

Graduate Texts in Physics

Massimiliano Bonamente

Statistics and Analysis of Scientific Data

Third Edition

MOREMEDIA



Springer

Graduate Texts in Physics

Series Editors

Kurt H. Becker, NYU Polytechnic School of Engineering, Brooklyn, NY, USA

Jean-Marc Di Meglio, Matière et Systèmes Complexes, Bâtiment Condorcet,
Université Paris Diderot, Paris, France

Sadri Hassani, Department of Physics, Illinois State University, Normal, IL, USA

Morten Hjorth-Jensen, Department of Physics, Blindern, University of Oslo, Oslo,
Norway

Bill Munro, NTT Basic Research Laboratories, Atsugi, Japan

Richard Needs, Cavendish Laboratory, University of Cambridge, Cambridge, UK

William T. Rhodes, Department of Computer and Electrical Engineering and
Computer Science, Florida Atlantic University, Boca Raton, FL, USA

Susan Scott, Australian National University, Acton, Australia

H. Eugene Stanley, Center for Polymer Studies, Physics Department, Boston
University, Boston, MA, USA

Martin Stutzmann, Walter Schottky Institute, Technical University of Munich,
Garching, Germany

Andreas Wipf, Institute of Theoretical Physics, Friedrich-Schiller-University Jena,
Jena, Germany

Graduate Texts in Physics publishes core learning/teaching material for graduate- and advanced-level undergraduate courses on topics of current and emerging fields within physics, both pure and applied. These textbooks serve students at the MS- or PhD-level and their instructors as comprehensive sources of principles, definitions, derivations, experiments and applications (as relevant) for their mastery and teaching, respectively. International in scope and relevance, the textbooks correspond to course syllabi sufficiently to serve as required reading. Their didactic style, comprehensiveness and coverage of fundamental material also make them suitable as introductions or references for scientists entering, or requiring timely knowledge of, a research field.

More information about this series at <https://link.springer.com/bookseries/8431>

Massimiliano Bonamente

Statistics and Analysis of Scientific Data

Third Edition



Springer

Massimiliano Bonamente
University of Alabama in Huntsville
Huntsville, AL, USA

ISSN 1868-4513

Graduate Texts in Physics

ISBN 978-981-19-0364-9

<https://doi.org/10.1007/978-981-19-0365-6>

ISSN 1868-4521 (electronic)

ISBN 978-981-19-0365-6 (eBook)

1st & 2nd editions: © Springer Science+Business Media New York 2013, 2017

3rd edition: © The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Portrait of John Duns Scotus by J. van Ghent, oil on panel, circa 1472–1476, Galleria Nazionale delle Marche

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721,
Singapore

Numbers don't lie

Foreword

It is now the era of Big Data. Huge troves of quantitative information reside everywhere and can be accessed from anywhere in the world in a heartbeat. While our tools for the collection and analysis of data have rocketed forward over a single human lifetime, we remain human. As such, we can only hold a few simple numbers and concepts in our poor, analog brains. We refer to those concepts as our understanding.

Statistical analysis of data allows us to make quantitative assessments of our concepts, hopefully keeping our preconceptions and misconceptions at bay. But using statistical methods is never just “turning a crank”. We must apply judgment. We must be careful. Computers can easily lull one into a false sense of security. In his third edition, Professor Bonamente provides the clearest exposition of statistical analysis and the probability theory upon which it is based. He provides a roadmap to the proper use of statistics.

In the opening chapter of this book, the author discusses Bayesian versus Frequentist statistical approaches. The Frequentist approach can be useful if one has no underlying model for the mechanisms at play. But, in the Physical Sciences and Engineering, since the time of Isaac Newton, we have models that rely on parameters. And those parameters can distill the understanding we seek. How to measure those parameters and place error estimates upon the results is the central theme of this book.

Today, there are surprisingly few courses in statistics offered at universities. And there are even fewer textbooks available to support those classes. Many of the books on statistics are written for the biological and medical fields and are suitable for use in research arenas where the parameters and mechanisms are poorly understood. In this text, you will find the approaches suitable for the analysis of data in physical fields as opposed to the life sciences.

This third edition is the textbook we have been awaiting. Every individual who encounters data in a professional capacity should read it. The material ranges from the basics of probability to the serious analysis of data. Any individual who works with data daily should master the materials in this book, keeping a copy for reference.

I frequently teach our department’s course in data analysis and statistics. I have used Bonamente’s second edition to good effect, but the third edition is the textbook

I have awaited. I only wish I had had it back when I was a student. I am old enough that I had to use a slide rule as an undergraduate and did not gain access to a hand calculator until graduate school. Computers were run on decks of punch cards. How far our hardware tools have come! Yet the statistical tools employed today have hardly changed.

Even the data have changed. When I was young, astronomers employed film. Today we use solid-state detectors that count individual photons at light levels so low that most of the detector bins have counted no photons at all. The switch from film (analog) to photon-counting (digital) has completely changed the nature of the appropriate statistical analysis.

In Chaps. 15 and 16, the author presents an approach to quantitative statistical analysis of low-signal-count data. It is often the case today that many events are recorded, but the instruments taking the data have even higher resolution, i.e., more data bins than events. In these two chapters, the author discusses the application of Poisson Statistics to this low-count-per-bin data using the Cash Statistic. He not only presents the use of this Cash Statistic but develops it to a new level that should prove useful across a range of modern applications.

Yes, the Cash Statistic is named after me, the writer of this foreword. I was lucky enough to participate in the very early days of x-ray astronomy and, in particular, be involved with the very first x-ray telescopes back in the 1970s. X-ray sources have notoriously weak photon fluxes, but every photon packs a punch, so it was x-ray and gamma-ray astronomers who first had to reckon with the statistics of images and spectra where a significant fraction of the data bins were empty. The data were fundamentally of a Poisson distribution nature and could not be addressed with the usual assumption of a Gaussian distribution. How, for example, does one deal with an image that has 1024 bins with an average of .01 x-rays per bin, yet there is one bin with three x-rays in it, right where the target is supposed to be? The answer lays in the Cash Statistic and showing that it was distributed as a Chi-square. Professor Bonamente has now, in this text, broadened the scope of the applications and provided a better framework for the application of the statistic.

In Part III of this book, the author addresses Monte Carlo methods. These are of fundamental importance to much of modern modeling, particularly in highly complex models like ray tracing of optical systems. Again, this book takes the reader (student) into rigorous analysis techniques that are difficult, if not impossible, to find explained elsewhere.

Statistics can be a hard master. The math can be challenging. But in the age of easy access to computing, the challenge to the user is to sift from all the numbers the “understanding” that is sought. The math becomes secondary. Judgment comes to the fore. What statistic should I use? What do the results actually mean? It is easy to overuse the statistics, finding statistically significant patterns that are actually of no significance.

And, as time goes by, and one's attention moves elsewhere, one is certain to forget how to perform the statistical analyses. But the general knowledge of the range of statistical tools will stay with you. Then, when the need arises, pull this book off the shelf. Professor Bonamente's book will bring it all back.

June 2022

Webster Cash

Professor of Astrophysical and Planetary Sciences
University of Colorado, Boulder

Preface

With the third edition of *Statistics and Analysis of Scientific Data*, I have continued to pursue the goal of providing a textbook that is both mathematically sound and easy to use for practical applications. Data analysis straddles the rigorous fields of statistics and mathematics, and the messy world of experiments and data collection, in a way that makes it often unclear as to where to draw the line between what *needs* to be done to the data and what *should* be done according to the available theoretical tools. This textbook aims to treat both the theoretical and practical aspects of several key methods of data analysis, including the limitations of certain common practices, such as the use of the chi-squared statistic for non-Gaussian data.

Starting with the second edition, I decided to mark portions of the textbook that are primarily theoretical with a gray sidebar. These parts are not essential for the practice of data analysis, and they can be skipped by the reader who is not interested in its theoretical underpinnings, or by the reader who needs a fast reference for a method that they already know. I find that this is an effective way of presenting certain mathematically intensive topics, and I have continued with this practice in the third edition.

Certain key theoretical results that have a great influence on statistical methods, such as H. Cramér's theorem on the limiting distribution of the chi-squared statistic or S.S. Wilk's theorem on the distribution of the likelihood ratio, are now presented in a more systematic way. The mathematical theory behind some of these results is quite complex, and therefore I sometimes found it necessary to describe only the main results, and then provide a reference to the original work, instead of delving into too deep a mathematical treatment. The more theoretically inclined reader is pointed to the relevant references instead. At the same time, it is important that the data analysis practitioner is aware of these mathematical results and the limitations they impose. A cautionary tale of what can go wrong when there is a mismatch between the method of analysis and the underlying theory is in K. Pearson's use of the chi-squared statistic for contingency tables, which are now treated in detail in Chap. 10. Pearson was erroneously convinced that 2×2 contingency tables with fixed

margins resulted in a statistic with three degrees of freedom, while R. A. Fisher later showed that they only had one degree of freedom, according to a result that I decided to present as Fisher's theorem on contingency tables. There are other statistics or tests that are often misused by the data analyst, such as the chi-squared statistic for Poisson count data or the *F*-test for model components, that I have also addressed in more detail in the third edition of this textbook.

Scientists are often drawn to data analysis and statistics out of the necessity to analyze and draw conclusion from data in their fields. Since the first edition, I have therefore emphasized the use of data from what I refer to as *Classic experiments*. In addition to G. Mendel's 1865 experiments on plant hybridization (now significantly expanded and improved), the 1897 J. J. Thomson's measurements that led to the discovery of the electron, the 1903 K. Pearson's study of biometric characters, the 1931 E. Hubble and M. Humason's data on the expansion of the universe, and the 1935 R. A. Fisher and E. Anderson's study of iris characteristics, I have added another landmark experiments, the 1915 study of inoculation statistics by M. Greenwood and U. Yule. The latter dataset provides the means to study contingency tables and diagnostic tools, which are the subject of a new chapter (Chap. 10) that was not previously present in the textbook. In part prompted by the *COVID-19* pandemic, during which the majority of the third edition was prepared, I found it useful to present such concepts as contingency tables and associated tests for independence, and binary diagnostic tests that describe the use of true and false positives and negatives, sensitivity and specificity of diagnostic tests, and vaccine efficacy. The use of these concepts in medical practice can be understood using the relatively simple statistical methods that were already discussed in the textbook, and thus I found that this addition was both straightforward and rewarding. Moreover, I decided to use two datasets associated with the *COVID-19* disease. Given their recent origin, these data were not labeled as classic experiments, but they were nonetheless used in detail to show how to evaluate vaccine efficacy and their uncertainties.

The data from classic experiments and from other sources are used throughout the textbook to illustrate the practice of data analysis. For example, the Pearson measurements of biometric characters are used to illustrate linear regression in Chap. 11 and the correlation coefficient in Chap. 14, the iris data of Fisher and Anderson are applied to the multi-variable linear regression method in Chap. 13, and the J. J. Thomson data are analyzed via the Kolmogorov–Smirnov test in Chap. 19. These data are also featured in end-of-chapter problems, which provide a combination of analytical and numerical exercises with close ties to the material presented in the relevant chapter.

