solutions

$$\tilde{\mathbf{p}}_1 = D^{-1} \mathbf{g}_1 \quad , \quad (\text{A.12.11})$$

i.e.,

$$p_\ell = g_\ell / s_\ell \quad , \quad \ell = 1, 2, \dots, k \quad ,$$

for arbitrary  $\mathbf{p}_2$ . The solution with minimum absolute value is

$$\tilde{\mathbf{x}} = V \begin{pmatrix} \tilde{\mathbf{p}}_1 \\ 0 \end{pmatrix} \quad . \quad (\text{A.12.12})$$

The residual vector has the form

$$\mathbf{r} = \mathbf{b} - A\tilde{\mathbf{x}} = U \begin{pmatrix} 0 \\ \mathbf{g}_2 \end{pmatrix} \quad (\text{A.12.13})$$

and the absolute value

$$r = |\mathbf{g}_2| = \left( \sum_{i=k+1}^m g_i^2 \right)^{1/2} \quad . \quad (\text{A.12.14})$$

## A.13 Singular Value Analysis

The rank  $k$  of the matrix  $A$  plays a decisive role in finding the solution of  $\mathbf{Ax} \approx \mathbf{b}$ . Here  $k$  characterizes the transition from non-zero to vanishing singular values,  $s_k > 0$ ,  $s_{k+1} = 0$ . How should one judge the case of very small values of  $s_k$ , i.e.,  $s_k < \varepsilon$  for a given small  $\varepsilon$ ?

**Example A.2:** Almost vanishing singular values

As a simple example we consider the case  $m = n = 2$ ,  $U = V = I$ . One then has

$$Ax = U S V^T x = Sx = \begin{pmatrix} s_1 & 0 \\ 0 & s_2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \mathbf{b} \quad , \quad (\text{A.13.1})$$

that is,

$$x_1 = b_1/s_1 \quad , \quad x_2 = b_2/s_2 \quad . \quad (\text{A.13.2})$$

If we now take  $s_2 \rightarrow 0$  in (A.13.2), then  $|x_2| \rightarrow \infty$ . At first glance one obtains a completely different picture if one sets  $s_2 = 0$  in (A.13.1) directly. This gives

$$s_1 x_1 = b_1 \quad , \quad 0 \cdot x_2 = b_2 \quad . \quad (\text{A.13.3})$$

Thus  $x_1 = b_1/s_1$  as in (A.13.2), but  $x_2$  is completely undetermined. The solution  $\tilde{\mathbf{x}}$  of minimum absolute value is obtained by setting  $x_2 = 0$ . The question is now, “What is right?  $x_2 = \infty$  or  $x_2 = 0$ ?” The answer is that one should set

$$x_2 = \begin{cases} b_2/s_2, & s_2 \geq \varepsilon \\ 0, & s_2 < \varepsilon \end{cases}$$

and choose the parameter  $\varepsilon$  such that the expression  $b_2/\varepsilon$  is still numerically well-defined. This means that one must have  $\varepsilon/|b_2| \gg 2^{-m}$  if  $m$  binary digits are available for the representation of a floating point number (cf. Sect. 4.2). Thus a finite value of  $b_2$  is computed as long as the numerical determination of this value is reasonable. If this is not the case then one approaches the situation  $s_2 = 0$ , where  $b_2$  is completely undetermined, and one sets  $b_2 = 0$ . ■

**Example A.3:** Point of intersection of two almost parallel lines

We consider the two lines shown in Fig. A.6, which are described by

$$\begin{aligned} -\alpha x_1 + x_2 &= 1 - \alpha \quad , \\ \alpha x_1 + x_2 &= 1 + \alpha \quad . \end{aligned}$$

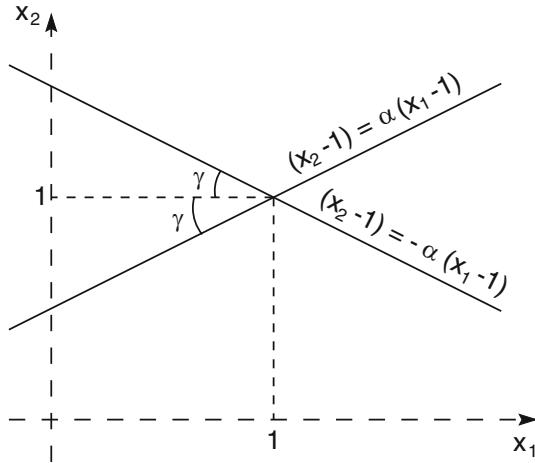
For the vector  $\mathbf{x}$  of the intersection point one has  $A\mathbf{x} = \mathbf{b}$  with

$$A = \begin{pmatrix} -\alpha & 1 \\ \alpha & 1 \end{pmatrix} \quad , \quad \mathbf{b} = \begin{pmatrix} 1 - \alpha \\ 1 + \alpha \end{pmatrix} \quad .$$

One can easily check that  $A = U S V^T$  holds with

$$U = \frac{1}{\sqrt{2}} \begin{pmatrix} -1 & -1 \\ -1 & 1 \end{pmatrix} \quad , \quad S = \sqrt{2} \begin{pmatrix} 1 & 0 \\ 0 & \alpha \end{pmatrix} \quad , \quad V = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \quad ,$$

i.e.,  $s_1 = \sqrt{2}$ ,  $s_2 = \alpha\sqrt{2}$ . Using



**Fig.A.6:** Two lines intersect at the point  $(1,1)$  with an angle  $2\gamma = 2 \arctan \alpha$ .

$$\mathbf{g} = U^T \mathbf{b} = \sqrt{2} \begin{pmatrix} -1 \\ \alpha \end{pmatrix} = S \mathbf{p} = \sqrt{2} \begin{pmatrix} p_1 \\ \alpha p_2 \end{pmatrix} , \quad \mathbf{p} = \begin{pmatrix} -1 \\ 1 \end{pmatrix}$$

one obtains

$$\mathbf{x} = V \mathbf{p} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

independent of  $\alpha$ .

If, however, one has  $s_2 = \alpha \sqrt{2} < \varepsilon$  and then sets  $s_2 = 0$ , then one obtains

$$\mathbf{p} = \begin{pmatrix} -1 \\ 0 \end{pmatrix} , \quad \mathbf{x} = \begin{pmatrix} 0 \\ 1 \end{pmatrix} .$$

From Fig. A.6 one can see that for  $\alpha \rightarrow 0$  the two lines come together to a single line described by  $x_2 = 1$ . The  $x_1$ -coordinate of the “intersection point” is completely undetermined. It is set equal to zero, since the solution vector  $\mathbf{x}$  has minimum length [cf. (A.5.14)].

As in the case of “indirect measurements” in Chap. 9 we now assume that the vector  $\mathbf{b}$  is equal to the vector of measurements  $\mathbf{y}$ , which characterize the two lines, and that their measurement errors are given by the covariance matrix  $C_y = G_y^{-1}$ . In the simplest case of equal uncorrelated errors one has

$$C_y = G_y^{-1} = \sigma^2 I .$$

The covariance matrix for the unknowns  $\tilde{\mathbf{x}}$  is according to (9.2.27)

$$C_x = (A^T G_y A)^{-1} ,$$

and thus in the case of uncorrelated measurement errors,

$$C_x = \sigma^2(A^T A)^{-1} = \sigma^2 C \quad .$$

Up to the factor  $\sigma^2$  it is thus equal to the unweighted covariance matrix  $C = (A^T A)^{-1}$  for the matrix  $A$ . For our matrix  $A$  one has

$$C = (A^T A)^{-1} = \begin{pmatrix} 1/2\alpha^2 & 0 \\ 0 & 1/2 \end{pmatrix} \quad .$$

The corresponding ellipse has the half-diameters  $\mathbf{e}_1/\alpha\sqrt{2}$  and  $\mathbf{e}_2/\sqrt{2}$ . The covariance ellipse of  $\tilde{\mathbf{x}}$  then has for the case of equal uncorrelated measurements equal half-diameters, multiplied, however, by the factor  $\sigma$ . They have the lengths  $\sigma_{x_1} = \sigma/\alpha\sqrt{2}$ ,  $\sigma_{x_2} = \sigma/\sqrt{2}$ . Clearly one then sets  $x_1 = 0$  if the inequality  $x_1 \ll \sigma_{x_1}$  would hold for a finite fixed  $x_1$ , i.e.,  $\alpha\sqrt{2} \ll \sigma$ . ■

The decision as to whether a small singular value should be set equal to zero thus depends on numerical considerations and on a consideration of the measurement errors. The following fairly general procedure of *singular value analysis* has proven to be useful in practice.

1. With a computer program one carries out the singular value decomposition, which yields, among other things, the ordered singular values

$$s_1 \geq s_2 \geq \cdots \geq s_k \quad .$$

2. Depending on the problem at hand one chooses a positive factor  $f \ll 1$ .
3. All singular values for which  $s_i < fs_1$  are set equal to zero. In place of  $k$  one has thus  $\ell \leq k$  such that  $s_i = 0$  for  $i > \ell$ .
4. With the replacements described above ( $k \rightarrow \ell$ ,  $s_{\ell+1} = \cdots = s_k = 0$ ) the formulas of Sect. A.12 retain their validity.

In place of the residual (A.12.14) one obtains a somewhat larger value, since in the sum in the expression

$$r = \left( \sum_{i=\ell+1}^m g_i^2 \right)^{1/2} \quad (\text{A.13.4})$$

one has more terms than in (A.12.14).

The procedure implies that in some cases one has an effective reduction of the rank  $k$  of the matrix  $A$  to a value  $\ell < k$ . If  $A$  has its full rank, then this can be reduced. This has the great advantage that numerical difficulties with small singular values are avoided. In contrast to, for example, the Gaussian or Cholesky procedures, the user does not have to worry about whether  $G = A^T A$  is singular or nearly singular.

Although the singular value analysis always gives a solution of minimum absolute value  $\tilde{\mathbf{x}}$  for the problem  $A\mathbf{x} \approx \mathbf{b}$ , caution is recommended for the case  $\ell < n$  (regardless of whether  $k < n$  or  $\ell < k = n$ ). In such a case one has an (almost) linear dependence of the unknowns  $x_1, \dots, x_n$ . The solution  $\tilde{\mathbf{x}}$  is not the only solution of the problem, but rather is simply the solution with the smallest absolute value out of the many possible solutions.

We have already remarked in Sect. A.5 that the singular value decomposition is advantageous also with respect to numerical precision compared to other methods, especially to the method of normal equations. A detailed discussion (see, e.g., [18]) is beyond the scope of this book. Here we limit ourselves to giving an example.

**Example A.4:** Numerical superiority of the singular value decomposition compared to the solution of normal equations

Consider the problem  $A\mathbf{x} \approx \mathbf{b}$  for

$$A = \begin{pmatrix} 1 & 1 \\ \beta & 0 \\ 0 & \beta \end{pmatrix} .$$

The singular values of  $A$  are the square roots of the eigenvalues of

$$G = A^T A = \begin{pmatrix} 1 + \beta^2 & 1 \\ 1 & 1 + \beta^2 \end{pmatrix}$$

and are determined by (A.11.16) to be

$$s_1 = \sqrt{2 + \beta^2} , \quad s_2 = |\beta| .$$

This was done with singular value decomposition without using the matrix  $G$ . If the computing precision is  $\varepsilon$  and if  $\beta^2 < \varepsilon$  but  $\beta > \varepsilon$ , then one obtains

$$s_1 = \sqrt{2} , \quad s_2 = |\beta| ,$$

i.e., both singular values remain non-zero. If instead of the singular value decomposition one uses the normal equations

$$A^T A \mathbf{x} = G \mathbf{x} = A^T \mathbf{b} ,$$

then the matrix  $G$  appears explicitly. With  $\beta^2 < \varepsilon$  it is numerically represented as

$$G = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} .$$

This matrix is singular,  $\det G = 0$ , and cannot be inverted, as foreseen in (A.5.17). This is also reflected in its singular values

$$s_1 = \sqrt{2} , \quad s_2 = 0 . \blacksquare$$

## A.14 Algorithm for Singular Value Decomposition

### A.14.1 Strategy

We will now give an algorithm for the singular value decomposition and follow for the most part the presentation of LAWSON and HANSON [18], which is based on the work of GOLUB and KAHN [19], BUSINGER and GOLUB [20], and GOLUB and REINSCH [21]. The strategy of the algorithm is based on the successive application of orthogonal transformations.

In the *first step*, the matrix  $A$  is transformed into a bidiagonal matrix  $C$ ,

$$A = Q C H^T \quad . \quad (\text{A.14.1})$$

The orthogonal matrices  $Q$  and  $H$  are themselves products of Householder="transformation matrices.

In *step 2* the matrix  $C$  is brought into diagonal form by means of an iterative procedure:

$$C = U' S' V'^T \quad . \quad (\text{A.14.2})$$

Here the matrices  $U'$  and  $V'$  are given by products of Givens-transformation matrices and if necessary reflection matrices. The latter ensure that all diagonal elements are non-negative.

In *step 3*, further orthogonal matrices  $U''$  and  $V''$  are determined. They are products of permutation matrices and ensure that the diagonal elements of

$$S = U''^T S' V'' \quad (\text{A.14.3})$$

are ordered so as to be non-increasing as in (A.12.3). As the result of the first three steps one obtains the matrices

$$U = Q U' U'' \quad , \quad V = H V' V'' \quad , \quad (\text{A.14.4})$$

which produce the singular value decomposition (A.12.1) as well as the diagonal elements  $s_1, s_2, \dots, s_k$ .

In *step 4* the singular value analysis is finally carried out. Stated simply, all singular values are set to zero for which

$$s_i < s_{\min} \quad . \quad (\text{A.14.5})$$

Thus one can finally give the solution vector  $\mathbf{x}$  to the equation

$$A\mathbf{x} \approx \mathbf{b} \quad . \quad (\text{A.14.6})$$

In practice, Eq. (A.14.6) is often solved simultaneously for various right-hand sides, which can then be arranged in an  $m \times \ell$  matrix,

$$B = (\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_\ell) \quad .$$

Instead of the solution vector  $\tilde{\mathbf{x}}$  we thus obtain the  $n \times \ell$  solution matrix

$$\tilde{X} = (\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, \dots, \tilde{\mathbf{x}}_\ell)$$

of the equation

$$A X \approx B \quad . \tag{A.14.7}$$

The method `DatanMatrix.singularValueDecomposition` computes this solution. It merely consists of four calls to other methods that carry out the four steps described above. These are explained in more detail in the following sections.

Usually one is only interested in the solution matrix  $\tilde{X}$  of the problem  $A X \approx B$  and possibly in the number of singular values not set to zero. Sometimes, however, one would like explicit access to the matrices  $U$  and  $V$  and to the singular values. This is made possible by the method `DatanMatrix.pseudoInverse`.

## A.14.2 Bidiagonalization

The method `DatanMatrix.sv1` performs the procedure described below. (It is declared as private within the class `DatanMatrix` as are the further methods referred to in Sect. A.14; of course their source code can be studied.) The goal is to find the  $m \times n$  matrix  $C$  in (A.14.1). Here  $C$  is of the form

$$C = \begin{pmatrix} C' \\ 0 \end{pmatrix} \quad , \tag{A.14.8}$$

and  $C'$  is an  $n \times n$  bidiagonal matrix

$$C' = \begin{pmatrix} d_1 & e_2 & & & & & \\ & d_2 & e_3 & & & & \\ & & \ddots & \ddots & & & \\ & & & \ddots & \ddots & & \\ & & & & \ddots & e_n & \\ & & & & & d_n & \end{pmatrix} \quad . \tag{A.14.9}$$

The goal is achieved by multiplication of the matrix  $A$  with appropriate Householder matrices (alternating on the right and left):

$$C = Q_n \cdots ((Q_1 A) H_2) \cdots H_n = Q^T A H \quad . \tag{A.14.10}$$

The matrix  $Q_1$  is computed from the elements of the first column of  $A$  and is applied to all columns of  $A$  (and  $B$ ). It results in only the element

(1, 1) of  $A$  remaining non-zero. The matrix  $H_2$  is computed from the first row of the matrix  $Q_1 A$ . It results in the element (1, 1) remaining unchanged, the element (1, 2) being recomputed and the elements (1, 3), ..., (1,  $n$ ) being set equal to zero. It is applied to all rows of the matrix ( $Q_1 A$ ). One thus obtains

$$Q_1 A = \begin{pmatrix} \bullet & \cdot & \cdot & \cdot \\ 0 & \cdot & \cdot & \cdot \end{pmatrix}, \quad Q_1 A H_2 = \begin{pmatrix} \bullet & \bullet & 0 & 0 \\ 0 & \cdot & \cdot & \cdot \end{pmatrix}.$$

Here “•” denotes an element of the final bidiagonal matrix  $C'$ , and “.” an element that will be changed further. Now  $Q_2$  is determined from the second column of  $Q_1 A H_2$  such that upon application to this column the element 1 remains unchanged and the element 2 and all others are changed such that only element 2 remains non-zero, and the elements 3 through  $m$  become zero.

The procedure can be summarized in the following way. The matrix  $Q_i$  is applied to the column vectors, it leaves the elements 1 through  $i - 1$  unchanged and changes the elements  $i$  through  $m$ . It produces an orthogonal transformation in the subspace of the components  $i$  through  $m$ . At the time of applying  $Q_i$ , however, these elements in the columns 1 through  $i - 1$  are already zero, so that  $Q_i$  must only be applied explicitly to columns  $i$  through  $n$ . The corresponding considerations hold for the matrix  $H_i$ . It acts on the row vectors, leaves elements 1 through  $i - 1$  unchanged, and produces an orthogonal transformation in the subspace of the components  $i$  through  $n$ . When it is applied, these components in the rows 1 through  $i - 2$  are all zero. The matrix  $H_i$  must only be applied explicitly to the rows  $i - 1, \dots, m - 1$ . Since only  $n - 1$  rows of  $A$  must be processed, the matrix  $H_n$  in (A.14.10) is the unit matrix;  $Q_n$  is the unit matrix only if  $m = n$ .

In addition to the transformed matrix  $Q^T A H$ , the matrix  $H = H_2 H_3 \cdots H_{n-1}$  is also stored. For this the information about each matrix  $H_i$  is saved. As we determined at the end of Sect. A.3, it is sufficient to store the quantities defined there,  $u_p$  and  $b$ , and the elements  $i + 1$  through  $n$  of the  $(i - 1)$ th row vectors defining the matrix  $H_i$ . This can be done, however, in the array elements of these row vectors themselves, since it is these elements that will be transformed to zero and which therefore do not enter further into the calculation. One needs to declare additional variables only for the quantities  $u_p$  and  $b$  of each of the matrices  $H_i$ . If all of the transformations have been carried out, then the diagonal elements  $d_i$  and the next-to-diagonal elements  $e_i$  are transferred to the arrays  $\mathbf{e}$  and  $\mathbf{d}$ . Finally, the product matrix  $H = H_2 H_3 \cdots H_{n-1} I$  is constructed in the first  $n$  rows of the array of the original matrix  $A$ , in the order  $H_{n-1} I, H_{n-2}(H_{n-1} I), \dots$ . Here as well, the procedure is as economical as possible, i.e., the Householder matrix is only applied to those columns of

the matrix to the right for which it would actually produce a change. The unit matrix  $I$  is constructed to the extent that it is needed, row-by-row in the array of the original matrix  $A$  starting from below. For this there is exactly the right amount of space available, which up to this point was necessary for storing information about the Householder transformations that were just applied.

### A.14.3 Diagonalization

This step is implemented in `DatanMatrix.sv2`. The bidiagonal matrix  $C$ , whose non-vanishing elements are stored in the arrays  $\mathbf{d}$  and  $\mathbf{e}$ , is now brought into diagonal form by appropriately chosen Givens transformations. The strategy is chosen such that the lowest non-diagonal element vanishes first and the non-diagonal elements always move up and to the left, until  $C$  is finally diagonal. All of the transformations applied to  $C$  from the left are also applied to the matrix stored in the array  $\mathbf{b}$ , and all transformations that act from the right are also applied to the matrix in  $\mathbf{a}$ . (We denote the matrix to be diagonalized during each step by  $C$ .)

Only the upper-left submatrix  $C_k$  with  $k$  rows and columns is not yet diagonal and must still be considered. The index  $k$  is determined such that  $e_k \neq 0$  and  $e_j = 0$ ,  $j > k$ . This means that the program runs through the loop  $k = n, n - 1, \dots, 2$ . Before the lower non-diagonal element  $e_k$  is systematically made zero by means of an iterative procedure, one checks for two special cases, which allow a shortening of the computation by means of special treatment.

*Special case 1*,  $d_k = 0$  (handled in `DatanMatrix.s21`): A Givens matrix  $W$  is applied from the right, which also causes  $e_k$  to vanish. The matrix  $w$  is the product  $W_{k-1} W_{k-2} \cdots W_1$ . Here  $W_i$  acts on the columns  $i$  and  $k$  of  $C_k$ , but of course only on those rows where at least one of these columns has a non-vanishing element.  $W_{k-1}$  acts on the row  $k - 1$ , annihilates the element  $e_k = C_{k-1,k}$ , and changes  $C_{k-1,k-1}$ . In addition,  $W_{k-1}$  acts on the row  $k - 2$ , changes the element  $C_{k-2,k-1}$ , and produces a non-vanishing element  $H = C_{k-2,k}$  in column  $k$ . Now the matrix  $W_{k-2}$  is applied, which annihilates exactly this element, but produces a new element in row  $k - 3$  and column  $k$ . When the additional element finally makes it to row 1, it can then be annihilated by the transformation  $W_1$ . As a result of this treatment of special case 1,  $C_k$  decomposes into a  $(k - 1) \times (k - 1)$  submatrix  $C_{k-1}$  and a  $1 \times 1$  null matrix.

*Special case 2*,  $C_k$  decomposes into submatrices (handled in `DatanMatrix.s22`): If  $e_\ell = 0$  for any value  $\ell$ ,  $2 \leq \ell \leq k$ , then the matrix

$$C_k = \begin{pmatrix} d_1 & e_2 & & & \\ & \ddots & & & \\ & & d_{\ell-1} & 0 & \\ & & & d_\ell & e_{\ell+1} \\ & & & & \ddots \\ & & & & & d_{k-1} & e_k \\ & & & & & & d_k \end{pmatrix} = \begin{pmatrix} C_{\ell-1} & 0 \\ 0 & \bar{C} \end{pmatrix}$$

can be decomposed into two matrices  $C_{\ell-1}$  and  $\bar{C}$ . One can first diagonalize  $\bar{C}$  and then  $C_{\ell-1}$ . In particular, if  $\ell = k$ , then one obtains

$$C_k = \begin{pmatrix} C_{k-1} & 0 \\ 0 & d_k \end{pmatrix} .$$

Here  $d_k$  is the singular value, and the loop index can only be decreased by one,  $k \rightarrow k - 1$ . First, however, the case  $d_k < 0$  must be treated; see “*change of sign*” below. If  $d_{\ell-1} = 0$ , but  $e_\ell \neq 0$ , then  $C$  can still be decomposed. For this one uses the transformation matrix  $T = T_k T_{k-1}, \dots, T_{\ell+1}$  acting on the left, where  $T_i$  acts on the rows  $\ell$  and  $i$ . In particular,  $T_{\ell+1}$  annihilates the element  $e_\ell = C_{\ell,\ell+1}$  and creates an element  $H = C_{\ell,\ell+2}$ .  $T_{\ell+2}$  annihilates this element and creates in its place  $H = C_{\ell,\ell+3}$ . Finally  $T_k$  annihilates the last created element  $H = C_{\ell k}$  by a transformation to  $C_{kk} = d_k$ .

After it has been checked whether one or both of the special cases were present and such an occurrence has been treated accordingly, it remains only to diagonalize the matrix  $\bar{C}$ . This consists of the rows and columns  $\ell$  through  $k$  of  $C$ . If special case 2 was not present, then one has  $\ell = 1$ . The problem is solved iteratively with the QR-algorithm.

*QR-algorithm* (carried out in DatanMatrix.s23): First we denote the square output matrix  $\bar{C}$  by  $C_1$ . We then determine orthogonal matrices  $U_i, V_i$  and carry out transformations

$$C_{i+1} = U_i^T C_i V_i \quad , \quad i = 1, 2, \dots ,$$

which lead to a diagonal matrix  $S$ ,

$$\lim_{i \rightarrow \infty} C_i = S \quad .$$

The following prescription is used for determining  $U_i$  and  $V_i$ :

- (A) One determines the eigenvalues  $\lambda_1, \lambda_2$  of the lower-right  $2 \times 2$  submatrix of  $C_i^T C_i$ . Here  $\sigma_i$  is the eigenvalue closest to the lower-right element of  $C_i^T C_i$ .

- (B) The matrix  $V_i$  is determined such that  $V_i^T(C_i^T C_i - \sigma_i I)$  has upper triangular form.
- (C) The matrix  $U_i$  is determined such that  $C_{i+1} = U_i^T C_i V_i$  is again bidiagonal.

The matrix  $V_i$  from step (B) exists, according to a theorem by FRANCIS [22], if

- (a)  $C_i^T C_i$  is tridiagonal with non-vanishing subdiagonal elements,
- (b)  $V_i$  is orthogonal,
- (c)  $\sigma_i$  is an arbitrary scalar,
- (d)  $V_i^T(C_i^T C_i)V_i$  is tridiagonal, and
- (e) The first column of  $V_i^T(C_i^T C_i - \sigma_i I)$  except the first element vanishes.

The requirement (a) is fulfilled, since  $C_i$  is bidiagonal and the special case 2 has been treated if necessary; (b) is also fulfilled by constructing  $V_i$  as the product of Givens matrices. This is done in such a way that simultaneously (d) and (e) are fulfilled. In particular one has

$$V_i = R_1 R_2 \cdots R_{n-1} , \quad U_i^T = L_{n-1} L_n \cdots L_1 ,$$

where

- $R_j$  acts on the columns  $j$  and  $j + 1$  of  $C$ ,
- $L_j$  acts on the rows  $j$  and  $j + 1$  of  $C$ ,
- $R_1$  is determined such that requirement (e) is fulfilled,
- $L_1, R_2, L_2, \dots$  are determined such that (e) is fulfilled without violating (d).

For  $\sigma_i$  one obtains

$$\sigma_i = d_n^2 + e_n \left( e_n - \frac{d_{n-1}}{t} \right)$$

with

$$\begin{aligned} t &= f + \sqrt{1 + f^2} , & f \geq 0 , \\ t &= f - \sqrt{1 + f^2} , & f < 0 , \end{aligned}$$

and

$$f = \frac{d_n^2 - d_{n-1}^2 + e_n^2 - e_{n-1}^2}{2e_n d_{n-1}} .$$

The first column of the matrix  $(C_i^T C_i - \sigma_i I)$  is

$$\begin{pmatrix} d_1^2 - \sigma_i \\ d_1 e_2 \\ 0 \\ \vdots \\ 0 \end{pmatrix} .$$

One determines the matrix  $R_1$ , which defines a Givens transformation, such that all elements of the first column of  $R_1^T(C_i^T C_i - \sigma_i I)$  except the first vanish. Application of  $R_1$  on  $C_i$  produces, however, an additional element  $H = C_{21}$ , so that  $C_i$  is no longer bidiagonal,

$$C_i = \begin{pmatrix} \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \end{pmatrix}, \quad C_i R_1 = \begin{pmatrix} \cdot & \cdot & \cdot & \cdot \\ H & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \end{pmatrix} .$$

By application of  $L_1$ , this element is projected onto the diagonal. In its place a new element  $H = C_{13}$  is created,

$$L_1 C_i R_1 = \begin{pmatrix} \cdot & \cdot & H & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \end{pmatrix} .$$

By continuing the procedure, the additional element is moved further down and to the right, and can be completely eliminated in the last step:

$$T_{n-1} \begin{pmatrix} \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ H & \cdot & \cdot & \cdot & \cdot \end{pmatrix} = \begin{pmatrix} \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \end{pmatrix} = C_{i+1} .$$

If the lower non-diagonal element of  $C_{i+1}$  is already zero, then the lower diagonal element is already a singular value. Otherwise the procedure is repeated, whereby it is first checked whether now one of the two special cases is present. The procedure typically converges in about  $2k$  steps; ( $k$  is the rank of the original matrix  $A$ ). If convergence has still not been reached after  $10k$  steps, then the algorithm is terminated without success.

*Change of sign.* If a singular value  $d_k$  has been found, i.e., if  $e_k = 0$ , then it is checked whether it is negative. If this is the case, then a simple orthogonal transformation is applied that multiplies the element  $d_k = C_{kk}$  of

$C$  and all elements in the  $k$ th column of the matrix contained in  $\mathbf{a}$  by  $-1$ . The index  $k$  can then be reduced by one. Corresponding to (A.12.6) the matrix  $B$  is multiplied on the left by  $U^T$ . This is done successively with the individual factors making up  $U^T$ .

#### A.14.4 Ordering of the Singular Values and Permutation

By the method `DatanMatrix.sv3` the singular values are put into non-increasing order. This is done by a sequence of permutations of neighboring singular values, carried out if the singular value that follows is larger than the preceding. The matrices stored in  $\mathbf{a}$  and  $\mathbf{b}$  are multiplied by a corresponding sequence of permutation matrices; cf. (A.14.4).

#### A.14.5 Singular Value Analysis

In the last step the singular value analysis is carried out as described in Sect. A.13 by the method `DatanMatrix.sv4`. For a given factor  $f \ll 1$  a value  $\ell \leq k$  is determined such that  $s_i < f s_1$  for  $i > \ell$ . The columns of the array  $\mathbf{B}$ , which now contains the vectors  $\mathbf{g}$ , are transformed in their first  $\ell$  elements into  $\tilde{\mathbf{p}}_1$  according to (A.12.11), and the elements  $\ell + 1, \dots, n$  are set equal to zero. Then the solution vectors  $\tilde{\mathbf{x}}$ , which make up the columns of the solution matrix  $\tilde{\mathbf{X}}$ , are computed according to (A.12.12).

### A.15 Least Squares with Weights

Instead of the problem (A.5.3),

$$r^2 = (\mathbf{Ax} - \mathbf{b})^T (\mathbf{Ax} - \mathbf{b}) = \min \quad , \quad (\text{A.15.1})$$

one often encounters a similar problem that in addition contains a positive-definite symmetric weight-matrix  $G_{m \times m}$ ,

$$r^2 = (\mathbf{Ax} - \mathbf{b})^T G (\mathbf{Ax} - \mathbf{b}) = \min \quad . \quad (\text{A.15.2})$$

In (A.15.1) one simply has  $G = I$ . Using the Cholesky decomposition (A.9.1) of  $G$ , i.e.,  $G = U^T U$ , one has

$$r^2 = (\mathbf{Ax} - \mathbf{b})^T U^T U (\mathbf{Ax} - \mathbf{b}) = \min \quad . \quad (\text{A.15.3})$$

With the definitions

$$\mathbf{A}' = U \mathbf{A} \quad , \quad \mathbf{b}' = U \mathbf{b} \quad (\text{A.15.4})$$

Eq. (A.15.3) takes on the form

$$r^2 = (A'\mathbf{x} - \mathbf{b}')^T(A'\mathbf{x} - \mathbf{b}') = \min \quad . \quad (\text{A.15.5})$$

After the replacement (A.15.4), the problem (A.15.2) is thus equivalent to the original one (A.15.1).

In Sect. A.11 we called the  $n \times n$  matrix

$$C = (A^T A)^{-1} \quad (\text{A.15.6})$$

the unweighted covariance matrix of the unknowns  $\mathbf{x}$  in the Problem (A.15.1). In Problem (A.15.2), the *weighted covariance matrix*

$$C_x = (A'^T A')^{-1} = (A^T G A)^{-1} \quad (\text{A.15.7})$$

appears in its place.

## A.16 Least Squares with Change of Scale

The goal in solving a problem of the type

$$(A\mathbf{x} - \mathbf{b})^T(A\mathbf{x} - \mathbf{b}) = \min \quad (\text{A.16.1})$$

is the most accurate numerical determination of the solution vector  $\tilde{\mathbf{x}}$  and the covariance matrix  $C$ . A change of scale in the elements of  $A$ ,  $\mathbf{b}$ , and  $\mathbf{x}$  can lead to an improvement in the numerical precision.

Let us assume that the person performing the problem already has an approximate idea of  $\tilde{\mathbf{x}}$  and  $C$ , which we call  $\mathbf{z}$  and  $K$ . The matrix  $K$  has the Cholesky decomposition  $K = L^T L$ . By defining

$$A' = A L \quad , \quad \mathbf{b}' = \mathbf{b} - A \mathbf{z} \quad , \quad \mathbf{x}' = L^{-1}(\mathbf{x} - \mathbf{z}) \quad , \quad (\text{A.16.2})$$

Eq. (A.16.1) becomes

$$(A'\mathbf{x}' - \mathbf{b}')^T(A'\mathbf{x}' - \mathbf{b}') = \min \quad .$$

The meaning of the new vector of unknowns  $\mathbf{x}'$  is easily recognizable for the case where  $K$  is a diagonal matrix. We write

$$K = \begin{pmatrix} \sigma_1^2 & & & \\ & \sigma_2^2 & & \\ & & \ddots & \\ & & & \sigma_n^2 \end{pmatrix} \quad , \quad L = \begin{pmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_n \end{pmatrix} \quad ,$$

where the quantities  $\sigma_i^2$  are the estimated values of the variances of the unknowns  $x_i$ . Thus the  $i$ th component of the vector  $\mathbf{x}'$  becomes

$$x'_i = \frac{x_i - z_i}{\sigma_i} .$$

If in fact one has  $\tilde{x}_i = z_i$  and the corresponding variance  $\sigma_i^2$ , then  $x'_i = 0$  and has a variance of one. If the estimates are at least of the correct order of magnitude, the  $x'_i$  are close to zero and their variances are of order unity. In addition, in case the full matrix is, in fact, estimated with sufficient accuracy, the components of  $\mathbf{x}'$  are not strongly correlated with each other.

In practice, one carries out the transformation (A.16.2) only in exceptional cases. One must take care, however, that “reasonable” variables are chosen for (A.16.2). This technique is applied in the graphical representation of data. If it is known, for example, that a voltage  $U$  varies in the region  $\sigma = 10 \text{ mV}$  about the value  $U_0 = 1 \text{ V}$ , then instead of  $U$  one would plot the quantity  $U' = (U - U_0)/\sigma$ , or some similar quantity.

## A.17 Modification of Least Squares According to Marquardt

Instead of the problem

$$(A\mathbf{x} - \mathbf{b})^T(A\mathbf{x} - \mathbf{b}) = \min , \quad (\text{A.17.1})$$

which we have also written in the shorter form

$$A\mathbf{x} - \mathbf{b} \approx 0 , \quad (\text{A.17.2})$$

let us now consider the modified problem

$$\frac{m}{n} \left\{ \underbrace{\begin{pmatrix} A \\ \lambda I \end{pmatrix}}_n \mathbf{x} \approx \begin{pmatrix} \mathbf{b} \\ 0 \end{pmatrix} \right\} \frac{m}{n} . \quad (\text{A.17.3})$$

Here  $I$  is the  $n \times n$  unit matrix and  $\lambda$  is a non-negative number. The modified problem is of considerable importance for fitting nonlinear functions with the method of least squares (Sect. 9.5) or for minimization (Sect. 10.15). For  $\lambda = 0$  Eq. (A.17.3) clearly becomes (A.17.2). If, on the other hand,  $\lambda$  is very large, or more precisely, if it is large compared to the absolute values of the elements of  $A$  and  $\mathbf{b}$ , then the last “row” of (A.17.3) determines the solution  $\tilde{\mathbf{x}}$ , which is the null vector for  $\lambda \rightarrow \infty$ .

We first ask which direction the vector  $\tilde{\mathbf{x}}$  has for large values of  $\lambda$ . The normal equations corresponding to (A.17.3), cf. (A.5.16), are

$$(A^T, \lambda I) \begin{pmatrix} A \\ \lambda I \end{pmatrix} \mathbf{x} = (A^T A + \lambda^2 I) \mathbf{x} = (A^T, \lambda I) \begin{pmatrix} \mathbf{b} \\ 0 \end{pmatrix} = A^T \mathbf{b}$$

with the solution

$$\tilde{\mathbf{x}} = (A^T A + \lambda^2 I)^{-1} A^T \mathbf{b} .$$

For large  $\lambda$  the second term in parentheses dominates, and one obtains simply

$$\tilde{\mathbf{x}} = \lambda^{-2} A^T \mathbf{b} .$$

That is, for large values of  $\lambda$ , the solution vector tends toward the direction of the vector  $A^T \mathbf{b}$ .

We will now show that for a given  $\lambda$  the solution  $\mathbf{x}^{(\lambda)}$  to (A.17.3) can easily be found with the singular value decomposition simultaneously with the determination of the solution  $\tilde{\mathbf{x}}$  of (A.17.2). The singular value decomposition (A.12.1) of  $A$  is

$$A = U \begin{pmatrix} S \\ 0 \end{pmatrix} V^T .$$

By substituting into (A.17.3) and multiplying on the left we obtain

$$\begin{pmatrix} U^T & 0 \\ 0 & V^T \end{pmatrix} \begin{pmatrix} U & 0 \\ 0 & \lambda I \end{pmatrix} \begin{pmatrix} S \\ 0 \end{pmatrix} V^T \mathbf{x} = \begin{pmatrix} U^T & 0 \\ 0 & V^T \end{pmatrix} \begin{pmatrix} \mathbf{b} \\ 0 \end{pmatrix}$$

or

$$\begin{pmatrix} S \\ 0 \\ \lambda I \end{pmatrix} \mathbf{p} = \begin{pmatrix} \mathbf{g} \\ 0 \end{pmatrix} , \quad (\text{A.17.4})$$

where, using the notation as in Sect. A.12,

$$\mathbf{p} = V^T \mathbf{x} , \quad \mathbf{g} = U^T \mathbf{b} .$$

By means of Givens transformations, the matrix on the left-hand side of (A.17.4) can be brought into diagonal form. One obtains

$$\begin{pmatrix} S^{(\lambda)} \\ 0 \\ 0 \end{pmatrix} \mathbf{p} = \begin{pmatrix} \mathbf{g}^{(\lambda)} \\ \mathbf{h}^{(\lambda)} \end{pmatrix}$$

with

$$g_i^{(\lambda)} = \frac{g_i s_i}{s_i^{(\lambda)}} , \quad i = 1, \dots, n ,$$

$$g_i^{(\lambda)} = g_i , \quad i = n+1, \dots, m ,$$

$$h_i^{(\lambda)} = -\frac{g_i^\lambda}{s_i^{(\lambda)}} , \quad i = 1, \dots, n ,$$

$$s_i^{(\lambda)} = \sqrt{s_i^2 + \lambda^2}$$

and thus

$$\begin{aligned} p_i^{(\lambda)} &= \frac{g_i s_i}{s_i^2 + \lambda^2} = p_i^{(0)} \frac{s_i^2}{s_i^2 + \lambda^2} \quad , \quad i = 1, \dots, k \quad , \\ p_i^{(\lambda)} &= 0 \quad , \quad i = k+1, \dots, n \quad . \end{aligned}$$

The solution  $\tilde{\mathbf{x}}^{(\lambda)}$  of (A.17.3) is then

$$\tilde{\mathbf{x}}^{(\lambda)} = V \mathbf{p}^{(\lambda)} \quad .$$

The method `DatanMatrix.marquardt` computes these solution vectors for two values of  $\lambda$ . It proceeds mostly as `DatanMatrix.singularValueDecomposition`; only in step 4 instead of `Datanmatrix.sv4` the method `DatanMatrix.svm` is used which is adapted to the Marquardt problem.

## A.18 Least Squares with Constraints

One often encounters the problem (A.5.3)

$$r^2 = (\mathbf{Ax} - \mathbf{b})^2 = \min \quad (\text{A.18.1})$$

with the *constraint*

$$E\mathbf{x} = \mathbf{d} \quad . \quad (\text{A.18.2})$$

Here  $A$  is as before an  $m \times n$  matrix and  $E$  is an  $\ell \times n$  matrix. We will restrict ourselves to the only case that occurs in practice,

$$\text{Rang } E = \ell < n \quad . \quad (\text{A.18.3})$$

The determination of an extreme value with constraints is usually treated in analysis textbooks with the method of Lagrange multipliers. Here as well we rely on orthogonal transformations. The following method is due to LAWSON and HANSON [18]. It uses a basis of the null space of  $E$ . First we carry out an orthogonal decomposition of  $E$  as in (A.5.5),

$$E = H R K^T \quad . \quad (\text{A.18.4})$$

Here we regard the orthogonal  $n \times n$  matrix  $K$  as being constructed out of an  $n \times \ell$  matrix  $K_1$  and an  $n \times (n - \ell)$  matrix  $K_2$ ,

$$K = (K_1, K_2) \quad . \quad (\text{A.18.5})$$

According to (A.5.6) and (A.5.7), all solutions of (A.18.2) have the form

$$\widehat{\mathbf{x}} = K_1 \widetilde{\mathbf{p}}_1 + K_2 \mathbf{p}_2 = \widetilde{\mathbf{x}} + K_2 \mathbf{p}_2 . \quad (\text{A.18.6})$$

Here  $\widetilde{\mathbf{x}}$  is the unique solution of minimum absolute value of (A.18.2). For brevity we will write this in the form  $\widetilde{\mathbf{x}} = E^+ \mathbf{d}$ ; cf. (A.10.3).  $\mathbf{p}_2$  is an arbitrary  $(n - \ell)$ -vector, since the vectors  $K_2 \mathbf{p}_2$  form the null space of  $E$ ,

$$E K_2 \mathbf{p}_2 = 0 . \quad (\text{A.18.7})$$

The constraint (A.18.2) thus says that the vector  $\mathbf{x}$  for which (A.18.1) is a minimum must come from the set of all vectors  $\widehat{\mathbf{x}}$ , i.e.,

$$\begin{aligned} (A \widehat{\mathbf{x}} - \mathbf{b})^2 &= (A(\widetilde{\mathbf{x}} + K_2 \mathbf{p}_2) - \mathbf{b})^2 \\ &= (A K_2 \mathbf{p}_2 - (\mathbf{b} - A \widetilde{\mathbf{x}}))^2 = \min . \end{aligned} \quad (\text{A.18.8})$$

This relation is a least-squares problem without constraints, from which the  $(n - \ell)$ -vector  $\mathbf{p}_2$  can be determined. We write its solution using (A.10.3) in the form

$$\widetilde{\mathbf{p}}_2 = (A K_2)^+ (\mathbf{b} - A \widetilde{\mathbf{x}}) . \quad (\text{A.18.9})$$

By substitution into (A.18.6) we finally obtain

$$\mathbf{x} = \widetilde{\mathbf{x}} + A^+ (\mathbf{b} - A \widetilde{\mathbf{x}}) = E^+ \mathbf{d} + K_2 (A K_2)^+ (\mathbf{b} - A E^+ \mathbf{d}) \quad (\text{A.18.10})$$

as the solution of (A.18.1) with the constraint (A.18.2).

The following prescription leads to the solution (A.18.10). Its starting point is the fact that one can set  $H = I$  because of (A.18.3).

*Step 1:* One determines an orthogonal matrix  $K = (K_1, K_2)$  as in (A.18.5) such that

$$E K = (E K_1, E K_2) = (\widetilde{E}_1, 0)$$

and such that  $\widetilde{E}_1$  is a lower triangular matrix. In addition one computes

$$A K = (A K_1, A K_2) = (\widetilde{A}_1, \widetilde{A}_2) .$$

*Step 2:* One determines the solution  $\widetilde{\mathbf{p}}_1$  of

$$\widetilde{E}_1 \mathbf{p}_1 = \mathbf{d} .$$

This is easy, since  $\widetilde{E}_1$  is a lower triangular matrix of rank  $\ell$ . Clearly one has  $\widetilde{\mathbf{x}} = K_1 \widetilde{\mathbf{p}}_1$ .

*Step 3:* One determines the vector

$$\bar{\mathbf{b}} = \mathbf{b} - \widetilde{A}_1 \widetilde{\mathbf{p}}_1 = \mathbf{b} - A K_1 K_1^T \widetilde{\mathbf{x}} = \mathbf{b} - A \widetilde{\mathbf{x}} .$$

*Step 4:* One determines the solution  $\widetilde{\mathbf{p}}_2$  to the least-squares problem (A.18.8) (without constraints)

$$(\tilde{A}_2 \mathbf{p}_2 - \bar{\mathbf{b}})^2 = \min \quad .$$

*Step 5:* From the results of steps 2 and 4 one finds the solution to (A.18.1) with the constraint (A.18.2)

$$\mathbf{x} = K \begin{pmatrix} \tilde{\mathbf{p}}_1 \\ \tilde{\mathbf{p}}_2 \end{pmatrix} = \begin{pmatrix} K_1 \tilde{\mathbf{p}}_1 \\ K_2 \tilde{\mathbf{p}}_2 \end{pmatrix} \quad .$$

We will now consider a simple example that illustrates both the least-squares problem with constraints as well as the method of solution given above.

### Example A.5: Least squares with constraints

Suppose the relation (A.18.1) has the simple form

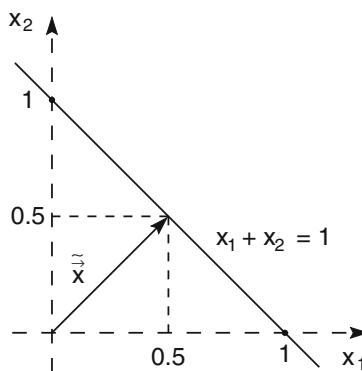
$$r^2 = \mathbf{x}^2 = \min$$

for  $n = 2$ . One then has  $m = n = 2$ ,  $A = I$ , and  $\mathbf{b} = 0$ . Suppose the constraint is

$$x_1 + x_2 = 1 \quad ,$$

i.e.,  $\ell = 1$ ,  $E = (1, 1)$ ,  $\mathbf{d} = 1$ .