Perhaps the most important guiding principle that I followed for the revisions leading to the third edition is how to make the textbook as easy as possible to use for students and teachers in the classroom and as a reference manual. To this end, first I decided to structure the 22 chapters that comprise this edition (up from 16 in the previous edition) into three parts:

Part I (Chaps. 1–8) features the treatment of probability, random variables, and an introduction of statistics as random variables based on measurements. Part I of

this textbook is therefore suitable as a textbook on elementary tools for the theory of probability, with an introduction to statistics.

Part II (Chaps. 9–19) covers the core of statistical tools for hypothesis testing, regression, and parameter estimation. Emphasis is placed on traditional Gaussian-based statistics, such as the chi-squared statistic, but also on Poisson-based statistics (such as the Cash statistic) that are essential for a more accurate treatment of data from counting experiments. This part includes also multi-variable regression, regression with errors in both variables, and statistical tests such as F -tests and Kolmogorov–Smirnov tests. Part II of the textbook can therefore be used as the basis for a course on traditional statistical methods for data analysis.

Part III (Chaps. 20–22) is a relatively short treatment of Monte Carlo methods, with emphasis on the popular Markov chain Monte Carlo methods (MCMC, Chap. 22), including a number of practical topics such as the construction of MCMC using the Metropolis–Hastings or Gibbs algorithms, and convergence tests. MCMC methods are so widespread that they are rapidly becoming a must for data analysts in many fields.

In my data analysis course at the *University of Alabama in Huntsville*, which is a one-semester graduate course, I usually opt for a short overview of key concepts from Part I, while spending the majority of the time in Part II, followed by a necessarily shortened treatment of Markov chain Monte Carlo methods. The reason for a clear demarcation among the different parts is to guide both the student and the teacher to make the most effective choice for their learning or teaching objectives. My personal teaching experience suggests that the material in this textbook would be best covered in two one-semester courses, with the first course covering Part I of the textbook, and the second using Parts II and III. The level of mathematics assumed in this textbook is elementary calculus and combinatorial mathematics, and therefore it is aimed at both upper level undergraduates and graduate students. Students familiar with the foundations of the theory of probability and basic concepts of statistics could proceed directly with Parts II and III of the textbook. Part II of the textbook is also structured in such a way that Chap. 10 (contingency tables and diagnostic tests) and Chaps. 15 and 16 (statistical methods using Poisson data and the Cash statistic) are conveniently isolated so that interested readers can easily find these topics, or they can be skipped by those readers who are not interested.

Finally, the widespread availability of computing resources convinced me that it was wise to complement this textbook with a complete set of computer codes that reproduce all examples and results in the textbook, including all end-of-chapter problems. For this task, I chose the `python` language, which has the advantage of being popular and easy to use even to the non-expert, so that people familiar with other languages might still find them useful. Codes provided with this textbook are described in Chap. 23. Moreover, I have also written a solutions manual for all end-of-chapter problems. These written solutions, together with the numerical codes that reproduce the results, should be of great help to the teacher and the student alike.

As the third edition of this book heads for the press, I near the 20-year mark as a teacher, and as a student of statistics. What I find especially fascinating about statistics is that it lies quite uniquely at the intersection of logic and mathematics, in

a way that it is impossible to separate the numbers from the words needed to interpret the numbers. So I shall lead you to the reading of this textbook with one word of advice: first, *write* down the question, if you want to have a *chance* at finding the right answer.

Huntsville, AL, USA

Massimiliano Bonamente

Acknowledgments

In my early postdoc years, I was struggling to solve a complex data analysis problem. My longtime colleague and good friend Marshall Joy one day walked down to my office and said something like “Max, I have a friend in Chicago who told me that there is a method that maybe can help us with our problem. I don’t understand any of it, but here’s a paper that talks about Monte Carlo Markov chains. See if it can help us.” That conversation led to the appreciation of one of statistics and data analysis’ most powerful tools and opened the door for virtually all the research papers that I wrote ever since. For over a decade, Marshall taught me how to be careful in the analysis of data and interpretation of results—and always used a red felt-tip marker to write comments on my drafts.

The journey leading to this book started in the early 2000s when Prof. A. Gordon Emslie and I decided to offer a new course in data analysis and statistics for graduate students in our department. Gordon’s uncanny ability to solve virtually any problem presented to him—and likewise make even the experienced scientist stumble with his questions—has been a great source of inspiration for this book.

Some of the material presented in this book is derived from Prof. Kyle Siegrist’s lectures on probability and stochastic processes at the University of Alabama in Huntsville. Kyle reinforced my love for mathematics and motivated my desire to emphasize both mathematics and applications for the material presented in this book. His textbook on probability and statistics, *Random*, freely available online, is my go-to resource of choice.

I am deeply indebted to two colleagues of mine at the University of Alabama in Huntsville, Prof. Roger Cruz-Vera and Prof. Louise O’Keefe, who provided a number of insightful comments. Louise was instrumental in checking and commenting on the new material pertaining to diagnostic testing, while Roger patiently went through my treatment of G. Mendel’s experiment and provided a number of very useful suggestions. I am also very grateful for all the comments and suggestions provided by Prof. Dmytro Inosov of the Technische Universität Dresden, especially for his review of the Gehrels approximation and for suggestions on p -values and hypothesis testing. A special thanks goes also to Prof. Adrian Raftery of the University of Washington,

who kindly agreed to review the section on his namesake MCMC convergence diagnostic. Adrian's suggestions also led to a more thorough treatment of certain topics of relevance to Markov chains, such as binary chains.

Finally, it has been my sincere pleasure to converse with Prof. Webster Cash of the University of Colorado Boulder on Poisson statistics and some ideas on how to treat counting data. Webster's work on Poisson data and his ability to make the related theoretical work accessible to the data analyst have been very influential to the development of the material for Chaps. 15 and 16 of the third edition of this textbook.

Huntsville, AL, USA

Massimiliano Bonamente

Contents

Part I Probability, Random Variables and Statistics

1	Theory of Probability	3
1.1	Experiments and Events	3
1.2	Probability of Events	4
1.2.1	The Kolmogorov Axioms	5
1.2.2	Frequentist or Classical Method	6
1.2.3	Bayesian or Empirical Method	6
1.2.4	Fundamental Properties of Probability	7
1.3	The Conditional Probability	8
1.4	Statistical Independence	9
1.5	A Classic Experiment: Mendel's Experiments on Plant Hybridization	11
1.6	The Total Probability Theorem and Bayes' Theorem	16
2	Random Variables and Their Distributions	21
2.1	Random Variables	21
2.2	Probability Distribution Functions	22
2.3	Expectations and Moments of a Distribution Function	24
2.3.1	The Mean and the Sample Mean	25
2.3.2	The Law of Large Numbers	26
2.3.3	The Variance and the Sample Variance	27
2.4	A Classic Experiment: J.J. Thomson's Discovery of the Electron	28
2.5	Covariance and Correlation Between Random Variables	30
2.5.1	Joint Distribution and Moments of Two Random Variables	31
2.5.2	Statistical Independence of Random Variables	33
2.6	The Expectation of the Sample Variance and Sample Covariance	35
2.7	A Classic Experiment: Pearson's Collection of Data on Biometric Characteristics	38

3 Three Fundamental Distributions: Binomial, Gaussian, and Poisson	43
3.1 The Binomial Distribution	43
3.1.1 Derivation of the Binomial Distribution	44
3.1.2 Moments of the Binomial Distribution	46
3.2 The Gaussian Distribution	47
3.2.1 Derivation of the Gaussian Distribution from the Binomial Distribution	48
3.2.2 Moments and Properties of the Gaussian Distribution	50
3.3 The Poisson Distribution	52
3.3.1 Derivation of the Poisson Distribution	53
3.3.2 Moments and Properties of the Poisson Distribution	53
3.3.3 The Poisson Distribution and the Poisson Process	55
3.4 Comparison of the Binomial, Gaussian, and Poisson Distributions	56
4 The Distribution of Functions of Random Variables	63
4.1 Functions of Random Variables	63
4.2 Linear Combination of Random Variables	64
4.2.1 Mean and Variance Formulas	64
4.2.2 Independent Measurements and the $1/\sqrt{N}$ Factor	66
4.3 The Moment Generating Function	66
4.3.1 Properties of the Moment Generating Function	67
4.3.2 Moment-Generating Functions of Selected Distributions	68
4.4 The Central Limit Theorem	69
4.4.1 The Distribution of the Sample Mean of Gaussian Measurements	71
4.4.2 The Distribution of the Sum of Standard Uniform Random Variables	71
4.4.3 Certain Limitations of the Central Limit Theorem	73
4.5 The Distribution of Functions of Random Variables	74
4.5.1 The Method of Change of Variables	74
4.5.2 Direct Method Using the Distribution Function	76
5 Error Propagation and Simulation of Random Variables	81
5.1 The Mean of Functions of Random Variables	81
5.2 The Variance of Functions of Random Variables and Error Propagation Formulas	83
5.2.1 Sum and Product of a Constant	85
5.2.2 Weighted Sum of Two Variables	85
5.2.3 Product and Division of Two Random Variables	86
5.2.4 Power of a Random Variable	88
5.2.5 Exponential of a Random Variable	88

5.2.6	Logarithm of a Random Variable	88
5.3	The Quantile Function and Simulation of Random Variables	89
5.3.1	General Method to Simulate a Variable	91
5.3.2	Simulation of a Gaussian Variable	92
6	Maximum Likelihood and Other Methods to Estimate Variables	97
6.1	Estimating Random Variables with Data	97
6.2	The Maximum-Likelihood Method	98
6.2.1	Maximum-Likelihood Methods for a Gaussian Variable	99
6.2.2	Maximum-Likelihood Estimate of the Gaussian Mean for Non-uniform Uncertainties	100
6.3	The Maximum-Likelihood Method for the Poisson and Other Distributions	102
6.4	Method of Moments	103
6.5	Method of Maximum Entropy	105
7	Methods of Inference and Confidence Intervals of Random Variables	113
7.1	Quantiles and Confidence Intervals	113
7.2	Fiducial Inference	114
7.3	Confidence Intervals for a Gaussian Variable	116
7.4	Upper and Lower Limits for a Gaussian Variable	118
7.5	Confidence Intervals for the Mean of a Poisson Variable	121
7.6	The Gehrels Approximation for Poisson Upper and Lower Limits	123
7.7	Bayesian Methods of Inference	126
7.7.1	Bayesian Expectation of the Poisson Mean	127
7.7.2	Bayesian Confidence Intervals for a Poisson Variable	129
8	Average Values of Random Variables	133
8.1	Point Estimates and Average Values	133
8.2	Linear and Weighted Averages	134
8.3	The Median	135
8.4	The Logarithmic Average and Fractional Errors	137
8.4.1	The Log-Normal Distribution	138
8.4.2	The Weighted Logarithmic Average	140
8.4.3	The Relative-Error Weighted Average	142
Part II Hypothesis Testing, Regression and Parameter Estimation		
9	Hypothesis Testing and Fundamental Statistics	147
9.1	Statistics and Hypothesis Testing	147
9.2	The <i>P</i> -Value of a Statistical Analysis	151