The problem has been chosen such that it can be solved by inspection without mathematical complications. The function  $z = \mathbf{x}^2 = x_1^2 + x_2^2$  corresponds to a paraboloid in the  $(x_1, x_2, z)$  space, whose minimum is at  $x_1 = x_2 = 0$ . We want, however, to find not the minimum in the entire  $(x_1, x_2)$  plane, but rather only on the line  $x_1 + x_2 = 1$ , as shown in Fig. A.7. It clearly lies at the point where the line has its smallest distance from the origin, i.e., at  $x_1 = x_2 = 1/2$ .



**Fig. A.7:** The solution  $\tilde{\mathbf{x}}$  to Example A.5 lies on the line given by the constraint  $x_1 + x_2 = 1$ .

Of course one obtains the same result with the algorithm. With

$$K = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}$$

we obtain  $\tilde{E}_1 = \sqrt{2}$ ,  $\tilde{\mathbf{p}}_1 = 1/\sqrt{2}$ ,

$$\tilde{A}_1 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad (\tilde{A}_2) = \frac{1}{\sqrt{2}} \begin{pmatrix} -1 \\ 1 \end{pmatrix}, \quad \bar{\mathbf{b}} = -\frac{1}{2} \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

We solve the problem  $(\tilde{A}_2 \mathbf{p}_2 - \bar{\mathbf{b}})^2 = \min$  with the normal equations

$$\tilde{\mathbf{p}}_2 = (\tilde{A}_2^T \tilde{A}_2)^{-1} \tilde{A}_2^T \bar{\mathbf{b}} = (\sqrt{2})^{-1} \cdot 0 = 0.$$

The full solution is then

$$\mathbf{x} = K \begin{pmatrix} \tilde{\mathbf{p}}_1 \\ \tilde{\mathbf{p}}_2 \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 1/\sqrt{2} \\ 0 \end{pmatrix} = \begin{pmatrix} 1/2 \\ 1/2 \end{pmatrix}. \blacksquare$$

The method `DatanMatrix.leastSquaresWithConstraints` solves the problem of least squares  $(A\mathbf{x} - \mathbf{b})^2 = \min$  with the linear constraint  $E\mathbf{x} = \mathbf{d}$ .

## A.19 Java Classes and Example Programs

### Java Classes for Vector and Matrix Operations

`DatanVector` contains methods for vector operations.

`DatanMatrix` contains methods for matrix operations.

**Example Program A.1:** The class `E1Mtx` demonstrates simple operations of matrix and vector algebra

At first the matrices

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 3 \end{pmatrix}, \quad B = \begin{pmatrix} 2 & 3 & 1 \\ 1 & 5 & 4 \end{pmatrix}, \quad C = \begin{pmatrix} 1 & 5 \\ 3 & 4 \\ 2 & 3 \end{pmatrix}$$

and the vectors

$$\mathbf{u} = \begin{pmatrix} 0 \\ 3 \\ 4 \end{pmatrix}, \quad \mathbf{v} = \begin{pmatrix} 3 \\ 1 \\ 2 \end{pmatrix}, \quad \mathbf{w} = \begin{pmatrix} 5 \\ 2 \end{pmatrix}$$

are defined. Then, with the appropriate methods, simple operations are performed with these quantities. Finally, the resulting matrices and vectors are displayed. The operations are

$$R = A, R = A + B, R = A - B, R = AC, S = AB^T, T = A^T B, R = I, R = 0.5 A, \\ R = A^T, \mathbf{z} = \mathbf{w}, \mathbf{x} = \mathbf{u} + \mathbf{v}, \mathbf{x} = \mathbf{u} - \mathbf{v}, d = \mathbf{u} \cdot \mathbf{v}, s = |\mathbf{u}|, \mathbf{x} = 0.5 \mathbf{u}, \mathbf{x} = 0.$$

**Example Program A.2:** The class E2Mtx demonstrates the handling of submatrices and subvectors.

The matrices

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 3 \end{pmatrix}, \quad D = \begin{pmatrix} 0 & 2 \\ 1 & 3 \end{pmatrix}$$

and the vectors

$$\mathbf{u} = \begin{pmatrix} 0 \\ 3 \\ 4 \end{pmatrix}, \quad \mathbf{w} = \begin{pmatrix} 5 \\ 2 \end{pmatrix}$$

are defined.

Next a submatrix  $S$  is taken from  $A$ , and a submatrix of  $A$  is overwritten by  $D$ . A column vector and a row vector are taken from  $A$  and inserted into  $A$ . Finally some elements are taken according to a list from the vector  $\mathbf{u}$  and assembled in a vector  $\mathbf{z}$ , and elements of  $\mathbf{w}$  are put into positions (defined by list) of the vector  $\mathbf{u}$ .

**Example Program A.3:** The class E3Mtx demonstrates the performance of Givens transformations

Fist the two vectors

$$\mathbf{u} = \begin{pmatrix} 3 \\ 4 \end{pmatrix}, \quad \mathbf{w} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

are defined. Next, by the use of `DatanMatrix.defineGivensTransformation` transformation parameters  $c$  and  $s$  for the vector  $\mathbf{u}$  are computed and displayed. The Givens transformation of  $\mathbf{u}$  with these parameters is performed with `DatanMatrix.applyGivensTransformation` yielding

$$\mathbf{u}' = \begin{pmatrix} 5 \\ 0 \end{pmatrix}.$$

Finally, by calling `DatanMatrix.applyGivensTransformation` parameters  $c$  and  $s$  are computed for the vector  $\mathbf{w}$  and the transformation is applied to this vector.

**Example Program A.4:** The class E4Mtx demonstrates the performance of Householder transformations

First the two vectors

$$\mathbf{v} = \begin{pmatrix} 1 \\ 2 \\ 0 \\ 4 \\ 3 \\ 4 \end{pmatrix}, \quad \mathbf{c} = \begin{pmatrix} 1 \\ 2 \\ 0 \\ 4 \\ 3 \\ 4 \end{pmatrix}$$

are defined. Moreover, the indices  $n$ ,  $p$  are set to  $\ell$   $n = 6$ ,  $p = 3$ ,  $\ell = 5$ . By calling `DatanMatrix.defineHouseholderTransformation` the Householder transformation defined by these indices and the vector  $\mathbf{v}$  is initialized. The application of this transformation to the vectors  $\mathbf{v}$  and  $\mathbf{c}$  is performed by two calls of `DatanMatrix.defineHouseholderTransformation`. The results are displayed alphanumerically.

**Example Program A.5:** The class E5Mtx demonstrates the Gaussian algorithm for the solution of matrix equations

The matrix

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & -2 \\ 1 & 1 & 2 \end{pmatrix}$$

and the  $3 \times 3$  unit matrix  $B$  are defined. A call of Datanmatrix.matrixEquation solves the matrix equation  $AX = B$ , i.e.,  $X$  is the inverse of  $A$ . The matrix  $A$  is identical to the one chosen in Example A.1. In that example the algorithm is shown step by step.

**Example Program A.6:** The class E6Mtx demonstrates Cholesky decomposition and Cholesky inversion

First, the matrices

$$B = \begin{pmatrix} 1 & 7 & 13 \\ 3 & 9 & 17 \\ 5 & 11 & 19 \end{pmatrix}, \quad C = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{pmatrix}$$

are defined and the symmetric, positive-definite matrix  $A = B^T B$  is constructed. By calling DatanMatrix.choleskyDecompostion the Cholesky decomposition  $A = U^T U$  is performed and the triangular matrix  $U$  is displayed. Multiplication of  $U^T$  by  $U$  yields in fact  $U^T U = S$ . The method DatanMatrix.cholesky-Multiply is then used to compute  $R = UC$ . Finally, by Cholesky inversion with DatanMatrix.choleskyInversion , the inverse  $S^{-1}$  of  $S$  is cōputed. Multiplication with the original matrix yields  $SS^{-1} = I$ .

**Example Program A.7:** The class E7Mtx demonstrates the singular value decomposition

The program first operates on the same matrix as E5Mtx. However, the matrix inversion is now performed with DatanMatrix.pseudoInverse

Next, the matrix is replaced by

$$A = \begin{pmatrix} 1 & 2 & 2 \\ 2 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix} .$$

This matrix, having two identical columns, is singular. With another call of DatanMatrix.pseudo\_Inversenot only the psudoinverse matrix but also the residuals, the diagonal matrix  $D$ , and the two orthogonal matrices  $U$  and  $V$  are determined.

**Example Program A.8:** The class E8Mtx demonstrates the solution of matrix equations by singular value decomposition for 9 different cases

In the framework of our programs, in particular that of least-squares and minimization problems, matrix equations are solved nearly exclusively by singular value decomposition. In Sect. A.5 we have listed the different cases of the matrix equation  $Ax \approx b$ . If  $A$  is an  $m \times n$  matrix then first we must distinguish between the cases  $m = n$  (case 1),  $m > n$  (case 2), and  $m < n$  (case 3). A further subdivision is brought about by the rank of  $A$ . If the rank  $k$  of  $A$  is  $k = \min(m, n)$  then we have the case 1a (or 2a or 3a). If, however,  $k < \min(m, n)$  then we are dealing with case 1b (or 2b or 3b). The rank of a matrix is equal to its number of non-vanishing singular values. In numerical calculations, which are always performed with finite accuracy, one obviously has to define more precisely the meaning of “non-vanishing”. For this definition we use the method of singular value analysis (Sect. A.13) and set a singular value equal to zero if it is smaller than a fraction  $f$  of the largest singular value. The number of finite singular values remaining in this analysis is called the *pseudorank*. In addition to the cases mentioned above we consider as well the cases 1c, 2c, and 3c, in which the matrix  $A$  has full rank but not full pseudorank.

The program consists of two nested loops. The outer loop runs through the cases 1, 2, 3, the inner loop through the subcases a, b, c. For each case the matrix  $A$  is composed of individual vectors. In the subcase b two of these vectors are chosen to be identical. In the subcase c they are identical except for one element, which in one vector differs by  $\varepsilon = 10^{-12}$  compared to the value in the other vector. In case 3 the system of linear equations symbolized by the matrix equation  $Ax = b$  has less equations ( $m$ ) than unknowns ( $n$ ). This case does not appear in practice and, therefore, is not included in the programs. It is simulated here in the following way. In case 3 with  $m = 2$  and  $n = 3$ , the matrix  $A$  is extended to become a  $3 \times 3$  matrix by addition of another row the elements, of which are all set to zero, and correspondingly  $m$  is set to 3.

If the singular value analysis shows that one or several singular values are smaller than the fraction  $f$  of the largest singular value, then they are set to zero. In our example program for each of the 9 cases the analysis is performed twice, first for  $f = 10^{-15}$  and then for  $f = 10^{-10}$ . For  $f = 10^{-15}$  in our example cases 1c, 2c, 3c the matrix  $A$  has full rank, in spite of the small value of  $\varepsilon = 10^{-12}$ . The singular value analysis with  $f = 10^{-10}$  reduces the number of singular values. Note that the elements of the solution matrix differ as the choice of  $f$  changes for cases 1c, 2c, 3c. The unwieldly numerical values in the case of  $f = 10^{-15}$  show that we are near the limits of numerical stability.

**Example Program A.9:** The class E9Mtx demonstrates the use of the method `DatanMatrix.marquardt`

The method `DatanMatrix.marquardt` will be rarely called directly. It was written to be used in LsqMar and MinMar. For completeness we demonstrate it with a short class. It solves the problem  $Ax \approx b$  – modified according to (A.17.3) – for given  $A$ ,  $b$ , and  $\lambda$  and displays the results  $\mathbf{x}_1$  and  $\mathbf{x}_2$ .

**Example Program A.10:** The class E10Mtx demonstrates solution of the least squares problem with constraints by the method  
DatanGraphics leastSquaresWithConstraints

The problem of Examples 9.9 and 9.10 is solved, i.e., the measurement  $x_1 = 89$ ,  $x_2 = 31$ ,  $x_3 = 62$  of the three angles of a triangle and the evaluation of these measurements using the constraint  $x_1 + x_2 + x_3 = 180$ . The evaluation requires the solution of  $(\mathbf{Ax} - \mathbf{b})^2 = \min$  with the constraint  $E\mathbf{x} = \mathbf{d}$  with

$$A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} 89 \\ 31 \\ 62 \end{pmatrix}, \quad E = (1, 1, 1), \quad \mathbf{d} = 180.$$

In the program the matrices and vectors  $A$ ,  $\mathbf{b}$ ,  $E$ ,  $\mathbf{d}$  are provided. The solution is computed by calling DatanGraphics leastSquaresWithConstraints . It is, of course, identical to the results found in the previously mentioned examples.

## B. Combinatorics

Consider  $n$  distinguishable objects  $a_1, a_2, \dots, a_n$ . We ask for the number of possible ways  $P_n^k$ , in which one can place  $k$  of them in a given order. Such orderings are called *permutations*. For the example  $n = 4, k = 2$  these permutations are

$$\begin{array}{lll} a_1a_2 & , & a_1a_3 & , & a_1a_4 & , \\ a_2a_1 & , & a_2a_3 & , & a_2a_4 & , \\ a_3a_1 & , & a_3a_2 & , & a_3a_4 & , \\ a_4a_1 & , & a_4a_2 & , & a_4a_3 & , \end{array}$$

i.e.,  $P_4^2 = 12$ . The answer for the general problem can be derived from the following scheme. There are  $n$  different possible ways to occupy the first place in a sequence. When one of these ways has been chosen, however, there are only  $n - 1$  objects left, i.e., there remain  $n - 1$  ways to occupy the second place, and so forth. One therefore has

$$P_k^n = n(n-1)(n-2)\cdots(n-k+1) . \quad (\text{B.1})$$

The result can also be written in the form

$$P_k^n = \frac{n!}{(n-k)!} , \quad (\text{B.2})$$

where

$$n! = 1 \cdot 2 \cdots n ; \quad 0! = 1 , \quad 1! = 1 . \quad (\text{B.3})$$

Often one is not interested in the order of the  $k$  objects within a permutation (the same  $k$  objects can be arranged in  $k!$  different ways within the sequence), but rather one only considers the number of different ways of choosing of  $k$  objects out of a total of  $n$ . Such a choice is called a *combination*. The number of possible combinations of  $k$  elements out of  $n$  is then

$$C_k^n = \frac{P_k^n}{k!} = \frac{n!}{k!(n-k)!} = \binom{n}{k} . \quad (\text{B.4})$$

| $n$ | $\binom{n}{k}$ |   |    |    |    |    |    |   |     |
|-----|----------------|---|----|----|----|----|----|---|-----|
| 0   |                |   |    |    |    |    |    |   | 1   |
| 1   |                |   |    |    |    |    | 1  | 1 |     |
| 2   |                |   |    | 1  | 2  | 1  |    |   |     |
| 3   |                |   | 1  | 3  | 3  | 1  |    |   |     |
| 4   |                | 1 | 4  | 6  | 4  | 1  |    |   |     |
| 5   | 1              | 5 | 10 | 10 | 5  | 1  |    |   |     |
| 6   | 1              | 6 | 15 | 20 | 15 | 6  | 1  |   |     |
| 7   | 1              | 7 | 21 | 35 | 35 | 21 | 7  | 1 |     |
| 8   | 1              | 8 | 28 | 56 | 70 | 56 | 28 | 8 | 1   |
| :   |                |   |    |    |    |    | :  |   |     |
| 0   | ↗              | ↗ | ↗  | ↗  | ↗  | ↗  | ↗  | ↗ | $k$ |

**Fig.B.1:** Pascal's triangle.

For the *binomial coefficients*  $\binom{n}{k}$  one has the simple recursion relation

$$\binom{n-1}{k} + \binom{n-1}{k-1} = \binom{n}{k}, \quad (\text{B.5})$$

which can easily be proven by computation:

$$\begin{aligned} & \frac{(n-1)!}{k!(n-k-1)!} + \frac{(n-1)!}{(k-1)!(n-k)!} \\ &= \frac{(n-k)(n-1)! + k(n-1)!}{k!(n-k)!} \\ &= \frac{n!}{k!(n-k)!}. \end{aligned}$$

The recursion formula is the basis for the famous *Pascal's triangle*, which, because of its beauty, is shown in Fig. B.1. The name “binomial coefficient” comes from the well-known *binomial theorem*,

$$(a+b)^n = \sum_{k=0}^n \binom{n}{k} a^k b^{n-k}, \quad (\text{B.6})$$

the proof of which (by induction) is left to the reader. We use the theorem in order to derive a very important property of the coefficient  $\binom{n}{k}$ . For this we write it in the simple form for  $b = 1$ , i.e.,

$$(a+1)^n = \sum_{k=0}^n \binom{n}{k} a^k$$

and we then apply it a second time,

$$(a+1)^{n+m} = (a+1)^n(a+1)^m \quad ,$$

$$\sum_{\ell=0}^{n+m} \binom{n+m}{\ell} a^\ell = \sum_{j=0}^n \binom{n}{j} a^j \sum_{k=0}^m \binom{m}{k} a^k \quad .$$

If we consider only the term with  $a^\ell$ , then by comparing coefficients we find

$$\binom{n+m}{\ell} = \sum_{j=0}^{\ell} \binom{n}{j} \binom{m}{\ell-j} \quad . \quad (\text{B.7})$$

## C. Formulas and Methods for the Computation of Statistical Functions

### C.1 Binomial Distribution

We present here two function subprograms for computing the binomial distribution (5.1.3)

$$W_k^n = \binom{n}{k} p^k (1-p)^{n-k} \quad (\text{C.1.1})$$

and the distribution function

$$P(k < K) = \sum_{k=0}^{K-1} W_k^n \quad (\text{C.1.2})$$

are computed by the methods `StatFunct.binomial` and `StatFunct.-cumulativeBinomial`, respectively. For reasons of numerical stability the logarithm of Euler's gamma function is used in the computation.

### C.2 Hypergeometric Distribution

The hypergeometric distribution (5.3.1)

$$W_k = \binom{K}{k} \binom{N-K}{n-k} / \binom{N}{n} \quad , \quad n \leq N, k \leq K \quad , \quad (\text{C.2.1})$$

and the corresponding distribution function

$$P(k < k') = \sum_{k=0}^{k'-1} W_k \quad (\text{C.2.2})$$

are computed by `StatFunct.hypergeometric` and `StatFunct.-cumulativeHypergeometric`, respectively.

### C.3 Poisson Distribution

The Poisson distribution (5.4.1)

$$f(k; \lambda) = \frac{\lambda^k}{k!} e^{-\lambda} \quad (\text{C.3.1})$$

and the corresponding distribution function

$$P(k < K) = \sum_{k=0}^{K-1} f(k; \lambda) \quad (\text{C.3.2})$$

are computed with the methods `StatFunct.poisson` and `StatFunct.-cumulative.Poisson` respectively.

The quantities  $f(k; \lambda)$  and  $F(K; \lambda)$  depend not only on the values of the discrete variables  $k$  and  $K$ , but also on the continuous parameter  $\lambda$ . For a given  $P$  there is a certain parameter value  $\lambda_P$  that fulfills Eq. (C.3.2). We can denote this as the quantile

$$\lambda = \lambda_P(K) \quad (\text{C.3.3})$$

of the Poisson distribution. It is computed by the method `StatFunct.-quantilePoisson`.

### C.4 Normal Distribution

The probability density of the *standard normal distribution* is

$$\phi_0(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2/2) \quad . \quad (\text{C.4.1})$$

It is computed by the method `StatFunc.standardNormal`.

The *normal distribution* with mean  $x_0$  and variance  $\sigma^2$ ,

$$\phi(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-x_0)^2}{2\sigma^2}\right) \quad , \quad (\text{C.4.2})$$

can easily be expressed in terms of the standardized variable

$$u = \frac{x - x_0}{\sigma} \quad (\text{C.4.3})$$

using (C.4.1) to be

$$\phi(x) = \frac{1}{\sigma} \phi_0(u) \quad . \quad (\text{C.4.4})$$

It is computed by the method StatFunct.normal.

The distribution function of the standard normal distribution

$$\psi_0(x) = \int_{-\infty}^x \phi_0(x) dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x \exp\left(-\frac{x^2}{2}\right) dx \quad (\text{C.4.5})$$

is an integral that cannot be computed in closed form. We can relate it, however, to the incomplete gamma function described in Sect. D.5.

The distribution function of a normal distribution with mean  $x_0$  and variance  $\sigma^2$  is found from (C.4.5) to be

$$\psi(x) = \psi_0(u) \quad , \quad u = \frac{x - x_0}{\sigma} \quad . \quad (\text{C.4.6})$$

We now introduce the *error function*

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt \quad , \quad x \geq 0 \quad . \quad (\text{C.4.7})$$

Comparing with the definition of the incomplete gamma function (D.5.1) gives

$$\begin{aligned} \operatorname{erf}(x) &= \frac{2}{\Gamma(\frac{1}{2})} \int_{t=0}^{t=x} e^{-t^2} dt = \frac{1}{\Gamma(\frac{1}{2})} \int_{u=0}^{u=x^2} e^{-u} u^{-1/2} du \quad , \\ \operatorname{erf}(x) &= P\left(\frac{1}{2}, x^2\right) \quad . \end{aligned} \quad (\text{C.4.8})$$

On the other hand there is a more direct connection between (C.4.6) and (C.4.7),

$$\psi_{x_0=0, \sigma=1/\sqrt{2}}(x) = \psi_0(u = \sqrt{2}x) = \frac{1}{2}[1 + \operatorname{sign}(x)\operatorname{erf}(|x|)]$$

or

$$\psi_0(u) = \frac{1}{2} \left[ 1 + \operatorname{sign}(u) \operatorname{erf}\left(\frac{|u|}{\sqrt{2}}\right) \right] \quad ,$$

i.e.,

$$\psi_0(u) = \frac{1}{2} \left[ 1 + \operatorname{sign}(u) P\left(\frac{1}{2}, \frac{u^2}{2}\right) \right] \quad . \quad (\text{C.4.9})$$

The methods StatFunct.cumulativeStandardNormal and StatFunct.cumulativeNormal yield the distribution functions (C.4.4) and (C.4.5), respectively.

Finally we compute the quantiles of the standard normal distribution (using the method StatFunct.quantileStandardnormal). For a given probability  $P$ , the quantile  $x_p$  is defined by the relation

$$P = \psi_0(x_p) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x_p} e^{-x^2/2} dx . \quad (\text{C.4.10})$$

We determine this by finding the zero of the function

$$k(x, P) = P - \psi_0(x) \quad (\text{C.4.11})$$

using the procedure of Sect. E.2.

For the quantile  $x_P$  for a probability  $P$  of the normal distribution with mean  $x_0$  and standard deviation  $\sigma$  (computed with StatFunct.quantileNormal) one has

$$P = \psi_0(u_P) , \quad x_P = x_0 + \sigma u_P . \quad (\text{C.4.12})$$

## C.5 $\chi^2$ -Distribution

The probability density (6.6.10) of the  $\chi^2$ -distribution for  $n$  degrees of freedom,

$$f(\chi^2) = \frac{1}{2^\lambda \Gamma(\lambda)} (\chi^2)^{\lambda-1} e^{-\frac{1}{2}\chi^2} , \quad \lambda = \frac{n}{2} , \quad (\text{C.5.1})$$

is computed by the method StatFunct.chiSquared.

The distribution function

$$\begin{aligned} F(\chi^2) &= \frac{1}{\Gamma(\lambda)} \int_{u=0}^{u=\chi^2} \frac{1}{2} \left(\frac{u}{2}\right)^{\lambda-1} e^{-\frac{1}{2}u} du \\ &= \frac{1}{\Gamma(\lambda)} \int_{t=0}^{t=\chi^2/2} e^{-t} t^{\lambda-1} dt \end{aligned} \quad (\text{C.5.2})$$

is seen from (D.5.1) to be an incomplete gamma function

$$F(\chi^2) = P\left(\lambda, \frac{\chi^2}{2}\right) = P\left(\frac{n}{2}, \frac{\chi^2}{2}\right) \quad (\text{C.5.3})$$

and computed by StatFunct.cumulativeChiSquared.

The quantile  $\chi_P^2$  of the  $\chi^2$ -distribution for a given probability  $P$ , which is given by

$$h(\chi_P^2) = P - F(\chi_P^2) = 0 , \quad (\text{C.5.4})$$

is computed as the zero of the function  $h(\chi)$  with StatFunct.quantileChiSquared

## C.6 *F*-Distribution

The probability density (8.2.3) of the *F*-distribution with  $f_1$  and  $f_2$  degrees of freedom,

$$f(F) = \left(\frac{f_1}{f_2}\right)^{\frac{1}{2}f_1} \frac{\Gamma(\frac{1}{2}(f_1 + f_2))}{\Gamma(\frac{1}{2}f_1)\Gamma(\frac{1}{2}f_2)} F^{\frac{1}{2}f_1-1} \left(1 + \frac{f_1}{f_2}F\right)^{-\frac{1}{2}(f_1+f_2)}, \quad (\text{C.6.1})$$

is computed with StatFunct.fDistribution.

The distribution function

$$F(F) = \frac{\Gamma(\frac{1}{2}(f_1 + f_2))}{\Gamma(\frac{1}{2}f_1)\Gamma(\frac{1}{2}f_2)} \left(\frac{f_1}{f_2}\right)^{\frac{1}{2}f_1} \int_0^F F^{\frac{1}{2}f_1-1} \left(1 + \frac{f_1}{f_2}F\right)^{-\frac{1}{2}(f_1+f_2)} dF \quad (\text{C.6.2})$$

can be rearranged using

$$t = \frac{f_2}{f_2 + f_1 F}, \quad |dt| = \frac{f_1 f_2}{(f_2 + f_1 F)^2} |dF|$$

to be

$$\begin{aligned} F(F) &= \frac{1}{B(\frac{1}{2}f_1, \frac{1}{2}f_2)} \int_{t=\frac{f_2}{f_2+f_1F}}^{t=1} (1-t)^{\frac{1}{2}f_1-1} t^{\frac{1}{2}f_2-1} dt \quad (\text{C.6.3}) \\ &= 1 - I_{f_2/(f_2+f_1F)}\left(\frac{1}{2}f_2, \frac{1}{2}f_1\right), \end{aligned}$$

i.e., it is related to the incomplete beta function; cf. (D.6.1). We compute it with the method StatFunct.cumulativeFDistribution.

The quantile  $F_P$  of the *F*-distribution for a given probability  $P$  is given by the zero of the function

$$h(F) = P - F(F). \quad (\text{C.6.4})$$

It is computed by StatFunct.quantileFDistribution.

## C.7 *t*-Distribution

The probability density (8.3.7) of Student's *t*-distribution with  $f$  degrees of freedom,

$$\begin{aligned} f(t) &= \frac{\Gamma(\frac{1}{2}(f+1))}{\Gamma(\frac{1}{2}f)\Gamma(\frac{1}{2})\sqrt{f}} \left(1 + \frac{t^2}{f}\right)^{-\frac{1}{2}(f+1)} \quad (\text{C.7.1}) \\ &= \frac{1}{B(\frac{1}{2}, \frac{f}{2})\sqrt{f}} \left(1 + \frac{t^2}{f}\right)^{-\frac{1}{2}(f+1)}, \end{aligned}$$

is computed with by StatFunct.Student.

By using the substitution

$$u = \frac{f}{f+t^2} \quad ,$$

the distribution function of the  $t$ -distribution can be expressed in terms of the incomplete beta function (D.6.1),

$$F(t) = \frac{1}{B(\frac{1}{2}, \frac{f}{2})\sqrt{f}} \int_{-\infty}^t \left(1 + \frac{t^2}{f}\right)^{-\frac{1}{2}(f+1)} dt \quad (\text{C.7.2})$$

$$= \frac{1}{2} + \frac{\text{sign}(t)}{B(\frac{1}{2}, \frac{f}{2})\sqrt{f}} \int_0^{|t|} \left(1 + \frac{t^2}{f}\right)^{-\frac{1}{2}(f+1)} dt$$

$$= \frac{1}{2} + \frac{\text{sign}(t)}{B(\frac{1}{2}, \frac{f}{2})\sqrt{f}} \frac{1}{2} \int_{u=f/(f+t^2)}^{u=1} u^{\frac{f}{2}-1} (1-u)^{\frac{1}{2}} du \quad ,$$

$$F(t) = \frac{1}{2} \left\{ 1 + \text{sign}(t) \left[ 1 - I_{f/(f+t^2)} \left( \frac{f}{2}, \frac{1}{2} \right) \right] \right\} \quad . \quad (\text{C.7.3})$$

It is computed by the method StatFunct.cumulativeStudent.

The quantile  $t_P$  of the  $t$ -distribution for a given probability  $P$  is computed by finding the zero of the function

$$h(t) = P - F(t) \quad (\text{C.7.4})$$

with the method StatFunct.quantileStudent.

## C.8 Java Class and Example Program

### Java Class for the Computation of Statistical Functions

StatFunct contains all methods mentioned in this Appendix.

**Example Program C.1:** The class FunctionsDemo demonstrates all methods mentioned in this Appendix

The user first selects a family of functions and then a function from that family. Next the parameters, needed in the chosen case, are entered. After the Go button is clicked, the function value is computed and displayed.

## D. The Gamma Function and Related Functions: Methods and Programs for Their Computation

### D.1 The Euler Gamma Function

Consider a real number  $x$  with  $x + 1 > 0$ . We define the Euler *gamma function* by

$$\Gamma(x+1) = \int_0^\infty t^x e^{-t} dt . \quad (\text{D.1.1})$$

Integrating by parts gives

$$\int_0^\infty t^x e^{-t} dt = [-t^x e^{-t}]_0^\infty + x \int_0^\infty t^{x-1} e^{-t} dt = x \int_0^\infty t^{x-1} e^{-t} dt .$$

Thus one has the relation

$$\Gamma(x+1) = x \Gamma(x) . \quad (\text{D.1.2})$$

This is the so-called *recurrence relation* of the gamma function. From (D.1.1) it follows immediately that

$$\Gamma(1) = 1 .$$

With (D.1.2) one then has generally that

$$\Gamma(n+1) = n! , \quad n = 1, 2, \dots . \quad (\text{D.1.3})$$

We now substitute  $t$  by  $\frac{1}{2}u^2$  (and  $dt$  by  $u du$ ) and get

$$\Gamma(x+1) = \left(\frac{1}{2}\right)^x \int_0^\infty u^{2x+1} e^{-\frac{1}{2}u^2} du .$$

If we now choose in particular  $x = -\frac{1}{2}$ , we obtain

$$\Gamma(\frac{1}{2}) = \sqrt{2} \int_0^\infty e^{-\frac{1}{2}u^2} du = \frac{1}{\sqrt{2}} \int_{-\infty}^\infty e^{-\frac{1}{2}u^2} du . \quad (\text{D.1.4})$$

The integral can be evaluated in the following way. We consider

$$A = \int_{-\infty}^\infty \int_{-\infty}^\infty e^{-\frac{1}{2}(x^2+y^2)} dx dy = \int_{-\infty}^\infty e^{-\frac{1}{2}x^2} dx \int_{-\infty}^\infty e^{-\frac{1}{2}y^2} dy = 2\{\Gamma(\frac{1}{2})\}^2 .$$

The integral  $A$  can be transformed into polar coordinates:

$$A = \int_0^{2\pi} \int_0^\infty e^{-\frac{1}{2}r^2} r dr d\phi = \int_0^{2\pi} d\phi \int_0^\infty e^{-\frac{1}{2}r^2} r dr = 2\pi \Gamma(1) = 2\pi .$$

Setting the two expressions for  $A$  equal gives

$$\Gamma(\frac{1}{2}) = \sqrt{\pi} . \quad (\text{D.1.5})$$

Using (D.1.2) we can thus determine the value of the gamma function for half-integral arguments.

For arguments that are not positive integers or half-integers, the integral (D.1.1) cannot be evaluated in closed form. In such cases one must rely on approximations. We discuss here the approximation of LANCZOS [17], which is based on the analytic properties of the gamma function. We first extend the definition of the gamma function to negative arguments by means of the reflection formula

$$\Gamma(1-x) = \frac{\pi}{\Gamma(x)\sin(\pi x)} = \frac{\pi x}{\Gamma(1+x)\sin(\pi x)} . \quad (\text{D.1.6})$$

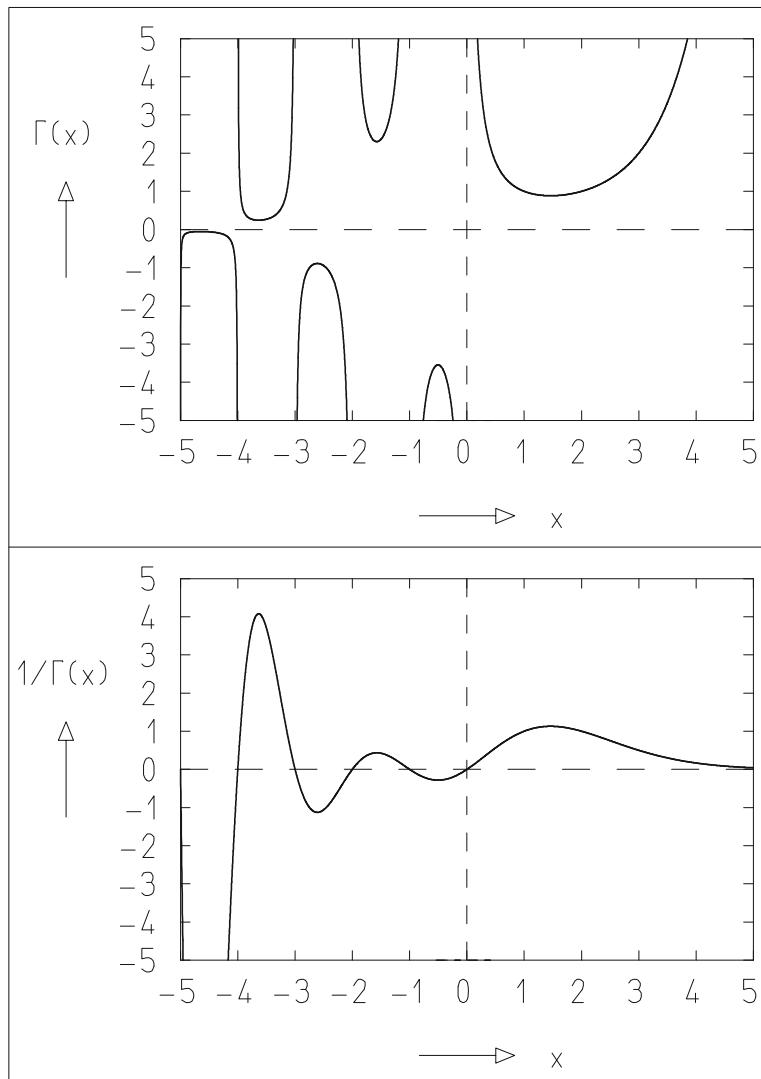
(By relations (D.1.1) and (D.1.6) the gamma function is also defined for an arbitrary complex argument if  $x$  is complex.) One sees immediately that the gamma function has poles at zero and at all negative integer values. The approximation of LANCZOS [17],

$$\Gamma(x+1) = \sqrt{2\pi} \left( x + \gamma + \frac{1}{2} \right)^{x+\frac{1}{2}} \exp\left(-x - \gamma - \frac{1}{2}\right) (A_\gamma(x) + \varepsilon) , \quad (\text{D.1.7})$$

takes into account the first few of these poles by the form of the function  $A_\gamma$ ,

$$A_\gamma(x) = c_0 + \frac{c_1}{x+1} + \frac{c_2}{x+2} + \cdots + \frac{c_{\gamma+1}}{x+\gamma+1} . \quad (\text{D.1.8})$$

For  $\gamma = 5$  one has for the error  $\varepsilon$  the approximation  $|\varepsilon| < 2 \cdot 10^{-10}$  for all points  $x$  in the right half of the complex plane. The method `Gamma.gamma` yields Euler's gamma function.



**Fig.D.1:** The functions  $\Gamma(x)$  and  $1/\Gamma(x)$ .

The gamma function is plotted in Fig. D.1. For large positive arguments the gamma function grows so quickly that it is difficult to represent its value in a computer. In many expressions, however, there appear ratios of gamma functions which have values in a problematic region. In such cases it is better to use the logarithm of the gamma function which is computed by the method `Gamma.logGamma`.

## D.2 Factorial and Binomial Coefficients

The expression

$$n! = 1 \cdot 2 \cdots n \quad (\text{D.2.1})$$

can either be directly computed as a product or as a gamma function using (D.1.3).

When computing binomial coefficients,

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} = \frac{n}{k} \cdot \frac{n-1}{k-1} \cdots \frac{n-k+1}{1} \quad , \quad (\text{D.2.2})$$

the expression on the right-hand side is preferable for numerical reasons to the expression in the middle and used in the method `Gamma.Binomial`.

## D.3 Beta Function

The *beta function* has two arguments and is defined by

$$B(z, w) = \int_0^1 t^{z-1} (1-t)^{w-1} dt \quad . \quad (\text{D.3.1})$$

The integral can be written in a simple way in terms of gamma functions,

$$B(z, w) = B(w, z) = \frac{\Gamma(z)\Gamma(w)}{\Gamma(z+w)} \quad . \quad (\text{D.3.2})$$

In this way the method `Gamma.beta` computes the beta function. Figure D.2 shows it as a function of  $w$  for several fixed values of  $z$ .

## D.4 Computing Continued Fractions

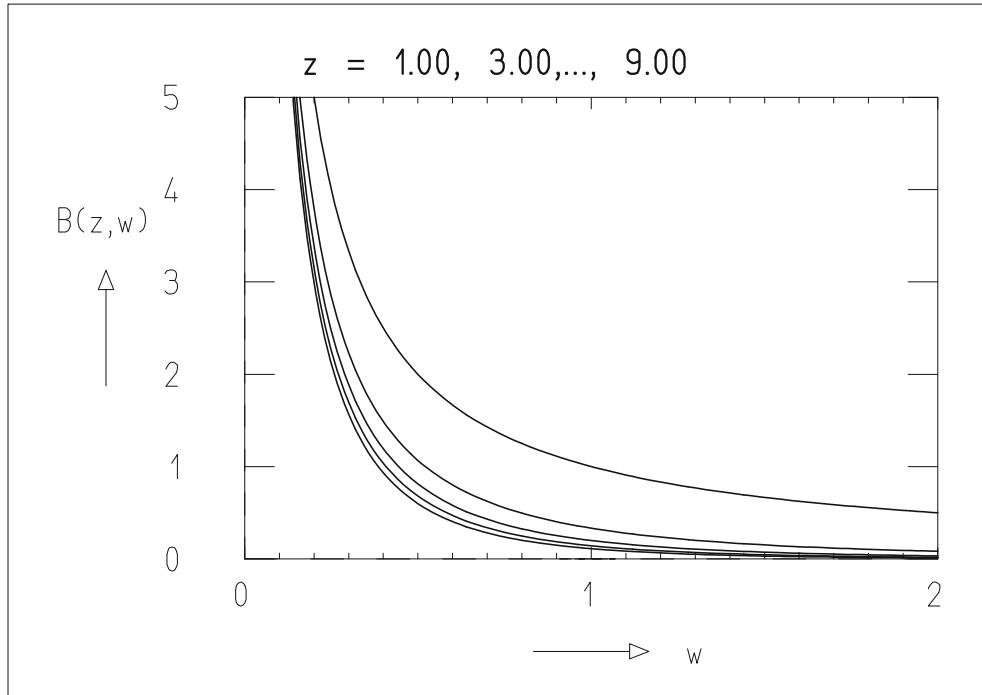
In the next two sections there appear *continued fractions*, i.e., expressions of the type

$$f = b_0 + \cfrac{a_1}{b_1 + \cfrac{a_2}{b_2 + \cfrac{a_3}{b_3 + \dots}}} \quad , \quad (\text{D.4.1})$$

that can also be written in the typographically simpler form

$$f = b_0 + \cfrac{a_1}{b_1 +} \cfrac{a_2}{b_2 +} \cfrac{a_3}{b_3 +} \cdots \quad . \quad (\text{D.4.2})$$

If we denote by  $f_n$  the value of the fraction (D.4.1) truncated after a finite number of terms up to the coefficients  $a_n$  and  $b_n$ , then one has



**Fig.D.2:** The beta function. For increasing  $z$  the curves  $B(z, w)$  are shifted to the left.

$$f_n = \frac{A_n}{B_n} . \quad (\text{D.4.3})$$

The quantities  $A_n$  and  $B_n$  can be obtained from the following recursion relation,

$$A_{-1} = 1 , \quad B_{-1} = 0 , \quad A_0 = b_0 , \quad B_0 = 1 , \quad (\text{D.4.4})$$

$$A_j = b_j A_{j-1} + a_j A_{j-2} , \quad (\text{D.4.5})$$

$$B_j = b_j B_{j-1} + a_j B_{j-2} . \quad (\text{D.4.6})$$

Since the relations (D.4.5) and (D.4.6) are linear in  $A_{j-1}$ ,  $A_{j-2}$  and  $B_{j-1}$ ,  $B_{j-2}$ , respectively, and since in (D.4.3) only the ratio  $A_n/B_n$  appears, one can always multiply the coefficients  $A_j$ ,  $A_{j-1}$ ,  $A_{j-2}$  and  $B_j$ ,  $B_{j-1}$ ,  $B_{j-2}$  by an arbitrary normalization factor. One usually chooses for this factor  $1/B_j$  and avoids in this way numerical difficulties from very large or very small numbers, which would otherwise occur in the course of the recursion. For steps in which  $B_j = 0$ , the normalization is not done.

Continued fractions, in a way similar to series expansions, appear in approximations of certain functions. In a region where the approximation for the continued fraction converges, the values of  $f_{n-1}$  and  $f_n$  for sufficiently

large  $n$  do not differ much. One can therefore use the following *truncation criterion*. If for a given  $\varepsilon \ll 1$  the inequality

$$\left| \frac{f_n - f_{n-1}}{f_n} \right| < \varepsilon$$

holds, then  $f_n$  is a sufficiently good approximation of  $f$ .

## D.5 Incomplete Gamma Function

The *incomplete gamma function* is defined for  $a > 0$  by the expression

$$P(a, x) = \frac{1}{\Gamma(a)} \int_0^x e^{-t} t^{a-1} dt . \quad (\text{D.5.1})$$

It can be expressed as a series expansion

$$P(a, x) = x^a e^{-x} \sum_{n=0}^{\infty} \frac{x^n}{\Gamma(a+n+1)} = \frac{1}{\Gamma(a)} x^a e^{-x} \sum_{n=0}^{\infty} \frac{\Gamma(a)}{\Gamma(a+n+1)} x^n . \quad (\text{D.5.2})$$

The sum converges quickly for  $x < a + 1$ . One uses the right-hand, not the middle form of (D.5.2), since the ratio of the two gamma functions reduces to

$$\frac{\Gamma(a)}{\Gamma(a+n+1)} = \frac{1}{a} \frac{1}{a+1} \cdots \frac{1}{a+n+1} .$$

In the region  $x > a + 1$  we use the continued fraction

$$1 - P(a, x) = \frac{1}{\Gamma(a)} e^{-x} x^a \left( \frac{1}{x+} \frac{1-a}{1+} \frac{1}{x+} \frac{2-a}{1+} \frac{2}{x+} \cdots \right) . \quad (\text{D.5.3})$$

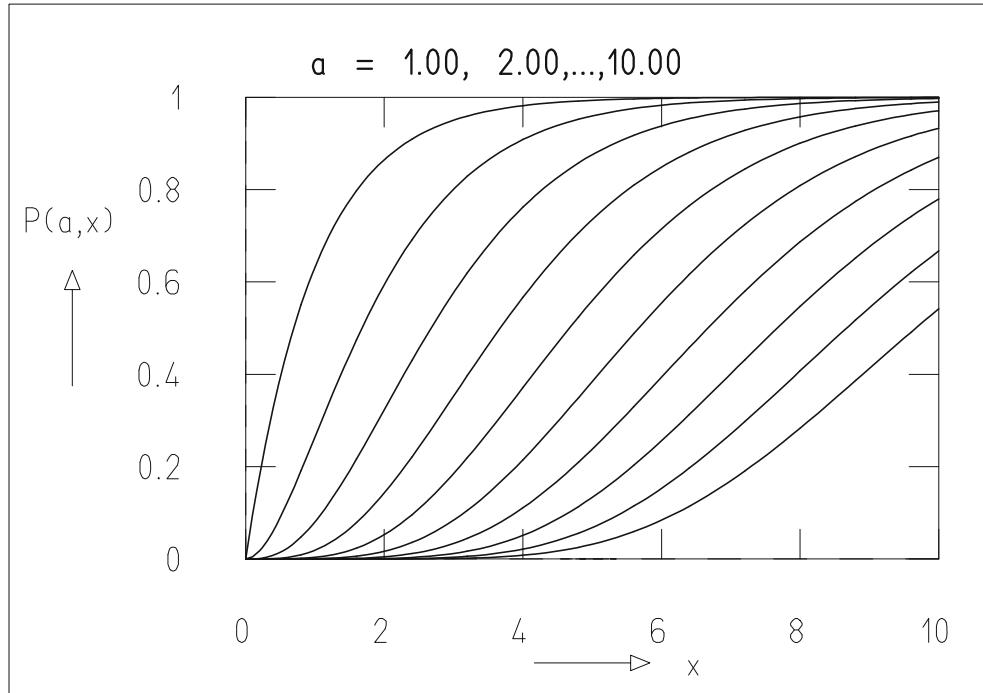
The method Gamma Incompletegamma yields the incomplete gamma function. It is shown in Fig. D.3 for several values of  $a$ . From the figure one sees immediately that

$$P(a, 0) = 0 , \quad (\text{D.5.4})$$

$$\lim_{x \rightarrow \infty} P(a, x) = 1 . \quad (\text{D.5.5})$$

## D.6 Incomplete Beta Function

The *incomplete beta function* is defined for  $a > 0, b > 0$  by the relation



**Fig.D.3:** The incomplete gamma function. With increasing  $a$  the graphs  $P(a, x)$  move to the right.

$$I_x(a, b) = \frac{1}{B(a, b)} \int_0^x t^{a-1} (1-t)^{b-1} dt \quad , \quad x \leq 0 \leq 1 \quad . \quad (\text{D.6.1})$$

The function obeys the symmetry relation

$$I_x(a, b) = 1 - I_{1-x}(b, a) \quad . \quad (\text{D.6.2})$$

The expression (D.6.1) can be approximated by the following continued fraction:

$$I_x(a, b) = \frac{x^a (1-x)^b}{a B(a, b)} \left\{ \frac{1}{1+} \frac{d_1}{1+} \frac{d_2}{1+} \dots \right\} \quad (\text{D.6.3})$$

with

$$\begin{aligned} d_{2m+1} &= -\frac{(a+m)(a+b+m)}{(a+2m)(a+2m-1)} x \quad , \\ d_{2m} &= \frac{m(b-m)}{(a+2m-1)(a+2m)} x \quad . \end{aligned}$$

The approximation converges quickly for

$$x > \frac{a+1}{a+b+1} \quad . \quad (\text{D.6.4})$$

If this requirement is not fulfilled, then  $1 - x$  is greater than the right-hand side of (D.6.4). In this case one computes  $I_{1-x}$  as a continued fraction and then uses (D.6.2).

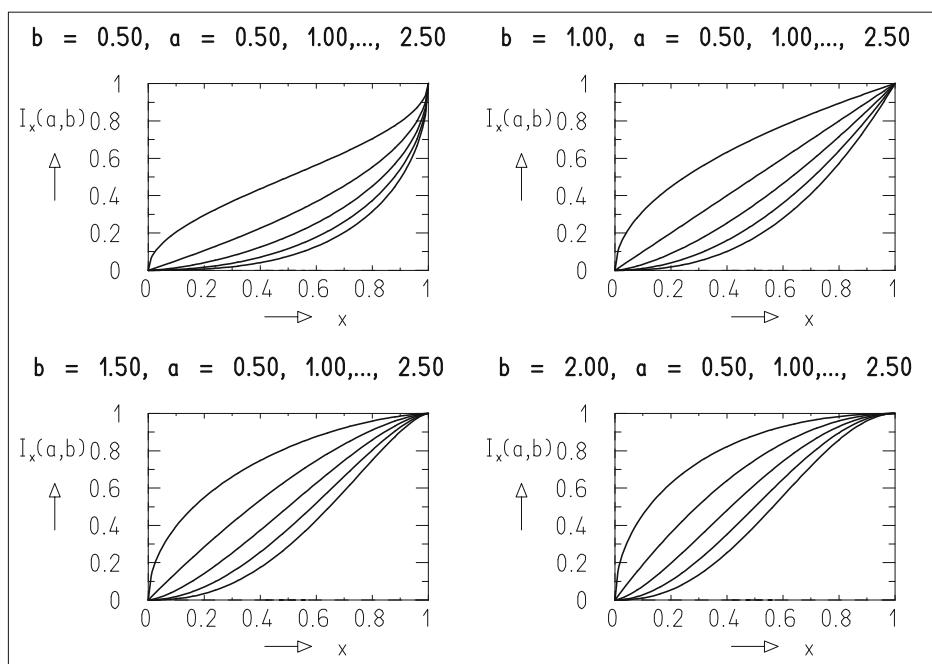
The method `Gamma.incompleteBezta` computes the incomplete beta function. In Fig. D.4 it is displayed for various values of the parameters  $a$  and  $b$ . Regardless of these parameters one has

$$I_0(a, b) = 0 \quad , \quad I_1(a, b) = 1 \quad . \quad (\text{D.6.5})$$

## D.7 Java Class and Example Program

### Java Class for the Computation of the Gamma Function and Related Functions

`Gamma` contains all methods mentioned in this Appendix.



**Fig. D.4:** The incomplete beta function. With increasing  $a$  the curves  $I_x(a, b)$  move further to the right.

**Example Program D.1:** The class `FunctionsDemo` demonstrates not only the methods of Appendix C but also all methods mentioned in the present Appendix

The user first selects a family of functions and the a function from that family. Next the parameters, needed in the chosen case, are entered. After the Go button is clicked, the function value is computed and displayed.

# E. Utility Programs

## E.1 Numerical Differentiation

The derivative  $d f(x)/dx$  of a function  $f(x)$  at the point  $x$  is given by the limit

$$f'(x) = \lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h} .$$

Obviously one can approximate  $f'(x)$  by

$$\frac{f(x+h) - f(x)}{h}$$

for a small finite value of  $h$ . In fact, for this the symmetrical difference ratio

$$\delta(h) = \frac{f(x+h) - f(x-h)}{2h} \quad (\text{E.1.1})$$

is more appropriate. This can be seen from the Taylor expansions at the point  $x$  for  $f(x+h)$  and  $f(x-h)$ , which give

$$\delta(h) = f'(x) + \frac{h^2}{3!} f'''(x) + \frac{h^4}{5!} f^{(5)}(x) + \dots ,$$

in which the leading additional term is already quadratic in  $h$ . Nevertheless, the choice of  $h$  is still critical, since for very small values of  $h$  there occur large rounding errors, and for larger values the approximation may not be valid.

One can compute  $\delta(h)$  for a monotonic sequence of values  $h = h_0, h_1, h_2, \dots$ . If the sequence  $\delta(h_0), \delta(h_1), \dots$  is monotonic (rising or falling) then this is a sign of convergence of the series to  $f'(x)$ . According to RUTISHAUSER [27], from the series  $\delta(h_0), \delta(h_1)$  one can also obtain others that converge more quickly. The method was modeled after the ROMBERG procedure [28] for numerical integration. Starting from  $h_0 = a$  one first chooses the sequence

$$h_0, h_1, \dots = a, 3a/4, a/2, 3a/8, a/4, \dots ,$$

sets

$$T_0^{(k)} = \delta(h_k) \quad ,$$

and computes the additional quantities  $T_m^{(k)}$

$$\begin{aligned} T_m^{(k)} &= \frac{2^m T_{m-1}^{(k+1)} - 1.125 T_{m-1}^{(k)}}{2^m - 1.125} \quad ; \quad m \text{ odd, } k \text{ even} \quad , \\ T_m^{(k)} &= \frac{2^m \cdot 1.125 T_{m-1}^{(k+1)} - T_{m-1}^{(k)}}{2^m \cdot 1.125 - 1} \quad ; \quad m \text{ odd, } k \text{ odd} \quad , \\ T_m^{(k)} &= \frac{2^m T_{m-1}^{(k+1)} - T_{m-1}(k)}{2^m - 1} \quad ; \quad m \text{ even} \quad . \end{aligned}$$

Arranging the quantities  $T_m^{(k)}$  in the form of a triangle,

$$\begin{array}{ccccccc} & & T_0^{(0)} & & & & \\ & & & T_1^{(0)} & & & \\ & T_0^{(1)} & & & T_2^{(0)} & & \\ & & T_1^{(1)} & & & T_3^{(0)} & \\ T_0^{(2)} & & & T_2^{(1)} & & & \\ & T_1^{(2)} & & & & & \\ & & T_0^{(3)} & & & & \\ & & & \vdots & & & \end{array}$$

the first column contains the sequence of our original difference ratios. Not only does  $T_0^{(k)}$  converge to  $f'(x)$ , but one has in general

$$\lim_{k \rightarrow \infty} T_m^{(k)} = f'(x) \quad , \quad \lim_{m \rightarrow \infty} T_m^{(k)} = f'(x) \quad .$$

The practical significance of the procedure is based on the fact that the columns on the right converge particularly quickly.

In the class **AuxDer**, the sequence  $T_0^{(0)}, \dots, T_0^{(9)}$  is computed starting from  $a = 1$ . If it is not monotonic, then  $a$  is replaced by  $a/10$  and a new sequence is computed. After 10 tries without success the procedure is terminated. If, however, a monotonic sequence is found, the triangle scheme is computed and  $T_0^{(0)}$  is given as the best approximation for  $f'(x)$ . The class **AuxDer** is similar to the program of KOELBIG [29] with the exception of minor changes in the termination criteria.

This program requires considerable computing time. For well behaved functions it is often sufficient to replace the differential ratio by the difference ratio (E.1.1). To compute second derivatives the procedure of difference ratios is extended correspondingly. The classes **AuxDri**, **AuxGrad** and **AuxHesse** therefore operate on the basis of difference ratios.

## E.2 Numerical Determination of Zeros

Computing the *quantile*  $x_P$  of a distribution function  $F(x)$  for a given probability  $P$  is equivalent to determining the zero of the function

$$k(x) = P - F(x) \quad . \quad (\text{E.2.1})$$

We treat the problem in two steps. In the first step we determine an interval  $(x_0, x_1)$  that contains the zero. In the second step we systematically reduce the interval such that its size becomes smaller than a given value  $\varepsilon$ .

In the first step we make use of the fact that  $k(x)$  is monotonic, since  $f(x)$  is monotonic. We begin with initial values for  $x_0$  and  $x_1$ . If  $f(x_0) \cdot f(x_1) < 0$ , i.e., if the function values have different signs, then the zero is contained within the interval. If this is not the case, then we enlarge the interval in the direction where the function has the smallest absolute value, and repeat the procedure with the new values of  $(x_0, x_1)$ .

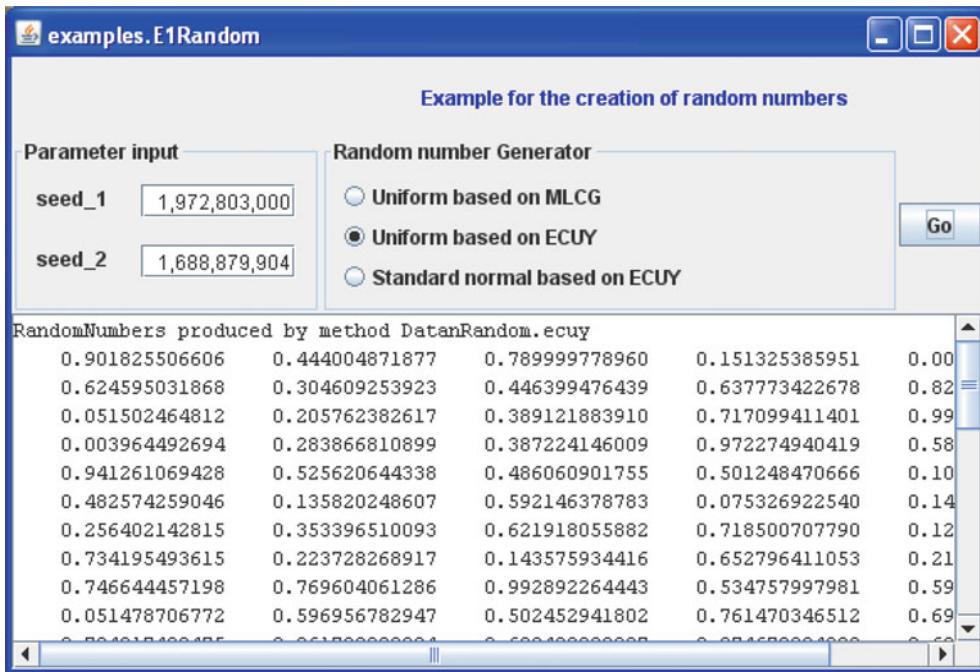
For localizing the zero within the initial interval  $(x_0, x_1)$  we use a comparatively slow but absolutely reliable procedure. The original interval is divided in half and replaced by the half for which the end points are of the opposite sign. The procedure is repeated until the interval width decreases below a given value.

This technique is implemented in the class AuxZero. It is also employed in several methods for the computation of quantiles in StatFunct in a direct way, i.e., without a call to AuxZero. An example for the application of AuxZero is the class E1MaxLike, see also Example Program 7.1.

## E.3 Interactive Input and Output Under Java

Java programs usually are *event driven*, i.e., while running they react to actions by the user. Thus an interaction between user and program is enabled. Its detailed design depends on the problem at hand and also on the user's taste. For our Example Programs four utility programs may suffice to establish simple interactions. They are explained in Fig. E.1.

It shows a screen window produced by the class DatanFrame. In its simplest form it consists only of a frame, a title line (here 'Example for the creation of random numbers') and an output region, into which the user's output can be written. The method DatanFra.add allows to add additional elements below the title line which will be arranged horizontally starting from the left. In Fig. E.1 these are an *input group*, a *radio-button group* and a *go button*. The input group is created by the class AuxJInputGroup and the radio-button group by AuxJRButtonGroup. Both (as well as DatanFra) make use of the standard Java Swing classes. The go button is



**Fig. E.1:** A window of the type `DatanFrame` with elements for interactive input, for starting the program, and for alphanumeric output of results.

directly created by a standard class. The input group itself is composed of an arbitrary number of *number-input regions*, arranged vertically one below the other, which are created by the class `AuxJNumberInput`. The detailed usage of these classes is summarized in the online documentation. we also recommend to study the source code of some of our Example Programs.

## E.4 Java Classes

`AuxDer` computes the derivative of a function using the Rutishauser method.

`AuxDri` computes the matrix  $A$  of derivatives required by `LsqNon` and `LsqMar`.

`AuxGrad` computes the gradient of a function at a given point.

`AuxHesse` computes the Hessian matrix of a function at a given point.

`AuxZero` finds the zero of a monotonic function.

`DatanFrame` creates a screen window with possibilities for interactive input and output.

AuxJInputGroup creates an input group within a screen window.

AuxJInputNumber creates a number-input region within an input group.

AuxJRButtonGroup creates a radio-button group within a screen window.

# F. The Graphics Class DatanGraphics

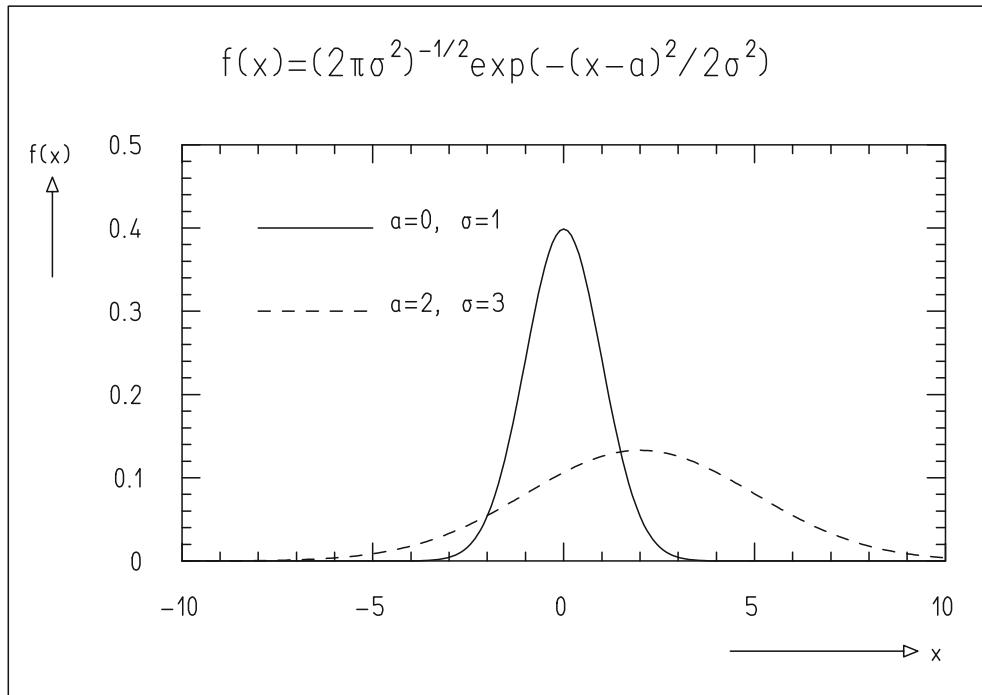
## F.1 Introductory Remarks

The graphical display of data and of curves of fitted functions has always been an important aid in data analysis. Here we present the class DatanGraphics comprising methods which produce graphics in screen windows and/or postscript files. We distinguish control, transformation, drawing, and auxilliary methods. All will be described in detail in this Appendix and their usage will be explained in a number of Example Programs. For many purposes, however, it is sufficient to use one of only five classes which, in turn, resort to DatanGraphics. These classes, by a single call, produce complete graphical structures; they are listed at the beginning of Sect. F.8.

## F.2 Graphical Workstations: Control Routines

As mentioned, a plot can be output either as in the form of a screen window or as a file in postscript format. The latter is easily embedded in digital documents or directly printed on paper, if necessary after conversion to another format such as pdf, by the use of a freely available program. For historical reasons we call both the screen window and the postscript file a *graphics workstation*.

The method DatanGraphics.openWorkstation “opens” a screen window or a file or both, i.e., it initializes buffers into which information is written by methods mentioned later. Only after the method DatanGraphics.closeWorkstation has been called, is the window presented on the screen and/or is the postscript file made available. In this way several graphics can be produced one after another. There windows can coexist on



**Fig.F.1:** Simple example of a plot produced with DatanGraphics.

the screen. They can be changed in size using the computer mouse; but their contents is not alterable.

## F.3 Coordinate Systems, Transformations and Transformation Methods

### F.3.1 Coordinate Systems

#### World Coordinates (WC)

Figure F.1 shows a plot made by DatanGraphics. Let us imagine for a moment that all of the graphical structures, including text, physically exist, e.g., that they are made out of bent wire. The coordinate system in which this wire structure is described is called the world coordinate system (WC). The coordinates of a point in the WC are denoted by  $(X, Y)$ .

## Computing Coordinates (CC)

If we consider the axes in Fig. F.1, we note that the axes designated with  $x$  and  $y$  have about the same length in world coordinates, but have very different lengths in terms of the numbers shown. The figure shows a plot of a function

$$y = f(x) \quad .$$

Each point  $(x, y)$  appears at the point  $(X, Y)$ . We call the coordinate system of the  $(x, y)$  points the computing coordinate system (CC). The transformation between WC and CC is given by Eq. (F.3.1).

## Device Coordinates (DC)

From the (fictitious) world coordinate system, the plot must be brought onto the working surface of a graphics device (terminal screen or paper). We call the coordinates  $(u, v)$  on this surface the device coordinates (DC).

### F.3.2 Linear Transformations: Window – Viewport

The concepts defined in this section and the individual transformations are illustrated in Fig. F.2.

Let us assume that the computing coordinates in  $x$  cover the range

$$x_a \leq x \leq x_b \quad .$$

The corresponding range in world coordinates is

$$X_a \leq X \leq X_b \quad .$$

A linear transformation  $x \rightarrow X$  is therefore defined by

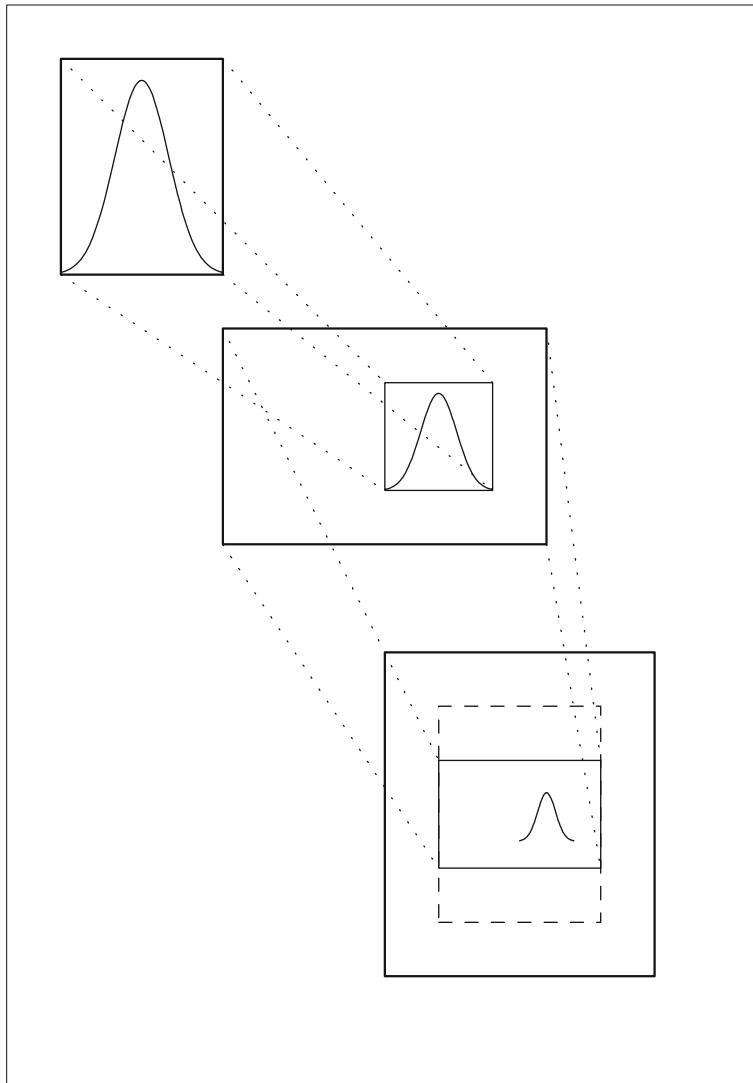
$$X = X_a + (x - x_a) \frac{X_b - X_a}{x_b - x_a} \quad . \quad (\text{F.3.1})$$

The transformation for  $y \rightarrow Y$  is defined in a corresponding way. One speaks of the mapping of the *window* in computing coordinates CC,

$$x_a \leq x \leq x_b \quad , \quad y_a \leq y \leq y_b \quad , \quad (\text{F.3.2})$$

onto the *viewport* in world coordinates WC,

$$X_a \leq X \leq X_b \quad , \quad Y_a \leq Y \leq Y_b \quad . \quad (\text{F.3.3})$$



**Fig. F.2:** The various coordinate systems. *Above*: window in computing coordinates. *Middle*: viewport (*small rectangle*) and window (*large rectangle*) in world coordinates. *Below*: preliminary viewport (*dashed rectangle*), final adjusted viewport (*small rectangle*), and border of the display surface (*large rectangle*) in device coordinates. The mappings from computing coordinates to world coordinates and from world coordinates to device coordinates are indicated by *dotted lines*.

The mapping is in general *distorted*, i.e., a square in CC becomes a rectangle in WC. It is undistorted only if the aspect ratios of the window and viewport are equal,

$$\frac{x_b - x_a}{y_b - y_a} = \frac{X_b - X_a}{Y_b - Y_a} . \quad (\text{F.3.4})$$

The user of DatanGraphics computes values first in CC that are to be plotted. By specifying the window and viewport he or she defines the mapping into WC. A possible linear distortion is perfectly acceptable in this step, since it provides a simple means of changing the scale.

Next a mapping onto the physically implemented device coordinates DC must be done. It can of course be defined by again providing a window (in WC) and a viewport (in DC). One wants, however, to avoid an additional distortion. We define, therefore, a viewport

$$u_a \leq u \leq u_b , \quad v_a \leq v \leq v_b \quad (\text{F.3.5})$$

and a window

$$X'_a \leq X \leq X'_b , \quad Y'_a \leq Y \leq Y'_b . \quad (\text{F.3.6})$$

The mapping from the window (F.3.6) onto the viewport (F.3.5) is only carried out if both have the same aspect ratio. Otherwise the viewport (F.3.6) is reduced symmetrically in width to the right and left or in height symmetrically above and below such that the viewport

$$u'_a \leq u \leq u'_b , \quad v'_a \leq v \leq v'_b \quad (\text{F.3.7})$$

has the same aspect ratio as the window (F.3.6). In this way a distortion free mapping is defined between the two.

## F.4 Transformation Methods

The user of DatanGraphics must define the transformations between the various coordinate systems by calling the appropriate routines. The transformations are then applied when drawing a graphical structure without any further intervention.

### Transformation CC → WC

This transformation is defined by calling the following two methods. Datan Graphics.setWindowInComputingCoordinates sets the window in computing coordinates, Datan Coordinates sets the viewport in world coordinates.

### Transformation WC → DC

This transformation is defined by calling the following two methods. Datan- Graphics.setWindowInWorldCoordinates sets the window

in world coordinates. DatanGraphics.setFormat defines the temporary viewport in device coordinates. Into that the final view port is fitted so that the width-to-height ratio is the same as for the window in world coordinates. If DatanGraphics.setFormat is not called at all, the format is taken to be A5 landscape. If the workstation is a screen window, then only the width-to-height ratio is taken into account. In the case of a postscript file the absolute size in centimeters is valid only if a plot of that size will fit on the paper in the printer. Otherwise the plot is demagnified until it just fits. In both cases the plot is centered on the paper. A call to the method DatanGraphics.setStandardPaperSizeAndBorders informs the program about the paper size. If it is not called, then A4 is assumed with a margin of 5 mm on all 4 sides.

In most cases, having defined the transformations in this way, the user will be interested only in computing coordinates.

## Clipping

The graphical structures are not completely drawn under certain circumstances. They are truncated if they extend past the boundary of the so-called *clipping region*. The structures are said to be *clipped*. For polylines, markers, data points, and contour lines (Sect. F.5) the clipping region is the window in computing coordinates; for text and graphics utility structures the clipping region is the window in world coordinates. These regions can be set explicitly using DatanGraphics.setSmallClippingWindow and DatanGraphics.setBigClippingWindow, respectively.

## F.5 Drawing Methods

### Colors and Line Widths

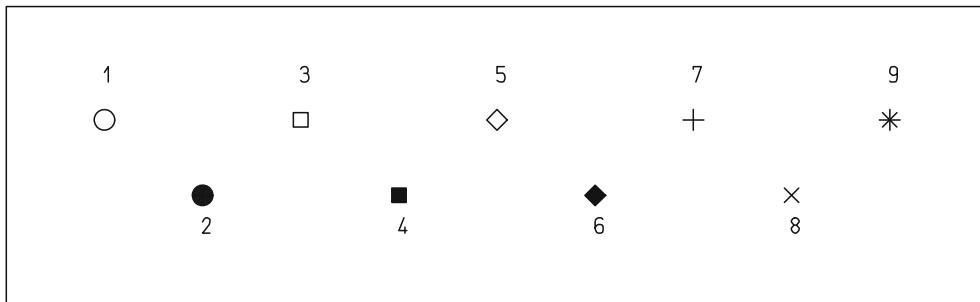
The methods mentioned up to this point carry out organizational tasks, but do not, however, produce any graphical structures on the workstation. All of the graphical structures created by DatanGraphics consist of lines. The lines possess a given color and width. The selection of these two attributes is done as follows. A pair of properties (color, linewidth) is assigned to each of set of 8 *color indices*. The set of 8 is different for screen window and postscript file. With the method DatanGraphics.chooseColor the user selects one particular color index. That then is valid until another color index is chosen. The user may assign his own choice of color and line width to a color index by the methods DatanGraphics.setScreenColor and/or DatanGraphics.setPSCColor. For the background of the screen

window (the standard is blue) another color can be chosen with `DatanGraphics.setScreenBackground`. For the postscript file the background is always transparent.

## Polyline

The concepts of a polyline and polymarker have been introduced for particularly simple graphics structures. A *polyline* defines a sequence of line segments from the point  $(x_1, y_1)$  through the points  $(x_2, y_2), (x_3, y_3), \dots$  to the point  $(x_n, y_n)$ . A polyline is drawn with the method `DatanGraphics.Polyline`.

A *polymarker* marks a plotting point with a graphical symbol. The polymarkers available in `DatanGraphics` are shown in Fig. F.3.



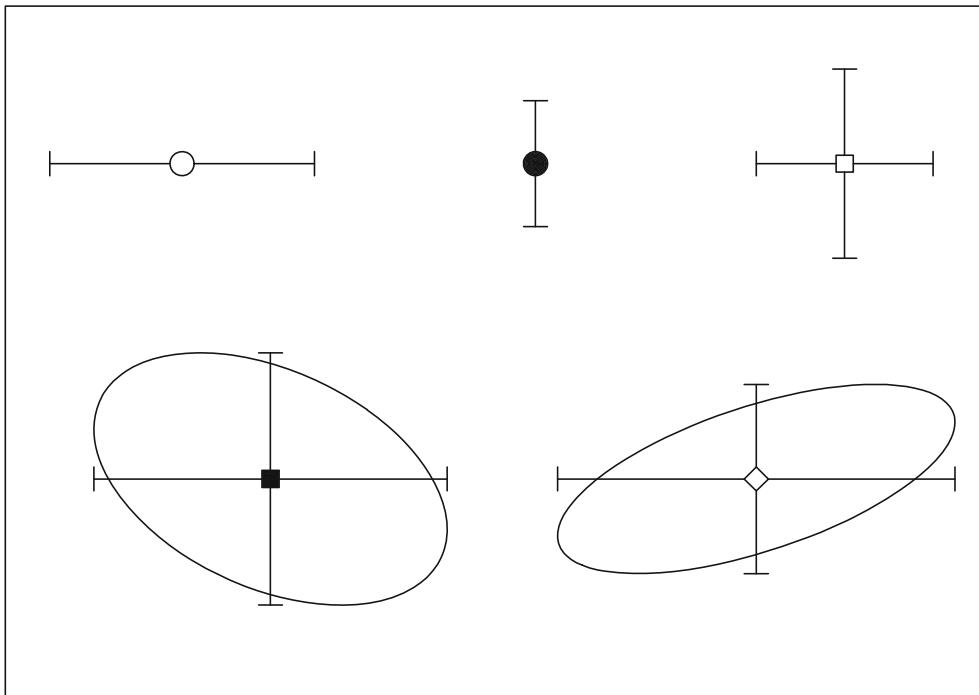
**Fig. F.3:** Polymarkers.

One can clearly achieve an arbitrarily good approximation of any graphical structure by means of polylines. For example, *graphs of functions* can be displayed with polylines as long as the individual points are sufficiently close to each other.

Sometimes one wants to draw a polyline not as a solid line but rather *dashed*, *dotted*, or *dot-dashed*. That is achieved by the method `DatanGraphics.drawBrokenPolyline`.

Polymarkers are especially suitable for marking data points. If a data point has error bars, then one would like to indicate these errors in one or both coordinates by means of *error bars*. In certain circumstances one would even like to show the complete covariance ellipse. This task is performed by the method `DatanGraphics.drawDatapoint`. Examples are shown in Fig. F.4.

An error bar in the  $x$  direction is only drawn if  $\sigma_x > 0$ , and in the  $y$  direction only if  $\sigma_y > 0$ . The covariance ellipse is only drawn if  $\sigma_x > 0, \sigma_y > 0$ , and  $\text{cov}(x, y) \neq 0$ . Error bars are not drawn if they would lie completely



**Fig.F.4:** Example for plotting data points.

within the polymarker itself. If part of the structure falls outside of the CC window ([F.3.2](#)), then this part is not drawn.

## Histogram

The method `DatanGraphics.drawHistogram` displays data in the form of a histogram.

## Contour Lines

A function  $f = f(x, y)$  defines a surface in a three dimensional  $(x, y, f)$  space. One can also get an idea of the function in the  $(x, y)$  plane by marking points for which  $f(x, y)$  is equal to a given constant  $c$ . The set of all such points forms the contour line  $f(x, y) = c$ . By drawing a set of such contour lines  $f(x, y) = c_1, c_2, \dots$  one obtains (as with a good topographical map) a rather good impression of the function.

Naturally it is impossible to compute the function for all points in the  $(x, y)$  plane. We restrict ourselves to a rectangular region in the  $(x, y)$  plane, usually the window in computing coordinates, and we break it into a total of

$N = n_x n_y$  smaller rectangles. The corner points of these smaller rectangles have  $x$  coordinates that are neighboring values in the sequence

$$x_0, x_0 + \Delta x, x_0 + 2\Delta x, \dots, x_0 + n_x \Delta x \quad .$$

The  $y$  coordinates of the corner points are adjacent points of the sequence

$$y_0, y_0 + \Delta y, y_0 + 2\Delta y, \dots, y_0 + n_y \Delta y \quad .$$

In each rectangle the contour line is approximated linearly. To do this one considers the function  $f(x, y) - c$  at the four corner points. If the function is of a different sign at two corner points that are the end points of an edge of the rectangle, then it is assumed that the contour lines intersect the edge. The intersection point is computed with linear interpolation. If the intersection points are on two edges of the rectangle, then they are joined by a line segment. If there are intersection points on more than two edges, then all pairs of such points are joined by line segments.

Clearly the approximation of the contour lines by line segments becomes better for a finer division into small rectangles. With a finer division, of course, the required computing time also becomes longer.

The method `DatanGraphics.drawContour` computes and draws a contour line. An example of a function represented by contour lines is shown in Fig. F.5.

## F.6 Utility Methods

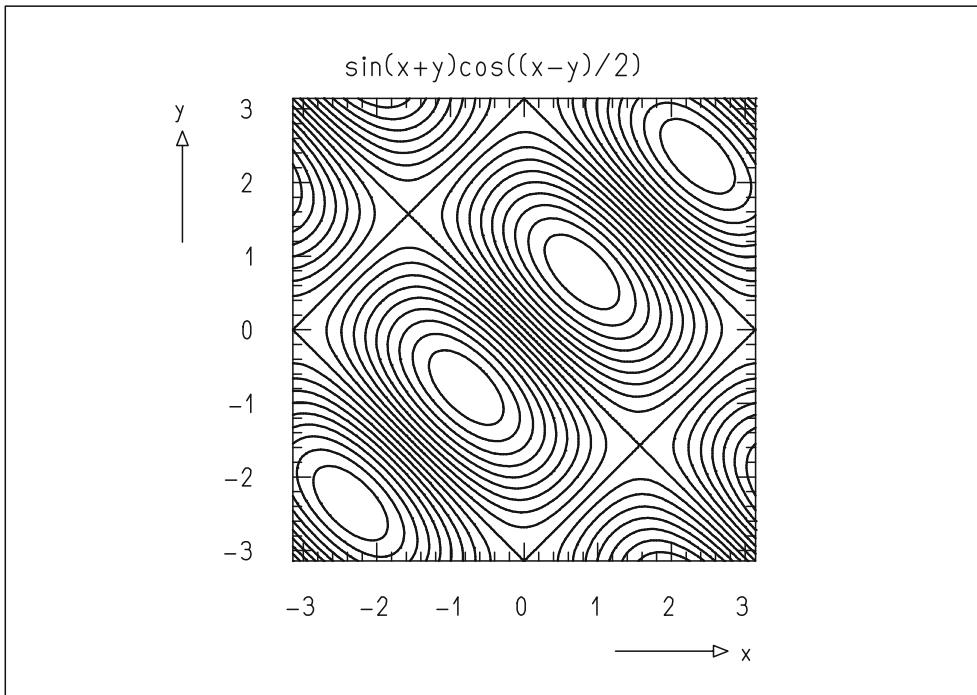
With the few methods described up to this point, a great variety of complicated plots can be produced. By using the methods of this section and the next, the tasks are made easier for the user, since they help to create graphical structures typically used in conjunction with the plots of data analysis, such as axes, coordinate crosses, and explanatory text.

### Frames

The method `DataGrahics.drawFrame` draws a frame around the plotted part of world coordinate system, i.e., the outer frame of the plots reproduced here. The method `DatanGraphics.drawBoundary`, on the other hand, draws a frame around the window of the computing coordinate system.

### Scales

The method `DatanGraphics.drawScaleX` draws a scale in  $x$  direction. Ticks appear at the upper and lower edge of the CC window pointing to the



**Fig.F.5:** Contour lines  $f(x, y) = -0.9, -0.8, \dots, 0.8, 0.9$  of the function  $f(x) = \sin(x + y) \cos([x - y]/2)$  in the  $(x, y)$  plane.

inside of the window. Below the lower edge numbers appear, marking some of these ticks. It is recommended to first call the method `DatanGraphics.drawBoundary`, to mark the edges themselves by lines. In addition an arrow with text can be drawn, showing in the direction of increasing  $x$  values. The method `DatanGraphics.drawScaleY` performs analogous tasks for an axis in  $y$  direction.

The creation of axis divisions and labels with these methods is usually done automatically without intervention of the user. Sometimes, however, the user will want to influence these operations. This can be done by using the method `DatanGraphics.setParametersForScale`. A call to this method influences only that scale which is generated by the very next call of `DatanGraphics.drawScaleX` or `DatanGraphics.drawScaleY`, respectively.

## Coordinate Cross

The method `DatanGraphics.drawCoordinateCross` draws a coordinate cross in the computing coordinate system. The axes of that system appear as broken lines inside the CC window.

## F.7 Text Within the Plot

Explanatory text makes plots easier to understand. The methods in this section create text superimposed on a plot which can be placed at any location.

The text must be supplied by the user as a character string. Before this text is translated into graphics characters, however, it must first be encoded. The simple encoding system used here allows the user to display simple mathematical formulas. For this there are *three character sets*: Roman, Greek, and mathematics, as shown in Table F.1. The character set is selected by *control characters*. These are the special characters

- @ for Roman,
- & for Greek,
- % for mathematics.

A control character in the text string causes all of the following characters to be produced with the corresponding character set, until another control symbol appears. The default character set is Roman.

In addition there exist the following *positioning symbols*:

- ^ for superscript (exponent),
- \_ for subscript (index),
- # for normal height,
- " for backspace.

All characters appear at normal height as long as no positioning symbol has appeared. One can move a maximum of two steps from normal height, e.g.,  $A_{\alpha\beta}$ ,  $A_{\alpha\beta}^{\beta}$ . The positioning symbols ^ and \_ remain in effect until the appearance of a #. The symbol " acts only on the character following it. This then appears over the previous character instead of after it. In this way one obtains, e.g.,  $A_{\alpha}^{\beta}$  instead of  $A_{\alpha}^{\beta}$ .

The method `DatanGraphics.drawCaption` draws a caption, centered slightly below the upper edge of the plotted section of the world coordinate system.

Sometimes the user wants to write text at a certain place in the plot, e.g., next to an individual curve or data point, and also to choose the text size. This is made possible by the method `DatanGraphics.drawText`.

## F.8 Java Classes and Example Programs

### Java Classes Producing Graphics

`DatanGraphics` contains the methods mentioned in this Appendix.

**Table F.1:** The various character sets for producing text.

| Input | Control characters |                      |          | Input | Control characters |                      |        |
|-------|--------------------|----------------------|----------|-------|--------------------|----------------------|--------|
|       | Roman              | Greek                | Math     |       | Roman              | Greek                | Math   |
| @     | &                  | %                    | @        | &     | %                  |                      |        |
| A     | A                  | A(ALPHA)             | Ä        | a     | a                  | $\alpha$ (alpha)     | ä      |
| B     | B                  | B(BETA)              | B        | b     | b                  | $\beta$ (beta)       | b      |
| C     | C                  | X(CHI)               | ⊓        | c     | c                  | $\chi$ (chi)         | c      |
| D     | D                  | $\Delta$ (DELTA)     | $\Delta$ | d     | d                  | $\delta$ (delta)     | d      |
| E     | E                  | E(EPSILON)           | E        | e     | e                  | $\epsilon$ (epsilon) | e      |
| F     | F                  | $\phi$ (PHI)         | F        | f     | f                  | $\varphi$ (phi)      | f      |
| G     | G                  | $\Gamma$ (GAMMA)     | ≠        | g     | g                  | $\gamma$ (gamma)     | g      |
| H     | H                  | H(ETA)               | H        | h     | h                  | $\eta$ (eta)         | h      |
| I     | I                  | I(IOTA)              | ʃ        | i     | i                  | $\iota$ (iota)       | i      |
| J     | J                  | I(IOTA)              | J        | j     | j                  | $\iota$ (iota)       | j      |
| K     | K                  | K(KAPPA)             | K        | k     | k                  | $\kappa$ (kappa)     | k      |
| L     | L                  | $\Lambda$ (LAMBDA)   |          | l     | l                  | $\lambda$ (lambda)   | l      |
| M     | M                  | M(MU)                | ±        | m     | m                  | $\mu$ (mu)           | m      |
| N     | N                  | N(NU)                | N        | n     | n                  | $\nu$ (nu)           | n      |
| O     | O                  | $\Omega$ (OMEGA)     | Ö        | o     | o                  | $\omega$ (omega)     | ö      |
| P     | P                  | $\Pi$ (PI)           | Ö        | p     | p                  | $\pi$ (pi)           | p      |
| Q     | Q                  | $\Theta$ (THETA)     | Q        | q     | q                  | $\vartheta$ (theta)  | q      |
| R     | R                  | R(RHO)               | ◦        | r     | r                  | $\rho$ (rho)         | r      |
| S     | S                  | $\Sigma$ (SIGMA)     | ß        | s     | s                  | $\sigma$ (sigma)     | s      |
| T     | T                  | T(TAU)               | -        | t     | t                  | $\tau$ (tau)         | t      |
| U     | U                  | O(OMICRON)           | Ü        | u     | u                  | $\circ$ (omicron)    | ü      |
| V     | V                  |                      | Ü        | v     | v                  |                      | v      |
| W     | W                  | $\Psi$ (PSI)         | √        | w     | w                  | $\psi$ (psi)         | w      |
| X     | X                  | $\Xi$ (XI)           | X        | x     | x                  | $\xi$ (xi)           | x      |
| Y     | Y                  | $\Upsilon$ (UPSILON) | Å        | y     | y                  | $\upsilon$ (upsilon) | y      |
| Z     | Z                  | Z(ZETA)              | Z        | z     | z                  | $\zeta$ (zeta)       | z      |
| ~     | ~                  | ~                    | ~        | -     | -                  | -                    | -      |
| !     | !                  | !                    | !        | =     | =                  | =                    | ≡      |
| \$    | \$                 | \$                   | \$       | {     | {                  | {                    | {      |
| *     | *                  | #                    | ×        | }     | }                  | }                    | }      |
| (     | (                  | ↑                    | ←        | -     | -                  | -                    | -      |
| )     | )                  | ↓                    | →        | [     | [                  | &                    | [      |
| +     | +                  | +                    | +        | ]     | ]                  | @                    | ]      |
| ,     | ,                  | ,                    | ,        | \     |                    |                      |        |
| 1     | 1                  | 1                    | 1        | :     | :                  | :                    | :      |
| 2     | 2                  | 2                    | 2        | ;     | ;                  | ;                    | ;      |
| 3     | 3                  | 3                    | 3        | ,     | ,                  | ,                    | ,      |
| 4     | 4                  | 4                    | 4        | <     | <                  | $\subset$            | $\leq$ |
| 5     | 5                  | 5                    | 5        | >     | >                  | $\supset$            | $\geq$ |
| 6     | 6                  | 6                    | 6        | ?     | ?                  | §                    | ~      |
| 7     | 7                  | 7                    | 7        | ,     | ,                  | ,                    | ,      |
| 8     | 8                  | 8                    | 8        | .     | .                  | .                    | .      |
| 9     | 9                  | 9                    | 9        | /     | /                  | \                    | %      |
| 0     | 0                  | 0                    | 0        |       |                    |                      |        |

`GraphicsWithHistogram` produces a complete plot with a histogram  
(an Example Program is `E2Sample`).

`GraphicsWith2DScatterDiagram` produces a complete plot with a  
two-dimensional scatter diagram (an Example Program is `E3Sample`).

`GraphicsWithHistogramAndPolylin` produces a complete graph-  
ics with a histogram and a polyline (an Example Program is `E6Gr`).

`GraphicsWithDataPointsAndPolyline` produces a complete plot  
with data points and one polyline (an Example Program is `E7Gr`).

`GraphicsWithDataPointAndMultiplePolylines` produces a  
complete plot with data points and several polylines (an Example Pro-  
gram is `E8Gr`).

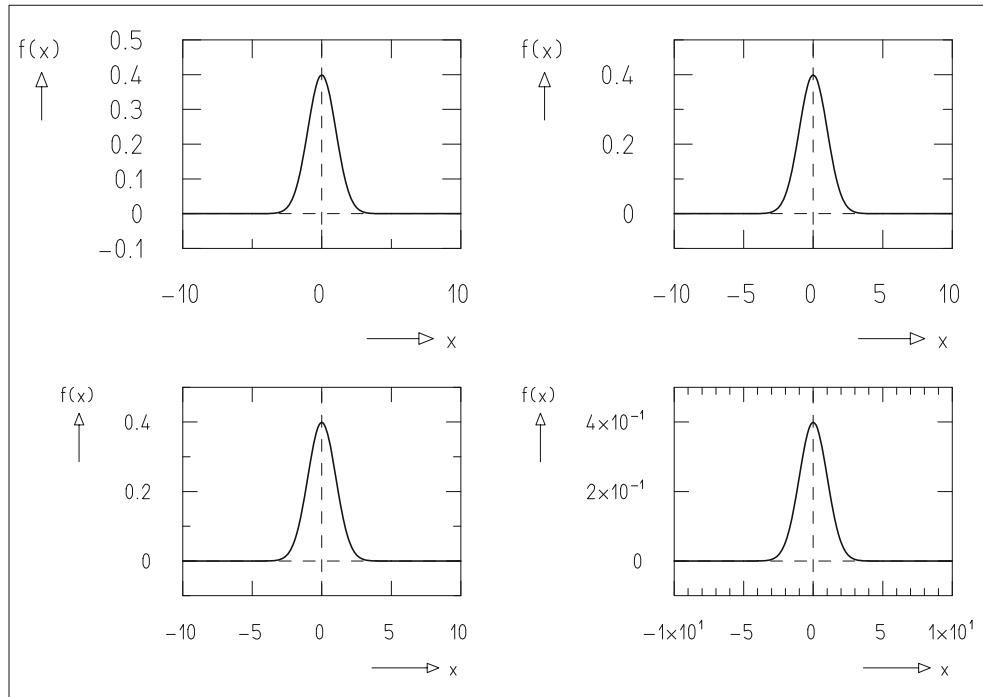
**Example Program F.1:** The class `E1Gr` demonstrates the use of the  
following methods of the class `DatanGraphics`:

`openWorkstation`, `closeWorkstation`,  
`setWindowInComputingCoordinates`,  
`setViewportInWorldCoordinates`,  
`setWindowInWorldCoordinates`, `setFormat`,  
`drawFrame`, `drawBoundary`, `chooseColor`,  
`drawPolyline`, `drawBrokenPolyline`, `drawScaleX`,  
`drawScaleY`, `drawCaption`, `drawText`

The program generates the simple plot of Fig. F.1. It opens the workstation and defines the different coordinate systems. The outer frame is drawn (enclosing the section of the world coordinate system to be displayed) and the inner frame (the boundary of the computing coordinate system). Next, the lettered scales for abscissa and ordinate are produced as is a caption for the plot. Now, the color index is changed. In a short loop a total of 201 coordinate pairs  $(x_i, y_i)$  are computed with  $x_i = -10, -9.9, -9.8, \dots, 10$  and  $y_i = f(x_i)$ . The function  $f(x)$  is the probability density of the standardized normal distribution. A polyline, defined by these pairs is drawn. In a second loop the points for a polyline are computed which correspond to a normal distribution with mean  $a = 2$  and standard deviation  $\sigma = 3$ . That polyline is represented as a broken line. Finally, two short straight lines are displayed in the upper left corner of the plot (one as a solid line and one as a dashed line). To the right each of these polylines a short text is displayed, indicating the parameters of the Gaussians displayed as solid and dashed curves, respectively. Before termination of the program the workstation is closed.