9.3	The χ^2 Statistic	155
9.3.1	The Probability Distribution Function	155
9.3.2	Moments and Other Properties	157
9.3.3	Hypothesis Testing	158
9.4	The Distribution of the Sample Variance	159
9.5	The F -Statistic	163
9.5.1	The Probability Distribution Function	163
9.5.2	Moments and Other Properties	165
9.5.3	Hypothesis Testing	165
9.6	The Sampling Distribution of the Mean and <i>Student's t</i> -Statistic	169
9.6.1	<i>Student's t</i> -Statistic for the Sample Mean	169
9.6.2	Hypothesis Testing with the <i>t</i> -Statistic	172
9.6.3	Comparison of Two Sample Means and Hypothesis Testing	175
10	Contingency Tables and Diagnostic Tests	181
10.1	A Classic Experiment: The 1915 Greenwood and Yule Inoculation Statistics	181
10.2	2×2 Contingency Tables	182
10.2.1	The χ^2 Test	184
10.2.2	χ^2 Test with the Yates Continuity Correction	186
10.2.3	The Fisher Exact Test for 2×2 Contingency Tables	187
10.2.4	Exact Tests Based on the Binomial Distribution	190
10.3	Higher Dimension $r \times c$ Contingency Tables	194
10.4	Binary Diagnostic Tests	195
10.4.1	Sensitivity, Specificity, and Likelihood Ratios	196
10.4.2	Posterior Probabilities: The Positive and Negative Predictive Values	197
10.4.3	Change in Posterior Probability with Repeated Testing	201
10.5	Vaccine Efficacy	204
11	Linear and Non-linear Regression for Gaussian Data	213
11.1	Measurement of Pairs of Variables and Regression	213
11.2	Regression Using Maximum Likelihood for Gaussian Data	214
11.3	Linear Regression with Gaussian Data	216
11.4	Multiple Linear Regression	219
11.5	Linear Regression with Uniform Variance	222
11.5.1	Alternative form of the Solution with Sample Moments	223
11.5.2	Choice of Independent Variable	225
11.6	A Classic Experiment: Edwin Hubble's Discovery of the Expansion of the Universe	225
11.7	Non-linear Regression	228

12 Goodness of Fit and Parameter Uncertainty for Gaussian Data	233
12.1 The χ^2_{\min} Goodness-of-Fit Statistic	233
12.2 Data with No Errors and the Model Sample Variance	236
12.3 The $\Delta\chi^2$ Statistic	238
12.4 Confidence Intervals of Model Parameters	239
12.5 Confidence Intervals on a Reduced Number of Parameters	241
13 Multi-variable Regression	247
13.1 Multi-variable Datasets	247
13.2 A Classic Experiment: The R.A. Fisher and E. Anderson Measurements of Iris Characteristics	248
13.3 The Multi-variable Linear Regression	250
13.4 Multi-variable Linear Regression with Uniform Variance	251
13.5 Goodness of Fit of Multi-variable Regression	253
13.6 Tests for the Significance of Multiple Regression Coefficients	254
13.6.1 t-Test for the Significance of Model Components	254
13.6.2 F-Test for the Significance of the a_1, \dots, a_m Parameters	256
13.6.3 The Coefficient of Determination	259
14 The Linear Correlation Coefficient	263
14.1 Linear Regression and Choice of the Independent Variable	263
14.2 The Linear Correlation Coefficient	266
14.3 Sampling Distribution of r and Hypothesis Testing	267
14.4 Distribution of the Coefficient of Determination R^2 and of r^2	271
15 Low-Count Poisson Data and the <i>Cash</i> Statistic	277
15.1 Poisson Data with Integer-Valued Variables	277
15.2 Likelihood of Poisson Data and the <i>Cash</i> Statistic	278
15.3 Distribution of the <i>Cash</i> Statistic for a Fully Specified Model	280
15.3.1 Asymptotic Values for the Mean and Variance	281
15.3.2 Analytical Approximations for the Mean and Variance	282
15.3.3 Other Useful Formulas for the Moments	283
15.4 Hypothesis Testing with the <i>C</i> Statistic	284
16 Maximum Likelihood Methods and Parameter Estimation with the <i>Cash</i> Statistic	291
16.1 Maximum Likelihood Methods for Poisson Data	291
16.2 Linear Regression with Poisson Data	292
16.2.1 The Standard Linear Model	293
16.2.2 A Factorized Linear Model with a Semi-Analytical Solution	293
16.2.3 An Extended Linear Model	296
16.2.4 Non-Uniform Bin Size and Gaps in the Data	299

16.3	Goodness of Fit and Hypothesis Testing with the <i>Cash</i> Statistic	301
16.3.1	The Wilks Theorem on the Likelihood Ratio	301
16.3.2	The Large-Count Regime	303
16.3.3	The Low-Count Regime	303
16.3.4	Approximate Methods in the Low-Count Regimes	306
16.4	Parameter Estimation with the <i>C</i> statistic	307
16.5	Biases Using χ^2 for Poisson Data in the Large-Count Limit	309
17	Systematic Errors and Intrinsic Scatter	315
17.1	What to Do When the Goodness-of-Fit Test Fails	315
17.2	Intrinsic Scatter and the Debiased Variance	316
17.2.1	Direct Calculation of the Intrinsic Scatter	316
17.2.2	Alternative Method for Gaussian Data	317
17.3	Systematic Errors	318
17.4	Estimate of Model Parameters with Systematic Errors or Intrinsic Scatter	320
18	Regression with Bivariate Errors	323
18.1	Two-Variable Data with Bivariate Errors	323
18.2	Least-Squares Linear Fit to Data with Bivariate Errors	324
18.3	Linear Fit Using Bivariate Errors in the χ^2 Statistic	329
19	Model and Data Comparison	333
19.1	The χ_{\min}^2 Statistic and the <i>F</i> -Test for Gaussian Data	333
19.2	<i>F</i> -Test for Two Independent χ^2 Measurements	334
19.3	<i>F</i> -Test for an Additional Model Component	336
19.4	Kolmogorov–Smirnov Tests	339
19.4.1	Comparison of Data to a Model	339
19.4.2	Two-Sample Kolmogorov–Smirnov Test	343
Part III Monte Carlo Methods		
20	Monte Carlo and Re-sampling Methods	351
20.1	What is a Monte Carlo Analysis?	351
20.2	Traditional Monte Carlo Integration	352
20.3	Hit-or-Miss Monte Carlo Methods	355
20.4	Simulation of Random Variables	357
20.5	Re-sampling Methods	358
20.6	The Jackknife Method	360
20.7	The Bootstrap Method	362
21	Introduction to Markov Chains	369
21.1	Stochastic Processes and Markov Chains	369
21.2	Mathematical Properties of Markov Chains	370
21.3	Recurrent and Transient States	372
21.4	Limiting Probabilities and Stationary Distribution	375
21.5	Ergodic Averages and Variance Estimates	379

Contents	xxiii
22 Markov Chain Monte Carlo	385
22.1 Introduction to Markov Chain Monte Carlo Methods	385
22.2 Markov Chain Monte Carlo for Regression Analysis	386
22.3 The Metropolis–Hastings MCMC	387
22.4 The Gibbs Sampler	393
22.5 Convergence of Markov Chain Monte Carlo	395
22.6 The Geweke z -Score Convergence Test	398
22.7 The Gelman–Rubin Convergence Test	399
22.8 The Raftery–Lewis Diagnostic	402
22.9 Inference with MCMC	406
23 Numerical Methods and <code>python</code> Codes	413
23.1 Analytical and Numerical Methods	413
23.2 Introduction to <code>python</code>	414
23.3 General Features of <code>python</code> Codes for this Textbook	416
23.3.1 Structure of the Codes	416
23.3.2 Functions, Library Import and Settings	418
23.3.3 Data Associated with the Codes	419
23.4 Description of Codes	420
23.5 Numerical Methods for Tables in Appendix	424
Appendix: Numerical Tables	427
References	475
Index	481

Part I

Probability, Random Variables and

Statistics

Chapter 1

Theory of Probability



Abstract The theory of probability is the mathematical framework for the study of the probability of occurrence of events. The first step is to establish a method to assign the probability of an event, for example, the probability that a coin lands heads up after a toss. The *frequentist* or empirical approach, and the *subjective* or Bayesian approach, are two methods that can be used to calculate probabilities. Once a method for assigning probabilities is established, the Kolmogorov axioms are introduced as the “rules” required to manipulate probabilities. Fundamental results known as Bayes’ theorem and the theorem of total probability are used to define and interpret the concepts of statistical independence and of conditional probability, which play a central role in much of the material presented in this book.

1.1 Experiments and Events

Every experiment has a number of possible outcomes. For example, the experiment consisting of the roll of a die can have six possible outcomes, according to the number that shows after the die lands. Assigning probabilities to possible outcomes is one of the key tasks of the theory of probability. The following section presents two methods to assign probabilities, the classical method based on the repetition of the experiment, and a method based on empirical knowledge of the experiment. The fact that there is more than one method available for this purpose should not be viewed as a limitation of the theory, but rather as the fact that for certain parts of the theory of probability, and even more so for statistics, there is an element of subjectivity that enters the analysis and the interpretation of the results. It is therefore the task of the statistician to keep track of any assumptions made in the analysis and to account for them in the interpretation of the results. Prior to describing the methods for assigning probabilities, it is necessary to develop the terminology required to describe experiments and their outcomes.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-981-19-0365-6_1.

The *sample space* Ω is defined as the set of all possible outcomes of the experiment. In the case of the roll of a die, the sample space can be written as the set of the six possible outcomes, $\Omega = \{1, 2, 3, 4, 5, 6\}$. An *event* A is a sub-set of Ω , $A \subset \Omega$, and it represents a number of possible outcomes for the experiment. For example, the event “even number” may be represented by $A = \{2, 4, 6\}$, and the event “odd number” as $B = \{1, 3, 5\}$. Different experiments will have different sample spaces that can be written in an equivalent way. For each experiment, two events always exist: the sample space itself which comprises all possible outcomes and the empty set that contains no outcomes represented as $A = \emptyset$ and called the *impossible event*.

Events are conveniently studied using elementary set theory. It is useful to review some of the properties of set theory that are of common use in probability and statistics. The complementary of an event A is indicated as \bar{A} , and it is the set of all possible outcomes except those in A . For example, the complementary of the event “odd number” is the event “even number”. Given two events A and B , the union $C = A \cup B$ is the event comprising all outcomes of A and those of B . In the roll of a die, the union of odd and even numbers is the sample space itself, consisting of all possible outcomes. The intersection of two events $C = A \cap B$ is the event comprising all outcomes of A that are also outcomes of B . When two events do not overlap, $A \cap B = \emptyset$, the events are said to be *mutually exclusive*. The union and intersection can be naturally extended to more than two events. Events are illustrated in Fig. 1.1.

Finally, a number of events A_i , for $i = 1, \dots, n$, are said to be a *partition* of the sample space if they satisfy the following two properties:

$$\begin{cases} A_i \cap A_j = \emptyset, & \text{when } i \neq j \\ \bigcup_{i=1}^n A_i = \Omega. \end{cases} \quad (1.1)$$

For example, the outcomes 1, 2, 3, 4, 5, and 6 for the roll of a die partition the sample space into a number of events that cover all possible outcomes, without any overlap among each other.

1.2 Probability of Events

The probability P of an event is a number that is intended to describe the odds of occurrence of an event in a single trial of the experiment. The modern theory of probability was developed over the course of the first half of the twentieth century with the contribution of a number of mathematicians, including S. Bernstein [9] and A. Kolmogorov [63]. All these contributions led to a framework that views the probability as a number between 0 and 1, where $P = 0$ corresponds to an impossible event, and $P = 1$ to a certain event. Therefore, the operation of “probability” can

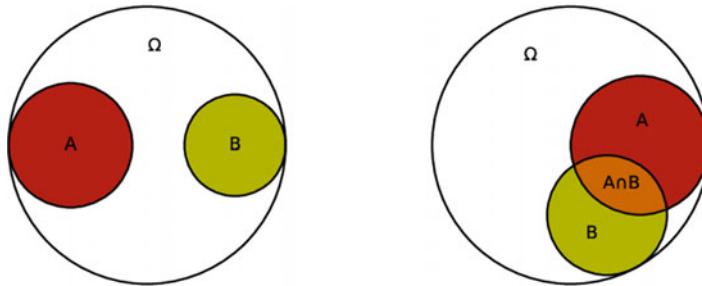


Fig. 1.1 Events can be represented as sub-sets of the sample space. The probability of the event $P(A \cup B)$ is the sum of the two individual probabilities, only if the two events are mutually exclusive. This property enables the interpretation of probability as the “area” of a given event within the sample space

be thought of as a function that transforms each possible event into a real number between 0 and 1.

1.2.1 The Kolmogorov Axioms

The first step towards determining the probability of an event is to establish a number of basic rules that capture the meaning of probability. The probability of an event is required to satisfy three axioms defined by Kolmogorov [63]¹:

1. The probability of an event A is a non-negative number, $P(A) \geq 0$;
2. The probability of all possible outcomes, or sample space, is normalized to the value of unity, $P(\Omega) = 1$;
3. If A and B are two mutually exclusive events, then

$$P(A \cup B) = P(A) + P(B). \quad (1.2)$$

Figure 1.1 illustrates the last property using set or Venn diagrams. Events in the left panel are mutually exclusive, and the probability of the union is represented by the area of $A \cup B$, or the sum of the two individual areas. For events that are not mutually exclusive, such as those in the right panel, this property does not apply.

These axioms should be regarded as the basic “ground rules” of the theory of probability, but they provide no guidance on how event probabilities should be assigned. For this purpose, there are two major avenues available. One is based on the repetition of the experiments a large number of times under the same conditions, and goes under the name of the frequentist or classical method. The other is based on a more theo-

¹ In his book *Foundations of the Theory of Probability* [63], Kolmogorov lists a larger number of axioms, due to the need of ensuring certain mathematical properties of probability.

retical knowledge of the experiment, but without the experimental requirement of a large number of repetitions, and is referred to as the Bayesian or empirical method.

1.2.2 Frequentist or Classical Method

Consider performing an experiment a large number N of times, under the same experimental conditions. The occurrence of the event A is indicated as the number $N(A)$. The probability of event A is given by

$$P(A) = \lim_{N \rightarrow \infty} \frac{N(A)}{N}; \quad (1.3)$$

that is, the probability is the relative frequency of occurrence of a given event from many repetitions of the same experiment. The obvious limitation of this definition is the need to perform the experiment a large number of times. This requirement is not only time consuming but also requires that the experiment be repeatable in the first place, which may or may not be possible. The limitation of this method is evident by considering a coin toss: no matter the number of tosses, the occurrence of heads up will never be exactly 50%, which is what one would expect based on an empirical knowledge of the experiment at hand.

1.2.3 Bayesian or Empirical Method

Another method to assign probabilities is to use knowledge of the experiment, both theoretical and experimental, but without the need for extensive experimental data. The probability assigned to an event represents the *degree of belief* that the event will occur in a given try of the experiment, and it implies an element of subjectivity which will become more evident with Bayes' theorem (see Sect. 1.6). The Bayesian probability is assigned based on a quantitative understanding of the nature of the experiment, and in accord with the Kolmogorov axioms. It is sometimes referred to as *empirical* probability, in recognition of the fact that sometimes the probability of an event is assigned based upon a practical knowledge of the experiment, although without the classical requirement of repeating the experiment a large number of times. This method is named after the Rev. Thomas Bayes, who pioneered the development of the theory of probability [8].

Example 1.1 (*Coin toss experiment*) In the coin toss experiment, the determination of the empirical probability for events “Heads up” or “Tails up” relies on the knowledge that the coin is unbiased, and that therefore it must be true that $P(T) = P(H)$. This empirical statement signifies the use of the Bayesian method to determine probabilities. With this information, we can then simply use the

Kolmogorov axioms to state that $P(T) + P(H) = 1$, and therefore obtain the intuitive result that $P(T) = P(H) = 1/2$. \diamond

1.2.4 Fundamental Properties of Probability

The following properties are useful to assign and manipulate event probabilities. They are somewhat intuitive, but it is nonetheless instructive to derive them from the Kolmogorov axioms.

1. The probability of the null event is zero, $P(\emptyset) = 0$.

This property can be derived starting with the mutually exclusive events \emptyset and Ω . Since their union is Ω , it follows from the third axiom that $P(\Omega) = P(\Omega) + P(\emptyset)$. From the second axiom, it is known that $P(\Omega) = 1$, and from this, it follows that $P(\emptyset) = 0$. The following property is a generalization of this property.

2. The probability of the complementary event \bar{A} satisfies the property

$$P(\bar{A}) = 1 - P(A). \quad (1.4)$$

By definition, it is true that $A \cup \bar{A} = \Omega$, and that A, \bar{A} are mutually exclusive. Using the second and third axioms, $P(A \cup \bar{A}) = P(A) + P(\bar{A}) = 1$, from which it follows that $P(\bar{A}) = 1 - P(A)$.

3. The probability of the union of two events satisfies the general property that

$$P(A \cup B) = P(A) + P(B) - P(A \cap B). \quad (1.5)$$

This property generalizes the third Kolmogorov axiom and can be interpreted as the fact that outcomes in the overlap region of the two events should be counted only once, as illustrated in Fig. 1.1. First, realize that the event $A \cup B$ can be written as the union of three mutually exclusive sets,

$$A \cup B = (A \cap \bar{B}) \cup (B \cap \bar{A}) \cup (A \cap B),$$

see Fig. 1.1. Therefore, using the third axiom,

$$P(A \cup B) = P(A \cap \bar{B}) + P(B \cap \bar{A}) + P(A \cap B).$$

Then, notice that for any event A or B , it is true that $A = (A \cap \bar{B}) \cup (A \cap B)$, since $\{B, \bar{B}\}$ is a partition of Ω . This implies that $P(A) = P(A \cap B) + P(A \cap \bar{B})$ due to the fact that the two sets are again mutually exclusive, with a similar equation for event B . It thus follows that $P(A \cup B) = P(A) + P(B) - P(A \cap B) + P(A \cap B) = P(A) + P(B) - P(A \cap B)$, which proves the property.

Example 1.2 An experiment consists of drawing a number between 1 and 100 at random. The event of interest C is “drawing either a number greater than 50 or an odd number, in a given try.” The sample space for this experiment is the set of numbers $i = 1, \dots, 100$ and the probability of drawing number i is $P(A_i) = 1/100$, since each number has the same probability of being drawn. With A the event consisting of all numbers greater than 50 and B the event with all odd numbers, it is clear that $P(A) = 0.5$ and $P(B) = 0.5$. The two events overlap, and event $A \cap B$ contains all odd numbers greater than 50, with $P(A \cap B) = 0.25$. Using Eq. 1.5, the probability of drawing either a number greater than 50, or an odd number, is

$$P(C) = P(A \cup B) = 0.75.$$

This result can be easily confirmed by a direct count of the possible outcomes. \diamond

1.3 The Conditional Probability

The *conditional* probability describes the occurrence of an event A , knowing that another event B has also occurred. Conditioning plays a prominent role in probability and statistics, since the probability of an event may depend on another event that is known to have occurred. The conditional probability is indicated as $P(A/B)$ or A given B . The following relationship defines the conditional probability:

$$P(A \cap B) = P(A/B) \cdot P(B), \quad (1.6)$$

and it can be equivalently expressed as

$$P(A/B) = \begin{cases} \frac{P(A \cap B)}{P(B)} & \text{if } P(B) \neq 0 \\ 0 & \text{if } P(B) = 0. \end{cases} \quad (1.7)$$

A justification for this definition is that the occurrence of B means that the probability of occurrence of A is proportional to the probability of occurrence of $A \cap B$. Furthermore, the denominator of the conditional probability is $P(B)$, instead of unity, because B is the set of all possible outcomes that are known to have happened. The situation is also illustrated in the right-hand panel of Fig. 1.1.

Example 1.3 (*Illustration of the conditional probability with dice*) Calculate the probability of obtaining 8 as the sum of two rolls of a die, given that the first roll was a 3. To this end, it is useful to define the following two events:

$$A = \{\text{The sum of two rolls is 8}\},$$

$$B = \{\text{The first roll shows 3}\}.$$

Event A is given by outcomes $(2,6)$, $(3,5)$, $(4,4)$, $(5,3)$, $(6,2)$, and since each combination has a probability of $1/36$, $P(A) = 5/36$. The probability of event B is $P(B) = 1/6$ since it relates to the outcomes of just one roll of the die. Also, the event $A \cap B$ occurs if the first roll is a 3 and the sum is 8, which can clearly occur only if a sequence of $(3,5)$ takes place, thus with probability $P(A \cap B) = 1/36$. According to the definition of conditional probability, the probability of interest is

$$P(A/B) = P(A \cap B)/P(B) = \frac{1/36}{1/6} = \frac{1}{6},$$

and in fact only combination $(5,3)$ —of the six available with 3 as the outcome of the second toss—gives rise to a sum of 8. The occurrence of 3 in the first roll has therefore increased the probability of A from $P(A) = 5/36$ to $P(A/B) = 1/6$, since not all outcomes of the first roll would be equally conducive to a sum of 8 in two rolls. \diamond

1.4 Statistical Independence

The concept of *statistical independence* among events means that the occurrence of one event has no influence on the occurrence of other events. Consider, for example, rolling two dice, one after the other: the outcome of one die is independent of the other and the two tosses are said to be statistically independent. On the other hand, consider rolling two dice, and being interested in the following pair of events: the first is the outcome of the roll of die 1 and the second is the sum the rolls of die 1 and die 2. It is clear that the outcome of the second event—e.g., the sum of both dice—depends on the first toss and the two events are not independent.