**Example Program F.2:** The class `E2Gr` demonstrates the use of the  
method `DatanGraphics.drawMark`.

The short program generates the plot of Fig. F.3, showing the different polymarkers, which can be drawn with `DatanGraphics.drawMark`.



**Fig.F.6:** Four versions of the same plot with different types of scales.

**Example Program F.3:** The class E3Gr demonstrates the use of the method `DatanGraphics.drawDataPoint`.

The program produces the plot of Fig. F.4, which contains examples for the different ways to present data points with errors.

**Example Program F.4:** The class E4Gr demonstrates the use of the method `DatanGraphics.drawContour`

A window of computing coordinates  $-\pi \leq x \leq \pi$ ,  $-\pi \leq y \leq \pi$  and a square viewport in world coordinates are selected. After creating scales and the caption, input parameters for `DatanGraphics.drawContour` are prepared. Next by successive calls of this method in a loop, contours of the function  $f(x, y) = \sin(x + y)\cos((x - y)/2)$  are drawn. The result is a plot corresponding to Fig. F.5.

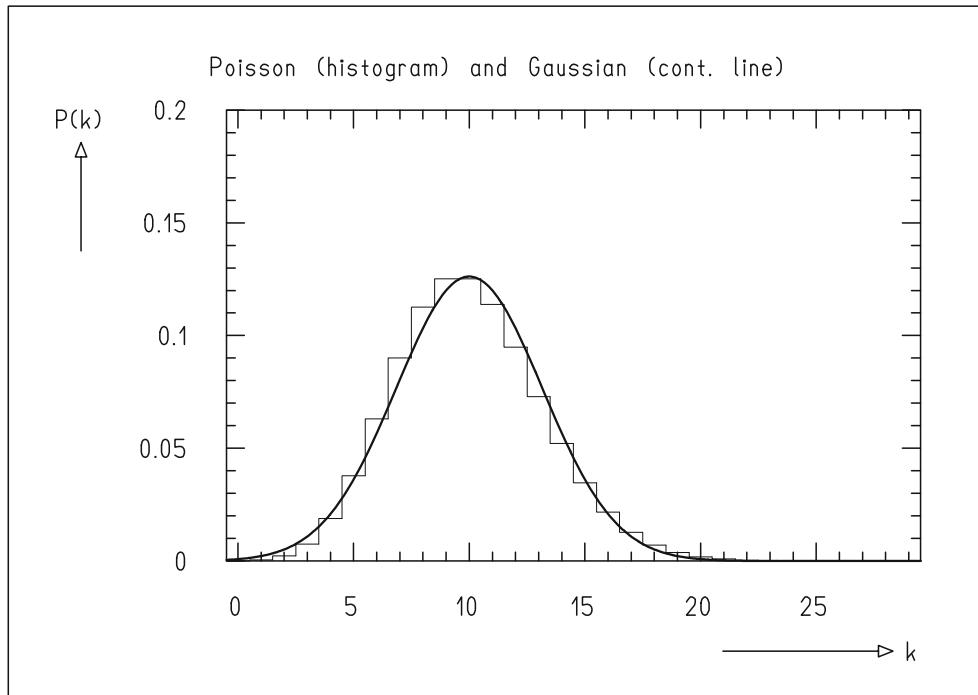
**Suggestions:** Extend the program such that the parameters `ncont` and `nstep` defining the number of contours and the number of intervals in  $x$  and  $y$  can be set interactively by the user. Study the changes in the plot resulting from very small values of `nstep`.

**Example Program F.5:** The class E5Gr demonstrates the methods

`DatanGraphics.setParametersForSale`

and `DatanGraphics.drawCoordinateCross`

The program generates the plots shown in Fig. F.6. It contains four plots which differ only by the design of their scales. The plots are generated in a loop where different



**Fig. F.7:** A plot generated with `GraphicsWithHistogramAndPolyline` containing a histogram and a polyline.

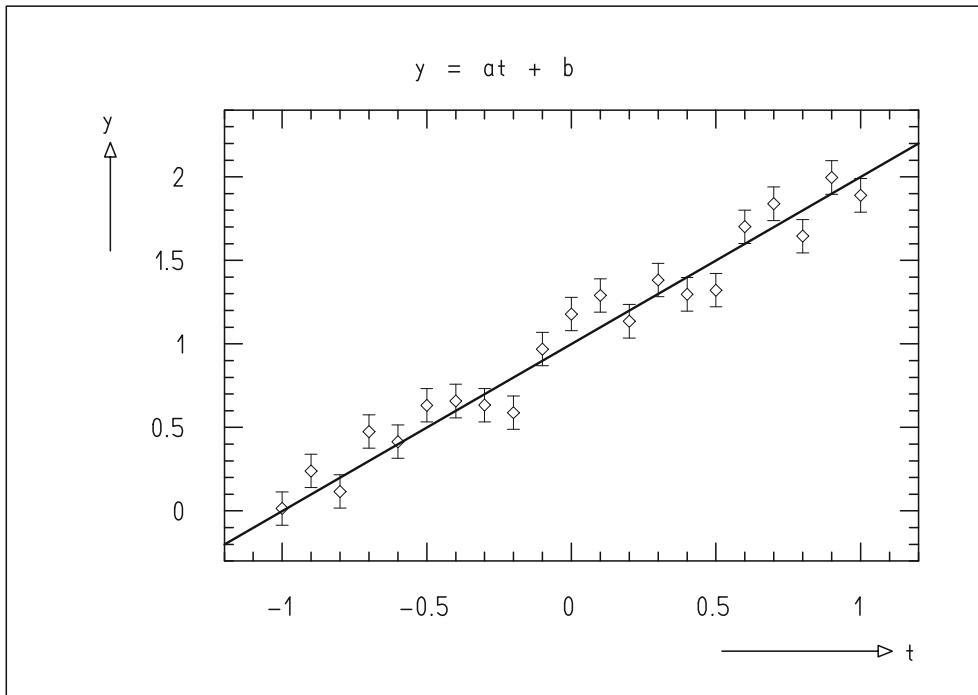
viewports in world coordinates are chosen in each step, so that the plots correspond to the upper-left, upper-right, lower-left, and lower-right quadrant of the window in world coordinates. For the upper-left plot the default values for the scale design are used. In the upper-right plot the number of ticks and the lettering of the scale is predefined. In the lower-left plot the size of the symbols used in the lettering of the scales is changed. In the lower-right plot the numbers are written in exponential notation. All plots contain a coordinate cross, which is generated by calling `DatanGraphics.drawCoordinateCross` and a curve corresponding to a Gaussian.

**Example Program F.6:** The class `E6Gr` demonstrates the use of the class `GraphicsWithHistogramAndPolyline`

The program first sets up a histogram which for each bin  $k$  contains the Poisson probability  $f(k; \lambda)$  for the parameter  $\lambda = 10$ . Next, points on a polyline are computed corresponding to the probability density of a normal distribution with mean  $\lambda$  and variance  $\lambda$ . Finally the text strings for the plot are defined and the complete plot is displayed by a call of `GraphicsWithHistogramAndPolyline` (Fig. F.7).

**Example Program F.7:** The class `E7Gr` demonstrates the use of the class `GraphicsWithDataPointsAndPolyline`

First, by calling `DatanRandom.line`, data points are generated which lie on a straight line  $y = at + b$  within the simulated errors. Next, the errors to be presented



**Fig. F.8:** A plot with data points and a polyline generated by calling `GraphicsWithDataPointsAndPolyline`.

in the directions of the horizontal and vertical axes and their covariance are defined. The latter two quantities in our example are equal to zero. The polyline defining the straight line consists of only two points. Their computation is trivial. After the definition of the axis labels and the caption, the plot is displayed by calling `GraphicsWithDataPointsAndPolyline` (Fig. F.8).

**Example Program F.8:** The class E8Gr demonstrates the use of the class `GraphicsWithDataPointsAndMultiplePolylines`

The program generates 21 data points which lie within the simulated errors on a Gaussian curve with zero mean and standard deviation  $\sigma = 1$ , and which span the abscissa region  $-3 \leq x \leq 3$ . Next, points on three polylines are computed corresponding to Gaussian curves with means of zero and standard deviations  $\sigma = 0.5$ ,  $\sigma = 1$ , and  $\sigma = 1.5$ . The polylines span the abscissa region  $-10 \leq x \leq 10$ . They are displayed in different colors. One polyline is shown as a continuous line, the other two as dashed lines. Three plots are produced: The first displays only the data points, the second only the polylines, and the third shows the data points together with the polylines. In this way the automatic choice of the scales in the different cases is demonstrated.

# G. Problems, Hints and Solutions, and Programming Problems

## G.1 Problems

### Problem 2.1: Determination of Probabilities through Symmetry Considerations

There are  $n$  students in a classroom. What is the probability for the fact that at least two of them have their birthday on the same day? Solve the problem by working through the following questions:

- (a) What is the number  $N$  of possibilities to distribute the  $n$  birthdays over the year (365 days)?
- (b) How large is the number  $N'$  of possibilities for which all  $n$  birthdays are different?
- (c) How large then is the probability  $P_{\text{diff}}$  that the birthdays are different?
- (d) How large finally is the probability  $P$  that at least two birthdays are not different?

### Problem 2.2: Probability for Non-exclusive Events

The probabilities  $P(A)$ ,  $P(B)$ , and  $P(AB) \neq 0$  for non-exclusive events  $A$  and  $B$  are given. How large is the probability  $P(A + B)$  for the observation of  $A$  or  $B$ ? As an example compute the probability that a playing card which was drawn at random out of a deck of 52 cards is either an ace or a diamond.

### Problem 2.3: Dependent and Independent Events

Are the events  $A$  and  $B$  that a playing card out of a deck is an ace or a diamond independent

- (a) If an ordinary deck of 52 cards is used,
- (b) If a joker is added to the deck?

**Problem 2.4:** Complementary Events

Show that  $\bar{A}$  and  $\bar{B}$  are independent if  $A$  and  $B$  are independent. Use the result of Problem 2.2 to express  $P(\bar{A}\bar{B})$  by  $P(A)$ ,  $P(B)$ , and  $P(AB)$ .

**Problem 2.5:** Probabilities Drawn from Large and Small Populations

A container holds a large number ( $>1000$ ) of coins. They are divided into three types  $A$ ,  $B$ , and  $C$ , which make up 20, 30, and 50% of the total.

- (a) What are the probabilities  $P(A)$ ,  $P(B)$ ,  $P(C)$  of picking a coin of type  $A$ ,  $B$ , or  $C$  if one coin is taken at random? What are the probabilities  $P(AB)$ ,  $P(AC)$ ,  $P(BC)$ ,  $P(AA)$ ,  $P(BB)$ ,  $P(CC)$ ,  $P(2 \text{ identical coins})$ ,  $P(2 \text{ different coins})$  for picking 2 coins?
- (b) What are the probabilities if 10 coins (2 of type  $A$ , 3 of type  $B$ , and 5 of type  $C$ ) are in the container?

**Problem 3.1:** Mean, Variance, and Skewness of a Discrete Distribution

The throwing of a die yields as possible results  $x_i = 1, 2, \dots, 6$ . For an ideally symmetric die one has  $p_i = P(x_i) = 1/6$ ,  $i = 1, 2, \dots, 6$ . Determine the expectation value  $\hat{x}$ , the variance  $\sigma^2(x) = \mu_2$ , and the skewness  $\gamma$  of the distribution,

- (a) For an ideally symmetric die,
- (b) For a die with

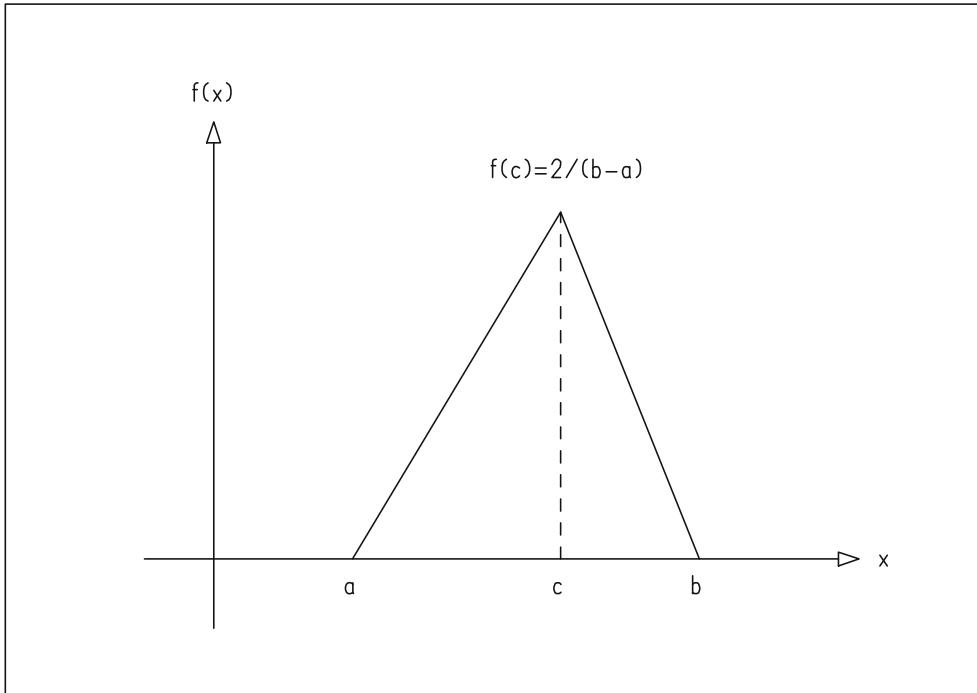
$$\begin{aligned} p_1 &= \frac{1}{6} \quad , & p_2 &= \frac{1}{12} \quad , & p_3 &= \frac{1}{12} \quad , \\ p_4 &= \frac{1}{6} \quad , & p_5 &= \frac{3}{12} \quad , & p_6 &= \frac{3}{12} \quad . \end{aligned}$$

**Problem 3.2:** Mean, Mode, Median, and Variance  
of a Continuous Distribution

Consider the probability density  $f(x)$  of a *triangular distribution* of the form shown in Fig. G.1, given by

$$\begin{aligned} f(x) &= 0 \quad , \quad x < a \quad , \quad x \geq b \quad , \\ f(x) &= \frac{2}{(b-a)(c-a)}(x-a) \quad , \quad a \leq x < c \quad , \\ f(x) &= \frac{2}{(b-a)(b-c)}(b-x) \quad , \quad c \leq x < b \quad . \end{aligned}$$

Determine the mean  $\hat{x}$ , the mode  $x_m$ , the median  $x_{0.5}$ , and the variance  $\sigma^2$  of the distribution. For simplicity choose  $c = 0$  (which corresponds to the substitution of  $x$  by  $x' = x - c$ ). Give explicit results for the symmetric case  $a = -b$  and for the case  $a = -2b$ .



**Fig.G.1:** Triangular probability density.

### Problem 3.3: Transformation of a Single Variable

In Appendix D it is shown that

$$\int_{-\infty}^{\infty} \exp(-x^2/2) dx = \sqrt{2\pi} \quad .$$

Use the transformation  $y = x/\sigma$  to show that

$$\int_{-\infty}^{\infty} \exp(-x^2/2\sigma^2) dx = \sigma\sqrt{2\pi} \quad .$$

### Problem 3.4: Transformation of Several Variables

A “normal distribution of two variables” (see Sect. 5.10) can take the form

$$f(x, y) = \frac{1}{2\pi\sigma_x\sigma_y} \exp\left(-\frac{1}{2}\frac{x^2}{\sigma_x^2} - \frac{1}{2}\frac{y^2}{\sigma_y^2}\right) \quad .$$

- (a) Determine the marginal probability densities  $f(x)$ ,  $f(y)$  by using the results of Problem 3.3.
- (b) Are  $x$  and  $y$  independent?

- (c) Transform the distribution  $f(x, y)$  to the variables

$$u = x \cos \phi + y \sin \phi \quad , \quad v = y \cos \phi - x \sin \phi \quad .$$

(The  $u, v$  coordinate system has the same origin as the  $x, y$  coordinate system, but it is rotated with respect to the latter by an angle  $\phi$ .)

Hint: Show that the transformation is orthogonal and use (3.8.12).

- (d) Show that  $u$  and  $v$  are independent variables only if  $\phi = 0^\circ, 90^\circ, 180^\circ, 270^\circ$  or if  $\sigma_x = \sigma_y = \sigma$ .
- (e) Consider the case  $\sigma_x = \sigma_y = \sigma$ , i.e.,

$$f(x) = \frac{1}{2\pi\sigma^2} \exp\left[-\frac{1}{2\sigma^2}(x^2 + y^2)\right] \quad .$$

Transform the distribution to polar coordinates  $r, \phi$ , determine the marginal probability densities  $g(r)$  and  $g(\phi)$  and show that  $r$  and  $\phi$  are independent.

### Problem 3.5: Error Propagation

The period  $T$  of a pendulum is given by  $T = 2\pi\sqrt{\ell/g}$ . Here  $\ell$  is the length of the pendulum and  $g$  is the gravitational acceleration. Compute  $g$  and  $\Delta g$  using the measured values  $\ell = 99.8$  cm,  $\Delta\ell = 0.3$  cm,  $T = 2.03$  s,  $\Delta T = 0.05$  s and assuming that the measurements of  $\ell$  and  $T$  are uncorrelated.

### Problem 3.6: Covariance and Correlation

We denote the mass and the velocity of an object by  $m$  and  $v$  and their measurement errors by  $\Delta m = \sqrt{\sigma^2(m)}$  and  $\Delta v = \sqrt{\sigma^2(v)}$ . The measurements are assumed to be independent, i.e.,  $\text{cov}(m, v) = 0$ . Furthermore, the relative errors of measurement are known, i.e.,

$$\Delta m/m = a \quad , \quad \Delta v/v = b \quad .$$

- (a) Consider the momentum  $p = mv$  and the kinetic energy  $E = \frac{1}{2}mv^2$  of the object and compute  $\sigma^2(p)$ ,  $\sigma^2(E)$ ,  $\text{cov}(p, E)$ , and the correlation  $\rho(p, E)$ . Discuss  $\rho(p, E)$  for the special cases  $a = 0$  and  $b = 0$ . Hint: Form the vectors  $\mathbf{x} = (m, v)$  and  $\mathbf{y} = (p, E)$ . Then approximate  $\mathbf{y} = \mathbf{y}(\mathbf{x})$  by a linear transformation and finally compute the covariance matrix.
- (b) For the case where the measured values of  $E$ ,  $p$  and the covariance matrix are known, compute the mass  $m$  and its error by error propagation. Use the results from (a) to verify your result. Note that you will obtain the correct result only if  $\text{cov}(p, E)$  is taken into account in the error propagation.

### Problem 5.1: Binomial Distribution

- (a) Prove the recursion formula

$$W_{k+1}^n = \frac{n-k}{k+1} \frac{p}{q} W_k^n .$$

- (b) It may be known for a certain production process that a fraction  $q = 0.2$  of all pieces produced are defective. This means that in 5 pieces produced the expected number of non-defective pieces is  $np = n(1-q) = 5 \cdot 0.8 = 4$ . What is the probability  $P_2$  and the probability  $P_3$  that at most 2 or at most 3 pieces are free from defects? Use relation (a) to simplify the calculation.
- (c) Determine the value  $k_m$  for which the binomial distribution is maximum, i.e.,  $k_m$  is the most probable value of the distribution. Hint: Since  $W_k^n$  is not a function of a continuous variable  $k$ , the maximum cannot be found by looking for a zero in the derivative. Therefore, one has to study finite differences  $W_k^n - W_{k-1}^n$ .
- (d) In Sect. 5.1 the binomial distribution was constructed by considering the random variable  $\mathbf{x} = \sum_{i=1}^n \mathbf{x}_i$ . Here  $\mathbf{x}_i$  was a random variable that took only the values 0 and 1 with the probabilities  $P(\mathbf{x}_i = 1) = p$  and  $P(\mathbf{x}_i = 0) = q$ .

The binomial distribution

$$f(x) = f(k) = W_k^n = \binom{n}{k} p^k q^{n-k}$$

was then obtained by considering in detail the probability to have  $k$  cases of  $\mathbf{x}_i = 1$  in a total of  $n$  observations of the variable  $\mathbf{x}_i$ . Obtain the binomial distribution in a more formal way by constructing the characteristic function  $\varphi_{\mathbf{x}_i}$  of the variable  $\mathbf{x}_i$ . From the  $n$ th power of  $\varphi_{\mathbf{x}_i}$  you obtain the characteristic function of  $\mathbf{x}$ . Hint: Use the binomial theorem (B.6).

### Problem 5.2: Poisson Distribution

In a certain hospital the doctor on duty is called on the average three times per night. The number of calls may be considered to be Poisson distributed. What is the probability for the doctor to have a completely quiet night?

### Problem 5.3: Normal Distribution

The resistance  $R$  of electrical resistors produced by a particular machine may be described by a normal distribution with mean  $R_m$  and standard deviation  $\sigma$ .

The production cost for one resistor is  $C$ , and the price is  $5C$  if  $R = R_0 \pm \Delta_1$ , and  $2C$  if  $R_0 - \Delta_2 < R < R_0 - \Delta_1$  or  $R_0 + \Delta_1 < R < R_0 + \Delta_2$ . Resistors outside these limits cannot be sold.

- (a) Determine the profit  $P$  per resistor produced for  $R_m = R_0$ ,  $\Delta_1 = a_1 R_0$ ,  $\Delta_2 = a_2 R_0$ ,  $\sigma = b R_0$ . Use the distribution function  $\psi_0$ .

- (b) Use Table I.2 to compute the numerical values of  $P$  for  $a_1 = 0.01$ ,  $a_2 = 0.05$ ,  $b = 0.05$ .
- (c) Show that the probability density (5.7.1) has points of inflection (i.e., second derivative equal to zero) at  $x = a \pm b$ .

### Problem 5.4: Multivariate Normal Distribution

A planar  $xy$  coordinate system is used as target. The probability to observe a hit in the plane may be given by the normal distribution of Problem 3.4 (e). Use the result of that problem to determine

- (a) The probability  $P(R)$ , to observe a hit within a given radius  $R$  around the origin,
- (b) The radius  $R$ , within which a hit is observed with a given probability. Compute as a numerical example the value of  $R$  for  $P = 90\%$  and  $\sigma = 1$ .

### Problem 5.5: Convolution

- (a) Prove the relation (5.11.11). Begin with (5.11.9) and use the expression (5.11.10) for  $f_y(y)$ . In the intervals  $0 \leq u < 1$  and  $2 \leq u < 3$  relations (5.11.10a) and (5.11.10b) hold, since in these intervals one always has  $y < 1$  and  $y > 1$ , respectively. In the interval  $1 \leq u < 2$  the resulting distribution  $f(u)$  must be constructed as sum of two integrals of the type (5.11.9) of which each contains one of the possible expression for  $f_y(y)$ . In this case particular care is necessary in the determination of the limits of integration. They are given by the limits of the intervals in which  $u$  and  $f_y(y)$  are defined.
- (b) Prove the relation (5.11.15) by performing the integration (5.11.5) for the case that  $f_x$  and  $f_y$  are normal distributions with means 0 and standard deviations  $\sigma_x$  and  $\sigma_y$ .

### Problem 6.1: Efficiency of Estimators

Let  $x_1, x_2, x_3$  be the elements of a sample from a continuous population with unknown mean  $\hat{x}$ , but known variance  $\sigma^2$ .

- (a) Show that the following quantities are unbiased estimators of  $\hat{x}$ ,

$$\begin{aligned} S_1 &= \frac{1}{4}x_1 + \frac{1}{4}x_2 + \frac{1}{2}x_3 \quad , \\ S_2 &= \frac{1}{5}x_1 + \frac{2}{5}x_2 + \frac{2}{5}x_3 \quad , \\ S_3 &= \frac{1}{6}x_1 + \frac{1}{3}x_2 + \frac{1}{2}x_3 \quad . \end{aligned}$$

*Hint:* It is simple to show that  $S = \sum_{i=1}^n a_i x_i$  is unbiased if  $\sum_{i=1}^n a_i = 1$  holds.

(b) Determine the variances  $\sigma^2(\mathbf{S}_1)$ ,  $\sigma^2(\mathbf{S}_2)$ ,  $\sigma^2(\mathbf{S}_3)$  using (3.8.7) and the assumption that the elements  $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$  are independent.

(c) Show that the arithmetic mean  $\bar{\mathbf{x}} = \frac{1}{3}\mathbf{x}_1 + \frac{1}{3}\mathbf{x}_2 + \frac{1}{3}\mathbf{x}_3$  has the smallest variance of all estimators of the type  $\mathbf{S} = \sum_{i=1}^3 a_i \mathbf{x}_i$  fulfilling the requirement  $\sum_{i=1}^3 a_i = 1$ .

*Hint:* Minimize the variance of  $\mathbf{S} = a_1 \mathbf{x}_1 + a_2 \mathbf{x}_2 + (1 - a_1 - a_2) \mathbf{x}_3$  with respect to  $a_1$  and  $a_2$ . Compute this variance and compare it with the variances which you found (b).

### Problem 6.2: Sample Mean and Sample Variance

Compute the sample mean  $\bar{\mathbf{x}}$ , the sample variance  $\mathbf{s}^2$ , and an estimate for the variance of the sample mean  $\mathbf{s}_{\bar{\mathbf{x}}}^2 = (1/n)\mathbf{s}^2$  for the following sample:

$$18, 21, 23, 19, 20, 21, 20, 19, 20, 17.$$

Use the method of Example 6.1.

### Problem 6.3: Samples from a Partitioned Population

An opinion poll is performed on an upcoming election. In our (artificially constructed) example the population is partitioned into three subpopulations, and from each subpopulation a preliminary sample of size 10 is drawn. Each element of the sample can have the values 0 (vote for party A) or 1 (vote for party B). The samples are

$$\begin{aligned} i = 1 \quad (p_i = 0.1) : x_{ij} &= 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, \\ i = 2 \quad (p_i = 0.7) : x_{ij} &= 0, 0, 1, 1, 0, 1, 0, 1, 1, 0, \\ i = 3 \quad (p_i = 0.2) : x_{ij} &= 0, 1, 1, 1, 1, 0, 1, 1, 1, 0, \end{aligned}$$

- (a) Use these samples to form an estimator for the result of the election and its variance  $\mathbf{s}_{\bar{\mathbf{x}}}^2$ . Does this result show a clear advantage for one party?
- (b) Use these samples to determine for a much larger sample of size  $n$  the sizes  $n_i$  of the subsamples in such a way that  $\bar{\mathbf{x}}$  has the smallest variance (cf. Example 6.6).

### Problem 6.4: $\chi^2$ -distribution

- (a) Determine the skewness  $\gamma = \mu_3/\sigma^3$  of the  $\chi^2$ -distribution using (5.5.7). Begin by expressing  $\mu_3$  by  $\lambda_3$ ,  $\widehat{x}$ , and  $E(\mathbf{x}^2)$ .
- (b) Show that  $\gamma \rightarrow 0$  for  $n \rightarrow \infty$ .

**Problem 6.5:** Histogram

Construct a histogram from the following measured values:

26.02, 27.13, 24.78, 26.19, 22.57, 25.16, 24.39, 22.73, 25.35, 26.13, 23.15, 26.29, 24.89, 23.76, 26.04, 22.03, 27.48, 25.42, 24.27, 26.58, 27.17, 24.22, 21.86, 26.53, 27.09, 24.73, 26.12, 28.04, 22.78, 25.36.

Use a bin size of  $\Delta x = 1$ .

*Hint:* For each value draw a cross of width  $\Delta x$  and height 1. In this way you do not need to order the measured values, since each cross is drawn within a bin on top of a preceding cross. In this way the bars of the histogram grow while drawing.

**Problem 7.1:** Maximum-Likelihood Estimates

- (a) Suppose it is known that a random variable  $x$  follows the uniform distribution  $f(x) = 1/b$  for  $0 < x < b$ . The parameter  $b$  is to be estimated from a sample. Show that  $S = \tilde{b} = x_{\max}$  is the maximum-likelihood estimator of  $b$ . (Hint: This result cannot be obtained by differentiation, but from a simple consideration about the likelihood function).
- (b) Write down the likelihood equations for the two parameters  $a$  and  $\Gamma$  of the Lorentz distribution (see Example 3.5). Show that these do not necessarily have unique solutions. You can, however, easily convince yourself that for  $|x^{(j)} - a| \ll \Gamma$  the arithmetic mean  $\bar{x}$  is an estimator of  $a$ .

**Problem 7.2:** Information

- (a) Determine the information  $I(\lambda)$  of a sample of size  $N$  that was obtained from a normal distribution of known variance  $\sigma^2$  but unknown mean  $\lambda = a$ .
- (b) Determine the information  $I(\lambda)$  of a sample of size  $N$  which was drawn from a normal distribution of known mean  $a$  but unknown variance  $\lambda = \sigma^2$ . Show that the maximum-likelihood estimator of  $\sigma^2$  is given by

$$S = \frac{1}{N} \sum_{j=1}^N (x^{(j)} - a)^2$$

and that the estimator is unbiased, i.e.,  $B(S) = E(S) - \lambda = 0$ .

**Problem 7.3:** Variance of an Estimator

- (a) Use the information inequality to obtain a lower limit on the variance of the estimator of the mean in Problem 7.2 (a). Show that this limit is equal to the minimum variance that was determined in Problem 6.1 (c).
- (b) Use the information inequality to obtain a lower limit on the variance of  $S$  in Problem 7.2 (b).
- (c) Show using Eq. (7.3.12) that  $S$  is a minimum variance estimator with the same variance the lower limit found in (b).

**Problem 8.1:** *F*-Test

Two samples are given:

$$(1) \quad 21, 19, 14, 27, 25, 23, 22, 18, 21, \quad (N_1 = 9), \\ (2) \quad 16, 24, 22, 21, 25, 21, 18, \quad (N_2 = 7).$$

Does sample (2) have a smaller variance than sample (1) at a significance level of  $\alpha = 5\%$ ?

**Problem 8.2:** Student's Test

Test the hypothesis that the 30 measurements of Problem 6.5 were drawn from a population with mean 25.5. Use a level of significance of  $\alpha = 10\%$ . Assume that the population is normally distributed.

**Problem 8.3:**  $\chi^2$ -Test for Variance

Use the likelihood-ratio method to construct a test of the hypothesis  $H_0(\sigma^2 = \sigma_0^2)$  that a sample stems from a normal distribution with unknown mean  $a$  and unknown variance  $\sigma_0^2$ . The parameters are  $\lambda = (a, \sigma)$ . In  $\omega$  one has  $\tilde{\lambda}^{(\omega)} = (\bar{x}, \sigma_0)$ , in  $\Omega$ :  $\tilde{\lambda}^{(\Omega)} = (\bar{x}, s)$ .

- (a) Form the likelihood ratio  $T$ .
- (b) Show that instead of  $T$ , the test statistic  $T' = Ns'^2/\sigma_0^2$  can be used as well.
- (c) Show that  $T'$  follows a  $\chi^2$ -distribution with  $N - 1$  degrees of freedom so that the test can be performed with the help of Table I.7.

**Problem 8.4:**  $\chi^2$ -Test of Goodness-of-Fit

- (a) Determine the mean  $\tilde{a}$  and the variance  $\tilde{\sigma}^2$  for the histogram of Fig. 6.1b, i.e.,

|       |     |     |     |     |     |     |     |     |     |     |
|-------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| $x_k$ | 193 | 195 | 197 | 199 | 201 | 203 | 205 | 207 | 209 | 211 |
| $n_k$ | 1   | 2   | 9   | 12  | 23  | 25  | 11  | 9   | 6   | 2   |

Use the result of Example 7.8 to construct the estimators. Give the estimators explicitly as functions of  $n_k$  and  $x_k$ .

- (b) Perform (at a significance level of  $\alpha = 10\%$ ) a  $\chi^2$ -test on the goodness-of-fit of a normal distribution with mean  $\tilde{a}$  and variance  $\tilde{\sigma}^2$  to the histogram. Use only those bins of the histogram for which  $np_k \geq 4$ . Determine  $p_k$  from the difference of two entries in Table I.2. Give a formula for  $p_k$  as function of  $x_k$ ,  $\Delta x$ ,  $\tilde{a}$ ,  $\tilde{\sigma}^2$ , and  $\psi_0$ . Construct a table for the computation of  $\chi^2$  containing columns for  $x_k$ ,  $n_k$ ,  $np_k$ , and  $(n_k - np_k)^2/np_k$ .

**Problem 8.5:** Contingency Table

- (a) In an immunology experiment [taken from SOKAL and ROHLF, Biometry (Freeman, San Francisco, 1969)] the effect of an antiserum on a particular type of bacteria is studied. 57 mice received a certain dose of bacteria and antiserum, whereas 54 mice received bacteria only. After some time the mice of both groups were counted and the following contingency table constructed.

|                           | Dead | Alive | Sum       |
|---------------------------|------|-------|-----------|
| Bacteria and<br>antiserum | 13   | 44    | 57        |
| Only bacteria             | 25   | 29    | 54        |
| Sum                       | 38   | 73    | Total 111 |

Test (at a significance level of  $\alpha = 10\%$ ) the hypothesis that the antiserum has no influence on the survival probability.

- (b) In the computation of  $\chi^2$  in (a) you will have noticed that the numerators in (8.8.1) all have the same value. Show that generally for  $2 \times 2$  contingency tables the following relation holds,

$$n_{ij} - n \tilde{p}_i \tilde{q}_j = (-1)^{i+j} \frac{1}{n} (n_{11}n_{22} - n_{12}n_{21}) \quad .$$

**G.2 Hints and Solutions****Problem 2.1**

$$(a) N = 365^n .$$

$$(b) N' = 365 \cdot 364 \cdots (365 - n + 1) .$$

$$\begin{aligned} (c) \quad P_{\text{diff}} = N'/N &= \frac{364}{365} \cdots \frac{365 - n + 1}{365} \\ &= \left(1 - \frac{1}{365}\right) \left(1 - \frac{2}{365}\right) \cdots \left(1 - \frac{n-1}{365}\right) . \end{aligned}$$

$$(d) P = 1 - P_{\text{diff}} .$$

Putting in numbers one obtains  $P \approx 0.5$  for  $n = 23$  and  $P \approx 0.99$  for  $n = 57$ .

**Problem 2.2**

$$P(A + B) = P(A) + P(B) - P(AB) \quad ,$$

$$\begin{aligned} P(\text{ace or diamond}) &= P(\text{ace}) + P(\text{diamond}) - P(\text{ace} + \text{diamond}) \\ &= \frac{4}{52} + \frac{13}{52} - \frac{1}{52} = \frac{4}{13} \quad . \end{aligned}$$

**Problem 2.3**

(a)  $P(A) = \frac{4}{52}$ ,  $P(B) = \frac{13}{52}$ ,  $P(AB) = \frac{1}{52}$ , i.e.,  $P(AB) = P(A)P(B)$ .

(b)  $P(A) = \frac{4}{53}$ ,  $P(B) = \frac{13}{53}$ ,  $P(AB) = \frac{1}{53}$ , i.e.,  $P(AB) \neq P(A)P(B)$ .

**Problem 2.4**

$$P(\overline{AB}) = 1 - P(A + B) = 1 - P(A) - P(B) + P(AB) .$$

For  $A$  and  $B$  independent one has  $P(AB) = P(A)P(B)$ . Therefore,

$$\begin{aligned} P(\overline{AB}) &= 1 - P(A) - P(B) + P(A)P(B) = (1 - P(A))(1 - P(B)) \\ &= P(\bar{A})P(\bar{B}) . \end{aligned}$$

**Problem 2.5**

(a)  $P(A) = 0.2$ ,  $P(B) = 0.3$ ,  $P(C) = 0.5$ ,  
 $P(AB) = 2 \cdot 0.2 \cdot 0.3$ ,  $P(AC) = 2 \cdot 0.2 \cdot 0.5$ ,  
 $P(BC) = 2 \cdot 0.3 \cdot 0.5$ ,  $P(AA) = 0.2^2$ ,  $P(BB) = 0.3^2$ ,  
 $P(CC) = 0.5^2$ .

(b)  $P(A) = 2/10 = 0.2$ ,  $P(B) = 3/10 = 0.3$ ,  $P(C) = 5/10 = 0.5$ ,

$$P(AB) = \frac{2}{10} \cdot \frac{3}{9} + \frac{3}{10} \cdot \frac{2}{9}, \quad P(AC) = \frac{2}{10} \cdot \frac{5}{9} + \frac{5}{10} \cdot \frac{2}{9},$$

$$P(BC) = \frac{3}{10} \cdot \frac{5}{9} + \frac{5}{10} \cdot \frac{3}{9},$$

$$P(AA) = \frac{2}{10} \cdot \frac{1}{9}, \quad P(BB) = \frac{3}{10} \cdot \frac{2}{9}, \quad P(CC) = \frac{5}{10} \cdot \frac{4}{9}.$$

For (a) and (b) it holds that

$$\begin{aligned} P(2 \text{ identical coins}) &= P(AA) + P(BB) + P(CC), \\ P(2 \text{ different coins}) &= P(AB) + P(AC) + P(BC) \\ &= 1 - P(2 \text{ identical coins}). \end{aligned}$$

**Problem 3.1**

$$\begin{aligned}
 \text{(a)} \quad \hat{x} &= \sum_{i=1}^6 x_i p_i = \frac{1}{6} \sum_{i=1}^6 i = \frac{21}{6} = 3.5 \quad , \\
 \sigma^2(x) &= \sum_{i=1}^6 (x_i - \hat{x})^2 p_i \\
 &= \frac{1}{6} (2.5^2 + 1.5^2 + 0.5^2 + 0.5^2 + 1.5^2 + 2.5^2) \\
 &= \frac{2}{6} (6.25 + 2.25 + 0.25) = 2.92 \quad , \\
 \mu_3 &= \sum_{i=1}^6 (x_i - \hat{x})^3 p_i = \frac{1}{6} \sum_{i=1}^6 (i - 3.5)^3 = 0 \quad , \\
 \gamma &= \mu_3 / \sigma^3 = 0 \quad .
 \end{aligned}$$

$$\begin{aligned}
 \text{(b)} \quad \hat{x} &= \frac{1}{12} (2 + 2 + 3 + 8 + 15 + 18) = 4 \quad , \\
 \sigma^2(x) &= \frac{1}{12} (2 \cdot 3^2 + 1 \cdot 2^2 + 1 \cdot 1^2 + 3 \cdot 1^2 + 3 \cdot 2^2) = \frac{38}{12} = 3.167 \quad , \\
 \mu_3 &= \frac{1}{12} (-2 \cdot 3^3 - 1 \cdot 2^3 - 1 \cdot 1^3 + 3 \cdot 1^3 + 3 \cdot 2^3) = -3 \\
 \gamma &= \mu_3 / \sigma^3 = -3 / 3.167^{3/2} = -0.533 \quad .
 \end{aligned}$$

**Problem 3.2**

$$\text{For } a = -b: \quad \hat{x} = 0 \quad , \quad x_{0.5} = 0 \quad , \quad \sigma^2(x) = \frac{b^2}{6} \quad .$$

$$\text{For } a = -2b: \quad \hat{x} = -\frac{b}{3} = -0.33b \quad , \quad x_{0.5} = b(\sqrt{3} - 2) = -0.27b \quad , \quad \sigma^2(x) = \frac{7}{18}b^2 \quad .$$

**Problem 3.3**

$$g(y) = \exp\left(-\frac{y^2}{2}\right) \quad , \quad y(x) = \frac{x}{\sigma} \quad ; \quad f(x) = \frac{dy(x)}{dx} g(y(x)) = \frac{1}{\sigma} \exp\left(-\frac{x^2}{2\sigma^2}\right) \quad .$$

**Problem 3.4**

$$\text{(a)} \quad f_x(x) = \frac{1}{\sqrt{2\pi}\sigma_x} \exp\left(-\frac{1}{2}\frac{x^2}{\sigma_x^2}\right), \quad f_y(y) = \frac{1}{\sqrt{2\pi}\sigma_y} \exp\left(-\frac{1}{2}\frac{y^2}{\sigma_y^2}\right).$$

(b) Yes, since (3.4.6) is fulfilled.

(c) The transformation can be written in the form

$$\begin{pmatrix} u \\ v \end{pmatrix} = R \begin{pmatrix} x \\ y \end{pmatrix} \quad , \quad R = \begin{pmatrix} \cos \phi & \sin \phi \\ -\sin \phi & \cos \phi \end{pmatrix} \quad .$$

It is orthogonal, since  $R^T R = I$ . Therefore,

$$\begin{aligned} g(u, v) &= f(x, y) = \frac{1}{2\pi\sigma_x\sigma_y} \exp\left(-\frac{x^2}{2\sigma_x^2} - \frac{y^2}{2\sigma_y^2}\right) \\ &= \frac{1}{2\pi\sigma_x\sigma_y} \exp\left(-\frac{(u \cos \phi - v \sin \phi)^2}{2\sigma_x^2} - \frac{(u \sin \phi + v \cos \phi)^2}{2\sigma_y^2}\right) \\ &= \frac{1}{2\pi\sigma_x\sigma_y} \exp\left(-\frac{u^2 \cos^2 \phi + v^2 \sin^2 \phi - 2uv \cos \phi \sin \phi}{2\sigma_x^2}\right. \\ &\quad \left.- \frac{u^2 \sin^2 \phi + v^2 \cos^2 \phi + 2uv \cos \phi \sin \phi}{2\sigma_y^2}\right) . \end{aligned}$$

- (d) For  $\phi = 90^\circ : \cos \phi = 0, \sin \phi = 1$  etc. the expression  $g(u, v)$  factorizes, i.e.,  $g(u, v) = g_u(u)g_v(v)$ .

For  $\sigma_x = \sigma_y = \sigma$ :

$$g(u, v) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{u^2 + v^2}{2\sigma^2}\right) = g(u)g(v) .$$

$$\begin{aligned} (e) \quad J &= \begin{vmatrix} \frac{\partial x}{\partial r} & \frac{\partial y}{\partial r} \\ \frac{\partial x}{\partial \phi} & \frac{\partial y}{\partial \phi} \end{vmatrix} = \begin{vmatrix} \cos \phi & \sin \phi \\ -r \sin \phi & r \cos \phi \end{vmatrix} = r , \\ g(r, \phi) &= rf(x, y) = \frac{r}{2\pi\sigma^2} \exp\left(-\frac{x^2 + y^2}{2\sigma^2}\right) \\ &= \frac{r}{2\pi\sigma^2} \exp\left(-\frac{r^2}{2\sigma^2}\right) ; \\ g_r(r) &= \int_0^{2x} g(r, \phi) d\phi = \frac{r}{\sigma^2} \exp\left(-\frac{r^2}{2\sigma^2}\right) , \\ g_\phi(\phi) &= \frac{1}{2\pi\sigma^2} \int_0^\infty r \exp\left(-\frac{r^2}{2\sigma^2}\right) dr \\ &= \frac{1}{4\pi\sigma^2} \int_0^\infty \exp\left(-\frac{u}{2\sigma^2}\right) du \\ &= -\frac{1}{2\pi} \left[ \exp\left(-\frac{u}{2\sigma^2}\right) \right]_0^\infty = \frac{1}{2\pi} ; \\ g(r, \phi) &= g_r(r)g_\phi(\phi) , \end{aligned}$$

therefore,  $r$  and  $\phi$  are independent.

**Problem 3.5**

$$\begin{aligned}
 g &= 4\pi^2 \frac{\ell}{T^2} = 4\pi^2 \frac{99.8}{2.03^2} \text{ cm s}^{-2} = 956.09 \text{ cm s}^{-2} , \\
 \frac{\partial g}{\partial \ell} &= \frac{4\pi^2}{T^2} = 9.58 \text{ s}^{-2} , \quad \frac{\partial g}{\partial T} = -\frac{8\pi^2 \ell}{T^3} = -942 \text{ cm s}^{-3} , \\
 (\Delta g)^2 &= \left( \frac{\partial g}{\partial \ell} \right)^2 (\Delta \ell)^2 + \left( \frac{\partial g}{\partial T} \right)^2 \Delta T^2 = 2226 \text{ cm}^2 \text{s}^{-4} , \\
 \Delta g &= 47.19 \text{ cm s}^{-2} .
 \end{aligned}$$

**Problem 3.6**

(a)  $y = Tx$ ,

$$\begin{aligned}
 T &= \begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} \end{pmatrix} = \begin{pmatrix} \frac{\partial p}{\partial m} & \frac{\partial p}{\partial v} \\ \frac{\partial E}{\partial m} & \frac{\partial E}{\partial v} \end{pmatrix} = \begin{pmatrix} v & m \\ \frac{1}{2}v^2 & mv \end{pmatrix} , \\
 C_y &= TC_xT^T = \\
 &= \begin{pmatrix} v & m \\ \frac{1}{2}v^2 & mv \end{pmatrix} \begin{pmatrix} a^2m^2 & 0 \\ 0 & b^2v^2 \end{pmatrix} \begin{pmatrix} v & \frac{1}{2}v^2 \\ m & mv \end{pmatrix} \\
 &= \begin{pmatrix} (a^2+b^2)m^2v^2 & (\frac{1}{2}a^2+b^2)m^2v^3 \\ (\frac{1}{2}a^2+b^2)m^2v^3 & (\frac{1}{4}a^2+b^2)m^2v^4 \end{pmatrix} , \\
 \rho(p, E) &= \frac{\text{cov}(p, E)}{\sigma(p)\sigma(E)} = \frac{(\frac{1}{2}a^2+b^2)m^2v^3}{\sqrt{(a^2+b^2)m^2v^2}\sqrt{(\frac{1}{4}a^2+b^2)m^2v^4}} \\
 &= \frac{\frac{1}{2}a^2+b^2}{\sqrt{(a^2+b^2)(\frac{1}{4}a^2+b^2)}} .
 \end{aligned}$$

For  $a = 0$  or  $b = 0$ :  $\rho = 1$ . (In this case either  $m$  or  $v$  are completely determined. Therefore, there is a strict relation between  $E$  and  $p$ .) If, however,  $a, b \neq 0$ , one has  $\rho \neq 1$ , e.g., for  $a = b$  one obtains  $\rho = 3/\sqrt{10}$ .