Two events A and B are said to be statistically independent if

$$P(A \cap B) = P(A) \cdot P(B). \quad (1.8)$$

This definition follows directly from Eq. 1.6. In fact, if A and B are statistically independent, then the conditional probability is $P(A/B) = P(A)$, i.e., the occurrence of B has no influence on the occurrence of A . Equation 1.8 is therefore a simple consequence of Eq. 1.6 when the two events are independent. A few examples illustrate the meaning of this definition.

Example 1.4 (*Illustration of statistical independence with dice*) Determine the probability of obtaining two 3s when rolling two dice. This event can be decomposed in two events:

$$A = \{\text{die 1 shows 3, and die 2 shows any number}\},$$

$$B = \{\text{die 2 shows 3, and die 1 shows any number}\}.$$

It is natural to assume that $P(A) = 1/6$, $P(B) = 1/6$, and state that the two events A and B are independent by nature, since each event involves a different die, which has no knowledge of the outcome of the other one; the same would be true also of the same die tossed two times. The event of interest is $C = A \cap B$, and the definition of probability of two statistically independent events leads to

$$P(C) = P(A \cap B) = P(A) \cdot P(B) = 1/36.$$

This result can be confirmed by a direct count of all possible outcomes in the toss of two dice, and the fact that there is only one combination out of 36 that gives rise to two consecutive 3s. \diamond

The example above highlights the importance of a proper, and sometimes extended, definition of an event. The more careful the description of the event and of the experiment that it is drawn from, the easier it is to make probabilistic calculations and the assessment of statistical independence.

Example 1.5 Determine whether the following events are statistically independent of each other:

$$\begin{cases} A = \{\text{die 1 shows 3 and die 2 shows any number}\}, \\ B = \{\text{the sum of the two dice is 9}\}. \end{cases}$$

The procedure is to calculate the probability of the two events, and then check whether they obey Eq. 1.8 or not. This calculation will illustrate that the two events are *not* statistically independent.

Event A has a probability $P(A) = 1/6$; in order to calculate the probability of event B , realize that a sum of 9 is given by the following combinations of outcomes of the two rolls: (3,6), (4,5), (5,4), and (6,3), and therefore $P(B) = 1/9$. The event $A \cap B$ is the situation in which *both* event A and B occur, which corresponds to the single combination (3,6); therefore, $P(A \cap B) = 1/36$. Since

$$P(A) \cdot P(B) = 1/6 \cdot 1/9 = 1/54 \neq P(A \cap B) = 1/36,$$

the two events are not statistically independent. This conclusion means that one event influences the other, since a 3 in the first toss has certainly an influence on the possibility of both tosses having a total of 9. \diamond

There are two important necessary (but not sufficient) conditions for statistical independence between two events. These properties can help identify whether two events are independent.

1. If $A \cap B = \emptyset$, A and B *cannot* be independent, unless one is the empty set. This property states that there must be some overlap between the two events, or else it is not possible for the events to be independent.

For A and B to be independent, it must be true that $P(A \cap B) = P(A) \cdot P(B)$,

which is zero by hypothesis. This can be true only if $P(A) = 0$ or $P(B) = 0$, which in turn means $A = \emptyset$ or $B = \emptyset$ as a consequence of the Kolmogorov axioms.

2. If $A \subset B$, then A and B *cannot* be independent, unless B is the entire sample space. This property states that the overlap between two events cannot be such that one event is included in the other, in order for statistical independence to be possible.

For A and B to be independent, $P(A \cap B) = P(A) \cdot P(B) = P(A)$, given that $A \subset B$. This can only be true if $B = \Omega$, since $P(\Omega) = 1$.

Example 1.6 Consider the Example 1.4 above of the roll of two dice. Notice how each event was formulated in terms of the outcome of both rolls to show that there was in fact overlap between two events that are independent of one another. \diamond

Example 1.7 Consider the following two events:

$$\begin{cases} A = \{\text{die 1 shows 3 and die 2 shows any number}\} \\ B = \{\text{die 1 shows 3 or 2 and die 2 shows any number}\}. \end{cases}$$

It is clear that $A \subset B$, $P(A) = 1/6$ and $P(B) = 1/3$. The event $A \cap B$ is thus identical to A and $P(A \cap B) = 1/6$. Therefore, $P(A \cap B) \neq P(A) \cdot P(B)$ and the two events are not statistically independent. This result can be easily explained by the fact that the occurrence of A implies the occurrence of B , which is a strong statement of dependence between the two events. The dependence between the two events can also be seen by the fact that the non-occurrence of B implies the non-occurrence of A . \diamond

1.5 A Classic Experiment: Mendel's Experiments on Plant Hybridization

The experiments performed in the nineteenth century by Gregor Mendel in the monastery of Brno led to the discovery of the main laws that determine the inheritance of characters. These experiments are now considered the foundation of modern genetics. They were performed over a span of several years, during which Mendel experimented with the reproduction of several species of pea plants [69].

Mendel started his experiments with a variety of pea plants that, year after year, resulted always in plants with the same characteristics. In particular, Mendel chose seven characteristics that were easy to observe:

1. the form of the ripe seed, or pea, which could be either round or wrinkled;
2. the difference in the color of the seed albumen, either yellow or green;
3. the color of the flowers, white or violet-red;
4. the shape of the ripe pod, which could be inflated or constricted;
5. the color of the unripe pod, green or yellow;

6. the position of the flowers, either terminal (near the top of the stem) or axial (along the main stem); and
7. the length of the stem, long (6–7 feet) or short (3/4 to 1 1/2 feet).

Hybrids from parents that differ in one character

Mendel began by crossing two pure lines of pea plants which differed in one single characteristic, by means of forced fertilization. These true or pure-breeding plants would be referred to as *homozygote*, in modern language. The results of this breeding yielded hybrid plants that displayed only one of the two characteristics, called the *dominant* character. For example, the hybrid plants all had round seed although they were bred from a population of pure round seed plants and one with wrinkled seed.

When the hybrids were allowed to self-fertilize among themselves, Mendel observed the data shown in Table 1.1 [69], in which each line corresponds to a separate experiment. This is the first generation of hybrids.

Mendel then continued to breed plants from each of the first generations of hybrids. He noticed that all second-generation hybrids followed a similar pattern in terms of the appearance of each character. The second-generation data for the round versus wrinkled seed experiment and for the yellow versus green seed experiment are reported in Table 1.2; data for the other five experiments are not reported in this table. An exact count of the number of seeds was not reported, only that in both cases a proportion of 3:1 was observed for seeds with the dominant character versus seeds with the recessive character.

Gregor Mendel summarizes these results with the following statement, as translated by W. Bateson:

The ratio 3:1, in accordance with which the distribution of the dominant and recessive characters results in the first generation, resolves itself therefore in all experiments into the ratio 2:1:1, if the dominant character be differentiated according to its significance as a hybrid-character or as a parental one.

Table 1.1 Data from G. Mendel's experiment for the first-generation (*F*1) from the hybrids

Character	No. plants	Characters of plants or seeds		
		Dominant	Recessive	Fraction
Round versus wrinkled seed	253	5,474	1,850	0.747
Yellow versus green seed	258	6,022	2,001	0.751
Violet-red versus white flower	929	705	224	0.759
Inflated versus constricted pod	1181	882	299	0.747
Green versus yellow unripe pod	580	428	152	0.738
Axial versus terminal flower	858	651	207	0.759
Long versus short stem	1064	787	277	0.740

Table 1.2 Data from G. Mendel's experiment for the second-generation (*F*2) from the hybrids, for the seed shape and seed color experiments. The numbers in parenthesis for hybrid plants (i.e., plants with seeds of both colors) is the proportion of the number of seeds with dominant to recessive character, as reported by Mendel. The final proportion is the ratio of hybrids (those displaying both characters in their seeds) to the pure dominant plants

Seed origin	No. plants	Characters of plants or seeds			Proportion
		Plants with only dom.	Hybrid plants (Seed properties)		
F1 round seeds	565	193	372 (3:1)		1.93:1
F1 yellow seed	519	166	353 (3:1)		2.13:1

Hybrids from parents that differ in two characters

Mendel also performed experiments in which two pure lines that differ by two characteristics were crossed. In particular, a line with yellow and round seeds was crossed with one that had green and wrinkled seeds. As in the previous case, the first generation of hybrid plants had a 100% occurrence of both dominant characters. Mendel let 15 of these hybrid plants self-fertilize and obtained 556 seeds. This first generation of the hybrids (*F*1) was distributed according to the data in Table 1.3.

Table 1.3 Data from G. Mendel's experiment for first-generation (*F*1) plants from the hybrids, with two different characters

	Yellow seed	Green seed
Round seed	315	108
Wrinkled seed	101	32

Mendel sowed all these seeds, and the resulting plants were left to self-fertilize to give new seeds. The second generation of the hybrids were plants with the characteristics in Table 1.4.

Table 1.4 Data from G. Mendel's experiment for second-generation (*F*2) plants from the hybrids, with two different characters. Possible genotypes are also listed

Round and yellow seed only	38	AA, BB
Round and yellow or green	65	AA, 2× Bb
Yellow and round or wrinkled	60	BB, 2× Aa
Round or wrinkled, and yellow or green	138	2× Aa, 2× Bb

These seeds were again sowed, and the resulting plants gave the seeds of the type reported in Table 1.5.

Table 1.5 Third-generation (F_3) plants with two different characters

Origin	Ensuing plants	Seed type
Wrinkled and yellow	96, of which: 28 all wrinkled and yellow 68 wrinkled yellow or green	aa, BB aa, Bb
Round and green	102, of which: 35 all round and green 67 green, round or wrinkled	AA, bb bb, Aa
Wrinkled and green	30, of which: 30 all wrinkled and green	aa, bb

Probability models for the experiments

These experiments can now be explained with the following principles:

(a) There are discrete units of inheritance, known as *genes*, that are passed on from parents to offspring. A gene is responsible for the appearance of a given trait or character, such as round versus wrinkled seed. Mendel identified that there were two factors (or *alleles*) for each basic trait and that each parent passed on one factor each to the offspring. For a given gene, Mendel used the notation A , a to refer to the two alleles. A plant with pure-breeding round seed would be AA (Mendel simply called this A), a plant with pure-breeding wrinkled seeds would be aa (or simply a in Mendel's notation), and a hybrid between the two would be Aa .

(b) The principle of *segregation*: during reproduction, the two factors (alleles) separate into separate reproductive cells (for plants, these would be the pollen and the ova). This makes it possible to select at random one allele from each parent.

(c) The principle of *independent assortment*: genes are inherited independently of one another. This principle is generally true for genes located on different chromosomes and it was true for the genes studied by Mendel. This principle explains the breeding of plants that differ by more than one character.