(b)  $m = \frac{1}{2}p^2/E$ ,  $v = E/2p$ ,

$$\begin{aligned}
 C_y &= \begin{pmatrix} (a^2+b^2)p^2 & (a^2+2b^2)Ep \\ (a^2+2b^2)Ep & (a^2+4b^2)E^2 \end{pmatrix} , \\
 m &= Ty , \\
 T &= \left( \frac{\partial m}{\partial y_1}, \frac{\partial m}{\partial y_2} \right) = \left( \frac{\partial m}{\partial p}, \frac{\partial m}{\partial E} \right) = \left( \frac{p}{E}, -\frac{p^2}{2E^2} \right) , \\
 C_m &= \sigma^2(m) = TC_yT^T \\
 &= \left( \frac{p}{E}, -\frac{p^2}{2E^2} \right) \begin{pmatrix} (a^2+b^2)p^2 & (a^2+2b^2)Ep \\ (a^2+2b^2)Ep & (a^2+4b^2)E^2 \end{pmatrix} \begin{pmatrix} \frac{p}{E} \\ -\frac{p^2}{2E^2} \end{pmatrix} \\
 &= a^2 \frac{p^4}{4E^2} = a^2 m^2 .
 \end{aligned}$$

**Problem 5.1**

$$(a) \quad W_{k+1}^n = \binom{n}{k+1} p^{k+1} q^{n-k-1} = \frac{n!}{(k+1)!(n-k-1)!} p^k q^{n-k} \frac{p}{q}$$

$$= \frac{n!}{k!(n-k)!} \frac{n-k}{k+1} p^k q^{n-k} \frac{p}{q} = W_k^n \frac{n-k}{k+1} \frac{p}{q}.$$

$$(b) \quad W_3^5 = \binom{5}{3} 0.8^3 \cdot 0.2^2 = 10 \cdot 0.512 \cdot 0.04 = 0.2048,$$

$$W_4^5 = W_3^5 \cdot \frac{2}{4} \cdot \frac{0.8}{0.2} = 0.2048 \cdot 2 = 0.4096,$$

$$W_5^5 = W_4^5 \cdot \frac{1}{5} \cdot \frac{0.8}{0.2} = 0.4096 \cdot 0.8 = 0.32768,$$

$$P_3 = 1 - W_4^5 - W_5^5 = 0.26272,$$

$$P_2 = P_3 - W_3^5 = 0.05792.$$

(c) Using the result of (a) we obtain

$$W_k^n - W_{k-1}^n = W_{k-1}^n \left( \frac{n-k+1}{k} \frac{p}{q} - 1 \right).$$

Thus the probability  $W_k^n$  increases as long as the expression in brackets is positive, i.e.,

$$\frac{(n-k+1)p}{kq} - 1 > 0.$$

Since  $k$  and  $q$  are positive we have

$$(n-k+1)p > kq = k(1-p), \quad k < (n+1)p.$$

The most probable value  $k_m$  is the largest value of  $k$  for which this inequality holds.

$$(d) \quad \varphi_{X_i}(t) = E\{e^{itX_i}\} = qe^{it \cdot 0} + pe^{it} = q + pe^{it};$$

$$\begin{aligned} \varphi_X = (q + pe^{it})^n &= \sum_{k=0}^m \binom{n}{k} q^{n-k} p^k e^{itk} = \sum_{k=0}^m f(k) e^{itk} = E\{e^{itk}\}, \\ f(k) &= W_k^n. \end{aligned}$$

**Problem 5.2**

For  $\lambda = 3$  one has  $f(0) = \frac{\lambda^0}{0!} e^{-\lambda} = e^{-\lambda} = 0.0498 \approx 0.05 = 5\%$ .

**Problem 5.3**

(a) The fraction

$$f_r = 2\psi_0\left(\frac{(R_0 - a_2 R_0) - R_0}{b R_0}\right) = 2\psi_0\left(-\frac{a_2}{b}\right)$$

is rejected, since  $R < R_0 - \Delta_2$  or  $R > R_0 + \Delta_2$ . Correspondingly the fractions

$$f_2 = 2\psi_0\left(-\frac{a_1}{b}\right) - f_r \quad \text{and} \quad f_5 = 1 - f_2 - f_r$$

give prices  $2C$  and  $5C$ , respectively. Therefore,

$$\begin{aligned} P &= 2f_2C + 5f_5C - C = C\{2f_2 + 5 - 5f_2 - 5f_r - 1\} \\ &= C\{4 - 3f_2 - 5f_r\} \\ &= C\left\{4 - 6\psi_0\left(-\frac{a_1}{b}\right) + 3f_r - 5f_r\right\} \\ &= C\left\{4 - 6\psi_0\left(-\frac{a_1}{b}\right) - 4\psi_0\left(-\frac{a_2}{b}\right)\right\}. \end{aligned}$$

$$(b) \psi_0(-0.2) = 0.421 ; \quad \psi_0(-1) = 0.159 ,$$

$$\text{i.e., } P = C\{4 - 2.526 - 0.636\} = 0.838C .$$

$$(c) \frac{d^2f}{dx^2} = \frac{1}{\sqrt{2\pi b^3}} \exp(-(x-a)^2/2b^2) \{(x-a)^2/b^2 - 1\} = 0$$

is fulfilled if the expression in the last set of brackets vanishes.

**Problem 5.4**

$$\begin{aligned} (a) \quad P(R) &= \int_0^R g(r) dr = \frac{1}{\sigma^2} \int_0^R r e^{-r^2/2\sigma^2} dr \\ &= \frac{1}{2\sigma^2} \int_0^{R^2} e^{-u/2\sigma^2} du \\ &= \left[ e^{-u/2\sigma^2} \right]_{R^2}^0 = 1 - e^{-R^2/2\sigma^2}. \end{aligned}$$

$$(b) \quad 1 - P = \exp(-R^2/2\sigma^2) , \quad R^2/2\sigma^2 = -\ln(1 - P) .$$

For  $\sigma = 1$ ,  $P = 0.9$  one obtains

$$R = \sqrt{-2 \ln 0.1} = \sqrt{4.61} = 2.15 .$$

**Problem 5.5**(a)  $0 \leq u < 1$ :

$$f(u) = \int_{u-1}^u f_1(y) dy = \int_0^u y dy = \frac{1}{2}u^2 .$$

 $1 \leq u < 2$ :

$$\begin{aligned}
f(u) &= \int_{u-1}^u f_1(y) dy + \int_{u-1}^u f_2(y) dy \\
&= \int_{u-1}^1 y dy + \int_1^u (2-y) dy \\
&= \frac{1}{2}(1-(u-1)^2) + \frac{1}{2}(1-(2-u)^2) \\
&= \frac{1}{2}(-3+6u-2u^2) .
\end{aligned}$$

$2 \leq u < 3$ :

$$f(u) = \int_{u-1}^u f_2(y) dy = \int_{u-1}^2 (2-y) dy = - \int_{3-u}^0 z dz = \frac{1}{2}(3-u)^2 .$$

$$(b) \quad f(u) = \frac{1}{2\pi\sigma_x\sigma_y} \int_{-\infty}^{\infty} \exp\left(-\frac{x^2}{2\sigma_x^2} - \frac{(u-x)^2}{2\sigma_y^2}\right) dx .$$

Completing the square yields for the exponent

$$-\frac{\sigma^2}{2\sigma_x^2\sigma_y^2} \left\{ \left( x - \frac{\sigma_x^2}{\sigma^2}u \right)^2 - \frac{\sigma_x^4}{\sigma^4}u^2 + \frac{\sigma_x^2}{\sigma^2}u^2 \right\} .$$

With the change of variables  $v = (\sigma/\sigma_x\sigma_y)(x - \sigma_x^2u/\sigma^2)$  we obtain

$$\begin{aligned}
f(u) &= \frac{1}{2\pi\sigma_x\sigma_y} \exp\left(\frac{\sigma_x^4 - \sigma^2\sigma_x^2}{2\sigma^2\sigma_x^2\sigma_y^2}u^2\right) \frac{\sigma_x\sigma_y}{\sigma} \\
&\quad \times \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}v^2\right) dv \\
&= \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{u^2}{2\sigma^2}\right) ,
\end{aligned}$$

since  $\sigma_x^4 = \sigma_x^2(\sigma^2 - \sigma_y^2)$ .

### Problem 6.1

- (a)  $E(\mathbf{S}) = E\{\sum a_i \mathbf{x}_i\} = \sum a_i E(\mathbf{x}_i) = \hat{x} \sum a_i = \hat{x}$   
if and only if  
 $\sum a_i = 1$ .

$$\begin{aligned}
(b) \quad \sigma^2 &= \sum \left( \frac{\partial \mathbf{S}}{\partial x_i} \right)^2 \sigma^2 = \sigma^2 \sum a_i^2 , \\
\sigma^2(\mathbf{S}_1) &= \sigma^2 \left( \frac{1}{16} + \frac{1}{16} + \frac{1}{4} \right) = \sigma^2 \cdot \frac{3}{8} = 0.375\sigma^2 , \\
\sigma^2(\mathbf{S}_2) &= \sigma^2 \left( \frac{1}{25} + \frac{4}{25} + \frac{4}{25} \right) = \sigma^2 \cdot \frac{9}{25} = 0.360\sigma^2 , \\
\sigma^2(\mathbf{S}_3) &= \sigma^2 \left( \frac{1}{36} + \frac{1}{9} + \frac{1}{4} \right) = \sigma^2 \cdot \frac{7}{18} = 0.389\sigma^2 .
\end{aligned}$$

$$\begin{aligned}
 (c) \quad \sigma^2(S) &= [a_1^2 + a_2^2 + (1 - (a_1 + a_2))^2] \sigma^2 , \\
 \frac{\partial \sigma^2(S)}{\partial a_1} &= [2a_1 - 2(1 - (a_1 + a_2))] \sigma^2 = 0 , \\
 \frac{\partial \sigma^2(S)}{\partial a_2} &= [2a_2 - 2(1 - (a_1 + a_2))] \sigma^2 = 0 , \\
 a_1 &= 1 - (a_1 + a_2) , \\
 a_2 &= 1 - (a_1 + a_2) , \quad a_1 = a_2 = \frac{1}{3} ; \\
 \sigma^2(S) &= \frac{1}{3} \sigma^2 = 0.333 \sigma^2 .
 \end{aligned}$$

### Problem 6.2

$$\begin{aligned}
 a &= 20 , \\
 \Delta &= \frac{1}{n} \sum \delta_i \\
 &= \frac{1}{10} (-2 + 1 + 3 - 1 + 0 + 1 + 0 - 1 + 0 - 3) = -0.2 , \\
 \bar{x} &= 19.8 , \\
 s &= \frac{1}{9} (1.8^2 + 1.2^2 + 3.2^2 + 0.8^2 + 0.2^2 + \\
 &\quad + 1.2^2 + 0.2^2 + 0.8^2 + 0.2^2 + 2.8^2) \\
 &= \frac{1}{9} \cdot 25.60 = 2.84 , \\
 s_x^2 &= 0.284 . \quad \text{Therefore } \bar{x} = 19.8 \pm 0.53 .
 \end{aligned}$$

### Problem 6.3

$$\begin{aligned}
 (a) \quad \bar{x}_1 &= 0.2, \bar{x}_2 = 0.5, \bar{x}_3 = 0.7 , \\
 \tilde{x} &= 0.02 + 0.35 + 0.14 = 0.51 , \\
 s_1^2 &= \frac{1}{9} (2 - 10 \cdot 0.2^2) = 0.178 , \\
 s_2^2 &= \frac{1}{9} (5 - 10 \cdot 0.5^2) = 0.278 , \\
 s_3^2 &= \frac{1}{9} (7 - 10 \cdot 0.7^2) = 0.233 , \\
 s_{\tilde{x}}^2 &= \frac{0.1^2}{10} 0.178 + \frac{0.7^2}{10} 0.278 + \frac{0.2^2}{10} 0.233 \\
 &= 0.001 \cdot 0.178 + 0.049 \cdot 0.278 + 0.004 \cdot 0.233 \\
 &= 0.0147 ,
 \end{aligned}$$

$$s_{\tilde{x}} = \sqrt{s_{\tilde{x}}^2} = 0.12 .$$

The result  $\tilde{x} = 0.51 \pm 0.12$  does not favor any one party significantly.

$$\begin{aligned} (b) \quad s_1 &= 0.422, s_2 = 0.527, s_3 = 0.483 , \\ p_1 s_1 &= 0.0422, p_2 s_2 = 0.369, p_3 s_3 = 0.0966 , \\ \sum p_i s_i &= 0.508 , \\ \frac{n_1}{n} &= 0.083, \frac{n_2}{n} = 0.726, \frac{n_3}{n} = 0.190 . \end{aligned}$$

### Problem 6.4

(a) For simplicity we write  $\chi^2 = u$ .

$$\begin{aligned} \mu_3 &= E\{(\mathbf{u} - \hat{\mathbf{u}})^3\} = E\{\mathbf{u}^3 - 3\mathbf{u}^2\hat{\mathbf{u}} + 3\mathbf{u}\hat{\mathbf{u}}^2 - \hat{\mathbf{u}}^3\} \\ &= E(\mathbf{u}^3) - 3\hat{\mathbf{u}}E(\mathbf{u}^2) + 3\hat{\mathbf{u}}^2E(\mathbf{u}) - \hat{\mathbf{u}}^3 \\ &= \lambda_3 - 3\hat{\mathbf{u}}E(\mathbf{u}^2) + 2\hat{\mathbf{u}}^3 , \\ E(\mathbf{u}^3) &= \lambda_3 = \frac{1}{i^3} \varphi'''(0) \\ &= i\varphi'''(0) = i(-\lambda)(-\lambda-1)(-\lambda-2)(-2i)^3 \\ &= 8\lambda(\lambda+1)(\lambda+2) = 8\lambda^3 + 24\lambda^2 + 16\lambda , \\ 3\hat{\mathbf{u}}E(\mathbf{u}^2) &= 6\lambda(4\lambda^2 + 4\lambda) = 24\lambda^3 + 24\lambda^2 , \\ 2\hat{\mathbf{u}}^3 &= 2 \cdot (2\lambda)^3 = 16\lambda^3 , \\ \mu_3 &= 16\lambda , \\ \gamma &= \mu_3/\sigma^3 = 16\lambda/8\lambda^{3/2} = 2\lambda^{-1/2} . \end{aligned}$$

(b) Obviously  $\gamma = 2/\sqrt{\lambda} \rightarrow 0$  holds for  $\lambda = \frac{1}{2}n \rightarrow \infty$ .

### Problem 6.5

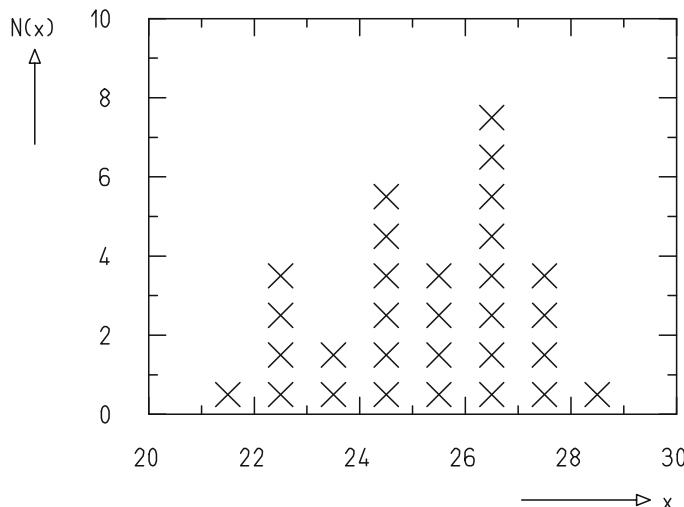
See Fig. G.2.

### Problem 7.1

$$(a) \quad L = \prod_{j=1}^N \frac{1}{b} = \frac{1}{b^N} ; \quad \text{obviously one has } L = \max \text{ for } b = x_{\max} .$$

(b)  $\lambda_1 = a, \lambda_2 = \Gamma$ ,

$$\begin{aligned} L &= \prod_{j=1}^N \frac{2}{\pi \lambda_2} \frac{\lambda_2^2}{4(x^{(j)} - \lambda_1)^2 + \lambda_2^2} \\ &= \left( \frac{2\lambda_2}{\pi} \right)^N \prod_{j=1}^N \frac{1}{4(x^{(j)} - \lambda_1)^2 + \lambda_2^2} , \end{aligned}$$



**Fig. G.2:** Histogram of the data of Problem 6.5.

$$\begin{aligned}
 \ell &= N(\ln 2 - \ln \pi + \ln \lambda_2) - \sum_{j=1}^N \ln[4(\mathbf{x}^{(j)} - \lambda_1)^2 + \lambda_2^2] \quad , \\
 \frac{\partial \ell}{\partial \lambda_1} &= \sum_{j=1}^N \frac{8(\mathbf{x}^{(j)} - \lambda_1)}{4(\mathbf{x}^{(j)} - \lambda_1)^2 + \lambda_2^2} = 0 \quad , \\
 \frac{\partial \ell}{\partial \lambda_2} &= \frac{N}{\lambda_2} - 2\lambda_2 \sum_{j=1}^N \frac{1}{4(\mathbf{x}^{(j)} - \lambda_1)^2 + \lambda_2^2} = 0 \quad .
 \end{aligned}$$

There is not necessarily a unique solution since the equations are not linear in  $\lambda_1$  and  $\lambda_2$ . For  $|\mathbf{x}^{(j)} - \lambda_1| \ll \lambda_2$ , however, we may write

$$\frac{8}{\lambda_2^2} \sum_{j=1}^N (\mathbf{x}^{(j)} - \lambda_1) = 0 \quad , \quad N\lambda_1 = \sum_{j=1}^N \mathbf{x}^{(j)} \quad , \quad \lambda_1 = a = \bar{\mathbf{x}} \quad .$$

### Problem 7.2

$$\begin{aligned}
 \text{(a)} \quad L &= \prod_{j=1}^N \frac{1}{\sqrt{2\pi}\sigma} \exp[-(\mathbf{x}^{(j)} - \lambda)^2/2\sigma^2] \\
 &= (\sqrt{2\pi}\sigma)^{-N} \prod_{j=1}^N \exp[-(\mathbf{x}^{(j)} - \lambda)^2/2\sigma^2] \quad ,
 \end{aligned}$$

$$\begin{aligned}
\ell &= -N \ln(\sqrt{2\pi}\sigma) - \frac{1}{2\sigma^2} \sum_{j=1}^N (\mathbf{x}^{(j)} - \lambda)^2 \quad , \\
\ell' &= \frac{1}{\sigma^2} \sum_{j=1}^N (\mathbf{x}^{(j)} - \lambda) \quad , \\
\ell'' &= -N/\sigma^2 \quad , \\
I(\lambda) &= -E(\ell'') = N/\sigma^2 \quad .
\end{aligned}$$

$$\begin{aligned}
(b) \quad L &= \prod_{j=1}^N \frac{1}{\sqrt{2\pi}\sqrt{\lambda}} \exp\left(-\frac{(\mathbf{x}^{(j)} - a)^2}{2\lambda}\right) \\
&= (2\pi)^{-N/2} \lambda^{-N/2} \prod_{j=1}^N \exp\left(-\frac{(\mathbf{x}^{(j)} - a)^2}{2\lambda}\right) \quad , \\
\ell &= -\frac{N}{2} \ln(2\pi) - \frac{N}{2} \ln \lambda - \frac{1}{2\lambda} \sum_{j=1}^N (\mathbf{x}^{(j)} - a)^2 \quad , \\
\ell' &= -\frac{N}{2\lambda} + \frac{1}{2\lambda^2} \sum_{j=1}^N (\mathbf{x}^{(j)} - a)^2 \quad , \\
\ell'' &= \frac{N}{2\lambda^2} - \frac{1}{\lambda^3} \sum_{j=1}^N (\mathbf{x}^{(j)} - a)^2 \quad , \\
I(\lambda) &= -E(\ell'') = -\frac{N}{2\lambda^2} + \frac{1}{\lambda^3} E \left\{ \sum_{j=1}^N (\mathbf{x}^{(j)} - a)^2 \right\} \\
&= -\frac{N}{2\lambda^2} + \frac{1}{\lambda^3} N\lambda \quad , \\
I(\lambda) &= \frac{1}{2\lambda^2} (-N + 2N) = \frac{N}{2\lambda^2} \quad , \\
E(\mathbf{S}) &= \frac{1}{N} E \left\{ \sum_{j=1}^N (\mathbf{x}^{(j)} - a)^2 \right\} = \sigma^2 = \lambda \quad .
\end{aligned}$$

### Problem 7.3

$$(a) \quad \sigma^2(\mathbf{S}) \geq \frac{1}{I(\lambda)} = \frac{\sigma^2}{N} \quad ,$$

In Problem 6.1 (c) we had  $\sigma^2(\bar{\mathbf{x}}) = 1/3\sigma^2$  for  $N = 3$ .

$$(b) \quad \sigma^2(\mathbf{S}) \geq \frac{1}{I(\lambda)} = \frac{2\lambda^2}{N} \quad .$$

(c) From Problem 7.2 we take

$$\ell' = -\frac{N}{2\lambda} + \frac{1}{2\lambda^2} N\mathbf{S} = \frac{N}{2\lambda^2} (\mathbf{S} - \lambda) = \frac{N}{2\lambda^2} (\mathbf{S} - E(\mathbf{S})) \quad .$$

$$\text{Thus, } \sigma^2(\mathbf{S}) = \frac{2\lambda^2}{N}.$$

### Problem 8.1

$$\begin{aligned}\bar{x}_1 &= 21.11 , & \bar{x}_2 &= 21.00 , \\ s_1^2 &= 14.86 , & s_2^2 &= 10.00 , \\ T &= s_1^2/s_2^2 = 1.486 , & F_{0.95}(8, 6) &= 4.15 .\end{aligned}$$

The variance of sample (2) is not significantly smaller.

### Problem 8.2

$$\begin{aligned}\bar{x} &= 25.142 , & s^2 &= 82.69/29 = 2.85 , & s &= 1.69 , \\ T &= \frac{\bar{x} - 25.5}{s/\sqrt{30}} = \frac{25.142 - 25.5}{1.69/\sqrt{30}} = -\frac{0.358}{0.309} = -1.16 , \\ |T| &< t_{0.95}(f = 29) = 1.70 .\end{aligned}$$

Therefore the hypothesis cannot be rejected.

### Problem 8.3

$$\begin{aligned}(a) \quad f(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(N)}, \tilde{\boldsymbol{\lambda}}^{(\Omega)}) &= \prod_{j=1}^N \frac{1}{\sqrt{2\pi}s'} \exp\left(-\frac{(\mathbf{x}^{(j)} - \bar{\mathbf{x}})^2}{2s'^2}\right) \\ &= (\sqrt{2\pi}s')^{-N} \exp\left(-\frac{1}{2s'^2} \sum_{j=1}^N (\mathbf{x}^{(j)} - \bar{\mathbf{x}})^2\right) , \\ f(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(N)}, \tilde{\boldsymbol{\lambda}}^{(\omega)}) &= (\sqrt{2\pi}\sigma_0)^{-N} \exp\left(-\frac{1}{2\sigma_0^2} \sum_{j=1}^N (\mathbf{x}^{(j)} - \bar{\mathbf{x}})^2\right) , \\ T &= \left(\frac{\sigma_0}{s'}\right)^N \exp\left[\frac{1}{2}Ns'^2\left(\frac{1}{\sigma_0^2} - \frac{1}{s'^2}\right)\right] , \\ \ln T &= N(\ln\sigma_0 - \ln s') + \frac{1}{2}Ns'^2\left(\frac{1}{\sigma_0^2} - \frac{1}{s'^2}\right) \\ &= N(\ln\sigma_0 - \ln s') + \frac{1}{2}N\frac{s'^2}{\sigma_0^2} - \frac{1}{2}N .\end{aligned}$$

- (b)  $T' = Ns'^2/\sigma_0^2$  is a monotonically increasing function of  $s'$  if  $s' > \sigma_0$ , otherwise it is monotonically decreasing. Therefore the test takes on the following form:

$$\begin{aligned} T' > T'_{1-\frac{1}{2}\alpha} &\quad , \quad T' < T'_{\frac{1}{2}\alpha} \quad \text{for} \quad H_0(\mathbf{s}' = \sigma_0) \quad , \\ T' > T'_{1-\alpha} &\quad \text{for} \quad H_0(\mathbf{s}' < \sigma_0) \quad , \\ T' < T'_\alpha &\quad \text{for} \quad H_0(\mathbf{s}' > \sigma_0) \quad . \end{aligned}$$

(c) The answer follows from (6.7.2) and  $\mathbf{s}'^2 = (N-1)\mathbf{s}^2/N$ .

### Problem 8.4

$$(a) \quad \tilde{a} = \frac{1}{n} \sum_k n_k \mathbf{x}_k = 202.36.$$

$$\tilde{\sigma}^2 = \frac{1}{n-1} \sum_k n_k (\mathbf{x}_k - \tilde{a})^2 = 13.40, \quad \tilde{\sigma} = 3.66.$$

$$(b) \quad p_k(\mathbf{x}_k) = \psi_0 \left( \frac{\mathbf{x}_k + \frac{1}{2} \Delta x - \tilde{a}}{\tilde{\sigma}} \right) - \psi_0 \left( \frac{\mathbf{x}_k - \frac{1}{2} \Delta x - \tilde{a}}{\tilde{\sigma}} \right) = \psi_{0+} - \psi_{0-}.$$

| $\mathbf{x}_k$   | $n_k$ | $\psi_{0+}$ | $\psi_{0-}$ | $np_k$ | $\frac{(n_k - np_k)^2}{np_k}$ |
|------------------|-------|-------------|-------------|--------|-------------------------------|
| 193              | 1     | 0.011       | 0.002       | (0.9)  | —                             |
| 195              | 2     | 0.041       | 0.011       | (3.0)  | —                             |
| 197              | 9     | 0.117       | 0.041       | 7.6    | 0.271                         |
| 199              | 12    | 0.260       | 0.117       | 14.3   | 0.362                         |
| 201              | 23    | 0.461       | 0.260       | 20.1   | 0.411                         |
| 203              | 25    | 0.673       | 0.463       | 21.2   | 0.679                         |
| 205              | 11    | 0.840       | 0.673       | 16.7   | 1.948                         |
| 207              | 9     | 0.938       | 0.840       | 9.8    | 0.071                         |
| 209              | 6     | 0.982       | 0.938       | 4.3    | 0.647                         |
| 211              | 2     | 0.996       | 0.982       | (1.4)  | —                             |
| $\chi^2 = 4.389$ |       |             |             |        |                               |

The number of degrees of freedom is  $7 - 2 = 5$ . Since  $\chi^2_{0.90}(5) = 9.24$ , the test does not reject the hypothesis.

### Problem 8.5

$$(a) \quad \tilde{p}_1 = \frac{1}{111}(13+44) = \frac{57}{111}, \quad \tilde{p}_2 = \frac{1}{111}(25+29) = \frac{54}{111},$$

$$\tilde{q}_1 = \frac{1}{111}(13+25) = \frac{38}{111}, \quad \tilde{q}_2 = \frac{1}{111}(44+29) = \frac{73}{111},$$

$$\chi^2 = \frac{\left(13 - \frac{57 \cdot 38}{111}\right)^2}{\frac{57 \cdot 38}{111}} + \frac{\left(44 - \frac{57 \cdot 73}{111}\right)^2}{\frac{57 \cdot 73}{111}}$$

$$\begin{aligned}
 & + \frac{\left(25 - \frac{54 \cdot 38}{111}\right)^2}{\frac{54 \cdot 38}{111}} + \frac{\left(29 - \frac{54 \cdot 73}{111}\right)^2}{\frac{54 \cdot 73}{111}} \\
 = & \frac{42.43}{19.51} + \frac{42.43}{37.49} + \frac{42.43}{18.49} + \frac{42.43}{35.51} = 6.78 \quad .
 \end{aligned}$$

Since  $\chi^2_{0.90} = 2.71$  for  $f = 1$ , the hypothesis of independence is rejected.

$$\begin{aligned}
 (b) \quad & n_{ij} - \frac{1}{n}(n_{i1} + n_{i2})(n_{1j} + n_{2j}) \\
 & = \frac{1}{n}[n_{ij}(n_{11} + n_{12} + n_{21} + n_{22}) - (n_{i1} + n_{i2})(n_{1j} + n_{2j})] \quad .
 \end{aligned}$$

One can easily show that the expression in square brackets takes the form  $(-1)^{i+j}(n_{11}n_{22} - n_{12}n_{21})$  for all  $i, j$ .

### G.3 Programming Problems

#### Programming Problem 4.1: Program to Generate Breit–Wigner-Distributed Random Numbers

Write a method with the following declaration

`double [] breitWignerNumbers(double a, double gamma, int n).`

It is to yield  $n$  random numbers, which follow a Breit–Wigner distribution having a mean of  $a$  and a FWHM of  $\Gamma$ . Make the method part of a class which allows for interactive input of  $n, a, \Gamma$  and numerical as well as graphical output of the random numbers in the form a histogram, Fig. G.3. (Example solution: S1Random)

#### Programming Problem 4.2: Program to Generate Random Numbers from a Triangular Distribution

Write a method with the declaration

`double [] triangularNumbersTrans(double a, double b, double c, int n).`

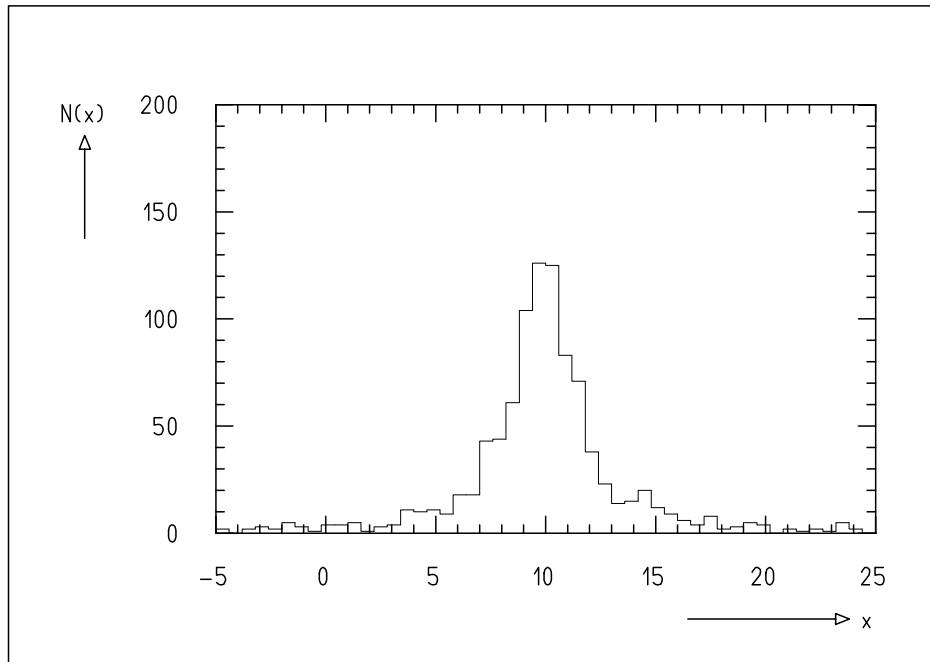
It is to yield  $n$  random numbers, following a triangular distribution with the parameters  $a, b, c$  generated by the transformation procedure of Example 4.3.

Write a second method with the declaration

`double [] triangularNumbersRej(double a, double b, double c, int n),`

which solves the same problem, but uses von Neumann's acceptance–rejection method. Which of the two programs is faster?

Write a class which allows to interactively choose either method. It should also allow for numerical and graphical output (as histogram) of the generated numbers, Fig. G.4. (Example solution: S2Random)



**Fig. G.3:** Histogram of 1000 random numbers following a Breit–Wigner distribution with  $a = 10$  and  $\Gamma = 3$ .

### Programming Problem 4.3: Program to Generate Data Points with Errors of Different Size

Write a method similar to `DatanRandom.line` which generates data points  $y = at + b + \Delta y$ . The errors  $\Delta y$ , however, are not to be taken for all data points  $y_i$  from the same uniform distribution with the width  $\sigma$ , but  $\Delta y_i$  is to be sampled from a normal distribution with the width  $\sigma_i$ . The widths  $\sigma_i$  are to be taken from a uniform distribution within the region  $\sigma_{\min} < \sigma_i < \sigma_{\max}$ .

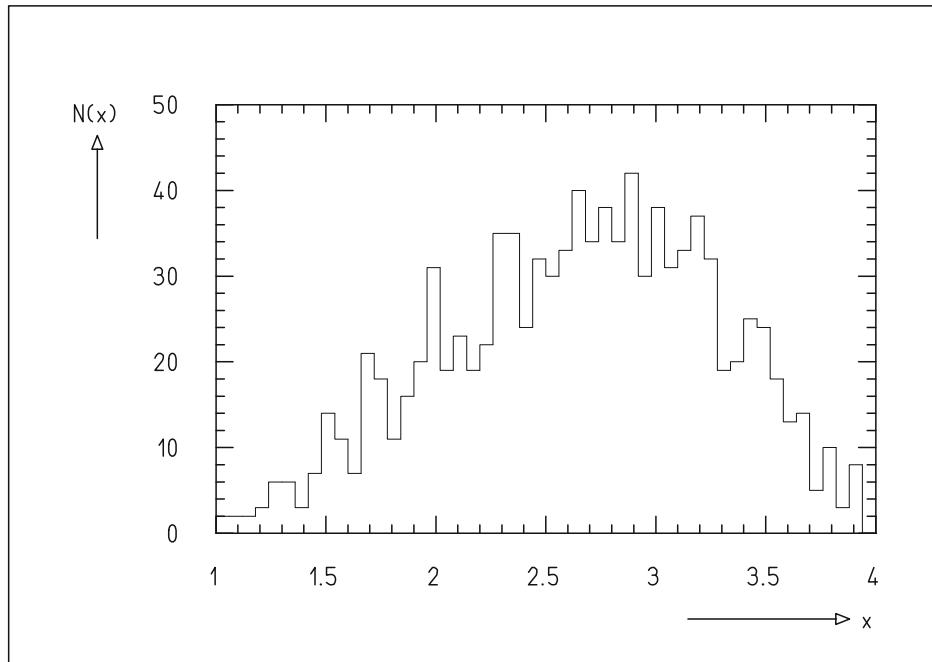
Write a class which calls this method and which displays graphically the straight line  $y = at + b$  as well as the simulated data points with error bars  $y_i \pm \Delta y_i$ , Fig. G.5. (Example solution: `S3Random`)

### Programming Problem 5.1: Convolution of Uniform Distributions

Because of the Central Limit Theorem the quantity  $\mathbf{x} = \sum_{i=1}^N \mathbf{x}_i$  follows in the limit  $N \rightarrow \infty$  the standard normal distribution if the  $\mathbf{x}_i$  come from an arbitrary distribution with mean zero and standard deviation  $1/\sqrt{N}$ . Choose for the  $\mathbf{x}_i$  the uniform distribution with the limits

$$a = -\sqrt{3/N} \quad , \quad b = -a \quad .$$

Perform a large number  $n_{\text{exp}}$  of Monte Carlo experiments, each giving a random value  $\mathbf{x}$ . Produce a histogram of the quantity  $\mathbf{x}$  and show in addition the distribution of  $\mathbf{x}$  as a continuous curve which you would expect from the standard normal distribution



**Fig. G.4:** Histogram of 1000 random numbers following a triangular distribution with  $a = 1$ ,  $b = 4$ ,  $c = 3$ .

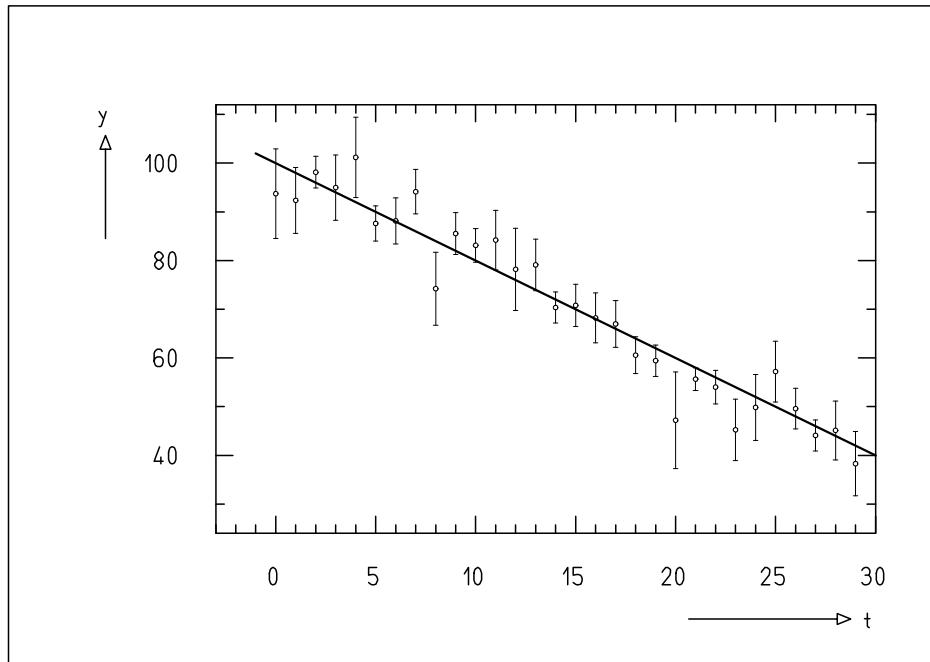
(Fig. G.6). (Use the class `GraphicsWithHistogramAndPolyline` for the simultaneous representation of histograms and curves.) Allow for interactive input of the quantities  $n_{\text{exp}}$  and  $N$ . (Example solution: `S1Distrib`)

### Programming Problem 5.2: Convolution of Uniform Distribution and Normal Distribution

If  $x$  is taken from a uniform distribution between  $a$  and  $b$  and if  $y$  is taken from a normal distribution with mean zero and width  $\sigma$ , then the quantity  $u = x + y$  follows the distribution (5.11.14). Perform a large number  $n_{\text{exp}}$  of Monte Carlo experiments, each resulting in a random number  $u$ . Display a histogram of the quantity  $u$  and show in addition a curve of the distribution you would expect from (5.11.14), Fig. G.7. Allow for the interactive input of the quantities  $n_{\text{exp}}$ ,  $a$ ,  $b$ , and  $\sigma$ .

### Programming Problem 7.1: Distribution of Lifetimes Determined from a Small Number of Radioactive Decays

In Example Program 7.1, an estimate  $\bar{t}$  of the mean lifetime  $\tau$  and its asymmetric errors  $\Delta_-$  and  $\Delta_+$  are found from a single small sample. In all cases the program yields  $\Delta_- < \Delta_+$ . Write a program that simulates a large number  $n_{\text{exp}}$  of experiments, in each of which  $N$  radioactive decays of mean lifetime  $\tau = 1$  are measured. Compute for each experiment the estimate  $\bar{t}$  and construct a histogram of the quantity  $\bar{t}$  for all experiments. Present this histogram  $N_i(\bar{t}_i \leq \bar{t} < \bar{t}_i + \Delta\bar{t})$  and also the cumulative frequency distribution  $h_i = (1/n_{\text{exp}}) \sum_{\bar{t} < \bar{t}_i} N_i$ . Allow for interactive input of  $n_{\text{exp}}$



**Fig. G.5:** Data points with errors of different size.

and  $N$ . Demonstrate that the distributions are asymmetric for small  $N$  and that they become symmetric for large  $N$ . Show that for small  $N$  the value  $\bar{t} = 1$  is not the most probable value, but that it is the expectation value of  $\bar{t}$ . Determine for a fixed value of  $N$ , e.g.,  $N = 4$ , limits  $\Delta_-$  and  $\Delta_+$  in such a way that  $\bar{t} < 1 - \Delta_-$  holds with the probability  $0.683/2$  and that with the same probability one has  $\bar{t} > 1 + \Delta_+$ . Compare the results found with a series of simulated experiments from the program E1MaxLike, Example Program 7.1. (Example solution: S1MaxLike)

### Programming Problem 7.2: Distribution of the Sample Correlation Coefficient

Modify the class E2MaxLike so that instead of numerical output, a histogram of the correlation coefficient  $r$  is presented (Fig. G.8). Produce histograms for  $\rho = 0$  and  $\rho = 0.95$ , each for  $n_{\text{pt}} = 5, 50, 500$ . Under what circumstances is the distribution asymmetric and why? Is this asymmetry in contradiction to the Central Limit theorem? (Example solution: S2MaxLike)

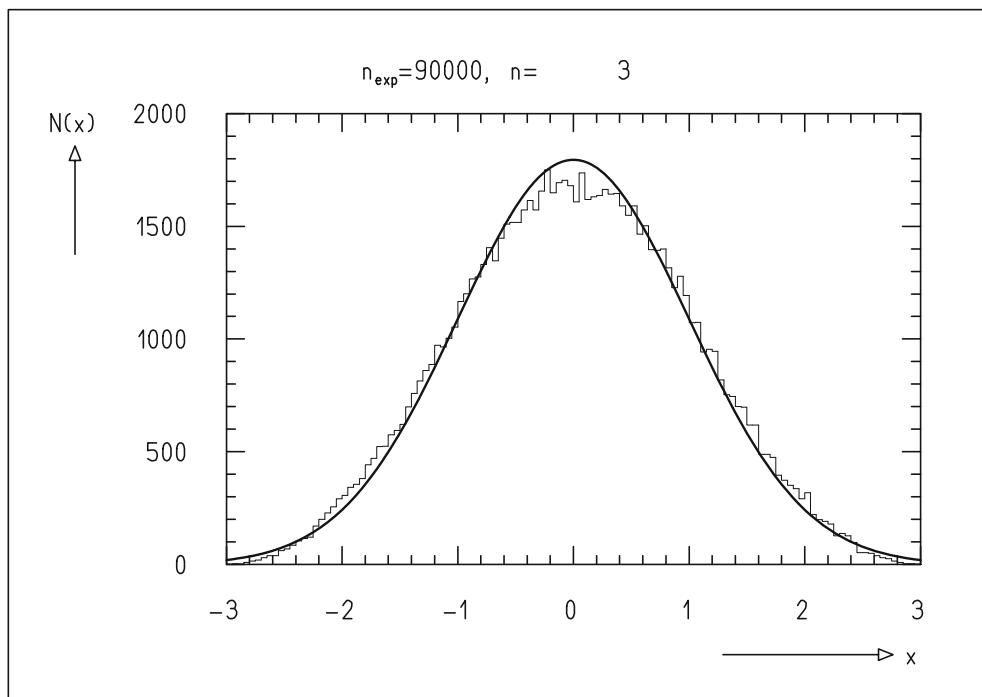
### Programming Problem 9.1: Fit of a First-Degree Polynomial to Data that Correspond to a Second-Degree Polynomial

In experimental or empirical studies one is often confronted with a large number of measurements or objects of the same kind (animals, elementary particle collisions, industrial products from a given production process, ...). The outcomes of the measurements performed on each object are described by some law. Certain assumptions are made about that law, which are to be checked by experiment.

Consider the following example. A series of measurements may contain  $n_{\text{exp}}$  experiments. Each experiment yields the measurements  $y_i = x_1 + x_2 t_i + x_3 t_i^2 + \varepsilon_i$  for 10 values  $t_i = 1, 2, \dots, 10$  of the controlled variable  $t$ . The  $\varepsilon_i$  are taken from a normal distribution with mean zero and width  $\sigma$ . In the analysis of the experiments it is assumed, however, that the true values  $\eta_i$  underlying the measurements  $y_i$  can be described by a first-degree polynomial  $\eta_i = x_1 + x_2 t$ . As result of the fit we obtain a minimum function  $M$  from which we can compute the “ $\chi^2$ -probability”  $P = 1 - F(M, n - r)$ . Here  $F(M, f)$  is the distribution function of a  $\chi^2$  distribution with  $f$  degrees of freedom,  $n$  is the number of data points, and  $r$  the number of parameters determined in the fit. If  $P < \alpha$ , then the fit of a first-degree polynomial to the data is rejected at a confidence level of  $\beta = 1 - \alpha$ .

Write a class performing the following steps:

- (i) Interactive input of  $n_{\text{exp}}, x_1, x_2, x_3, \sigma, \Delta y$ .
- (ii) Generation of  $n_{\text{exp}}$  sets of data  $(t_i, y_i, \Delta y)$ , fit of a first-degree polynomial to each set of data and computation of  $P$ . Entry of  $P$  into a histogram.
- (iii) Graphical representation of the histogram.



**Fig. G.6:** Histogram of 90 000 random numbers  $x$ , each of which is a sum of three uniformly distributed random numbers. The curve corresponds to the standard normal distribution. Significant differences between curve and histogram are visible only because of the very large number of random numbers used.

**Suggestions:** (a) Choose  $n_{\text{exp}} = 1000$ ,  $x_1 = x_2 = 1$ ,  $x_3 = 0$ ,  $\sigma = \Delta y = 1$ . As expected you will obtain a flat distribution for  $P$ . (b) Choose (keeping the other input quantities as above) different values  $x_3 \neq 0$ . You will observe a shift of the distribution towards small  $P$  values, cf. Fig. G.9. Determine approximately the smallest positive value of  $x_3$  such that the hypothesis of a first-degree polynomial is rejected at 90% confidence level in 95% of all experiments. (c) Choose  $x_3 = 0$ , but  $\sigma \neq \Delta y$ . You will again observe a shift in the distribution, e.g., towards larger  $P$  values for  $\Delta y > \sigma$ . (d) From the experience gained in (a), (b), and (c), one might conclude that if erroneously too large measurement errors are assumed ( $\Delta y > \sigma$ ) then a flat  $P$  distribution would result. In this way one would get the impression that a first-degree polynomial could describe the data. Begin with  $n_{\text{exp}} = 1000$ ,  $x_1 = x_2 = 1$ ,  $x_3 = 0.2$ ,  $\sigma = 1$ ,  $\Delta y = 1$  and increase  $\Delta y$  in steps of 0.1 up to  $\Delta y = 2$ . (Example solution: SILsq)

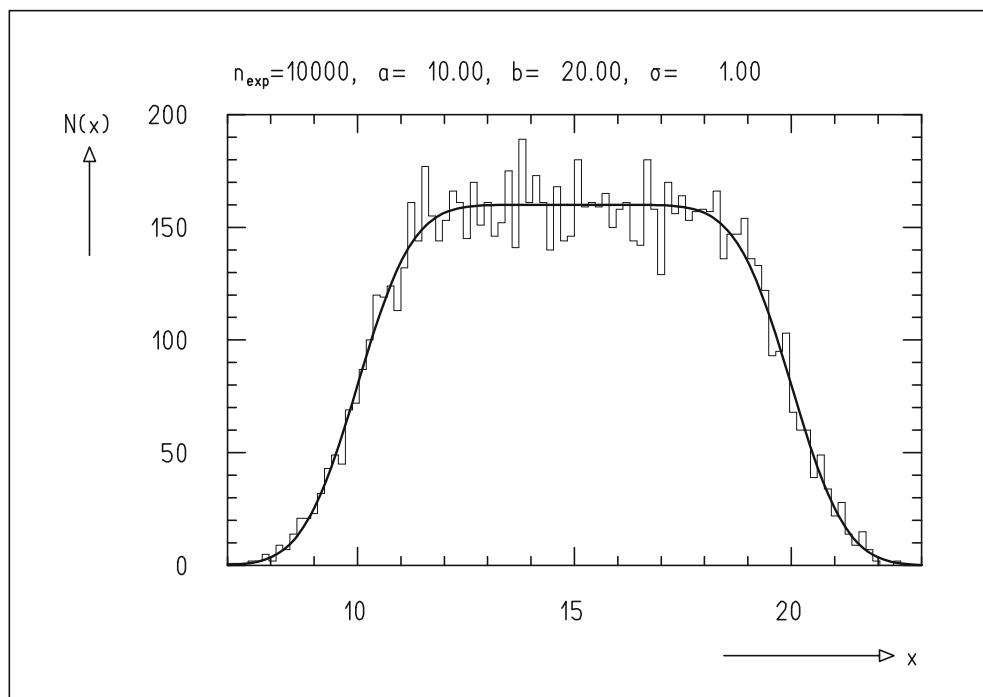
### Programming Problem 9.2: Fit of a Power Law (Linear Case)

A power law

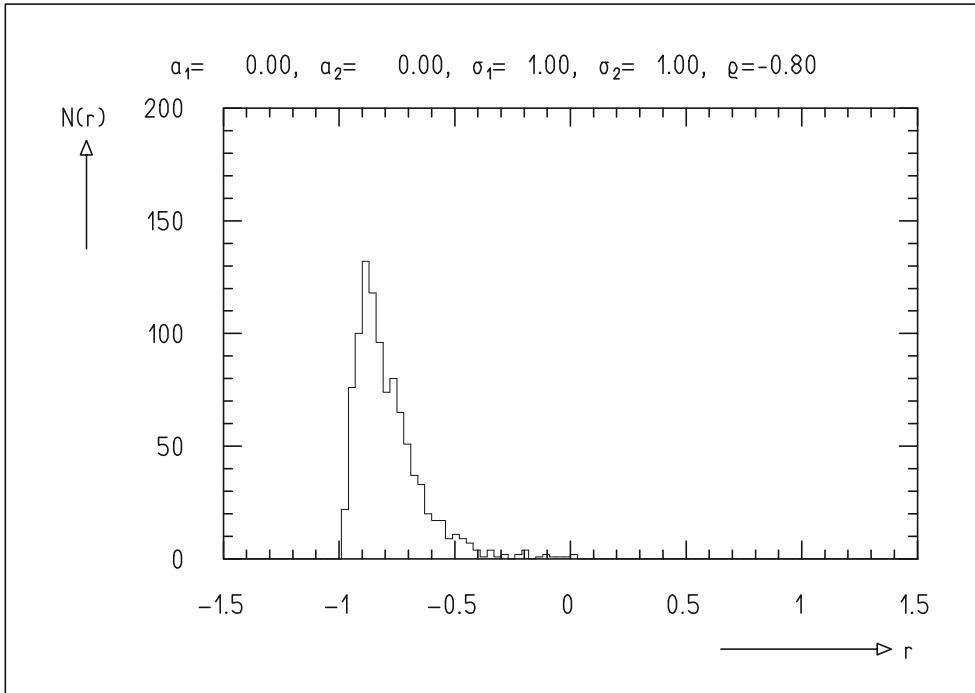
$$\eta = xt^w$$

is linear in the parameter  $x$  if  $w$  is a constant. This function is to be fitted to measurements  $(t_i, y_i)$  given by

$$t_i = t_0 + (i - 1)\Delta t, \quad i = 1, \dots, n,$$



**Fig. G.7:** A histogram of 10 000 random numbers, each of which is the sum of a uniformly distributed random number and a normally distributed random number. The curve corresponds to the convolution of a uniform and a normal distribution.



**Fig. G.8:** Histogram of the sample correlation coefficient computed for 1000 samples of size 10 from a bivariate Gaussian distribution with the correlation coefficient  $\rho = -0.8$ .

$$y_i = xt_i^w + \varepsilon_i .$$

Here the  $\varepsilon_i$  follow a normal distribution centered about zero with width  $\sigma$ .

Write a class performing the following steps:

- (i) Interactive input of  $n, t_0, \Delta t, x, w, \sigma$ .
- (ii) Generation of measured points.
- (iii) Fit of the power law.
- (iv) Graphical display of the data and the fitted function, cf. Fig. G.10.

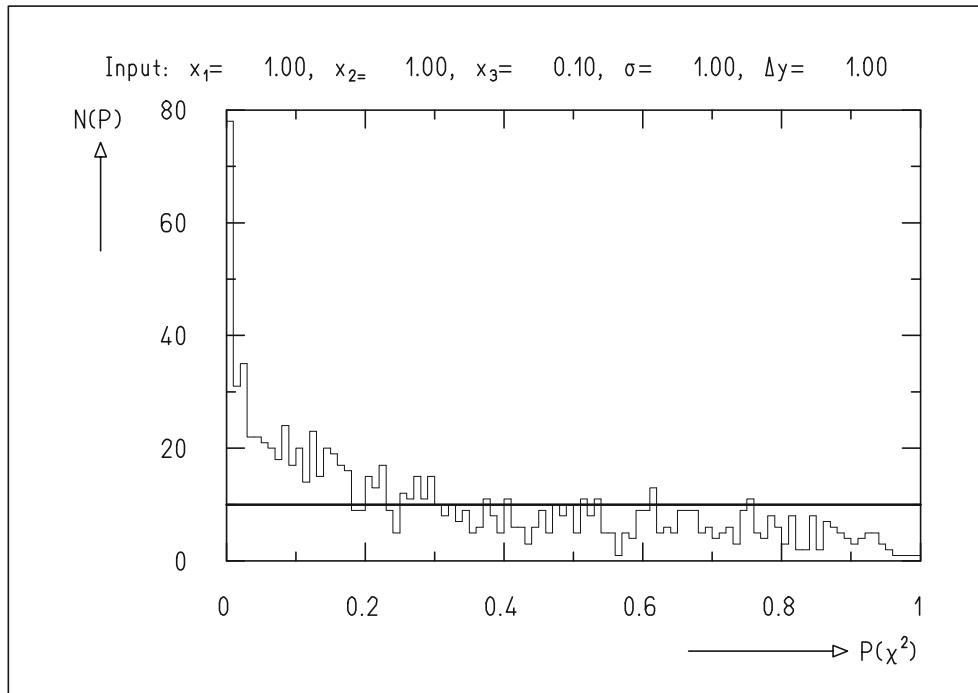
(Example solution: S2Lsq)

### Programming Problem 9.3: Fit of a Power Law (Nonlinear Case)

If the power law has the form

$$\eta = x_1 t^{x_2} ,$$

i.e., if the power itself is an unknown parameter, the problem becomes nonlinear. For the fit of a nonlinear function we have to start from a first approximation of the parameters. We limit ourselves to the case  $t_i > 0$  for all  $i$  which occurs frequently in practice. Then one has  $\ln \eta = \ln x_1 + x_2 \ln t$ . If instead of  $(t_i, y_i)$  we now use



**Fig. G.9:** Histogram of the  $\chi^2$ -probability for fits of a first-degree polynomial to 1000 data sets generated according to a second-degree polynomial.

$(\ln t_i, \ln y_i)$  as measured variables, we obtain a linear function in the parameters  $\ln x_1$  and  $x_2$ . However, in this transformation the errors are distorted so that they are no longer Gaussian. We simply choose all errors to be of equal size and use the result of the linear fit as the first approximation of a nonlinear fit to the  $(t_i, y_i)$ . We still have to keep in mind (in any case for  $x_1 > 0$ ) that one always has  $\eta > 0$  for  $t > 0$ . Because of measurement errors, however, measured values  $y_i < 0$  can occur. Such points of course must not be used for the computation of the first approximation.

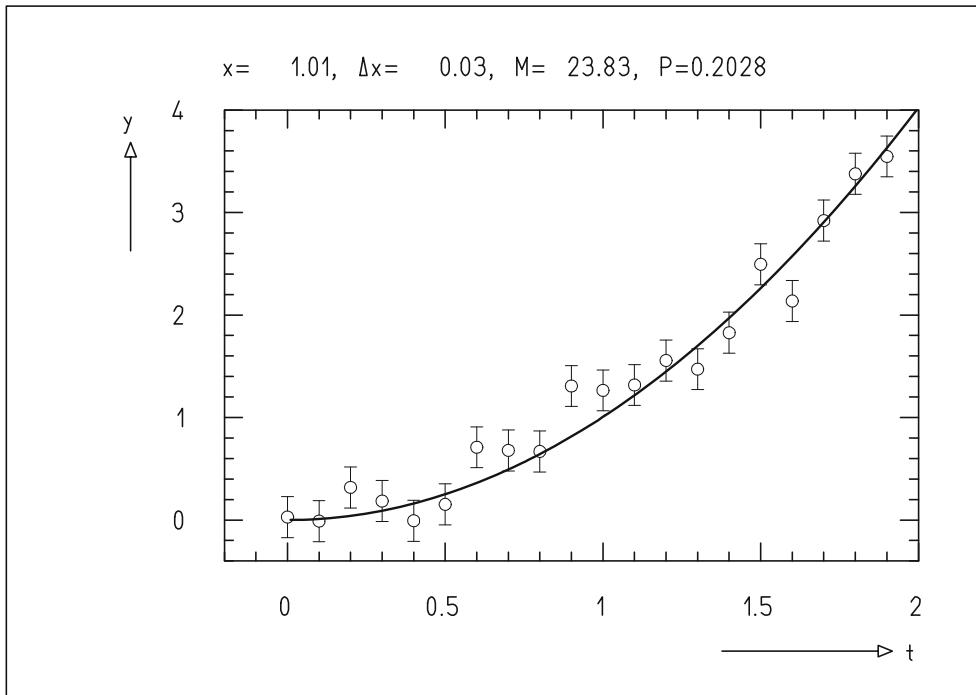
Write a class with the following steps:

- (i) Interactive input of  $n, t_0, \Delta t, x_1, x_2, \sigma$ .
- (ii) Generation of the  $n$  measured points

$$\begin{aligned} t_i &= t_0 + (i-1)\Delta t \quad , \quad i = 1, \dots, n \quad , \\ y_i &= x_1 t_i^{x_2} + \varepsilon_i \quad , \end{aligned}$$

where  $\varepsilon_i$  comes from a normal distribution centered about zero with width  $\sigma$ .

- (iii) Computation of first approximations  $x_1, x_2$  by fitting a linear function to  $(\ln t_i, \ln y_i)$  with LsqLin.



**Fig. G.10:** Result of the fit of a parabola  $y = xt^2$  to 20 measured points.

(iv) Fit of a power law to  $(t_i, y_i)$  with LsqNon.

(v) Graphical display of the results, cf. Fig. G.11.

(Example solution: S3Lsq)

#### Programming Problem 9.4: Fit of a Breit–Wigner Function to Data Points with Errors

For the  $N = 21$  values  $t_i = -3, -2.7, \dots, 3$  of the controlled variable the measured values

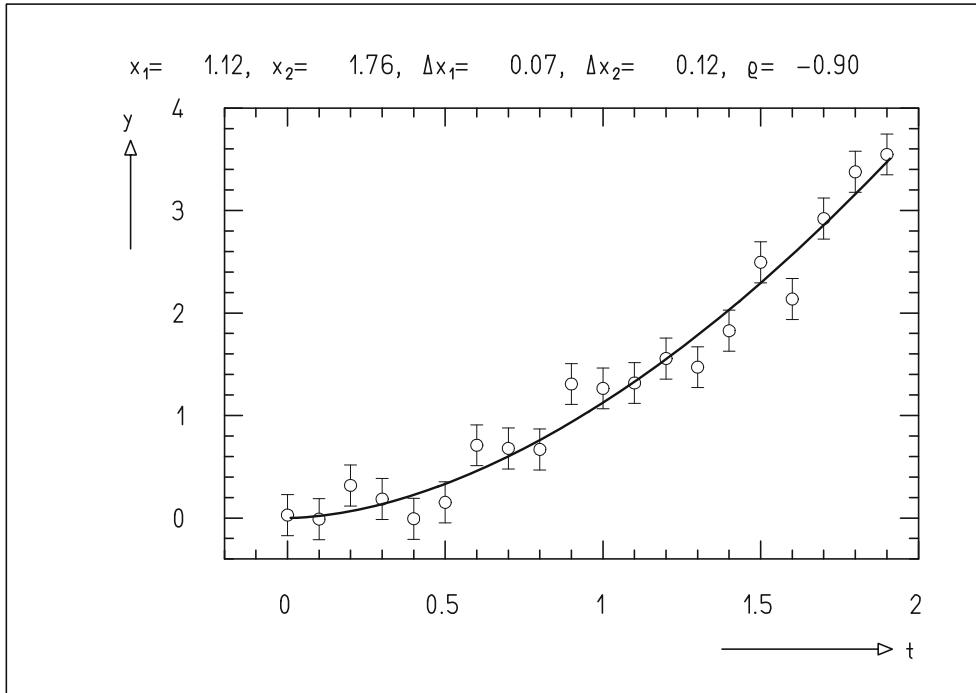
$$y_i = f(t_i) + \varepsilon_i \quad (\text{G.3.1})$$

are to be simulated. Here,

$$f(t) = \frac{2}{\pi x_2} \frac{x_2^2}{4(t - x_1)^2 + x_2^2} \quad (\text{G.3.2})$$

is the Breit–Wigner function (3.3.32) with  $a = x_1$  and  $\Gamma = x_2$ . The measurement errors  $\varepsilon_i$  are to be taken from a normal distribution around zero with width  $\sigma$ . Choose  $a = 0$  and  $\Gamma = 1$ . The points  $(t_i, y_i)$  scatter within the measurement errors around a bell-shaped curve with a maximum at  $t = a$ . A bell-shaped curve with a maximum at the same position, however, could be given by the Gaussian function,

$$f(t) = \frac{1}{x_2 \sqrt{2\pi}} \exp \left\{ -\frac{(t - x_1)^2}{2x_2^2} \right\} . \quad (\text{G.3.3})$$



**Fig. G.11:** Fit of a function  $y = x_1 t^{x_2}$  to 20 data points. The data points are identical to those in Fig. G.10.

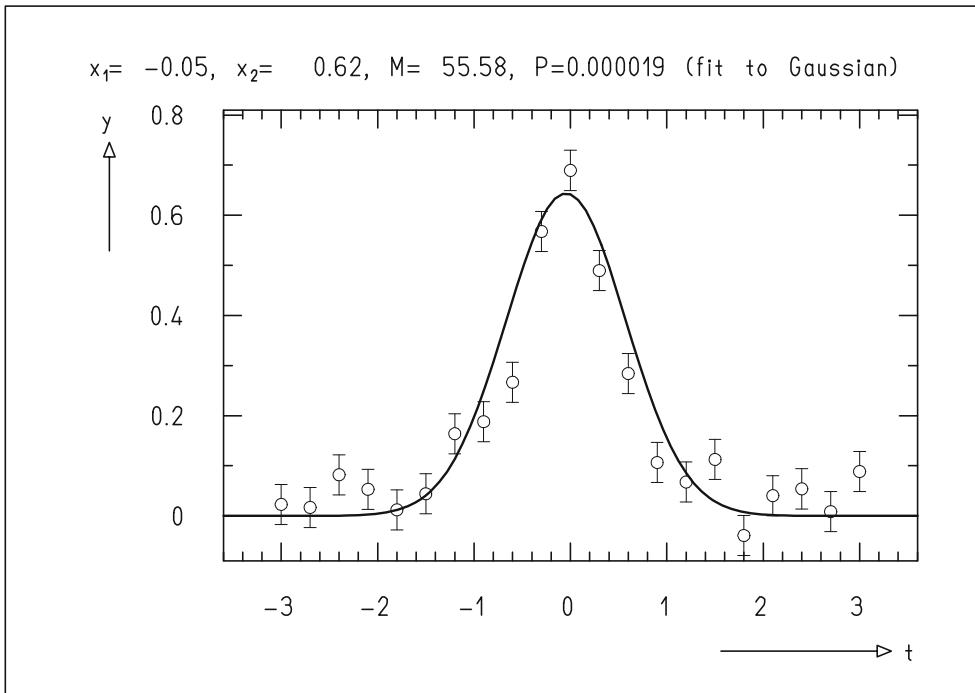
Write a class with the following properties:

- Interactive input of  $\sigma$  and possibility to choose whether a Breit–Wigner function or a Gaussian is to be fitted.
- Generation of the data, i.e., of the triplets of numbers  $(t_i, y_i, \Delta y_i = \varepsilon_i)$ .
- Fit of the Breit–Wigner function (G.3.2) or of the Gaussian (G.3.3) to the data and computation of the minimum function  $M$ .
- Graphical representation of the measured points with measurement errors and of the fitted function, cf. Fig. G.12.

Run the program using different values of  $\sigma$  and find out for which range of  $\sigma$  the data allow a clear discrimination between the Breit–Wigner and Gaussian functions. (Example solution: S4Lsq)

### Programming Problem 9.5: Asymmetric Errors and Confidence Region for the Fit of a Breit–Wigner Function

Supplement the solution of Programming Problem 9.4 such that it yields a graphical representation for the parameters, their errors, covariance ellipse, asymmetric errors, and confidence region similar to Fig. 9.11. Discuss the differences obtained when fitting a Breit–Wigner or a Gaussian, respectively. For each case try  $\sigma = 0.1$  and  $\sigma = 1$ . (Example solution: S5Lsq)



**Fig. G.12:** Fit of a Gaussian to data points that correspond to a Breit–Wigner function. The goodness-of-fit is poor.

### Programming Problem 9.6: Fit of a Breit–Wigner Function to a Histogram

In the Programming Problem 9.4 we started from measurements  $y_i = f(t_i) + \varepsilon_i$ . Here  $f(t)$  was a Breit–Wigner function (3.3.32), and the measurement errors  $\varepsilon_i$  corresponded to a Gaussian distribution centered about zero with width  $\sigma$ .

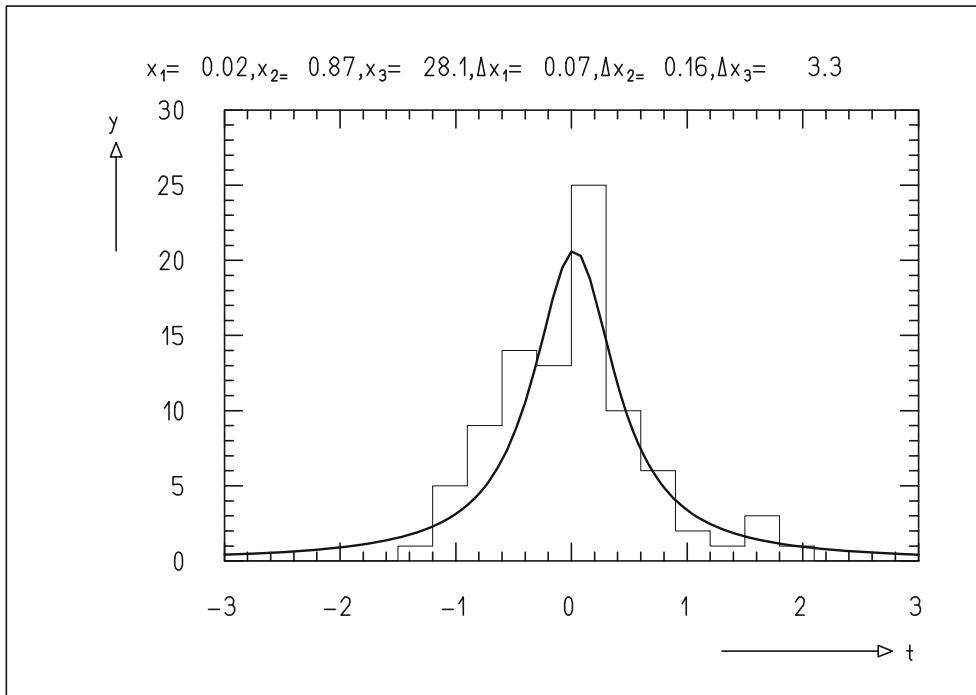
We now generate a sample of size  $n_{\text{ev}}$  from a Breit–Wigner distribution with mean  $a = 0$  and full width at half maximum  $\Gamma = 1$ . We represent the sample by a histogram that is again characterized by triplets of numbers  $(t_i, y_i, \Delta y_i)$ . Now  $t_i$  is the center of the  $i$ th bin  $t_i - \Delta t/2 \leq t < t_i + \Delta t/2$ , and  $y_i$  is the number of sample elements falling into this bin. For not too small  $y_i$  the corresponding statistical error is  $\Delta y_i = \sqrt{y_i}$ . For small values  $y_i$  this simple statement is problematic. It is completely wrong for  $y_i = 0$ . In the fit of a function to a histogram, care is therefore to be taken that empty bins (possibly also bins with few entries) are not to be considered as data points. The function to be fitted is

$$f(t) = x_3 \frac{2}{\pi x_2} \frac{x_2^2}{4(t - x_1)^2 + x_2^2} . \quad (\text{G.3.4})$$

The function is similar to (G.3.2) but there is an additional parameter  $x_3$ .

Write a class with the following steps:

- (i) Interactive input of  $n_{\text{ev}}$  (sample size) and  $n_t$  (number of histogram bins). The lower limit of the histogram is to be fixed at  $t = -3$ , the upper limit at  $t = 3$ .



**Fig. G.13:** Fit of a Breit–Wigner function to a histogram.

- (ii) Generation of the sample, cf. Programming Problem 4.1.
- (iii) Construction of the histogram.
- (iv) Construction of the triplets  $(t_i, y_i, \Delta y_i)$  to be used for the fit.
- (v) Fit of the function (G.3.4).
- (vi) Output of the results in numerical and graphical form, cf. Fig. G.13.

**Suggestions:** Perform consecutive fits for the same sample but different numbers of bins. Try to find an optimum for the number of bins. (Example solution: S6Lsq)

**Programming Problem 9.7:** Fit of a Circle to Points with Measurement Errors in Abscissa and Ordinate

A total of  $m$  data points  $(s_i, t_i)$  are given in the  $(s, t)$  plane. The measurement errors are defined by  $2 \times 2$  covariance matrices of the form

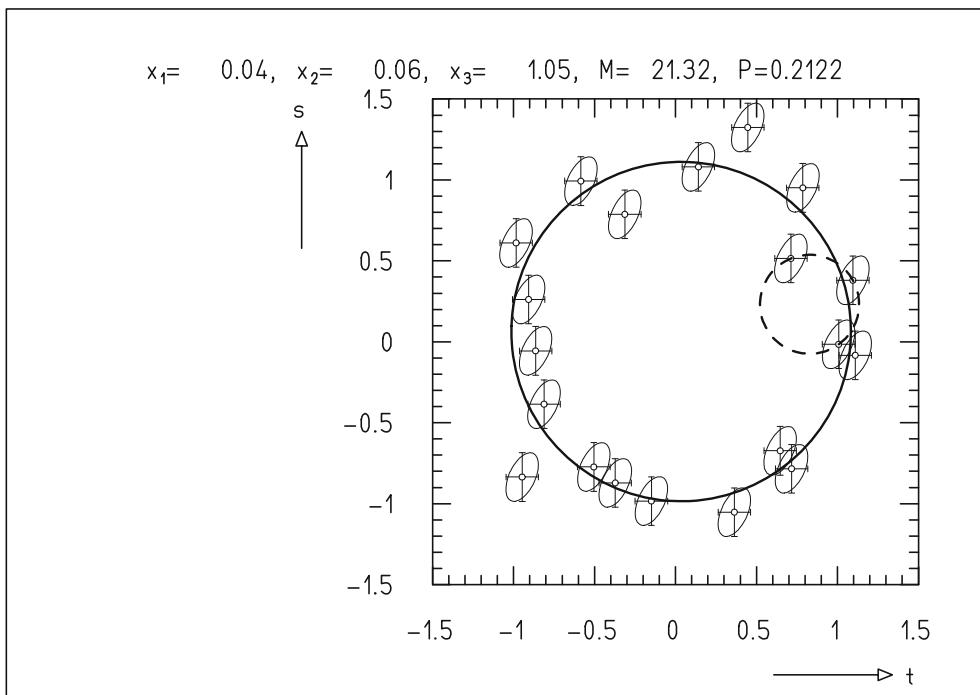
$$\begin{pmatrix} \Delta s_i^2 & c_i^2 \\ c_i^2 & \Delta t_i^2 \end{pmatrix} \quad (\text{G.3.5})$$

as in Example 9.11. Here  $c_i = \Delta s_i \Delta t_i \rho_i$  and  $\rho_i$  is the correlation coefficient between the measurement errors  $\Delta s_i$ ,  $\Delta t_i$ . As in Example 9.11, construct the vector  $\mathbf{y}$  of measurements from the  $s_i$  and  $t_i$  and construct the covariance matrix  $C_y$ . Set up

the equations  $f_k(x, \eta) = 0$  assuming that the true positions underlying the measured points lie on a circle with center  $(x_1, x_2)$  and radius  $x_3$ . Write a program with the following steps:

- (i) Input of  $m$ ,  $\Delta t$ ,  $\Delta s$ ,  $\rho$ .
- (ii) Generation of  $m$  measured points  $(s_i, t_i)$  using bivariate normal distributions, the means of which are positioned at regular intervals on the unit circle ( $x_1 = x_2 = 0$ ,  $x_3 = 1$ ) and the covariance matrix of which is given by (G.3.5) with  $\Delta s_i = \Delta s$ ,  $\Delta t_i = \Delta t$ ,  $c = \Delta s \Delta t \rho$ .
- (iii) Determination of a first approximation for  $x_1$ ,  $x_2$ ,  $x_3$  by computation of the parameters of a circle through the first three measured points.
- (iv) Fit to all measured points using LsqGen and a user function specially written for this problem.
- (v) Graphical representation of the measured points and the fitted circle as in Fig. G.14.

(Example solution: S6Lsq)



**Fig. G.14:** Measured points with covariance ellipses, circle of the first approximation which is given by 3 measured points (broken line), and circle fitted to all measured points.

**Programming Problem 10.1:** Monte Carlo Minimization to Choose a Good First Approximation

For some functions in Example Program 10.1 the choice of the point  $\mathbf{x}_0$  defining the first approximation was decisive for success or failure of the minimization. If a function has several minima and if its value is smallest at one of them, that is if an “absolute minimum” exists, the following procedure will work. One uses the Monte Carlo method to determine a first approximation  $\mathbf{x}_0$  of the absolute minimum in a larger region of the parameter space by generating points  $\mathbf{x} = (x_1, x_2, \dots, x_n)$  according to a uniform distribution in that region and by choosing that point at which the function has the smallest value.

On the basis of E1Min write class that determines the absolute minimum of the function  $f_7(\mathbf{x})$  described in Sect. 10.1. A first approximation  $\mathbf{x}_0$  within the region

$$-10 < x_{0i} < 10 \quad , \quad i = 1, 2, 3 \quad ,$$

is to be determined by the Monte Carlo method. Perform the search for the first approximation with  $N$  points generated at random and allow for an interactive input of  $N$ . (Example solution: S1Min)

**Programming Problem 10.2:** Determination of the Parameters of a Breit–Wigner Distribution from the Elements of a Sample

By modifying a copy of E2Min produce a class that simulates a sample from a Breit–Wigner distribution with mean  $a$  and full width at half maximum  $\Gamma$ , and that subsequently determines the numerical values of these parameters by minimization of the negative log-likelihood function of the sample. Allow for interactive input of the sample size and the parameters  $a$  and  $\Gamma$  used in the simulation. (Example solution: S2Min)

**Programming Problem 11.1:** Two-Way Analysis of Variance with Crossed Classification

The model (11.2.16) for the data in an analysis of variance with crossed classification is

$$x_{ijk} = \mu + a_i + b_j + (ab)_{ij} + \varepsilon_{ijk} \quad .$$

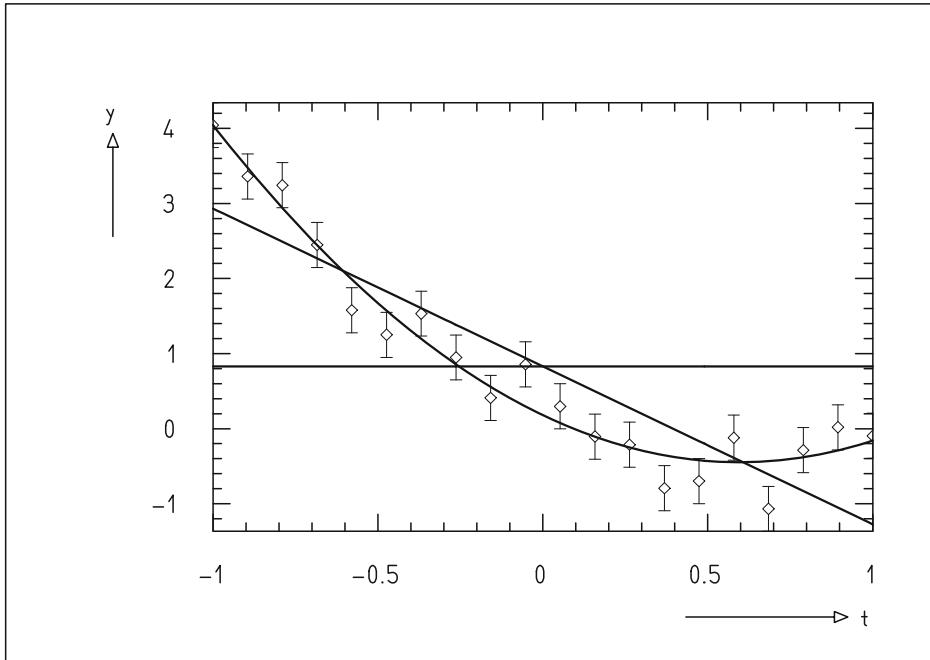
The analysis of variance tests the null hypothesis

$$a_i = b_j = (ab)_{ij} = 0 \quad .$$

Data corresponding to this null hypothesis are generated by the program E2Anova, Sect. 11.2, and are subsequently analyzed.

Write a program similar to E2Anova which generates data  $x_{ijk}$  according to the above formula with

$$\begin{aligned} a_i &= \left( i - \frac{I+1}{2} \right) a \quad , \\ b_j &= \left( j - \frac{J+1}{2} \right) b \quad , \\ (ab)_{ij} &= \text{signum}(a_i) \text{ signum}(b_j) ab \quad . \end{aligned}$$



**Fig. G.15:** Data points with errors and regression polynomials of different degrees.

These relations fulfill the requirements (11.2.7). The  $\varepsilon_{ijk}$  as in E2Anova are to be drawn from a normal distribution with mean zero and standard deviation  $\sigma$ . Allow for interactive input of the quantities  $a, b, ab, \sigma, \mu$ . Perform an analysis of variance on the simulated data. Study different cases, e.g.,  $a = 0, b \neq 0, ab = 0; a \neq 0, b = 0, ab = 0; a = 0, b = 0, ab \neq 0$ ; etc. (Example solution: S1Anova)

### Programming Problem 11.2: Two-Way Analysis of Variance with Nested Classification

Modify Programming Problem 11.1 for the treatment of a nested classification with data of the form (11.2.22), e.g.,

$$x_{ijk} = \mu + a_i + b_{ij} + \varepsilon_{ijk} \quad ,$$

and use the relations

$$a_i = \left( i - \frac{I+1}{2} \right) a \quad , \quad b_{ij} = \left( j - \frac{J+1}{2} \right) b \quad .$$

(Example solution: S2Anova)

### Programming Problem 12.1: Simulation of Data and Plotting Regression Polynomials of Different Degrees

Write a class which generates  $n$  data points  $(t_i, y_i)$ . The  $t_i$  are to be spread equidistantly over the interval  $-1 \leq t \leq 1$ ; the  $y_i$  are to correspond to a polynomial with  $r$

terms and to have errors with standard deviation  $\sigma$ . A regression analysis is to be performed and a plot as in Fig. G.15 of the data and the regression polynomials to be produced. Your program may be largely based on the classes E4Reg and E2Reg. (Example solution: S1Reg)

### **Programming Problem 12.2:** Simulation of Data and Plotting the Regression Line with Confidence Limits

Extend the solution of Programming Problem 12.1 so that a regression polynomial of the desired degree together with confidence limits, corresponding to Sect. 12.2 is shown. (Example solution: S2Reg)

### **Programming Problem 13.1:** Extrapolation in a Time Series Analysis

In Sect. 13.3 we have stressed that one must be very cautious with the interpretation of the results of a time series analysis at the edge of a time series, and of the extrapolation in regions outside the time series. In particular, we found that the extrapolation yields meaningless results if the data are not at least approximately described by a polynomial. The degree of this polynomial must be smaller than or equal to the degree of the polynomial used for the time series analysis.

Study these statements by simulating a number of time series and analyzing them. Write a program – starting from E2TimSer – that for  $n = 200$ ,  $i = 1, 2, \dots, n$ , generates data of the form

$$y_i = t_i^m + \varepsilon_i \quad , \quad t_i = \frac{i - 100}{50} \quad .$$

Here the  $\varepsilon_i$  are to be generated according to a normal distribution with mean zero and standard deviation  $\sigma$ . Allow for the interactive input of  $m$ ,  $\sigma$ ,  $k$ ,  $\ell$ , and  $P$ . After generating the data perform a time series analysis and produce a plot as in E2TimSer. Study different combinations of  $m$ ,  $k$ , and  $\ell$ , and for each combination use small values of  $\sigma$  (e.g.,  $\sigma = 0.001$ ) and large values of  $\sigma$  (e.g.,  $\sigma = 0.1$ ). (Example solution: S1TimSer)

### **Programming Problem 13.2:** Discontinuities in Time Series

In the development of time series analysis it was assumed that the measurements, apart from their statistical fluctuations, are continuous functions of time. We therefore expect unreliable results in regions where the measurements are discontinuous. Write a program that generates the following three types of time series, analyzes them, and displays the results graphically. One of them is continuous; the other two contain discontinuities.

*Sine function:*

$$y_i = \sin(\pi t_i / 180) + \varepsilon_i \quad , \quad t_i = i \quad , \quad i = 1, 2, \dots, n \quad .$$

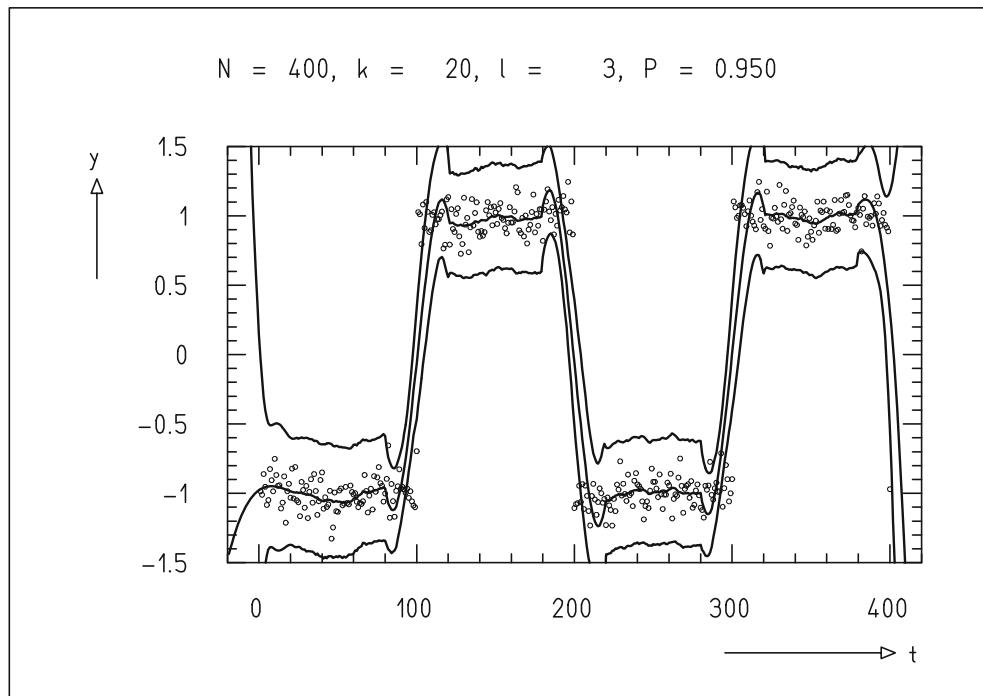
*Step function:*

$$y_i = \begin{cases} -1 + \varepsilon_i, & t_i \leq 100 \\ 1 + \varepsilon_i, & t_i > 100 \end{cases}, \quad t_i = i \bmod 200, \quad i = 1, 2, \dots, n.$$

*Sawtooth function:*

$$y_i = (t_i - 50)/100 + \varepsilon_i, \quad t_i = i \bmod 100, \quad i = 1, 2, \dots, n.$$

The  $\varepsilon_i$  are again to be generated according to normal distribution with mean zero and standard deviation  $\sigma$ . Allow for the choice of one of the functions and for the interactive input of  $n$ ,  $\sigma$ ,  $k$ ,  $\ell$ , and  $P$ . Study the time series using different values for the parameters and discuss the results. Figure G.16 shows an example. (Example solution: S2TimSer)



**Fig. G.16:** Time series corresponding to a step function with moving averages and confidence limits.

## H. Collection of Formulas

### Probability

$A, B, \dots$  are *events*;  $\bar{A}$  is the event “*not A*”.

$(A + B)$  and  $(AB)$  combine events with *logical “or”* or *logical “and”*.

$P(A)$  is the *probability* of the event  $A$ .

$P(B|A) = P(AB)/P(A)$  is the probability for  $B$  given the condition  $A$  (*conditional probability*).

The following rules hold:

For *every* event  $A$

$$P(\bar{A}) = 1 - P(A) \quad ,$$

for *mutually exclusive* events  $A, B, \dots, Z$

$$P(A + B + \dots + Z) = P(A) + P(B) + \dots + P(Z) \quad ,$$

for *independent* events  $A, B, \dots, Z$

$$P(AB \cdots Z) = P(A)P(B) \cdots P(Z) \quad .$$

### Single Random Variable

*Distribution function*:  $F(x) = P(x < x)$

*Probability density* (for  $F(x)$  differentiable):

$$f(x) = F'(x) = dF(x)/dx$$

*Moments of order  $\ell$* :

- (a) About the point  $c$ :  $\alpha_\ell = E\{(\mathbf{x} - c)^\ell\}$
- (b) About the origin (central moments):  $\lambda_\ell = E\{\mathbf{x}^\ell\}$
- (c) About the mean:  $\mu_\ell = E\{(\mathbf{x} - \hat{x})^\ell\}$

**Table H.1:** Expectation values for discrete and continuous distributions.

|   | $\mathbf{x}$ discrete   | $\mathbf{x}$ continuous;<br>$F(x)$ differentiable                            |
|---|---|--|
| Probability density                         | —   | $f(x) = F'(x) = \frac{dF(x)}{dx}$ ,<br>$\int_{-\infty}^{\infty} f(x) dx = 1$ |
| Mean of $\mathbf{x}$<br>(expectation value) | $\hat{x} = E(\mathbf{x})$<br>$= \sum_i x_i P(\mathbf{x} = x_i)$ | $\hat{x} = E(\mathbf{x})$<br>$= \int_{-\infty}^{\infty} x f(x) dx$           |
| Mean of the function $H(\mathbf{x})$        | $E\{H(\mathbf{x})\}$<br>$= \sum_i H(x_i) P(\mathbf{x} = x_i)$   | $E\{H(\mathbf{x})\}$<br>$= \int_{-\infty}^{\infty} H(x) f(x) dx$             |

Variance:  $\sigma^2(\mathbf{x}) = \text{var}(\mathbf{x}) = \mu_2 = E\{(\mathbf{x} - \hat{x})^2\}$

Standard deviation or error of  $\mathbf{x}$ :  $\Delta\mathbf{x} = \sigma(\mathbf{x}) = +\sqrt{\sigma^2(\mathbf{x})}$

Skewness:  $\gamma = \mu_3/\sigma^3$

Reduced variable:  $\mathbf{u} = (\mathbf{x} - \hat{\mathbf{x}})/\sigma(\mathbf{x})$  ;  $E(\mathbf{u}) = 0$  ,  $\sigma^2(\mathbf{u}) = 1$

Mode (most probable value)  $x_m$  defined by:  $P(\mathbf{x} = x_m) = \max$

Median  $x_{0.5}$  defined by:  $F(x_{0.5}) = P(\mathbf{x} < x_{0.5}) = 0.5$

Quantile  $x_q$  defined by:  $F(x_q) = P(\mathbf{x} < x_q) = q$  ;  $0 \leq q \leq 1$

## Several Random Variables

Distribution function:

$$F(\mathbf{x}) = F(x_1, x_2, \dots, x_n) = P(\mathbf{x}_1 < x_1, \mathbf{x}_2 < x_2, \dots, \mathbf{x}_n < x_n)$$

Joint probability density (only for  $F(\mathbf{x})$  differentiable with respect to all variables):

$$f(\mathbf{x}) = f(x_1, x_2, \dots, x_n) = \partial^n F(x_1, x_2, \dots, x_n) / \partial x_1 \partial x_2 \dots \partial x_n$$

Marginal probability density of the variable  $x_i$ :

$$g_i(x_i) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f(x_1, x_2, \dots, x_n) dx_1 dx_2 \dots dx_{i-1} dx_{i+1} \dots dx_n$$

*Expectation value of a function  $H(\mathbf{x})$ :*

$$E\{H(\mathbf{x})\} = \int H(\mathbf{x}) f(\mathbf{x}) d\mathbf{x}$$

*Expectation value of the variables  $x_i$ :*

$$\hat{x}_i = E(x_i) = \int x_i f(\mathbf{x}) d\mathbf{x} = \int_{-\infty}^{\infty} x_i g_i(x_i) dx_i$$

The variables  $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$  are *independent*, if

$$f(x_1, x_2, \dots, x_n) = g_1(x_1)g_2(x_2)\dots g_n(x_n)$$

*Moments of order  $\ell_1, \ell_2, \dots, \ell_n$ :*

(a) About  $\mathbf{c} = (c_1, c_2, \dots, c_n)$ :

$$\alpha_{\ell_1 \ell_2 \dots \ell_n} = E\{(\mathbf{x}_1 - c_1)^{\ell_1} (\mathbf{x}_2 - c_2)^{\ell_2} \dots (\mathbf{x}_n - c_n)^{\ell_n}\}$$

(b) About the origin:  $\lambda_{\ell_1 \ell_2 \dots \ell_n} = E\{\mathbf{x}_1^{\ell_1} \mathbf{x}_2^{\ell_2} \dots \mathbf{x}_n^{\ell_n}\}$

(c) About  $\hat{\mathbf{x}}$ :  $\mu_{\ell_1 \ell_2 \dots \ell_n} = E\{(\mathbf{x}_1 - \hat{x}_1)^{\ell_1} (\mathbf{x}_2 - \hat{x}_2)^{\ell_2} \dots (\mathbf{x}_n - \hat{x}_n)^{\ell_n}\}$

*Variance of  $\mathbf{x}_i$ :*  $\sigma^2(\mathbf{x}_i) = E\{(\mathbf{x}_i - \hat{x}_i)^2\} = c_{ii}$

*Covariance between  $\mathbf{x}_i$  and  $\mathbf{x}_j$ :*  $\text{cov}(\mathbf{x}_i, \mathbf{x}_j) = E\{(\mathbf{x}_i - \hat{x}_i)(\mathbf{x}_j - \hat{x}_j)\} = c_{ij}$

For  $\mathbf{x}_i, \mathbf{x}_j$  independent:  $\text{cov}(\mathbf{x}_i, \mathbf{x}_j) = 0$

*Covariance matrix:*  $C = E\{(\mathbf{x} - \hat{\mathbf{x}})(\mathbf{x} - \hat{\mathbf{x}})^T\}$

*Correlation coefficient:*

$$\rho(\mathbf{x}_1, \mathbf{x}_2) = \text{cov}(\mathbf{x}_1, \mathbf{x}_2) / \sigma(\mathbf{x}_1)\sigma(\mathbf{x}_2) ; -1 \leq \rho \leq 1$$

*Rules of computation:*

$$\sigma^2(c\mathbf{x}_i) = c^2\sigma^2(\mathbf{x}_i) ,$$

$$\sigma^2(a\mathbf{x}_i + b\mathbf{x}_j) = a^2\sigma^2(\mathbf{x}_i) + b^2\sigma^2(\mathbf{x}_j) + 2ab \text{cov}(\mathbf{x}_i, \mathbf{x}_j) ;$$

$a, b, c$  are constants

## Transformation of Variables

Original variables:  $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)$

Probability density:  $f(\mathbf{x})$

Transformed variables:  $\mathbf{y} = (y_1, y_2, \dots, y_n)$

Mapping:  $y_1 = y_1(\mathbf{x}), y_2 = y_2(\mathbf{x}), \dots, y_n = y_n(\mathbf{x})$

Probability density:  $g(\mathbf{y}) = |J| f(\mathbf{x})$

with the *Jacobian determinant*

$$J = J \left( \frac{x_1, x_2, \dots, x_n}{y_1, y_2, \dots, y_n} \right) = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_2}{\partial y_1} & \dots & \frac{\partial x_n}{\partial y_1} \\ \vdots & & & \\ \frac{\partial x_1}{\partial y_n} & \frac{\partial x_2}{\partial y_n} & \dots & \frac{\partial x_n}{\partial y_n} \end{vmatrix}$$

## Error Propagation

The original variables  $\mathbf{x}$  have the covariance matrix  $C_x$ . The *covariance matrix of the transformed variables*  $\mathbf{y}$  is

$$C_y = T C_x T^T \quad \text{with} \quad T = \begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \dots & \frac{\partial y_1}{\partial x_n} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & \dots & \frac{\partial y_2}{\partial x_n} \\ \vdots & & & \\ \frac{\partial y_m}{\partial x_1} & \frac{\partial y_m}{\partial x_2} & \dots & \frac{\partial y_m}{\partial x_n} \end{pmatrix} .$$

The formula is only exact for a linear relation between  $\mathbf{y}$  and  $\mathbf{x}$ , but is a good approximation for small deviations from linearity in a region around  $\hat{\mathbf{x}}$  of the magnitude of the standard deviation. *Only for vanishing covariances in  $C_x$  does one have*

$$\sigma(y_i) = \Delta y_i = \sqrt{\sum_{j=1}^m \left( \frac{\partial y_i}{\partial x_j} \right)^2 (\Delta x_j)^2} .$$

## The Law of Large Numbers

A total of  $n$  observations are carried out, which are characterized by the random variable  $x_i$  ( $= 1$ , if on the  $i$ th observation the event  $A$  occurs, otherwise  $= 0$ ). The *frequency* of  $A$  is

$$h = \frac{1}{n} \sum_{i=1}^n x_i .$$

For  $n \rightarrow \infty$  this frequency is equal to the probability  $p$  for the occurrence of  $A$ ,

$$E(h) = h = p , \quad \sigma^2(h) = \frac{1}{n} p(1-p) .$$

**Table H.2:** Distributions of discrete variables.

| Distribution                                   | Probability for observing $\mathbf{x} = \mathbf{k}$<br>$(x_1 = k_1, \dots, x_l = k_l)$ | Mean                      | Variance<br>(elements of covariance matrix)        |
|--|--|---------------------------|--|
| Binomial                                       | $\frac{W_k^n}{\binom{n}{k}} p^k (1-p)^{n-k}$   | $\hat{x} = np$            | $\sigma^2(\mathbf{x}) = np(1-p)$                   |
| Multinomial<br>$\sum_{j=1}^l p_j = 1$          | $\frac{W_{k_1, k_2, \dots, k_l}^n}{\prod_{j=1}^l k_j!} \prod_{j=1}^l p_j^{k_j}$        | $\hat{x}_j = np_j$        | $c_{ij} = np_i(\delta_{ij} - p_j)$                 |
| Hypergeometric<br>$L = N - K$ ,<br>$l = n - k$ | $W_k = \binom{K}{k} \binom{L}{\ell} : \binom{N}{n}$                                    | $\hat{x} = n \frac{K}{N}$ | $\sigma^2(\mathbf{x}) = \frac{nKL(N-n)}{N^2(N-1)}$ |
| Poisson  | $f(k) = \frac{\lambda^k}{k!} e^{-\lambda}$   | $\hat{x} = \lambda$       | $\sigma^2(\mathbf{x}) = \lambda$                   |

## Central Limit Theorem

If  $x_i$  are independent variables with mean  $a$  and variance  $\sigma^2$ , then  $(1/n) \sum_{i=1}^n x_i$  for  $n \rightarrow \infty$  follows a normal distribution with mean  $a$  and variance  $\sigma^2/n$ .