In probabilistic terms, the principle of segregation suggests that each plant has a probability $p = 1/2$ of drawing an allele (e.g., A or a) from each parent. For the first-generation hybrids, all plants will have an Aa configuration or *genotype*, and all plants show the dominant character. For the subsequent first generation of the hybrids (F_1), allowed to self-fertilize, the following probabilities are expected:

$$\begin{cases} P(AA) = P(a) \cdot P(A) = 1/4, \\ P(Aa) = P(A) \cdot P(a) + P(a) \cdot P(A) = 1/2, \\ P(aa) = P(a) \cdot P(a) = 1/4, \end{cases} \quad (1.9)$$

since the order in which alleles are chosen is irrelevant. Since the dominant allele determines the appearance of the characteristics, the observed 1:3 ratio in Table 1.1 is consistent with this explanation.

Probabilities for the occurrence of the dominant character in the second generation from the hybrids (F_2) can be established in a similar manner. Round seeds from the F_1 population were obtained from plants with the AA or Aa genotype, in a ratio of 1:2, or

$$\begin{cases} P(AA/F_1 \text{ round}) = 1/3 \\ P(Aa/F_1 \text{ round}) = 2/3. \end{cases}$$

Allowing this mixture to self-fertilize, yields these probabilities for the F_2 generation

$$P(AA) = 1/2 \times P(Aa/F_1 \text{ round}) = 1/3,$$

since this is the probability of selecting at random allele A from the Aa parent (the other parent can only provide A). These seeds will continue to breed true if self-fertilizing among themselves. This means that approximately $2/3$ of this F_2 population has the Aa configuration, for a 2:1 ratio of hybrids to only dominant, in agreement with Table 1.2. Allowed to self-fertilize, these hybrid plants behave in the same way as the F_1 generation of the hybrids, and yield a 3:1 ratio in the appearance of the dominant character.

For plants differing by two characters, there are four alleles to consider in each plant. The first generation of hybrids all had the Aa, Bb set of alleles, and all displayed the dominant characters. The first generation of the hybrids, Table 1.3, had the following possible genotypes:

556 total seeds	genotypes	Prob. and Expectations	
315 round and yellow	AA, (BB, $2 \times Bb$), or $2 \times Aa, (BB, 2 \times Bb)$	9/16 choices	(312.75)
101 wrinkled and yellow	aa, (BB, $2 \times Bb$)	3/16 choices	(104.25)
108 round and green	bb, (AA, $2 \times Aa$)	3/16 choices	(104.25)
32 wrinkled and green	aa, bb	1/16 choice	(34.75)

For example, with a total of 16 equally probable choices, the expected number of wrinkled and green seeds is 34.75, which is similar to the 32 observed.

Finally, the data in Table 1.5 can be explained by the following considerations. These are the possible genotypes of the seeds from the F_2 plants, including their probability of occurrence (Table 1.4):

F_2 plants yields	Number	Genotypes of seeds
All round, yellow	38	AABB
Round, yellow or green	65	$2 \times (AA \times (BB + 2Bb + bb))$
Yellow, round or wrinkled	60	$2 \times (BB \times (AA + 2Aa + aa))$
Round or wrinkled, yellow or green	138	$4 \times (AA + 2Aa + aa)$ $\times (BB + 2Bb + bb)$

The characters or *phenotypes* of these seeds can also be arranged accordingly:

Seed phenotype	Exp. Freq.	Genotypes available from F_2 seeds
Round and yellow	39	AABB(9), AaBb(16), AaBB(12), AABb(12)
Round and green	14	AAbb(6), Aabb(8)
Wrinkled and yellow	14	aaBB(6), aaBb(8)
wrinkled and green	4	aabb(4)

When these seeds are sowed, according to the seed phenotypes, the resulting plants will bear seeds according to their genotypes. Self-fertilization among the round and yellow seed plants will lead to plants all with the dominant characters. For the other combinations:

Seed phenotype	Genotypes available	F_3 Plant types
Round and green	$6 \times (AA \times bb)$ $+ 8bb \times (AA + 2Aa + aa)$	RG(6+8), RW/G(8×3)
Wrinkled and yellow	$6 \times (aa \times BB)$ $+ 8aa \times (BB + 2Bb + bb)$	WY(6+8), WY/G(8×3)

Here, one expects a proportion of 14:24, which is only approximately followed by the observed rates of 35 : 67 and 28 : 68 from the results of Table 1.5.

1.6 The Total Probability Theorem and Bayes' Theorem

This section describes two theorems that are of great importance in a number of practical situations.

Theorem 1.1 (Total Probability Theorem) *Given an event B and a set of events A_i that form a partition according to properties (1.1),*

$$P(B) = \sum_{i=1}^n P(B \cap A_i) = \sum_{i=1}^n P(B/A_i) \cdot P(A_i). \quad (1.10)$$

The first equation is immediately verified given that the $B \cap A_i$ are mutually exclusive events such that $B = \cup_i (B \cap A_i)$. The second equation derives from the application of the definition of conditional probability.

The total probability theorem is useful when the probability of an event B cannot be easily calculated and it is easier to calculate the conditional probabilities B/A_i .

Example 1.8 Consider the event B consisting of obtaining a sum of 8 in two consecutive tosses of a die. Each toss can be partitioned in 6 events A_i representing the die showing i , with $P(A_i) = 1/6$. It is clear that

$$P(B/A_i) = 1/6, \text{ only for } i = 2, \dots, 6$$

and null for $i = 1$, since there is no chance of a sum of 8 if the first toss is a 1. It follows that the sum in the total probability theorem leads to

$$P(B) = 5 \times \left(\frac{1}{6} \cdot \frac{1}{6} \right) = \frac{5}{36}.$$

◇

Theorem 1.2 (Bayes' Theorem) *Given an event B and a set of events A_i that forms a partition according to properties (1.1),*

$$P(A_i/B) = \frac{P(B/A_i)P(A_i)}{P(B)} = \frac{P(B/A_i)P(A_i)}{\sum_{i=1}^n P(B \cap A_i)}. \quad (1.11)$$

The proof is an immediate consequence of the definition of conditional probability, Eq. 1.6, and of the total probability theorem, Eq. 1.10.

Bayes' theorem is often written in a simpler form by taking into account two events only, $A_i = A$ and B . In this simplified situation, the theorem can be written as

$$P(A/B) = \frac{P(B/A)P(A)}{P(B)}. \quad (1.12)$$

In this form, Bayes' theorem is just a statement of how the order of conditioning between two events can be inverted. T. Bayes presented the first formulation of this theorem in his publication *An Essay towards solving a Problem in the Doctrine of Chances*, published posthumously by M. Price in 1763 [8]. Bayes was specifically interested in the problem of estimating the unknown probability p of a binary experiment that had resulted in x successes and $n - x$ failures. Equation 1.12 can be used to address this problem with the following interpretation of the events. The experiment B can be considered as the *data* collected in a given experiment; for Bayes' problem, it was the fact that x of n experiments were a success. The event A is a *model* that is used to describe the data, in Bayes' case the unknown probability p . Accordingly, the probabilities involved in Bayes' theorem can be interpreted as follows:

(a) $P(B/A)$ is the probability, or *likelihood* \mathcal{L} , of the data given the specified model. For Bayes' problem, this is the probability of having x successes, if p was known. Notice how $P(B/A)$ means that the model A is given, or known.

(b) $P(A)$ is the probability of the model A , without any knowledge of the data. This term is interpreted as a *prior probability*, or the degree of belief that the model is true before the measurements are made. For Bayes, this is the probability that a specific value of p is the correct one, before the experimental data (x successes out of n experiments) were collected. Prior probabilities should be based upon quantitative knowledge of the experiment, but can also reflect the subjective belief of the analyst.

This step in the interpretation of Bayes' theorem explicitly introduces an element of subjectivity that is characteristic of Bayesian statistics.

(c) $P(B)$ is the probability of collecting the dataset B . In practice, this probability acts as a normalization constant and its numerical value is typically of no practical consequence.

(d) Finally, $P(A/B)$ is the *posterior probability* of the model after the data have been collected. The posterior probability is the ultimate goal of the analysis since it describes the probability of the model based on the collection of data. For Bayes, this was the sought-after estimate of p based on the available data.

This interpretation of Bayes' theorem is the foundation of Bayesian statistics, and it can be summarized as

$$\text{Posterior probability} \propto \text{Likelihood} \times \text{Prior probability}.$$

Bayes' theorem provides a way to update the prior knowledge of model parameters given the measurements, leading to posterior estimates of parameters. One key feature of Bayesian statistics is that the calculation of probabilities is based on a prior probability, which may rely on a subjective interpretation of what is known about the experiment before any measurements are made. Therefore, great attention must be paid to the assignment of prior probabilities and the effect of priors on the final results of the analysis. H. Jeffreys' *Theory of Probability* [56] is a key reference for Bayesian statistics and the importance of priors.

Summary of Key Concepts for this Chapter

Event: A set of possible outcomes of an experiment.

Sample space: All possible outcomes of an experiment.

Probability of an Event: A number between 0 and 1 that follows the Kolmogorov axioms.

Frequentist or Classical approach: A method to determine the probability of an event based on many repetitions of the experiment.

Bayesian or Empirical approach: A method to determine probabilities that uses prior knowledge of the experiment.

Conditional probability: Probability of occurrence of an event given that another event is known to have occurred, $P(A/B) = P(A \cap B)/P(B)$.

Statistical independence: Two events are statistically independent when the occurrence of one has no influence on the occurrence of the other, $P(A \cap B) = P(A)P(B)$.

Total Probability theorem: A relationship among probabilities of events that form a partition of the sample space, $P(B) = \sum P(B/A_i)P(A_i)$.

Bayes' theorem: A relationship among conditional probabilities that enables the change in the order of conditioning of the events, $P(A/B) = P(B/A)P(A)/P(B)$.

Problems

- 1.1** Describe the sample space of the experiment consisting of flipping four coins simultaneously. Assign the probability to the event consisting of “two heads up and two tails up.” In this experiment, it is irrelevant to know which specific coin shows heads up or tails up.
- 1.2** An experiment consists of rolling two dice simultaneously and independently of one another. Find the probability of the event consisting of having either an odd number in the first roll or a total of 9 in both rolls.
- 1.3** For the roll of a die, find the probability of the event consisting of having either an even number or a number greater than 4.
- 1.4** An experiment consists of rolling two dice simultaneously and independently of one another. Show that the two events, “the sum of the two rolls is 8” and “the first roll shows 5” are not statistically independent.
- 1.5** An experiment consists of rolling two dice simultaneously and independently of one another. Show that the two events, “first roll is even” and “second roll is even” are statistically independent.
- 1.6** A box contains 5 balls, of which 3 are red and 2 are blue. Calculate (a) the probability of drawing two consecutive red balls and (b) the probability of drawing two consecutive red balls, given that the first draw is known to be a red ball. Assume that after each draw the ball is replaced in the box.
- 1.7** A box contains 10 balls that can be either red or blue. Of the first three draws, done with replacement, two result in the draw of a red ball. Calculate (a) the ratio of the probabilities that there are 2 red balls or just 1 red ball in the box, and (b) the ratio of probabilities that there are 5 red balls or just 1 red ball.
- 1.8** In the game of baseball a player at-bat either reaches base or is retired. Consider 3 baseball players: player A was at-bat 200 times and reached base a fraction 0.310 of times; player B was at-bat 250 times, with an on-base percentage of 0.296; player C was at-bat 300 times, with an on-base percentage 0.260. Find (a) the probabilities that when either player A, B, or C was at-bat, the player reached base, (b) the probabilities that, given that a player reached base, it was player A, B, or C.
- 1.9** An experiment consists of rolling two dice simultaneously and independently of one another. Calculate (a) the probability of the first roll being a 1, given that the sum of both rolls was 5, (b) the probability of the sum being 5, given that the first roll was a 1, and (c) the probability of the first roll being a 1 and the sum being 5. Finally, (d) verify your results with Bayes’ theorem.
- 1.10** Four coins labeled 1 through 4 are tossed simultaneously and independently of one another. Calculate (a) the probability of having an ordered combination of

heads–tails–heads–tails in the 4 coins, (b) the probability of having the same ordered combination given that any two coins are known to have landed heads-up, and (c) the probability of having two coins land heads up given that the sequence heads–tails–heads–tails has occurred.