## Convolutions of Distributions

The probability density of the sum  $u = x + y$  of two independent random variables  $x$  and  $y$  is

$$f_u(u) = \int_{-\infty}^{\infty} f_x(x) f_y(u-x) dx = \int_{-\infty}^{\infty} f_y(y) f_x(u-y) dy .$$

- The convolution of two Poisson distributions with the parameters  $\lambda_1$  and  $\lambda_2$  is a Poisson distribution with the parameter  $\lambda_1 + \lambda_2$ .
- The convolution of two Gaussian distributions with means  $a_1, a_2$  and variances  $\sigma_1^2, \sigma_2^2$  is a Gaussian distribution with mean  $a = a_1 + a_2$  and variance  $\sigma^2 = \sigma_1^2 + \sigma_2^2$ .

**Table H.3:** Distributions of continuous variables.

| Distribution                  | Probability density  | Mean                              | Variance<br>(covariance<br>matrix)                            |
|-------------------------------|--|-----------------------------------|---|
| Uniform                       | $\begin{cases} 0; x > a, x \geq b \\ \frac{1}{b-a}; a \leq x < b \end{cases}$  | $\frac{1}{2}(b+a)$                | $(b-a)^2/12$  |
| Gaussian                      | $\frac{1}{b\sqrt{2\pi}} \exp\left(-\frac{(x-a)^2}{2b^2}\right)$  | $a$                               | $b^2$   |
| Stand. Gaussian               | $\frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}x^2)$  | 0                                 | 1   |
| Gaussian of several variables | $k \exp\left\{-\frac{(\mathbf{x}-\mathbf{a})^T B (\mathbf{x}-\mathbf{a})}{2}\right\}$  | $\mathbf{a}$                      | $C = B^{-1}$  |
| $\chi^2$                      | $\frac{1}{\Gamma(\frac{1}{2}f)2^{\frac{1}{2}f}} (\chi^2)^{\frac{1}{2}f-1} \exp(-\frac{1}{2}\chi^2)$  | $f$                               | $2f$  |
| Fisher's $F$                  | $\left(\frac{f_1}{f_2}\right)^{\frac{1}{2}f_1} \frac{\Gamma(\frac{1}{2}(f_1+f_2))}{\Gamma(\frac{1}{2}f_1)\Gamma(\frac{1}{2}f_2)} \\ \times F^{\frac{1}{2}f_1-1} \left(1 + \frac{f_1}{f_2}F\right)^{-\frac{1}{2}(f_1+f_2)}$ | $\frac{f_2}{f_2-2},$<br>$f_2 > 2$ | $\frac{2f_2^2(f_1+f_2-2)}{f_1(f_2-2)^2(f_2-4)},$<br>$f_2 > 4$ |
| Student's $t$                 | $\frac{\Gamma(\frac{1}{2}(f+1))}{\Gamma(\frac{1}{2}f)\sqrt{\pi}\sqrt{f}} \left(1 + \frac{t^2}{f}\right)^{-\frac{1}{2}(f+1)}$   | 0                                 | $\frac{f}{f-2},$<br>$f > 2$                                   |

- The convolution of two  $\chi^2$ -distributions with  $f_1$  and  $f_2$  degrees of freedom is a  $\chi^2$ -distribution with  $f = f_1 + f_2$  degrees of freedom.

## Samples

*Population:* An infinite (in some cases finite) set of elements described by a discrete or continuously distributed random variable  $\mathbf{x}$ .

*(Random) sample of size  $N$ :* A selection of  $N$  elements  $(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(N)})$  from the population. (For the requirements for a sample to be random see Sect. 6.1.)

**Table H.4:** Samples from different populations.

|                           | Sample of size $n$<br>from a continuously<br>distributed population    | Sample of size $n$ from a<br>discrete population of size $N$ .<br>Variable of the population is $y$ ,<br>variable of the sample is $\mathbf{x}$ |
|---------------------------|--|---|
| Population mean           | $E(\mathbf{x}) = \hat{x}$  | $\bar{y} = \frac{1}{N} \sum_{j=1}^N y_j$  |
| Population variance       | $\sigma^2(\mathbf{x})$   | $\sigma^2(y) = \frac{1}{N-1} \sum_{j=1}^N (y_j - \bar{y})^2$  |
| Sample mean               | $\bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i$             | $\bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i$  |
| Sample variance<br>mean   | $\sigma^2(\bar{\mathbf{x}}) = \frac{1}{n} \sigma^2(\mathbf{x})$        | $\sigma^2(\bar{\mathbf{x}}) = \frac{\sigma^2(y)}{n} \left(1 - \frac{n}{N}\right)$   |
| Variance of<br>the sample | $s^2 = \frac{1}{n-1} \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})^2$ | $s^2 = \frac{1}{n-1} \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})^2$  |

*Distribution function of the sample:*  $W_n(x) = n_x/N$ ,  
where  $n_x$  is the number of elements in the sample for which  $\mathbf{x} < x$ .

*Statistic:* An arbitrary function of the elements of a sample

$$\mathbf{S} = \mathbf{S}(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(N)}) \quad .$$

*Estimator:* A statistic used to estimate a parameter  $\lambda$  of the population. An estimator is *unbiased*, if  $E(\mathbf{S}) = \lambda$  and *consistent*, if

$$\lim_{N \rightarrow \infty} \sigma(\mathbf{S}) = 0 \quad .$$

## Maximum-Likelihood Method

Consider a population described by the probability density  $f(\mathbf{x}, \boldsymbol{\lambda})$ , where  $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \dots, \lambda_p)$  is a set of parameters. If a sample  $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(N)}$  is obtained, then the *likelihood function* is  $L = \prod_{j=1}^N f(\mathbf{x}^{(j)}, \boldsymbol{\lambda})$  and the *log-likelihood function* is  $\ell = \ln L$ . In order to determine the unknown parameters

$\lambda$  from the sample the maximum-likelihood method prescribes the value of  $\lambda$  for which  $L$  (or  $\ell$ ) is a maximum. That is, one must solve the *likelihood equation*  $\partial\ell/\partial\lambda_i = 0; i = 1, 2, \dots, p$  or (for only one parameter)  $d\ell/d\lambda = \ell' = 0$ .

*Information of a sample:*  $I(\lambda) = E(\ell'^2) = -E(\ell'')$

*Information inequality:*  $\sigma^2(S) \geq \{1 - B'(\lambda)\}^2/I(\lambda)$

Here  $S$  is an *estimator* for  $\lambda$ , and  $B(\lambda) = E(S) - \lambda$  is its *bias*.

An estimator has *minimum variance* if  $\ell' = A(\lambda)(S - E(S))$ , where  $A(\lambda)$  does not depend on the sample.

The maximum-likelihood estimator  $\tilde{\lambda}$ , i.e., the solution of the likelihood equation, is *unique, asymptotically unbiased* (i.e., for  $N \rightarrow \infty$ ), and has *minimum variance*.

The *asymptotic form* of the likelihood function for one parameter  $\lambda$  is

$$L = \text{const} \cdot \exp\left(-\frac{(\lambda - \tilde{\lambda})^2}{2b^2}\right) ,$$

$$b^2 = \sigma^2(\tilde{\lambda}) = \frac{1}{E(\ell'^2(\tilde{\lambda}))} = -\frac{1}{E(\ell''(\tilde{\lambda}))}$$

and for several parameters

$$L = \text{const} \cdot \left\{ -\frac{1}{2}(\lambda - \tilde{\lambda})^T B(\lambda - \tilde{\lambda}) \right\}$$

with the covariance matrix  $C = B^{-1}$  and

$$B_{ij} = -E\left(\frac{\partial^2 \ell}{\partial \lambda_i \partial \lambda_j}\right)_{\lambda=\tilde{\lambda}} .$$

## Testing Hypotheses

*Null hypothesis*  $H_0(\lambda = \lambda_0)$ : Assumption of values for the parameters  $\lambda$  that determine the probability distribution  $f(\mathbf{x}, \lambda)$  of a population.

*Alternative hypotheses*  $H_1(\lambda = \lambda_1), H_2(\lambda = \lambda_2), \dots$ : Other possibilities for  $\lambda$ , against which the null hypothesis is to be tested by consideration of a sample  $X = (\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(N)})$  from the population.

A hypothesis is *simple* if the parameters are completely determined, e.g.,  $H_0(\lambda_1 = 1, \lambda_2 = 5)$ , otherwise it is *composite*, e.g.,  $H_1(\lambda_1 = 2, \lambda_2 < 7)$ .

*Test* of a hypothesis  $H_0$  with a *significance level*  $\alpha$  or *confidence level*  $1 - \alpha$ :  $H_0$  is rejected if  $X \in S_c$ , where  $S_c$  is the *critical region* in the sample space and

$$P(X \in S_c | H_0) = \alpha .$$

*Error of the first kind:* Rejection of  $H_0$ , although  $H_0$  is true. The probability of this error is  $\alpha$ .

*Error of the second kind:*  $H_0$  is not rejected although  $H_1$  is true. The probability of this error is  $P(X \notin S_c | H_1) = \beta$ .

*Power function:*

$$M(S_c, \boldsymbol{\lambda}) = P(X \in S_c | H) = P(X \in S_c | \boldsymbol{\lambda}) \quad .$$

*Operating characteristic function:*

$$L(S_c, \boldsymbol{\lambda}) = 1 - M(S_c, \boldsymbol{\lambda}) \quad .$$

*Most powerful test* of  $H_0$  with respect to  $H_1$  has  $M(S_c, \boldsymbol{\lambda}_1) = 1 - \beta = \max$ . A *uniformly most powerful test* is a most powerful test with respect to all possible  $H_1$ .

An *unbiased test* has  $M(S_c, \boldsymbol{\lambda}_1) \geq \alpha$  for all possible  $H_1$ .

**NEYMAN–PEARSON LEMMA:** A test of  $H_0$  with respect to  $H_1$  (both simple hypotheses) with the critical region  $S_c$  is a most powerful test if  $f(X|H_0)/f(X|H_1) \leq c$  for  $X \in S_c$  and  $\geq c$  for  $X \notin S_c$ , where  $c$  is a constant only depending on  $\alpha$ .

*Test statistic*  $T(X)$ : Scalar function of the sample  $X$ . By means of a mapping  $X \rightarrow T(X)$ ,  $S_c(X) \rightarrow U$  the question as to whether  $X \in S_c$  can be reformulated as  $T \in U$ .

*Likelihood-ratio test:* If  $\omega$  denotes the region in the parameter space corresponding to the null hypothesis and  $\Omega$  denotes the entire possible parameter region, then the test statistic

$$T = f(x; \tilde{\boldsymbol{\lambda}}^{(\Omega)})/f(x; \tilde{\boldsymbol{\lambda}}^{(\omega)})$$

is used. Here  $\tilde{\boldsymbol{\lambda}}^{(\Omega)}$  and  $\tilde{\boldsymbol{\lambda}}^{(\omega)}$  are the maximum-likelihood estimators in the regions  $\Omega$  and  $\omega$ .  $H_0$  is rejected if  $T > T_{1-\alpha}$  with  $P(T > T_{1-\alpha} | H_0) = \int_{T_{1-\alpha}}^{\infty} g(T) dT = \alpha$ ;  $g(T)$  is the conditional probability density of  $T$  for a given  $H_0$ .

**WILKS theorem** (holds for weak requirements concerning the probability density of the population): If  $H_0$  specifies  $p - r$  out of the  $p$  parameters, then  $-2 \ln T$  (where  $T$  is the likelihood-ratio test statistic) follows a  $\chi^2$ -distribution with  $f = p - r$  degrees of freedom in the limit  $n \rightarrow \infty$ .

## $\chi^2$ -Test for Goodness-of-Fit

*Hypothesis:* The  $N$  measured values  $y_i$  with normally distributed errors  $\sigma_i$  are described by given quantities  $f_i$ .

**Table H.5:** Frequently used statistical tests for a sample from a normal distribution with mean  $\lambda$  and variance  $\sigma^2$ . (Case 1:  $\sigma$  known; case 2:  $\sigma$  unknown (Student's test); case 3:  $\chi^2$ -test of the variance; case 4: Student's difference test of two samples of sizes  $N_1$  and  $N_2$  on the significance of  $s_{\Delta}^2$ , cf. (8.3.19); case 5:  $F$ -test of two samples).

| Case | Test statistic   | Null hypothesis   | Critical region for test statistic  | Number of degrees of freedom       |
|------|--|---|---|------------------------------------|
| 1    | $T = \frac{\bar{x} - \lambda_0}{\sigma/\sqrt{N}},$<br>$\bar{x} = \frac{1}{N} \sum_{j=1}^N x^{(j)}$               | $\lambda = \lambda_0$<br>$\lambda \leq \lambda_0$<br>$\lambda \geq \lambda_0$             | $ T  > \Omega(1 - \alpha/2)$<br>$T > \Omega(1 - \alpha)$<br>$T < \Omega(\alpha)$                | —                                  |
| 2    | $T = \frac{\bar{x} - \lambda_0}{s/\sqrt{N}},$<br>$s^2 = \frac{1}{N-1} \times \sum_{j=1}^N (x^{(j)} - \bar{x})^2$ | $\lambda = \lambda_0$<br>$\lambda \leq \lambda_0$<br>$\lambda \geq \lambda_0$             | $ T  > t_{1-\alpha/2}$<br>$T > t_{1-\alpha}$<br>$T < -t_{1-\alpha} = t_\alpha$                  | $N - 1$                            |
| 3    | $T = (N-1) \frac{s^2}{\sigma_0^2}$   | $\sigma^2 = \sigma_0^2$<br>$\sigma^2 \leq \sigma_0^2$<br>$\sigma^2 \geq \sigma_0^2$       | $\chi_{1-\alpha/2}^2 < T < \chi_{\alpha/2}^2$<br>$T > \chi_{1-\alpha}^2$<br>$T < \chi_\alpha^2$ | $N - 1$                            |
| 4    | $T = \frac{\bar{x}_1 - \bar{x}_2}{s_{\Delta}}$   | $\lambda_1 = \lambda_2$   | $ T  > t_{1-\frac{1}{2}\alpha}$   | $N_1 + N_2 - 2$                    |
| 5    | $T = s_1^2/s_2^2,$<br>$s_i^2 = \frac{1}{N_i-1} \times \sum_{j=1}^{N_i} (x_i^{(j)} - \bar{x}_i)^2$                | $\sigma_1^2 = \sigma_2^2$<br>$\sigma_1^2 \leq \sigma_2^2$<br>$\sigma_1^2 \geq \sigma_2^2$ | $F_{1-\alpha/2} < T < F_{\alpha/2}$<br>$T > F_{1-\alpha}$<br>$T < F_\alpha$                     | $f_1 = N_1 - 1$<br>$f_2 = N_2 - 1$ |

*Test function:*  $T = \sum_{i=1}^N (y_i - f_i)^2 / \sigma_i^2$ .

*Critical region:*  $T > \chi_{1-\alpha}^2$ .

*Number of degrees of freedom:*  $N$  (or  $N - p$ , if  $p$  parameters are determined from the measurements).

## The Method of Least Squares

One considers a set of  $m$  equations  $f_k(\mathbf{x}, \boldsymbol{\eta}) = 0$ ;  $k = 1, \dots, m$  relating the  $r$ -vector of the *unknowns*  $\mathbf{x} = (x_1, x_2, \dots, x_r)$  to the  $n$ -vector of *measurable quantities*  $\boldsymbol{\eta} = (\eta_1, \eta_2, \dots, \eta_n)$ . Instead of  $\boldsymbol{\eta}$ , the quantities  $\mathbf{y}$  are measured, which deviate from them by the *measurement errors*  $\boldsymbol{\varepsilon}$ , i.e.,  $\mathbf{y} = \boldsymbol{\eta} + \boldsymbol{\varepsilon}$ . The quantities  $\boldsymbol{\varepsilon}$  are assumed to be normally distributed about zero. This is expressed by the covariance matrix  $C_y = G_y^{-1}$ . In order to obtain the solution  $\tilde{\mathbf{x}}$ ,  $\tilde{\boldsymbol{\eta}}$ , one expands the  $f_k$  for the *first approximations*  $\mathbf{x}_0$ ,  $\boldsymbol{\eta}_0 = \mathbf{y}$ . Only the linear terms of the expansion are kept and the second approximation  $\mathbf{x}_1 = \mathbf{x}_0 + \boldsymbol{\xi}$ ,  $\boldsymbol{\eta}_1 = \boldsymbol{\eta} + \boldsymbol{\delta}$  is computed. The procedure is repeated iteratively until certain convergence criteria are met, for example, until a scalar function  $M$  no longer decreases. If the  $f_k$  are linear in  $\mathbf{x}$  and  $\boldsymbol{\eta}$ , then only one step is necessary. The method can be interpreted as a procedure to minimize  $M$ . The function  $M$  corresponding to the solution depends on the measurement errors. It is a random variable and follows a  $\chi^2$ -distribution with  $f = m - r$  degrees of freedom. It can therefore be used for a  $\chi^2$ -test of the goodness-of-fit or of other assumptions, in particular the assumption  $C_y = G_y^{-1}$ . If the errors  $\boldsymbol{\varepsilon}$  are not normally distribution, but still distributed symmetrically about zero, then the least-squares solution  $\tilde{\mathbf{x}}$  still has the smallest possible variance, and one has  $E(M) = m - r$  (*Gauss–Markov theorem*).

**Table H.6:** Least squares in the general case and in the case of constrained measurements.

|                                       | General case  | Constrained measurements  |
|---------------------------------------|---|---|
| Equations                             | $f_k(\mathbf{x}, \boldsymbol{\eta}) = 0, k = 1, \dots, m$   | $f_k(\boldsymbol{\eta}) = 0$  |
| First approximations                  | $\mathbf{x}_0, \boldsymbol{\eta}_0 = \mathbf{y}$<br>$\mathbf{f} = A\tilde{\boldsymbol{\xi}} + B\boldsymbol{\delta} + \mathbf{c} + \dots$  | $\boldsymbol{\eta}_0 = \mathbf{y}$<br>$\mathbf{f} = B\boldsymbol{\delta} + \mathbf{c} + \dots$  |
| Equations expanded                    | $\{A\}_{kl} = (\partial f_k / \partial x_l)_{\mathbf{x}_0, \boldsymbol{\eta}_0}$<br>$\{B\}_{kl} = (\partial f_k / \partial \eta_l)_{\mathbf{x}_0, \boldsymbol{\eta}_0}$<br>$\mathbf{c} = \mathbf{f}(\mathbf{x}_0, \boldsymbol{\eta}_0)$   | $\bar{\{B\}}_{kl} = (\partial f_k / \partial \eta_l)_{\boldsymbol{\eta}_0}$<br>$\mathbf{c} = \mathbf{f}(\boldsymbol{\eta}_0)$                                     |
| Covariance matrix of the measurements | $C_y = G_y^{-1}$  | $C_y = G_y^{-1}$  |
| Corrections                           | $\tilde{\boldsymbol{\xi}} = -(A^T G_B A)^{-1} A^T G_B \mathbf{c}$<br>$\tilde{\boldsymbol{\delta}} = -G_y^{-1} B^T G_B (A\tilde{\boldsymbol{\xi}} + \mathbf{c})$<br>$G_B = (B G_y^{-1} B^T)^{-1}$  | $\tilde{\boldsymbol{\delta}} = -G_y^{-1} B^T G_B \mathbf{c}$<br>$G_B = (B G_y^{-1} B^T)^{-1}$   |
| Next step                             | $\mathbf{x}_1 = \mathbf{x}_0 + \tilde{\boldsymbol{\xi}}, \boldsymbol{\eta}_1 = \boldsymbol{\eta}_0 + \tilde{\boldsymbol{\delta}}$ ,<br>new values for $A, B, \mathbf{c}, \tilde{\boldsymbol{\xi}}, \tilde{\boldsymbol{\delta}}$           | $\boldsymbol{\eta}_1 = \boldsymbol{\eta}_0 + \tilde{\boldsymbol{\delta}}$ ,<br>new values for $B, \mathbf{c}, \tilde{\boldsymbol{\delta}}$                        |
| Solution (after $s$ steps)            | $\tilde{\mathbf{x}} = \mathbf{x}_{s-1} + \tilde{\boldsymbol{\xi}}$ ,<br>$\tilde{\boldsymbol{\eta}} = \boldsymbol{\eta}_{s-1} + \tilde{\boldsymbol{\delta}}$ ,<br>$\tilde{\boldsymbol{\epsilon}} = \mathbf{y} - \tilde{\boldsymbol{\eta}}$ | $\tilde{\boldsymbol{\eta}} = \boldsymbol{\eta}_{s-1} + \tilde{\boldsymbol{\delta}}$ ,<br>$\tilde{\boldsymbol{\epsilon}} = \mathbf{y} - \tilde{\boldsymbol{\eta}}$ |
| Minimum function                      | $M = (B\tilde{\boldsymbol{\epsilon}})^T G_B (B\tilde{\boldsymbol{\epsilon}})$   | $M = (B\tilde{\boldsymbol{\epsilon}})^T G_B (B\tilde{\boldsymbol{\epsilon}})$   |
| Covariance matrices                   | $G_{\tilde{\boldsymbol{\xi}}}^{-1} = (A^T G_B A)^{-1}$<br>$G_{\tilde{\boldsymbol{\eta}}}^{-1} = G_y^{-1}$<br>$-G_y^{-1} B^T G_B B G_y^{-1}$<br>$+ G_y^{-1} B^T G_B A (A^T G_B A)^{-1}$<br>$\times A^T G_B B G_y^{-1}$                     | $G_{\tilde{\boldsymbol{\eta}}}^{-1} = G_y^{-1}$<br>$-G_y^{-1} B^T G_B B G_y^{-1}$   |

**Table H.7:** Least squares for indirect and direct measurements.

|   | Indirect measurements   | Direct measurements<br>of different<br>accuracy  |
|---|---|--|
| Equations                                   | $f_k = \eta_k - g_k(x) = 0$   | $f_k = \eta_k - x$   |
| First approx-<br>imations                   | $\mathbf{x}_0, \boldsymbol{\eta}_0 = \mathbf{y}$  | $x_0 = 0, \boldsymbol{\eta}_0 = \mathbf{y}$  |
|   |   | $\mathbf{f} = A\tilde{\boldsymbol{\xi}} + \boldsymbol{\epsilon} + \mathbf{c}$  |
| Equations<br>expanded                       | $\mathbf{f} = A\tilde{\boldsymbol{\xi}} - \boldsymbol{\epsilon} + \mathbf{c} + \dots$<br>$\{A\}_{kl} = \left( \frac{\partial f_k}{\partial x_l} \right)_{\mathbf{x}_0}$<br>$\mathbf{c} = \mathbf{y} - \mathbf{g}(\mathbf{x}_0)$ | $A = - \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$<br>$\mathbf{c} = \mathbf{y}$<br>$C_y = G_y^{-1}$<br>$= \begin{pmatrix} \sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_n^2 \end{pmatrix}$ |
| Covariance<br>matrix of the<br>measurements | $C_y = G_y^{-1}$  |  |
| Corrections                                 | $\tilde{\boldsymbol{\xi}} = -(A^T G_y A)^{-1} A^T G_y \mathbf{c}$   | –  |
| Next<br>step                                | $\mathbf{x}_1 = \mathbf{x}_0 + \tilde{\boldsymbol{\xi}},$<br>new values for $A, \mathbf{c}, \tilde{\boldsymbol{\xi}}$   | –  |
| Solution<br>(after $s$<br>steps)            | $\tilde{\mathbf{x}} = \mathbf{x}_{s-1} + \tilde{\boldsymbol{\xi}},$<br>$\tilde{\boldsymbol{\epsilon}} = A\tilde{\boldsymbol{\xi}} + \mathbf{c}$   | $\tilde{\boldsymbol{\xi}} = \tilde{\mathbf{x}} = \frac{\sum_k y_k / \sigma_k^2}{\sum_k 1 / \sigma_k^2},$<br>$\tilde{\boldsymbol{\epsilon}}_k = y_k - \tilde{x}$                                      |
| Minimum<br>function                         | $M = \tilde{\boldsymbol{\epsilon}}^T G_y \tilde{\boldsymbol{\epsilon}}$   | $M = \tilde{\boldsymbol{\epsilon}}^T G_y \tilde{\boldsymbol{\epsilon}}$  |
| Covariance<br>matrices                      | $G_{\tilde{x}}^{-1} = (A^T G_y A)^{-1}$<br>$G_{\tilde{\eta}}^{-1} = A(A^T G_y A)A^T$  | $G_{\tilde{x}}^{-1} = \sigma^2(\tilde{x})$<br>$= \left( \sum_k \frac{1}{\sigma_k^2} \right)^{-1}$<br>–   |

## Analysis of Variance

The influence of external variables on a measured random variable  $x$  is investigated. One tries to decide by means of appropriately constructed  $F$ -tests whether  $x$  is independent of the *external variables*. Various *models* are constructed depending on the number of external variables and the assumptions concerning their influence.

A simple model is the *crossed two-way classification* with multiple observations. Two external variables are used to classify the observations into the classes  $A_i, B_j$  ( $i = 1, 2, \dots, I; j = 1, 2, \dots, J$ ). Each class  $A_i, B_j$  contains  $K$  observations  $x_{ijk}$  ( $k = 1, 2, \dots, K$ ). One assumes the model

$$x_{ijk} = \mu + a_i + b_j + (ab)_{ij} + \varepsilon_{ijk} \quad ,$$

where the error of the observation  $\varepsilon_{ijk}$  is assumed to be normally distributed about zero and  $a_i, b_j$ , and  $(ab)_{ij}$  are the influences of the *classifications* in  $A, B$  and their *interactions*. Three *null hypotheses*

$$H_0^{(A)}(a_i = 0, i = 1, \dots, I) \quad , \quad H_0^{(B)}(b_j = 0, j = 1, \dots, J) \quad ,$$

$$H_0^{(AB)}((ab)_{ij} = 0, i = 1, \dots, I, j = 1, \dots, J)$$

can be tested with the ratios  $F^{(A)}, F^{(B)}, F^{(AB)}$ . They are summarized in an analysis of variance (ANOVA) table. For other models see Chap. 11.

## Polynomial Regression

*Problem:* The true values  $\eta(t)$ , for which one has  $N$  *measurements*  $y_i(t_i)$  with normally distributed measurement errors  $\sigma_i$ , are to be described by a *polynomial* of order  $r - 1$  in the *controlled variable*  $t$ . Instead of  $\eta(t) = x_1 + x_2 t + \dots + x_r t^{r-1}$  one writes

$$\eta(t) = x_1 f_1(t) + x_2 f_2(t) + \dots + x_r f_r(t) \quad .$$

Here the  $f_j$  are *orthogonal polynomials* of order  $j - 1$ ,

$$f_j(t) = \sum_{k=1}^j b_{jk} t^{k-1} \quad ,$$

whose coefficients  $b_{jk}$  are determined by the *orthogonality conditions*

$$\sum_{i=1}^N g_i f_j(t_i) f_k(t_i) = \delta_{jk} \quad , \quad g_i = 1/\sigma_i^2 \quad .$$

The *unknowns*  $x_j$  are obtained by least squares from

$$\sum_{i=1}^N g_i \left\{ y_i(t_i) - \sum_{j=1}^r x_j f_j(t_i) \right\}^2 = \min \quad .$$

The covariance matrix of the  $x_j$  is the  $r$ -dimensional unit matrix.

## Time Series Analysis

One is given a series of *measured values*  $y_i(t_i)$ ,  $i = 1, \dots, n$ , which (in an unknown way) depend on a *controlled variable*  $t$  (usually time). One treats the  $y_i$  as the sum of a *trend*  $\eta_i$  and an *error*  $\varepsilon_i$ ,  $y_i = \eta_i + \varepsilon_i$ . The measurements are carried out at regular time intervals, i.e.,  $t_i - t_{i-1} = \text{const}$ . In order to minimize the errors  $\varepsilon_i$ , a *moving average* is constructed for every  $t_i$  ( $i > k$ ,  $i \leq n - k$ ), by fitting a polynomial of order  $\ell$  to the  $2k + 1$  measurements situated symmetrically about measurement  $i$ . The result of the fit at the point  $t_i$  is the moving average

$$\tilde{\eta}_0(i) = a_{-k} y_{i-k} + a_{-k+1} y_{i-k+1} + \dots + a_k y_{i+k} \quad .$$

The coefficients  $a_{-k}, \dots, a_k$  are given in Table 13.1 for low values of  $k$  and  $\ell$ . For the beginning and end points  $t_i$  ( $i < k$ ,  $i > n - k$ ), the results of the fit can be used with caution also for points other than at the center of the interval of the  $2k + 1$  measurements.

# I. Statistical Tables

**Table I.1:** Quantiles  $\lambda_P(k)$  of the Poisson distribution.

| $k$ | $P$    |        |        |        |        |        |        |
|-----|--------|--------|--------|--------|--------|--------|--------|
|     | 0.0005 | 0.0010 | 0.0050 | 0.0100 | 0.0250 | 0.0500 | 0.1000 |
| 1   | 7.601  | 6.908  | 5.298  | 4.605  | 3.689  | 2.996  | 2.303  |
| 2   | 9.999  | 9.233  | 7.430  | 6.638  | 5.572  | 4.744  | 3.890  |
| 3   | 12.051 | 11.229 | 9.274  | 8.406  | 7.225  | 6.296  | 5.322  |
| 4   | 13.934 | 13.062 | 10.977 | 10.045 | 8.767  | 7.754  | 6.681  |
| 5   | 15.710 | 14.794 | 12.594 | 11.605 | 10.242 | 9.154  | 7.994  |
| 6   | 17.411 | 16.455 | 14.150 | 13.108 | 11.668 | 10.513 | 9.275  |
| 7   | 19.055 | 18.062 | 15.660 | 14.571 | 13.059 | 11.842 | 10.532 |
| 8   | 20.654 | 19.626 | 17.134 | 16.000 | 14.423 | 13.148 | 11.771 |
| 9   | 22.217 | 21.156 | 18.578 | 17.403 | 15.763 | 14.435 | 12.995 |
| 10  | 23.749 | 22.657 | 19.998 | 18.783 | 17.085 | 15.705 | 14.206 |
| 11  | 25.256 | 24.134 | 21.398 | 20.145 | 18.390 | 16.962 | 15.407 |
| 12  | 26.739 | 25.589 | 22.779 | 21.490 | 19.682 | 18.208 | 16.598 |
| 13  | 28.203 | 27.026 | 24.145 | 22.821 | 20.962 | 19.443 | 17.782 |
| 14  | 29.650 | 28.446 | 25.497 | 24.139 | 22.230 | 20.669 | 18.958 |
| 15  | 31.081 | 29.852 | 26.836 | 25.446 | 23.490 | 21.886 | 20.128 |
| 16  | 32.498 | 31.244 | 28.164 | 26.743 | 24.740 | 23.097 | 21.292 |
| 17  | 33.902 | 32.624 | 29.482 | 28.030 | 25.983 | 24.301 | 22.452 |
| 18  | 35.294 | 33.993 | 30.791 | 29.310 | 27.219 | 25.499 | 23.606 |
| 19  | 36.676 | 35.351 | 32.091 | 30.581 | 28.448 | 26.692 | 24.756 |
| 20  | 38.047 | 36.701 | 33.383 | 31.845 | 29.671 | 27.879 | 25.903 |
| 21  | 39.410 | 38.042 | 34.668 | 33.103 | 30.888 | 29.062 | 27.045 |
| 22  | 40.764 | 39.375 | 35.946 | 34.355 | 32.101 | 30.240 | 28.184 |
| 23  | 42.110 | 40.700 | 37.218 | 35.601 | 33.308 | 31.415 | 29.320 |
| 24  | 43.449 | 42.019 | 38.484 | 36.841 | 34.511 | 32.585 | 30.453 |
| 25  | 44.780 | 43.330 | 39.745 | 38.077 | 35.710 | 33.752 | 31.584 |

**Table I.1:** (continued)

| $k$  | $P$    |        |        |        |        |        |        |
|--|--------|--------|--------|--------|--------|--------|--------|
|  | 0.9000 | 0.9500 | 0.9750 | 0.9900 | 0.9950 | 0.9990 | 0.9995 |
| $P = \sum_{n=0}^{k-1} e^{-\lambda_P} \lambda_P^n / n!$ |        |        |        |        |        |        |        |
| 1  | 0.105  | 0.051  | 0.025  | 0.010  | 0.005  | 0.001  | 0.001  |
| 2  | 0.532  | 0.355  | 0.242  | 0.149  | 0.103  | 0.045  | 0.032  |
| 3  | 1.102  | 0.818  | 0.619  | 0.436  | 0.338  | 0.191  | 0.150  |
| 4  | 1.745  | 1.366  | 1.090  | 0.823  | 0.672  | 0.429  | 0.355  |
| 5  | 2.433  | 1.970  | 1.623  | 1.279  | 1.078  | 0.739  | 0.632  |
| 6  | 3.152  | 2.613  | 2.202  | 1.785  | 1.537  | 1.107  | 0.967  |
| 7  | 3.895  | 3.285  | 2.814  | 2.330  | 2.037  | 1.520  | 1.348  |
| 8  | 4.656  | 3.981  | 3.454  | 2.906  | 2.571  | 1.971  | 1.768  |
| 9  | 5.432  | 4.695  | 4.115  | 3.507  | 3.132  | 2.452  | 2.220  |
| 10   | 6.221  | 5.425  | 4.795  | 4.130  | 3.717  | 2.961  | 2.699  |
| 11   | 7.021  | 6.169  | 5.491  | 4.771  | 4.321  | 3.491  | 3.202  |
| 12   | 7.829  | 6.924  | 6.201  | 5.428  | 4.943  | 4.042  | 3.726  |
| 13   | 8.646  | 7.690  | 6.922  | 6.099  | 5.580  | 4.611  | 4.269  |
| 14   | 9.470  | 8.464  | 7.654  | 6.782  | 6.231  | 5.195  | 4.828  |
| 15   | 10.300 | 9.246  | 8.395  | 7.477  | 6.893  | 5.794  | 5.402  |
| 16   | 11.135 | 10.036 | 9.145  | 8.181  | 7.567  | 6.405  | 5.990  |
| 17   | 11.976 | 10.832 | 9.903  | 8.895  | 8.251  | 7.028  | 6.590  |
| 18   | 12.822 | 11.634 | 10.668 | 9.616  | 8.943  | 7.662  | 7.201  |
| 19   | 13.671 | 12.442 | 11.439 | 10.346 | 9.644  | 8.306  | 7.822  |
| 20   | 14.525 | 13.255 | 12.217 | 11.082 | 10.353 | 8.958  | 8.453  |
| 21   | 15.383 | 14.072 | 12.999 | 11.825 | 11.069 | 9.619  | 9.093  |
| 22   | 16.244 | 14.894 | 13.787 | 12.574 | 11.792 | 10.288 | 9.741  |
| 23   | 17.108 | 15.719 | 14.580 | 13.329 | 12.521 | 10.964 | 10.397 |
| 24   | 17.975 | 16.549 | 15.377 | 14.089 | 13.255 | 11.647 | 11.060 |
| 25   | 18.844 | 17.382 | 16.179 | 14.853 | 13.995 | 12.337 | 11.730 |

**Table I.2:** Normal distribution  $\psi_0(x)$ .

| $x$  | 0     | 1     | 2     | 3     | 4     | 5     | 6     | 7     | 8     | 9     |
|------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| -3.0 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 |
| -2.9 | 0.002 | 0.002 | 0.002 | 0.002 | 0.002 | 0.002 | 0.002 | 0.001 | 0.001 | 0.001 |
| -2.8 | 0.003 | 0.002 | 0.002 | 0.002 | 0.002 | 0.002 | 0.002 | 0.002 | 0.002 | 0.002 |
| -2.7 | 0.003 | 0.003 | 0.003 | 0.003 | 0.003 | 0.003 | 0.003 | 0.003 | 0.003 | 0.003 |
| -2.6 | 0.005 | 0.005 | 0.004 | 0.004 | 0.004 | 0.004 | 0.004 | 0.004 | 0.004 | 0.004 |
| -2.5 | 0.006 | 0.006 | 0.006 | 0.006 | 0.005 | 0.005 | 0.005 | 0.005 | 0.005 | 0.005 |
| -2.4 | 0.008 | 0.008 | 0.008 | 0.008 | 0.007 | 0.007 | 0.007 | 0.007 | 0.007 | 0.006 |
| -2.3 | 0.011 | 0.010 | 0.010 | 0.010 | 0.010 | 0.009 | 0.009 | 0.009 | 0.009 | 0.008 |
| -2.2 | 0.014 | 0.014 | 0.013 | 0.013 | 0.013 | 0.012 | 0.012 | 0.012 | 0.011 | 0.011 |
| -2.1 | 0.018 | 0.017 | 0.017 | 0.017 | 0.016 | 0.016 | 0.015 | 0.015 | 0.015 | 0.014 |
| -2.0 | 0.023 | 0.022 | 0.022 | 0.021 | 0.021 | 0.020 | 0.020 | 0.019 | 0.019 | 0.018 |
| -1.9 | 0.029 | 0.028 | 0.027 | 0.027 | 0.026 | 0.026 | 0.025 | 0.024 | 0.024 | 0.023 |
| -1.8 | 0.036 | 0.035 | 0.034 | 0.034 | 0.033 | 0.032 | 0.031 | 0.031 | 0.030 | 0.029 |
| -1.7 | 0.045 | 0.044 | 0.043 | 0.042 | 0.041 | 0.040 | 0.039 | 0.038 | 0.038 | 0.037 |
| -1.6 | 0.055 | 0.054 | 0.053 | 0.052 | 0.051 | 0.049 | 0.048 | 0.047 | 0.046 | 0.046 |
| -1.5 | 0.067 | 0.066 | 0.064 | 0.063 | 0.062 | 0.061 | 0.059 | 0.058 | 0.057 | 0.056 |
| -1.4 | 0.081 | 0.079 | 0.078 | 0.076 | 0.075 | 0.074 | 0.072 | 0.071 | 0.069 | 0.068 |
| -1.3 | 0.097 | 0.095 | 0.093 | 0.092 | 0.090 | 0.089 | 0.087 | 0.085 | 0.084 | 0.082 |
| -1.2 | 0.115 | 0.113 | 0.111 | 0.109 | 0.107 | 0.106 | 0.104 | 0.102 | 0.100 | 0.099 |
| -1.1 | 0.136 | 0.133 | 0.131 | 0.129 | 0.127 | 0.125 | 0.123 | 0.121 | 0.119 | 0.117 |
| -1.0 | 0.159 | 0.156 | 0.154 | 0.152 | 0.149 | 0.147 | 0.145 | 0.142 | 0.140 | 0.138 |
| -0.9 | 0.184 | 0.181 | 0.179 | 0.176 | 0.174 | 0.171 | 0.169 | 0.166 | 0.164 | 0.161 |
| -0.8 | 0.212 | 0.209 | 0.206 | 0.203 | 0.200 | 0.198 | 0.195 | 0.192 | 0.189 | 0.187 |
| -0.7 | 0.242 | 0.239 | 0.236 | 0.233 | 0.230 | 0.227 | 0.224 | 0.221 | 0.218 | 0.215 |
| -0.6 | 0.274 | 0.271 | 0.268 | 0.264 | 0.261 | 0.258 | 0.255 | 0.251 | 0.248 | 0.245 |
| -0.5 | 0.309 | 0.305 | 0.302 | 0.298 | 0.295 | 0.291 | 0.288 | 0.284 | 0.281 | 0.278 |
| -0.4 | 0.345 | 0.341 | 0.337 | 0.334 | 0.330 | 0.326 | 0.323 | 0.319 | 0.316 | 0.312 |
| -0.3 | 0.382 | 0.378 | 0.374 | 0.371 | 0.367 | 0.363 | 0.359 | 0.356 | 0.352 | 0.348 |
| -0.2 | 0.421 | 0.417 | 0.413 | 0.409 | 0.405 | 0.401 | 0.397 | 0.394 | 0.390 | 0.386 |
| -0.1 | 0.460 | 0.456 | 0.452 | 0.448 | 0.444 | 0.440 | 0.436 | 0.433 | 0.429 | 0.425 |
| 0.0  | 0.500 | 0.496 | 0.492 | 0.488 | 0.484 | 0.480 | 0.476 | 0.472 | 0.468 | 0.464 |