Chapter 2

Random Variables and Their Distributions



Abstract Random variables can be continuous or discrete, and they are described by a distribution function that measures the probability of occurrence of any value of the variable. The purpose of performing experiments and collecting data is to gain information on random variables of interest. Useful measures of the distribution of random variables are the mean, variance, and other higher-order moments. Measurements of the variable can be used to estimate these moments by calculating quantities such as the sample mean and the sample variance.

2.1 Random Variables

A random variable is a quantity of interest whose true value is unknown. To gain information on a random variable, it is necessary to design and conduct experiments. It is inherent to any experiment that the random variable of interest will never be known exactly. Instead, the variable will be characterized by a *probability distribution function*, which determines what is the probability that a given value of the random variable occurs. Repeating the measurement typically increases the knowledge of the distribution of the variable: this is the reason for wanting to measure the quantity as many times as possible.

Examples of random variables are the mass of the Earth, the height of the Eiffel tower, or the voltage from an electrical outlet. The random nature of virtually all quantities lies primarily in the fact that no quantity is known exactly to us without performing an experiment, and that no experiment is ever perfect because of practical or even theoretical limitations. Among the practical reasons are, for example, limitations in the precision of the measuring apparatus. Theoretical reasons depend on the nature of the variable. For example, the measurement of the position and

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-981-19-0365-6_2.

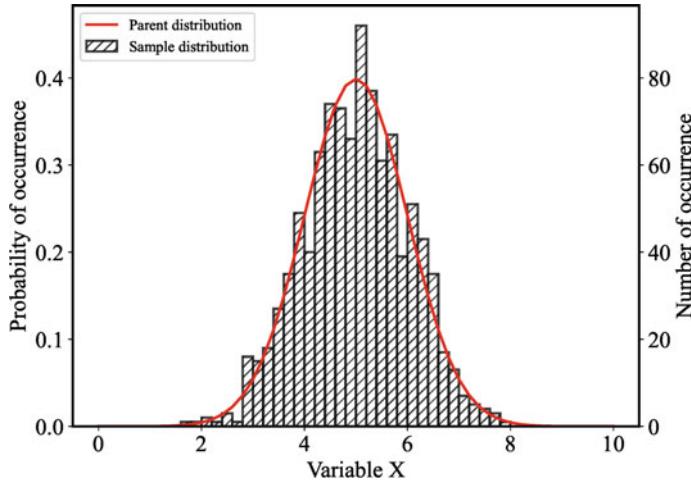


Fig. 2.1 The sample distribution of a random variable X from 500 measurements, obtained by combining the measurements into bins of equal width ($\Delta x = 0.2$). The shape of the parent distribution depends on the nature of the experiment and of the number of measurements, and it is represented in this example by the red curve

velocity of a subatomic particle is limited by the Heisenberg uncertainty principle [49], which forbids an exact knowledge of both quantities even in the presence of a perfect measuring apparatus.

The general method for gaining information on a random variable X starts with set of measurements x_i , ensuring that measurements are performed under the same experimental conditions. From these measurements, one obtains a distribution of the frequency of occurrence of all values of X known as the *sample distribution* of the variable, which describes the empirical distribution of values collected in the experiment (Fig. 2.1). A random variable is also expected to have a theoretical distribution, e.g., Gaussian, uniform, etc., according to the nature of the variable itself and the method of measurement. This theoretical distribution is referred to as the *parent distribution*, and it represents the belief that there is an ideal description of the random variable. As will be shown in later chapters, the sample distribution is expected to become the parent distribution if an infinite number of measurements are performed, in such a way that the randomness associated with a small number of measurements is eliminated.

2.2 Probability Distribution Functions

It is convenient to define a function that describes the probability of occurrence of the random variable. Discrete random variables are described by a *probability mass function* $f(x_i)$, where $f(x_i)$ represents the probability that the variable has an

exact value of x_i . Continuous variables are described by a *probability distribution function* $f(x)$, such that $f(x)dx$ is the probability that the variable is in the interval $[x, x + dx]$. For simplicity, most of the properties will be illustrated for continuous variables, with the understanding that they also apply to discrete variables with straightforward modifications, such as the replacement of an integral with a sum. In principle, it is also possible to have more complex random variables that are in part continuous and in part discrete. A more complete treatment of random variables can be found in any textbook dedicated to the subject, such as those of S. Ross [86] or K. Siegrist [92]. Probability distribution functions have the following key properties:

1. They are normalized to 1. For continuous variables, this means that

$$\int_{-\infty}^{+\infty} f(x)dx = 1. \quad (2.1)$$

For variables that are defined in a subset of the real numbers, e.g., only values $x \geq 0$ or in a finite interval, $f(x)$ is set to zero outside the domain of definition of the function. For discrete variables, the integral is replaced by a sum over all values that the variable can have.

2. The probability distribution can never be negative, $f(x) \geq 0$. This is a consequence of the Kolmogorov axiom that requires a probability to be non-negative.
3. The *cumulative distribution function* $F(x)$, or simply distribution function

$$F(x) = P(X \leq x) = \int_{-\infty}^x f(\tau)d\tau, \quad (2.2)$$

represents the probability that the variable has a value less or equal than x . $F(x)$ is a non-decreasing function of x that starts at zero and has its highest value of one. A related function is the *survival function* $S(x) = 1 - F(x)$, representing the probability that $X > x$.

Example 2.1 (*Distributions of the exponential random variable*) The *exponential random variable* follows the probability distribution function:

$$f(x) = \lambda e^{-\lambda x}, \quad x \geq 0, \quad (2.3)$$

where λ is a rate parameter that must be positive. The probability distribution function is therefore $f(x) = 0$ for negative values of the variable. The cumulative distribution function is given by

$$F(x) = 1 - e^{-\lambda x}. \quad (2.4)$$

In Fig. 2.2 are drawn the probability distribution function $f(x)$ and the cumulative distribution function $F(x)$ for an exponential variable with $\lambda = 0.5$. \diamond

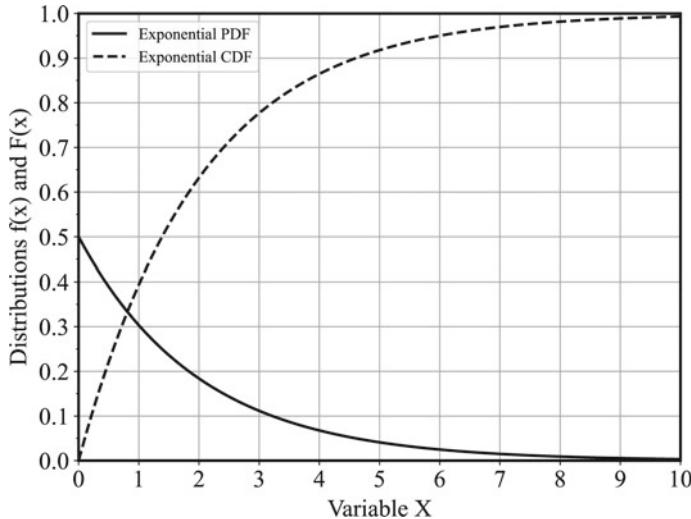


Fig. 2.2 The distribution function $f(x)$ (solid curve) and the cumulative distribution function $F(x)$ (dashed curve) for an exponential variable with $\lambda = 0.5$

2.3 Expectations and Moments of a Distribution Function

The probability distribution function $f(x)$ or the distribution function $F(x)$ provides a complete description of the random variable X . It is convenient to find a few quantities that describe the salient features of the distribution. The *expectation* of a deterministic function $g(x)$ of the random variable is defined by

$$E[g(X)] = \int g(x)f(x)dx. \quad (2.5)$$

The meaning of Eq. 2.5 is that each value $g(x)$ is weighed according to the probability of occurrence of x , which is $f(x)$. The expectation is a linear operator, i.e., it satisfies the property

$$E[ag_1(x) + bg_2(x)] = aE[g_1(x)] + bE[g_2(x)]$$

due to the property of linearity of the integral.

The *moment of order n* is defined as

$$\mu_n = E[X^n] = \int f(x)x^n dx. \quad (2.6)$$

The moment μ_n is therefore the *expectation* of the function $g(x) = x^n$. It is possible to demonstrate, although beyond the scope of this book, that the knowledge of moments of all orders is sufficient to determine uniquely the distribution function

[102]. This is an important fact, since it shifts the problem of determining the distribution function to that of determining at least some of its moments. Moreover, a number of distribution functions only have a few non-zero moments, and this renders the task even more manageable.

The moments of a distribution are theoretical quantities that can be calculated from the probability distribution $f(x)$, without any reference to measurements. In other words, they are *parent* quantities. The following describes two of two most commonly used moments, the mean and the variance, and the *sample* quantities that approximate them, the sample mean and the sample variance. Chapter 6 describes a method to justify the estimates of parent quantities via sample quantities.

2.3.1 The Mean and the Sample Mean

The moment of the first order is also known as the *mean* or *expectation* of the random variable,

$$\mu = E[X] = \int_{-\infty}^{+\infty} x f(x) dx. \quad (2.7)$$

The mean μ is a characteristic number representing an average value of X , obtained by weighting all possible values of the variable by its distribution function. The mean is therefore a very simple and convenient measure of the random variable. To estimate the mean of the random variable X , consider N measurements x_i , with $i = 1, \dots, N$. Each measurement x_i can be considered as a sample drawn at random from the parent distribution of X . The *sample mean* is defined as

$$\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i. \quad (2.8)$$

It is clear that different samples of size N will not give rise to the same value of \bar{x} , due to the randomness of the measurements. This indicates that \bar{x} is not a fixed number but is itself a random variable that ought to be written as

$$\bar{X} = \frac{1}{N} \sum_{i=1}^N X_i.$$

This new equation is formally identical to Eq. 2.8 except for the use of uppercase letters, indicating that the quantities are random variables, not just numbers. It is therefore reasonable to ask whether this new random variable \bar{X} has the same expectation as the parent variable X itself. This question can be easily answered considering that the X_i variables are identically distributed in the same way as X , since they represent random measurements from X . The expectation of \bar{X} is thus

$$\mathbb{E}[\bar{X}] = \mathbb{E}\left[\frac{1}{N} \sum_{i=1}^N X_i\right] = \frac{1}{N} \sum_{i=1}^N \mathbb{E}[X_i] = \frac{N\mu}{N} = \mu,$$

making use of the property of linearity of the expectation. The sample mean has the same expectation as the parent variable, and therefore \bar{X} is said to be an *unbiased* estimator of X . This is a very desirable property, given that the sample mean was designed exactly for the purpose of estimating the unknown parent variable X with its measurements. It is also expected that, as the number of measurements N increases, the sample mean will approximate the parent mean with increasing accuracy. This is shown in the following section.