**Table I.2:** (continued)

**Table I.3:** Normal distribution  $2\psi_0(x) - 1$ .

**Table I.4:** Quantiles  $x_P = \Omega(P)$  of the normal distribution.

| $P$ | 0         | 1     | 2     | 3     | 4     | 5     | 6     | 7     | 8     | 9     |
|-----|-----------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
|     | $-\infty$ | -2.33 | -2.05 | -1.88 | -1.75 | -1.64 | -1.55 | -1.48 | -1.41 | -1.34 |
| 0.1 | -1.28     | -1.23 | -1.17 | -1.13 | -1.08 | -1.04 | -0.99 | -0.95 | -0.92 | -0.88 |
| 0.2 | -0.84     | -0.81 | -0.77 | -0.74 | -0.71 | -0.67 | -0.64 | -0.61 | -0.58 | -0.55 |
| 0.3 | -0.52     | -0.50 | -0.47 | -0.44 | -0.41 | -0.39 | -0.36 | -0.33 | -0.31 | -0.28 |
| 0.4 | -0.25     | -0.23 | -0.20 | -0.18 | -0.15 | -0.13 | -0.10 | -0.08 | -0.05 | -0.03 |
| 0.5 | 0.00      | 0.03  | 0.05  | 0.08  | 0.10  | 0.13  | 0.15  | 0.18  | 0.20  | 0.23  |
| 0.6 | 0.25      | 0.28  | 0.31  | 0.33  | 0.36  | 0.39  | 0.41  | 0.44  | 0.47  | 0.50  |
| 0.7 | 0.52      | 0.55  | 0.58  | 0.61  | 0.64  | 0.67  | 0.71  | 0.74  | 0.77  | 0.81  |
| 0.8 | 0.84      | 0.88  | 0.92  | 0.95  | 0.99  | 1.04  | 1.08  | 1.13  | 1.17  | 1.23  |
| 0.9 | 1.28      | 1.34  | 1.41  | 1.48  | 1.55  | 1.64  | 1.75  | 1.88  | 2.05  | 2.33  |

$$P = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x_P} \exp(-x^2/2) dx$$

**Table I.5:** Quantiles  $x'_P = \Omega'(P)$  of the normal distribution.

| $P$   | 0     | 1     | 2     | 3     | 4     | 5     | 6     | 7     | 8     | 9     |
|---|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| $P = \frac{1}{\sqrt{2\pi}} \int_{-x'_P}^{x'_P} \exp(-x^2/2) dx$ |       |       |       |       |       |       |       |       |       |       |
| 0.0   | 0.000 | 0.013 | 0.025 | 0.038 | 0.050 | 0.063 | 0.075 | 0.088 | 0.100 | 0.113 |
| 0.1   | 0.126 | 0.138 | 0.151 | 0.164 | 0.176 | 0.189 | 0.202 | 0.215 | 0.228 | 0.240 |
| 0.2   | 0.253 | 0.266 | 0.279 | 0.292 | 0.305 | 0.319 | 0.332 | 0.345 | 0.358 | 0.372 |
| 0.3   | 0.385 | 0.399 | 0.412 | 0.426 | 0.440 | 0.454 | 0.468 | 0.482 | 0.496 | 0.510 |
| 0.4   | 0.524 | 0.539 | 0.553 | 0.568 | 0.583 | 0.598 | 0.613 | 0.628 | 0.643 | 0.659 |
| 0.5   | 0.674 | 0.690 | 0.706 | 0.722 | 0.739 | 0.755 | 0.772 | 0.789 | 0.806 | 0.824 |
| 0.6   | 0.842 | 0.860 | 0.878 | 0.896 | 0.915 | 0.935 | 0.954 | 0.974 | 0.994 | 1.015 |
| 0.7   | 1.036 | 1.058 | 1.080 | 1.103 | 1.126 | 1.150 | 1.175 | 1.200 | 1.227 | 1.254 |
| 0.8   | 1.282 | 1.311 | 1.341 | 1.372 | 1.405 | 1.440 | 1.476 | 1.514 | 1.555 | 1.598 |
| 0.9   | 1.645 | 1.695 | 1.751 | 1.812 | 1.881 | 1.960 | 2.054 | 2.170 | 2.326 | 2.576 |
| $P$   | 0     | 1     | 2     | 3     | 4     | 5     | 6     | 7     | 8     | 9     |
| 0.90  | 1.645 | 1.650 | 1.655 | 1.660 | 1.665 | 1.670 | 1.675 | 1.680 | 1.685 | 1.690 |
| 0.91  | 1.695 | 1.701 | 1.706 | 1.711 | 1.717 | 1.722 | 1.728 | 1.734 | 1.739 | 1.745 |
| 0.92  | 1.751 | 1.757 | 1.762 | 1.768 | 1.774 | 1.780 | 1.787 | 1.793 | 1.799 | 1.805 |
| 0.93  | 1.812 | 1.818 | 1.825 | 1.832 | 1.838 | 1.845 | 1.852 | 1.859 | 1.866 | 1.873 |
| 0.94  | 1.881 | 1.888 | 1.896 | 1.903 | 1.911 | 1.919 | 1.927 | 1.935 | 1.943 | 1.951 |
| 0.95  | 1.960 | 1.969 | 1.977 | 1.986 | 1.995 | 2.005 | 2.014 | 2.024 | 2.034 | 2.044 |
| 0.96  | 2.054 | 2.064 | 2.075 | 2.086 | 2.097 | 2.108 | 2.120 | 2.132 | 2.144 | 2.157 |
| 0.97  | 2.170 | 2.183 | 2.197 | 2.212 | 2.226 | 2.241 | 2.257 | 2.273 | 2.290 | 2.308 |
| 0.98  | 2.326 | 2.346 | 2.366 | 2.387 | 2.409 | 2.432 | 2.457 | 2.484 | 2.512 | 2.543 |
| 0.99  | 2.576 | 2.612 | 2.652 | 2.697 | 2.748 | 2.807 | 2.878 | 2.968 | 3.090 | 3.291 |
| $P$   | 0     | 1     | 2     | 3     | 4     | 5     | 6     | 7     | 8     | 9     |
| 0.990   | 2.576 | 2.579 | 2.583 | 2.586 | 2.590 | 2.594 | 2.597 | 2.601 | 2.605 | 2.608 |
| 0.991   | 2.612 | 2.616 | 2.620 | 2.624 | 2.628 | 2.632 | 2.636 | 2.640 | 2.644 | 2.648 |
| 0.992   | 2.652 | 2.656 | 2.661 | 2.665 | 2.669 | 2.674 | 2.678 | 2.683 | 2.687 | 2.692 |
| 0.993   | 2.697 | 2.702 | 2.706 | 2.711 | 2.716 | 2.721 | 2.727 | 2.732 | 2.737 | 2.742 |
| 0.994   | 2.748 | 2.753 | 2.759 | 2.765 | 2.770 | 2.776 | 2.782 | 2.788 | 2.794 | 2.801 |
| 0.995   | 2.807 | 2.814 | 2.820 | 2.827 | 2.834 | 2.841 | 2.848 | 2.855 | 2.863 | 2.870 |
| 0.996   | 2.878 | 2.886 | 2.894 | 2.903 | 2.911 | 2.920 | 2.929 | 2.938 | 2.948 | 2.958 |
| 0.997   | 2.968 | 2.978 | 2.989 | 3.000 | 3.011 | 3.023 | 3.036 | 3.048 | 3.062 | 3.076 |
| 0.998   | 3.090 | 3.105 | 3.121 | 3.138 | 3.156 | 3.175 | 3.195 | 3.216 | 3.239 | 3.264 |
| 0.999   | 3.291 | 3.320 | 3.353 | 3.390 | 3.432 | 3.481 | 3.540 | 3.615 | 3.719 | 3.891 |

**Table I.6:**  $\chi^2$ -distribution  $F(\chi^2)$ .

| $\chi^2$ | $f$   |       |       |       |       |       |       |       |       |       |
|----------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
|          | 1     | 2     | 3     | 4     | 5     | 6     | 7     | 8     | 9     | 10    |
| 0.1      | 0.248 | 0.049 | 0.008 | 0.001 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| 0.2      | 0.345 | 0.095 | 0.022 | 0.005 | 0.001 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| 0.3      | 0.416 | 0.139 | 0.040 | 0.010 | 0.002 | 0.001 | 0.000 | 0.000 | 0.000 | 0.000 |
| 0.4      | 0.473 | 0.181 | 0.060 | 0.018 | 0.005 | 0.001 | 0.000 | 0.000 | 0.000 | 0.000 |
| 0.5      | 0.520 | 0.221 | 0.081 | 0.026 | 0.008 | 0.002 | 0.001 | 0.000 | 0.000 | 0.000 |
| 0.6      | 0.561 | 0.259 | 0.104 | 0.037 | 0.012 | 0.004 | 0.001 | 0.000 | 0.000 | 0.000 |
| 0.7      | 0.597 | 0.295 | 0.127 | 0.049 | 0.017 | 0.006 | 0.002 | 0.000 | 0.000 | 0.000 |
| 0.8      | 0.629 | 0.330 | 0.151 | 0.062 | 0.023 | 0.008 | 0.003 | 0.001 | 0.000 | 0.000 |
| 0.9      | 0.657 | 0.362 | 0.175 | 0.075 | 0.030 | 0.011 | 0.004 | 0.001 | 0.000 | 0.000 |
| 1.0      | 0.683 | 0.393 | 0.199 | 0.090 | 0.037 | 0.014 | 0.005 | 0.002 | 0.001 | 0.000 |
| 2.0      | 0.843 | 0.632 | 0.428 | 0.264 | 0.151 | 0.080 | 0.040 | 0.019 | 0.009 | 0.004 |
| 3.0      | 0.917 | 0.777 | 0.608 | 0.442 | 0.300 | 0.191 | 0.115 | 0.066 | 0.036 | 0.019 |
| 4.0      | 0.954 | 0.865 | 0.739 | 0.594 | 0.451 | 0.323 | 0.220 | 0.143 | 0.089 | 0.053 |
| 5.0      | 0.975 | 0.918 | 0.828 | 0.713 | 0.584 | 0.456 | 0.340 | 0.242 | 0.166 | 0.109 |
| 6.0      | 0.986 | 0.950 | 0.888 | 0.801 | 0.694 | 0.577 | 0.460 | 0.353 | 0.260 | 0.185 |
| 7.0      | 0.992 | 0.970 | 0.928 | 0.864 | 0.779 | 0.679 | 0.571 | 0.463 | 0.363 | 0.275 |
| 8.0      | 0.995 | 0.982 | 0.954 | 0.908 | 0.844 | 0.762 | 0.667 | 0.567 | 0.466 | 0.371 |
| 9.0      | 0.997 | 0.989 | 0.971 | 0.939 | 0.891 | 0.826 | 0.747 | 0.658 | 0.563 | 0.468 |
| 10.0     | 0.998 | 0.993 | 0.981 | 0.960 | 0.925 | 0.875 | 0.811 | 0.735 | 0.650 | 0.560 |
| 11.0     | 0.999 | 0.996 | 0.988 | 0.973 | 0.949 | 0.912 | 0.861 | 0.798 | 0.724 | 0.642 |
| 12.0     | 0.999 | 0.998 | 0.993 | 0.983 | 0.965 | 0.938 | 0.899 | 0.849 | 0.787 | 0.715 |
| 13.0     | 1.000 | 0.998 | 0.995 | 0.989 | 0.977 | 0.957 | 0.928 | 0.888 | 0.837 | 0.776 |
| 14.0     | 1.000 | 0.999 | 0.997 | 0.993 | 0.984 | 0.970 | 0.949 | 0.918 | 0.878 | 0.827 |
| 15.0     | 1.000 | 0.999 | 0.998 | 0.995 | 0.990 | 0.980 | 0.964 | 0.941 | 0.909 | 0.868 |
| 16.0     | 1.000 | 1.000 | 0.999 | 0.997 | 0.993 | 0.986 | 0.975 | 0.958 | 0.933 | 0.900 |
| 17.0     | 1.000 | 1.000 | 0.999 | 0.998 | 0.996 | 0.991 | 0.983 | 0.970 | 0.951 | 0.926 |
| 18.0     | 1.000 | 1.000 | 1.000 | 0.999 | 0.997 | 0.994 | 0.988 | 0.979 | 0.965 | 0.945 |
| 19.0     | 1.000 | 1.000 | 1.000 | 0.999 | 0.998 | 0.996 | 0.992 | 0.985 | 0.975 | 0.960 |
| 20.0     | 1.000 | 1.000 | 1.000 | 1.000 | 0.999 | 0.997 | 0.994 | 0.990 | 0.982 | 0.971 |

**Table I.7:** Quantiles  $\chi_P^2$  of the  $\chi^2$ -distribution.

| $f$ | $P$     |         |         |         |         |
|-----|---------|---------|---------|---------|---------|
|     | 0.900   | 0.950   | 0.990   | 0.995   | 0.999   |
| 1   | 2.706   | 3.841   | 6.635   | 7.879   | 10.828  |
| 2   | 4.605   | 5.991   | 9.210   | 10.597  | 13.816  |
| 3   | 6.251   | 7.815   | 11.345  | 12.838  | 16.266  |
| 4   | 7.779   | 9.488   | 13.277  | 14.860  | 18.467  |
| 5   | 9.236   | 11.070  | 15.086  | 16.750  | 20.515  |
| 6   | 10.645  | 12.592  | 16.812  | 18.548  | 22.458  |
| 7   | 12.017  | 14.067  | 18.475  | 20.278  | 24.322  |
| 8   | 13.362  | 15.507  | 20.090  | 21.955  | 26.124  |
| 9   | 14.684  | 16.919  | 21.666  | 23.589  | 27.877  |
| 10  | 15.987  | 18.307  | 23.209  | 25.188  | 29.588  |
| 11  | 17.275  | 19.675  | 24.725  | 26.757  | 31.264  |
| 12  | 18.549  | 21.026  | 26.217  | 28.300  | 32.909  |
| 13  | 19.812  | 22.362  | 27.688  | 29.819  | 34.528  |
| 14  | 21.064  | 23.685  | 29.141  | 31.319  | 36.123  |
| 15  | 22.307  | 24.996  | 30.578  | 32.801  | 37.697  |
| 16  | 23.542  | 26.296  | 32.000  | 34.267  | 39.252  |
| 17  | 24.769  | 27.587  | 33.409  | 35.718  | 40.790  |
| 18  | 25.989  | 28.869  | 34.805  | 37.156  | 42.312  |
| 19  | 27.204  | 30.144  | 36.191  | 38.582  | 43.820  |
| 20  | 28.412  | 31.410  | 37.566  | 39.997  | 45.315  |
| 30  | 40.256  | 43.773  | 50.892  | 53.672  | 59.703  |
| 40  | 51.805  | 55.758  | 63.691  | 66.766  | 73.402  |
| 50  | 63.167  | 67.505  | 76.154  | 79.490  | 86.661  |
| 60  | 74.397  | 79.082  | 88.379  | 91.952  | 99.607  |
| 70  | 85.527  | 90.531  | 100.425 | 104.215 | 112.317 |
| 80  | 80.000  | 101.879 | 112.329 | 116.321 | 124.839 |
| 90  | 107.565 | 113.145 | 124.116 | 128.299 | 137.208 |
| 100 | 118.498 | 124.342 | 135.807 | 140.169 | 149.449 |

**Table I.8:** Quantiles  $F_P$  of the  $F$ -distribution.

| $0.900 = P = \int_0^{F_P} f(F; f_1, f_2) dF$ |       |       |       |       |       |       |       |       |       |       |
|--|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| $f_2$  | $f_1$ |       |       |       |       |       |       |       |       |       |
|  | 1     | 2     | 3     | 4     | 5     | 6     | 7     | 8     | 9     | 10    |
| 1  | 39.86 | 49.50 | 53.59 | 55.83 | 57.24 | 58.20 | 58.91 | 59.44 | 59.86 | 60.19 |
| 2  | 8.526 | 9.000 | 9.162 | 9.243 | 9.293 | 9.326 | 9.349 | 9.367 | 9.381 | 9.392 |
| 3  | 5.538 | 5.462 | 5.391 | 5.343 | 5.309 | 5.285 | 5.266 | 5.252 | 5.240 | 5.230 |
| 4  | 4.545 | 4.325 | 4.191 | 4.107 | 4.051 | 4.010 | 3.979 | 3.955 | 3.936 | 3.920 |
| 5  | 4.060 | 3.780 | 3.619 | 3.520 | 3.453 | 3.405 | 3.368 | 3.339 | 3.316 | 3.297 |
| 6  | 3.776 | 3.463 | 3.289 | 3.181 | 3.108 | 3.055 | 3.014 | 2.983 | 2.958 | 2.937 |
| 7  | 3.589 | 3.257 | 3.074 | 2.961 | 2.883 | 2.827 | 2.785 | 2.752 | 2.725 | 2.703 |
| 8  | 3.458 | 3.113 | 2.924 | 2.806 | 2.726 | 2.668 | 2.624 | 2.589 | 2.561 | 2.538 |
| 9  | 3.360 | 3.006 | 2.813 | 2.693 | 2.611 | 2.551 | 2.505 | 2.469 | 2.440 | 2.416 |
| 10   | 3.285 | 2.924 | 2.728 | 2.605 | 2.522 | 2.461 | 2.414 | 2.377 | 2.347 | 2.323 |

| $0.950 = P = \int_0^{F_P} f(F; f_1, f_2) dF$ |       |       |       |       |       |       |       |       |       |       |
|--|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| $f_2$  | $f_1$ |       |       |       |       |       |       |       |       |       |
|  | 1     | 2     | 3     | 4     | 5     | 6     | 7     | 8     | 9     | 10    |
| 1  | 161.4 | 199.5 | 215.7 | 224.6 | 230.2 | 234.0 | 236.8 | 238.9 | 240.5 | 241.9 |
| 2  | 18.51 | 19.00 | 19.16 | 19.25 | 19.30 | 19.33 | 19.35 | 19.37 | 19.38 | 19.40 |
| 3  | 10.13 | 9.552 | 9.277 | 9.117 | 9.013 | 8.941 | 8.887 | 8.845 | 8.812 | 8.786 |
| 4  | 7.709 | 6.944 | 6.591 | 6.388 | 6.256 | 6.163 | 6.094 | 6.041 | 5.999 | 5.964 |
| 5  | 6.608 | 5.786 | 5.409 | 5.192 | 5.050 | 4.950 | 4.876 | 4.818 | 4.772 | 4.735 |
| 6  | 5.987 | 5.143 | 4.757 | 4.534 | 4.387 | 4.284 | 4.207 | 4.147 | 4.099 | 4.060 |
| 7  | 5.591 | 4.737 | 4.347 | 4.120 | 3.972 | 3.866 | 3.787 | 3.726 | 3.677 | 3.637 |
| 8  | 5.318 | 4.459 | 4.066 | 3.838 | 3.687 | 3.581 | 3.500 | 3.438 | 3.388 | 3.347 |
| 9  | 5.117 | 4.256 | 3.863 | 3.633 | 3.482 | 3.374 | 3.293 | 3.230 | 3.179 | 3.137 |
| 10   | 4.965 | 4.103 | 3.708 | 3.478 | 3.326 | 3.217 | 3.135 | 3.072 | 3.020 | 2.978 |

**Table I.8:** (continued)

|       |       | $0.975 = P = \int_0^{F_P} f(F; f_1, f_2) dF$ |       |       |       |       |       |       |       |       |    |
|-------|-------|--|-------|-------|-------|-------|-------|-------|-------|-------|----|
|       |       | $f_1$  |       |       |       |       |       |       |       |       |    |
| $f_2$ |       | 1  | 2     | 3     | 4     | 5     | 6     | 7     | 8     | 9     | 10 |
| 1     | 647.8 | 799.5  | 864.2 | 899.6 | 921.8 | 937.1 | 948.2 | 956.7 | 963.3 | 968.6 |    |
| 2     | 38.51 | 39.00  | 39.17 | 39.25 | 39.30 | 39.33 | 39.36 | 39.37 | 39.39 | 39.40 |    |
| 3     | 17.44 | 16.04  | 15.44 | 15.10 | 14.88 | 14.73 | 14.62 | 14.54 | 14.47 | 14.42 |    |
| 4     | 12.22 | 10.65  | 9.979 | 9.605 | 9.364 | 9.197 | 9.074 | 8.980 | 8.905 | 8.844 |    |
| 5     | 10.01 | 8.434  | 7.764 | 7.388 | 7.146 | 6.978 | 6.853 | 6.757 | 6.681 | 6.619 |    |
| 6     | 8.813 | 7.260  | 6.599 | 6.227 | 5.988 | 5.820 | 5.695 | 5.600 | 5.523 | 5.461 |    |
| 7     | 8.073 | 6.542  | 5.890 | 5.523 | 5.285 | 5.119 | 4.995 | 4.899 | 4.823 | 4.761 |    |
| 8     | 7.571 | 6.059  | 5.416 | 5.053 | 4.817 | 4.652 | 4.529 | 4.433 | 4.357 | 4.295 |    |
| 9     | 7.209 | 5.715  | 5.078 | 4.718 | 4.484 | 4.320 | 4.197 | 4.102 | 4.026 | 3.964 |    |
| 10    | 6.937 | 5.456  | 4.826 | 4.468 | 4.236 | 4.072 | 3.950 | 3.855 | 3.779 | 3.717 |    |

|       |       | $0.990 = P = \int_0^{F_P} f(F; f_1, f_2) dF$ |       |       |       |       |       |       |       |       |    |
|-------|-------|--|-------|-------|-------|-------|-------|-------|-------|-------|----|
|       |       | $f_1$  |       |       |       |       |       |       |       |       |    |
| $f_2$ |       | 1  | 2     | 3     | 4     | 5     | 6     | 7     | 8     | 9     | 10 |
| 1     | 4052  | 5000   | 5403  | 5625  | 5764  | 5859  | 5928  | 5981  | 6022  | 6056  |    |
| 2     | 98.50 | 99.00  | 99.17 | 99.25 | 99.30 | 99.33 | 99.36 | 99.37 | 99.39 | 99.40 |    |
| 3     | 34.12 | 30.82  | 29.46 | 28.71 | 28.24 | 27.91 | 27.67 | 27.49 | 27.35 | 27.23 |    |
| 4     | 21.20 | 18.00  | 16.69 | 15.98 | 15.52 | 15.21 | 14.98 | 14.80 | 14.66 | 14.55 |    |
| 5     | 16.26 | 13.27  | 12.06 | 11.39 | 10.97 | 10.67 | 10.46 | 10.29 | 10.16 | 10.05 |    |
| 6     | 13.75 | 10.92  | 9.780 | 9.148 | 8.746 | 8.466 | 8.260 | 8.102 | 7.976 | 7.874 |    |
| 7     | 12.25 | 9.547  | 8.451 | 7.847 | 7.460 | 7.191 | 6.993 | 6.840 | 6.719 | 6.620 |    |
| 8     | 11.26 | 8.649  | 7.591 | 7.006 | 6.632 | 6.371 | 6.178 | 6.029 | 5.911 | 5.814 |    |
| 9     | 10.56 | 8.022  | 6.992 | 6.422 | 6.057 | 5.802 | 5.613 | 5.467 | 5.351 | 5.257 |    |
| 10    | 10.04 | 7.559  | 6.552 | 5.994 | 5.636 | 5.386 | 5.200 | 5.057 | 4.942 | 4.849 |    |

**Table I.9:** Quantiles  $t_P$  of Student's distribution.

| $f$  | $P$    |        |        |        |        |         |         |
|------|--------|--------|--------|--------|--------|---------|---------|
|      | 0.9000 | 0.9500 | 0.9750 | 0.9900 | 0.9950 | 0.9990  | 0.9995  |
| 1    | 3.078  | 6.314  | 12.706 | 31.821 | 63.657 | 318.309 | 636.619 |
| 2    | 1.886  | 2.920  | 4.303  | 6.965  | 9.925  | 22.327  | 31.599  |
| 3    | 1.638  | 2.353  | 3.182  | 4.541  | 5.841  | 10.215  | 12.924  |
| 4    | 1.533  | 2.132  | 2.776  | 3.747  | 4.604  | 7.173   | 8.610   |
| 5    | 1.476  | 2.015  | 2.571  | 3.365  | 4.032  | 5.893   | 6.869   |
| 6    | 1.440  | 1.943  | 2.447  | 3.143  | 3.707  | 5.208   | 5.959   |
| 7    | 1.415  | 1.895  | 2.365  | 2.998  | 3.499  | 4.785   | 5.408   |
| 8    | 1.397  | 1.860  | 2.306  | 2.896  | 3.355  | 4.501   | 5.041   |
| 9    | 1.383  | 1.833  | 2.262  | 2.821  | 3.250  | 4.297   | 4.781   |
| 10   | 1.372  | 1.812  | 2.228  | 2.764  | 3.169  | 4.144   | 4.587   |
| 11   | 1.363  | 1.796  | 2.201  | 2.718  | 3.106  | 4.025   | 4.437   |
| 12   | 1.356  | 1.782  | 2.179  | 2.681  | 3.055  | 3.930   | 4.318   |
| 13   | 1.350  | 1.771  | 2.160  | 2.650  | 3.012  | 3.852   | 4.221   |
| 14   | 1.345  | 1.761  | 2.145  | 2.624  | 2.977  | 3.787   | 4.140   |
| 15   | 1.341  | 1.753  | 2.131  | 2.602  | 2.947  | 3.733   | 4.073   |
| 16   | 1.337  | 1.746  | 2.120  | 2.583  | 2.921  | 3.686   | 4.015   |
| 17   | 1.333  | 1.740  | 2.110  | 2.567  | 2.898  | 3.646   | 3.965   |
| 18   | 1.330  | 1.734  | 2.101  | 2.552  | 2.878  | 3.610   | 3.922   |
| 19   | 1.328  | 1.729  | 2.093  | 2.539  | 2.861  | 3.579   | 3.883   |
| 20   | 1.325  | 1.725  | 2.086  | 2.528  | 2.845  | 3.552   | 3.850   |
| 30   | 1.310  | 1.697  | 2.042  | 2.457  | 2.750  | 3.385   | 3.646   |
| 40   | 1.303  | 1.684  | 2.021  | 2.423  | 2.704  | 3.307   | 3.551   |
| 50   | 1.299  | 1.676  | 2.009  | 2.403  | 2.678  | 3.261   | 3.496   |
| 60   | 1.296  | 1.671  | 2.000  | 2.390  | 2.660  | 3.232   | 3.460   |
| 70   | 1.294  | 1.667  | 1.994  | 2.381  | 2.648  | 3.211   | 3.435   |
| 80   | 1.292  | 1.664  | 1.990  | 2.374  | 2.639  | 3.195   | 3.416   |
| 90   | 1.291  | 1.662  | 1.987  | 2.368  | 2.632  | 3.183   | 3.402   |
| 100  | 1.290  | 1.660  | 1.984  | 2.364  | 2.626  | 3.174   | 3.390   |
| 200  | 1.286  | 1.653  | 1.972  | 2.345  | 2.601  | 3.131   | 3.340   |
| 500  | 1.283  | 1.648  | 1.965  | 2.334  | 2.586  | 3.107   | 3.310   |
| 1000 | 1.282  | 1.646  | 1.962  | 2.330  | 2.581  | 3.098   | 3.300   |

# List of Computer Programs\*

|   |                    |
|---|--------------------|
| AnalysisOfVariance, 318, 319              | E2Distrib, 107     |
| AuxDer, 426, 428                          | E2Gr, 443          |
| AuxDri, 230, 426, 428                     | E2Lsq, 263         |
| AuxGrad, 426, 428                         | E2MaxLike, 173     |
| AuxHesse, 426, 428                        | E2Min, 305         |
| AuxJInputGroup, 427, 429                  | E2Mtx, 400         |
| AuxJNumberInput, 428, 429                 | E2Random, 67       |
| AuxJRButtonGroup, 427, 429                | E2Reg, 329         |
| AuxZero, 427, 428                         | E2Sample, 150, 443 |
| DatanFrame, 427, 428                      | E2Test, 205        |
| DatanGraphics, 118, 431, 432, 435,<br>436 | E2TimSer, 340      |
| DatanMatrix, 348, 399                     | E3Distrib, 107     |
| DatanRandom, 67                           | E3Gr, 444          |
| DatanUserFunction, 230, 257, 275          | E3Lsq, 263         |
| DatanVector, 348, 399                     | E3Min, 306         |
| E10Mtx, 403                               | E3Mtx, 400         |
| E1Anova, 319                              | E3Random, 68       |
| E1Distrib, 106                            | E3Reg, 329         |
| E1Gr, 443                                 | E3Sample, 150, 443 |
| E1Lsq, 263                                | E3Test, 206        |
| E1MaxLike, 173, 427                       | E4Gr, 444          |
| E1Min, 305                                | E4Lsq, 264         |
| E1Mtx, 399                                | E4Min, 306         |
| E1Random, 67                              | E4Mtx, 400         |
| E1Reg, 329                                | E4Random, 68       |
| E1Sample, 150                             | E4Reg, 329         |
| E1Test, 205                               | E4Sample, 150      |
| E1TimSer, 340                             | E5Gr, 444          |
| E2Anova, 319                              | E5Lsq, 264         |
|   | E5Mtx, 401         |

\*The slanted numbers refer to the Appendix.

- E5Sample, 150  
E6Gr, 443, 445  
E6Lsq, 264  
E6Mtx, 401  
E6Sample, 151  
E7Gr, 443, 445  
E7Lsq, 264  
E7Mtx, 401  
E7Sample, 151  
E8Gr, 443, 446  
E8Lsq, 265  
E8Mtx, 402  
E9Lsq, 265  
E9Mtx, 402  
FunctionOnLine, 275, 303  
FunctionsDemo, 414, 422  
Gamma, 422, 441  
GraphicsWith2DScatterDiagram,  
120, 150, 443  
GraphicsWithDataPointsAndPolyline,  
443  
GraphicsWithDataPointsAnd-  
MultiplePolylines,  
443  
GraphicsWithHistogramAndPolyline,  
443  
GraphicsWithHistogram, 118, 150,  
443  
Histogram, 149, 150  
LsqAsg, 263, 265  
LsqAsm, 243, 262  
LsqAsn, 243, 262, 264  
LsqGE, 264  
LsqGen, 255, 257, 262, 264, 265  
LsqLin, 224, 261, 263  
LsqMar, 235, 243, 262, 264  
LsqNon, 230, 235, 243, 261, 263  
LsqPol, 223, 261, 263  
MinAsy, 298, 305, 306  
MinCjg, 291, 304, 305  
MinCombined, 280, 303  
MinCov, 297, 305  
MinDir, 280, 281, 303  
MinEnclose, 280, 303  
MinMar, 294, 304, 305  
MinParab, 303  
MinPow, 288, 304, 305  
MinQdr, 292, 304, 305  
MinSim, 283, 303, 305  
Regression, 329  
S1Anova, 484  
S1Distrib, 472  
S1Lsq, 475  
S1MaxLike, 473  
S1Min, 483  
S1Random, 470  
S1Reg, 485  
S1TimSer, 485  
S2Anova, 484  
S2Lsq, 476  
S2MaxLike, 473  
S2Min, 483  
S2Random, 470  
S2Reg, 485  
S2TimSer, 486  
S3Lsq, 478  
S3Random, 471  
S4Lsq, 479  
S5Lsq, 479  
S6Lsq, 481, 482  
Sample, 115, 149, 150  
SmallSample, 149, 150  
StatFunct, 414, 427  
TimeSeries, 340

# Index\*

- a posteriori probability, 153  
acceptance probability, 188  
addition of errors  
    quadratic, 105  
alternative hypothesis, 186, 494  
analysis of variance, 307, 500  
    model, 312  
    one-way, 307  
    table, 310  
    two-way, 311  
        crossed classification, 483  
        nested classification, 484  
and, 10  
ANOVA, 307  
    table, 310  
asymmetric error, 167, 298  
asymptotically unbiased estimator, 165  
average  
    moving, 332, 501  
  
background, 142  
backward substitution, 374  
basis, 351  
beta function, 418  
    incomplete, 420  
bias, 157, 494  
bidiagonal matrix, 350  
binomial coefficient, 406, 418  
binomial distribution, 70, 409, 451  
    parameter estimation, 163  
binomial theorem, 406  
bit, 42  
bracketing the minimum, 275  
  
Breit–Wigner distribution, 25, 57, 470  
byte, 43  
  
Cauchy distribution, 23  
central limit theorem, 90, 491  
characteristic equation, 378  
 $\chi^2$ -distribution, 130, 412, 453, 510  
    quantiles, 511  
 $\chi^2$ -test, 199, 206, 455  
Cholesky decomposition, 372, 401  
Cholesky inversion, 401  
classification, 307, 500  
    crossed, 313  
    nested, 313, 316  
    one-way, 311  
    two-way, 313  
clipping region, 436  
cofactor, 361  
color index, 436  
column space, 352  
column vector, 348  
combination, 405  
combinatorics, 405  
computing coordinates, 433  
confidence  
    ellipsoid, 100, 240, 297  
    interval, 326, 336  
    level, 133, 494  
    limits, 297, 298  
    region, 170, 241, 260, 297, 479  
conjugate directions, 285  
consistent estimator, 111  
constrained measurements, 243, 258  
constraint, 396, 403  
    equations, 244

\*The slanted numbers refer to the Appendix.

- contingency table, 203, 456  
 continued fraction, 418  
 contour line, 438  
 controlled variable, 218  
 convolution, 101, 452, 491  
     of uniform distribution and normal distribution, 472  
     of uniform distributions, 102, 471  
     with the normal distribution, 103  
 coordinate cross  
     in graphics, 440  
 correlation coefficient, 29, 489  
 correlation coefficient of a sample, 473  
 counted number, 137  
 covariance, 29, 32, 489  
     ellipse, 97, 306  
     ellipsoid, 99, 240  
     matrix, 297, 489  
     weighted, 393  
 critical region, 177, 187, 494  
  
 decile, 22  
 degrees of freedom, 128, 238  
 derivative  
     logarithmic, 156  
 determinant, 360  
 device coordinates, 433  
 diagonal matrix, 349  
 dice, 11  
 direct measurements  
     of different accuracy, 210  
     of equal accuracy, 209  
 direct sum, 353  
 dispersion, 19  
 distribution  
     binomial, 70, 409, 451  
     Breit–Wigner, 57, 470  
     Breit–Wigner, 25  
     Cauchy, 23  
      $\chi^2$ , 412, 453, 510  
         quantiles, 511  
     F, 413  
         quantiles, 512  
     frequency, 109  
     function, 16, 487  
         of sample, 110  
         of several variables, 488  
         of several variables, 30  
         of two variables, 25  
     Gaussian, 86  
     hypergeometric, 74, 409  
  
     Lorentz, 25  
     multivariate normal, 452  
     multivariate normal, 94  
     normal, 86, 410, 451, 505, 507  
         quantiles, 508, 509  
     of a continuous variable, 492  
     of a discrete variable, 491  
     Poisson, 78, 410, 451  
     Polya, 77  
     standard normal, 84, 410  
         quantiles, 508, 509  
     Student's, 182, 413  
         quantiles, 514  
     t, 413  
         quantiles, 514  
     triangular, 58, 448, 470  
     uniform, 22  
     unimodal, 21  
  
 efficient estimator, 111  
 eigenvalue, 376  
     equation, 376  
 eigenvector, 376  
 elements, 244  
 equations of constraint, 238  
 error  
     asymmetric, 167, 173, 242, 260, 298, 306, 479  
     combination, 164  
     of mean, 113  
     of sample variance, 113  
     of the first kind, 187, 495  
     of the second kind, 187, 495  
     standard, 89  
     statistical, 73, 106, 137  
     symmetric, 297, 306  
 error bars, 117  
 error function, 411  
 error model of Laplace, 92  
 error propagation, 37, 450, 490  
 E space, 186  
 estimation, 111  
 estimator, 111, 493, 494  
     asymptotically unbiased, 165, 494  
     consistent, 111, 493  
     efficiency of, 452  
     efficient, 111  
     minimum variance, 161  
     unbiased, 111, 493  
     unique, 494  
 event, 8, 487

- expectation value, 17, 31, 488, 489
- experiment, 7
- factorial, 418
- F*-distribution, 413
  - quantiles, 512
- fit
  - of a Breit–Wigner-function, 478
  - of a polynomial, 473
  - of a power law, 475
  - of an arbitrary linear function, 224
  - of an exponential function, 232
  - of a Breit–Wigner function, 480
  - of a circle to points with
    - measurement errors in abscissa and ordinate, 481
  - of a Gaussian, 231, 263
  - of a nonlinear function, 228
  - of a polynomial, 222, 263
  - of a proportionality, 224, 263
  - of a straight line, 218
  - of a straight line to points with
    - measurement errors in abscissa and ordinate, 264
  - of a sum of exponential functions, 233
  - of a sum of two Gaussians and a polynomial, 235, 264
- fluctuation
  - statistical, 106
- forward substitution, 374
- frequency, 106, 490
- frequency distribution, 109
- F*-test, 177, 205, 455
- full width at half maximum (FWHM), 119
- functional equation, 415
- FWHM, 119
- Galton's board, 107
- gamma function, 415
  - incomplete, 420
- Gauss–Markov theorem, 238
- Gaussian distribution, 86
  - multivariate, 94
- Gaussian elimination, 367
  - with pivoting, 369
- Givens transformation, 354
- go button, 427
- golden section, 278
- graphics, 431
  - class, 431
  - workstation, 431
- Hessian matrix, 271
- histogram, 117, 454
  - bin width, 117
- determination of parameters from, 306
- Householder transformation, 356, 400
- hypergeometric distribution, 74, 409
- hyperplane, 353
- hypothesis, 175, 186
  - alternative, 186, 494
  - composite, 186, 494
  - null, 186, 494
  - simple, 186, 494
  - test of, 494
- implementation, 50
- incomplete beta function, 420
- incomplete gamma function, 420
- independence
  - of events, 11
  - of random variables, 26, 31
- indirect measurements
  - linear case, 214
  - nonlinear case, 226
- information, 160, 454
  - inequality, 157, 161, 494
  - of a sample, 494
- input group, 427
- interaction, 315
- inverse matrix, 365
- Jacobian determinant, 35, 490
  - of an orthogonal transformation, 40
- kernel, 352
- Kronecker symbol, 128
- Lagrange function, 248
- Lagrange multipliers, 126, 247
- Laplace error model, 92
- law of large numbers, 73, 490
- LCG, 45

- least squares, 209, 362, 497
  - according to Marquardt, 394
  - constrained measurements, 243, 498
  - direct measurements, 209, 499
  - general case, 251, 498
  - indirect measurements, 214, 226, 499
  - properties of the solution, 236
  - with constraints, 396, 403
  - with change of scale, 393
  - with weights, 392
- likelihood
  - equation, 155, 494
  - function, 154, 493
    - logarithmic, 155, 493
    - ratio, 154
  - likelihood-ratio test, 194, 495
- linear combination, 351
- linear system of equations, 362
- Lorentz distribution, 25
- lotto, 12
- LR-decomposition, 369
- main diagonal, 349
- mantissa, 43
- mapping, 352
- marginal probability density, 26, 31, 488
- Marquardt minimization, 292
- matrix, 348
  - addition, 348
  - adjoint, 361, 366
  - antisymmetric, 350
  - bidiagonal, 350
  - diagonal, 349
  - equations, 362
  - inverse, 365
  - main diagonal, 349
  - multiplication
    - by a constant, 348
    - by a matrix, 348
  - norm, 350
  - null, 349
  - orthogonal, 354
  - product, 348
  - pseudo-inverse, 375
  - rank, 352
  - singular, 352, 361
  - subtraction, 348
  - symmetric, 350
  - transposition, 348
- triangular, 350
- tridiagonal, 350
- unit, 349
- maximum likelihood, 493
- maximum-likelihood estimates, 454
- mean, 17
  - error, 113
  - of sample, 111
- mean square, 128
- mean square deviation, 128
- median, 21, 488
- minimization, 267
  - combined method by BRENT, 280
  - along a direction, 280
  - along chosen directions, 287
  - along the coordinate directions, 284
  - bracketing the minimum, 275
  - by the method of POWELL, 287
  - choice of method, 295
  - errors in, 296
  - examples, 298
  - golden section, 277
  - in the direction of steepest descent, 271, 288
  - Marquardt procedure, 292
  - quadratic interpolation, 280
  - simplex method, 281
  - with the quadratic form, 292
- minimum variance
  - bound, 161
  - estimator, 161
- MLCG, 45
- mode, 21, 488
- moments, 18, 27, 31, 487, 489
  - about the mean, 18
- Monte Carlo method, 41
  - for simulation, 66
  - minimization, 483
  - of integration, 64
- moving average, 332, 501
- Neyman–Pearson lemma, 191, 495
- norm
  - Euclidian, 350
- normal distribution, 86, 410, 451, 505, 507
  - multivariate, 63, 94, 452
  - quantiles, 508, 509
  - standard, 84, 410
- normal equations, 365
- null hypothesis, 186, 494

- null matrix, 349
- null space, 352
- null vector, 349
- number-input region, 428
  
- one-sided test, 176
- operating characteristic function, 188, 495
- orthogonal, 349
- orthogonal complements, 353
- orthogonal matrix, 354
- orthogonal polynomials, 500
- orthogonal transformation, 354
  
- parabola through three points, 273
- Pascal's triangle, 92, 406
- permutation, 405
- permutation transformation, 359
- pivot, 369
- Poisson distribution, 78, 410, 451
  - parameter estimation, 162
  - quantiles, 503
- Polya distribution, 77
- polyline, 437
- polymarker, 437
- polynomial
  - orthogonal, 500
- polynomial regression, 500
- population, 109, 492
- portability, 50
- power function, 188, 495
- precision
  - absolute, 44
  - relative, 44
- primitive element, 46
- principal axes, 378
- principal axis transformation, 377
- probability, 487
  - a posteriori, 153
  - conditional, 10
  - density, 16, 487
    - conditional, 26
    - joint, 488
    - marginal, 26, 31, 488
    - of several variables, 31
    - of two variables, 26
  - frequency definition, 9
  - total, 10
- pseudo-inverse matrix, 375
  
- quadratic average of individual errors, 164
- quantile, 22, 488
- quartile, 22
  
- radio-button group, 427
- random component, 332
- random number generator, 44
  - linear congruential (LCG), 45
  - multiplicative linear congruential (MLCG), 45
- random numbers, 41
  - arbitrarily distributed, 55
  - acceptance-rejection technique, 59
  - transformation procedures, 56
  - normally distributed, 62
- random variable, 15, 487
  - continuous, 15
  - discrete, 15
- rank, 352
- ratio of small numbers of events, 144
  - with background, 147
- reduced variable, 20, 488
- regression, 321
  - curve, 325
  - polynomial, 325, 500
- representation of numbers in a computer, 42
- row space, 352
- row vector, 348
  
- sample, 74, 109, 492
  - correlation coefficient, 473
  - distribution function, 110
  - error
    - of variance, 113
    - of mean, 113
  - from a bivariate normal distribution, 173
  - from a continuous population, 111
  - from finite population, 127
  - from Gaussian distribution, 130
  - from partitioned population, 453
  - from subpopulation, 122
  - graphical representation, 115
  - information, 160
  - mean, 111, 150, 453
  - random, 110

- size, 109
- small, 136
  - with background, 142
- space, 7
- variance, 112, 150, 453
- scalar, 349
- scalar product, 349
- scale factor, 213
- scale in graphics, 439
- scatter diagram, 150
  - one-dimensional, 116
  - two-dimensional, 120
- seed, 52
- sign inversion, 359
- signal, 142
- significance level, 175, 494
- simplex, 282
- singular matrix, 352, 361
- singular value, 379
- singular value analysis, 380, 383
- singular value decomposition, 379, 385, 401, 402
- skewness, 20, 488
- small numbers of events, 136
  - ratio, 144
    - with background, 147
- small sample, 136
- span, 352
- standard deviation, 19, 89, 488
- standard normal distribution, 410
- statistic, 111, 493
  - test, 187
- statistical error, 73, 137
- statistical test, 175
- steepest descent, 288
- step diagram, 117
- Student's difference test, 184
- Student's distribution, 182, 413
  - quantiles, 514
- Student's test, 180, 205, 455
- subpopulation, 122
- subspace, 352
- sum of squares, 127, 308
- symmetric errors, 297
- system of equations
  - linear, 362
  - triangular, 368
  - underdetermined, 362
- t*-distribution, 182, 413
  - quantiles, 514
- test, 175, 494, 496
  - $\chi^2$ , 199, 206, 455, 495
  - likelihood-ratio, 194, 495
  - most powerful, 188, 495
  - one-sided, 176
  - statistic, 176, 187, 495
  - Student's, 205
  - two-sided, 176
  - unbiased, 188, 495
  - uniformly most powerful, 188
- text in plot, 441
- three-door game, 13
- time series, 331
  - analysis, 331, 501
  - extrapolation, 485
  - discontinuities, 485
- transformation
  - Givens, 354, 400
  - Householder, 356, 400
  - linear, 36
  - of a vector, 352
  - of variables, 33, 449, 489
  - orthogonal, 39, 354
  - permutation, 359
  - principal axis, 377
  - sign inversion, 359
- transposition, 348
- trend, 332, 501
- triangular distribution, 58, 448, 470
- triangular matrix, 350
- triangular system of equations, 368
- tridiagonal matrix, 350
- two-sided test, 176
- 2 × 2 table, 204
- 2 × 2 table test, 204
- underdetermined system of equations, 362
- uniform distribution, 22
- unit matrix, 349
- unit vector, 350
- variance, 27, 32, 488, 489
  - of sample, 112
  - of an estimator, 454
  - of a random variable, 18
- vector, 348
  - absolute value, 350
  - components, 348
  - norm, 350

- null, [349](#)
- row, [348](#)
- space, [351](#)
  - basis, [351](#)
  - closed, [351](#)
  - dimension, [351](#)
- transformation, [352](#)
- unit, [350](#)
- vectors
  - linearly dependent, [351](#)
- linearly independent, [351](#)
- orthonormal, [351](#)
- viewport, [433](#)
- weight matrix, [215](#)
- weighted covariance matrix, [393](#)
- width
  - full at half maximum, [119](#)
- Wilks theorem, [195, 495](#)
- window, [433](#)
- world coordinates, [432](#)