2.3.2 The Law of Large Numbers

The *law of large numbers* states that the sample mean converges to the parent mean as the sample size increases, and is one of the fundamental theorems of probability. Consider N random variables X_i that are identically distributed with μ their common mean. The law can be stated as

$$\lim_{N \rightarrow \infty} \frac{X_1 + \dots + X_N}{N} = \mu, \quad (2.9)$$

implying that the sample mean \bar{X} tends to the mean μ , which is a deterministic number and not a random variable. Equation (2.9) is a very strong statement because it shows that, asymptotically, the sum of random variables becomes a constant equal to the sample mean of the N variables, or N measurements. Although no indication is given towards establishing how large N should be in order to achieve this goal, it is nonetheless a key result to establish the asymptotic behavior of random variables. It is useful to point out that there are different versions of this law, according to method by which this convergence is established. Mathematical properties of this law and its derivation can be found in books of theory of probability, such as [63, 86].

This law has also important consequence for functions of random variables. Given a function $g(x)$, it would be convenient to estimate its expected value $\mathbb{E}[g(X)]$ from the N measurements of the variables X_i . According to the law of large numbers, it is possible to say that

$$\lim_{N \rightarrow \infty} \frac{g(X_1) + \dots + g(X_N)}{N} = \mathbb{E}[g(X)]. \quad (2.10)$$

Equation (2.10) states that a large number of measurements of the variables X_i can be used to measure the expectation of $\mathbb{E}[g(X)]$, entirely bypassing the probability distribution function of the function $g(X)$. This property will become useful when studying the distribution of functions of a random variable.

The law of large numbers and the convergence of the sample mean to the parent mean can be illustrated using a discrete variable. Assuming for simplicity that the random variable X has M possible values, the expectation is

$$\mathbb{E}[X] = \sum_{j=1}^M f(x_j)x_j = \mu,$$

where $f(x_j)$ is the probability mass function. According to the classical interpretation of the probability, this function is given by

$$f(x_j) = \lim_{N \rightarrow \infty} \frac{N(x_j)}{N},$$

where $N(x_j)$ is the number of times the value x_j occurred, with $j = 1, \dots, M$. Since $\sum N(x_j)x_j$ is the sum of the values obtained in N measurements, it is equivalent to $\sum x_i$, where $i = 1, \dots, N$ labels the N measurements. Therefore,

$$\lim_{N \rightarrow \infty} \bar{x} = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{i=1}^N x_i = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{j=1}^M N(x_j)x_j = \sum_{j=1}^M f(x_j)x_j = \mu,$$

showing that the sample mean will be identical to the parent mean in the limit of an infinite number of measurements, in accordance with the law of large numbers.

2.3.3 The Variance and the Sample Variance

The *variance* is the expectation of the square of the deviation of X from its mean,

$$\text{Var}(X) = \mathbb{E}[(X - \mu)^2] = \int_{-\infty}^{+\infty} (x - \mu)^2 f(x) dx \quad (2.11)$$

and it is usually indicated as $\text{Var}(X) = \sigma^2$. The square root of the variance is referred to as the *standard deviation* or *standard error* σ and it is a common measure of the average difference of a given measurement x_i from the mean of the random variable. The main reason for defining the average difference of a measurement from its mean in terms of a moment of the second order is that the expectation of the *deviation* $X - \mu$ is always zero, as can be immediately seen using the linearity property of the expectation. The deviation of a random variable is therefore not of common use in statistics, since its expectation is null. Notice that the physical dimensions of moments of the n -th order are those of the variable to the n -th power. For example, if X is measured in meters, the variance is measured in meters square (m^2), whereas the standard deviation has the same dimensions as the variable and its mean.

Using the linear property of the expectation, it is straightforward to show that the following property applies:

$$\text{Var}(X) = \mathbb{E}[X^2] - \mu^2. \quad (2.12)$$

This relationship is very convenient to calculate the variance from the moments of the first and second orders. Another useful property of the variance, which follows from the fact that the variance is a moment of the second order, is

$$\text{Var}(aX) = a^2 \text{Var}(X), \quad (2.13)$$

where a is a constant. The deviation and the variance are moments calculated with respect to the mean, and they are referred to as *central moments*.

The *sample variance* is defined as

$$s^2 = \frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})^2 \quad (2.14)$$

as a new random variable that is intended to estimate the parent variance σ^2 via N random measurements. The factor of $N-1$, and not N , at the denominator is required to ensure that the sample variance is an unbiased estimator of the parent variance. A proof of this property is provided in Sect. 2.6.

2.4 A Classic Experiment: J.J. Thomson's Discovery of the Electron

A set of experiments by J.J. Thomson in the late nineteenth century were aimed at the measurement of the ratio between the mass and charge of a new lightweight particle, which was later named *electron*. The experiment was truly groundbreaking not just for the method used but also because it revolutionized our understanding of physics and natural sciences by proving that the new particle was considerably lighter than the previously known charge carrier, the proton.

The experiment described in this book was reported by Thomson in [97]. It consists of measuring the deflection of negatively charged cathode rays by a magnetic field H in a tube. Thomson wanted to measure the mass m of the charged particles that constituted these cathode rays. The experiment is based on the measurement of the following quantities: W is the kinetic energy of the particles, $Q = Ne$ is the amount of electricity carried by the particles (N is the number of particles and e the charge of each particle), and $I = HR$, where R is the radius of curvature of the path of these rays in a magnetic field H . The measurements performed by Thomson were used to infer the ratio m/e and the speed v of the new lightweight particle according to

$$v = \frac{2W}{QI};$$

$$\frac{m}{e} = \frac{I^2 Q}{2W}. \quad (2.15)$$

For the purpose of the data analysis of this experiment, it is only necessary to know that W/Q and I are the primary quantities being measured, and inferences on the secondary quantities of interest are based on (2.15). For the proton, the mass-to-charge ratio was known to be approximately 1×10^{-4} g per electromagnetic (EMU) charge unit, where the EMU charge unit is equivalent to 10^{-10} electrostatic charge units, or ESU (a more common unit of measure for charge). In Thomson's units, the accepted value of the mass to charge ratio of the electron is now 5.7×10^{-8} . Some of the experimental data collected by Thomson are reported in Tables 2.1 and 2.2, in which "gas" refers to the gas used in the tubes he used for the experiment.

Some of Thomson's conclusions are reported here:

- (a) It will be seen from these tables that the value of m/e is independent of the nature of the gas.
- (b) the values of m/e were, however, the same in the two tubes.
- (c) for the first tube, the mean for air is 0.40×10^{-7} , for hydrogen 0.42×10^{-7} and for carbonic acid 0.4×10^{-7} .
- (d) for the second tube, the mean for air is 0.52×10^{-7} , for hydrogen 0.50×10^{-7} and for carbonic acid 0.54×10^{-7} .

Using the equations for sample mean and variance explained in Sect. 2.3, the sample means and variances in air for the two tubes are calculated as follows (the numbers need to be multiplied by 10^{-7}):

$$\text{Tube 1 (air): } \overline{m/e} = 0.423 \quad s_{m/e}^2 = 0.072, \quad \text{reported as } 0.423 \pm 0.085$$

$$\text{Tube 2 (air): } \overline{m/e} = 0.529 \quad s_{m/e}^2 = 0.044, \quad \text{reported as } 0.529 \pm 0.066$$

To make more quantitative statements on the statistical agreement between the two measurements, one needs to know the distribution of the sample mean. The test to determine whether the two measurements are consistent with each other will be explained in Sect. 9.6. For now, the fact that the range of the two measurements overlap is an indication of the likely statistical agreement between the two measurements.

Note: The three measurements marked with a star appear to have values of v or m/e that are inconsistent with the formulas to calculate them from W/Q and I . They may be typographical errors in the original publication or approximations. The first appears to be a typo in W/Q (6×10^{12} should be 6×10^{11}), the corrected value is assumed throughout this book. The second has an inconsistent value for v (should be 6.5×10^9 , not 7.5×10^9), the third has inconsistent values for both v and m/e , but no correction was applied in these cases to the data in the tables.

Table 2.1 Data from Thomson's measurements of Tube 1

Gas	W/Q	I	m/e	v
<i>Tube 1</i>				
Air	4.6×10^{11}	230	0.57×10^{-7}	4×10^9
Air	1.8×10^{12}	350	0.34×10^{-7}	1×10^{10}
Air	6.1×10^{11}	230	0.43×10^{-7}	5.4×10^9
Air	2.5×10^{12}	400	0.32×10^{-7}	1.2×10^{10}
Air	5.5×10^{11}	230	0.48×10^{-7}	4.8×10^9
Air	1×10^{12}	285	0.4×10^{-7}	7×10^9
Air	1×10^{12}	285	0.4×10^{-7}	7×10^9
Hydrogen* .	6×10^{12}	205	0.35×10^{-7}	6×10^9
Hydrogen ..	2.1×10^{12}	460	0.5×10^{-7}	9.2×10^9
Carbonic acid*	8.4×10^{11}	260	0.4×10^{-7}	7.5×10^9
Carbonic acid	1.47×10^{12}	340	0.4×10^{-7}	8.5×10^9
Carbonic acid	3.0×10^{12}	480	0.39×10^{-7}	1.3×10^{10}

See Note for meaning of \star **Table 2.2** Data from Thomson's measurements of Tube 2

Gas	W/Q	I	m/e	v
<i>Tube 2</i>				
Air	2.8×10^{11}	175	0.53×10^{-7}	3.3×10^9
Air*	2.8×10^{11}	175	0.47×10^{-7}	4.1×10^9
Air	3.5×10^{11}	181	0.47×10^{-7}	3.8×10^9
Hydrogen .	2.8×10^{11}	175	0.53×10^{-7}	3.3×10^9
Air	2.5×10^{11}	160	0.51×10^{-7}	3.1×10^9
Carbonic acid	2.0×10^{11}	148	0.54×10^{-7}	2.5×10^9
Air	1.8×10^{11}	151	0.63×10^{-7}	2.3×10^9
Hydrogen .	2.8×10^{11}	175	0.53×10^{-7}	3.3×10^9
Hydrogen .	4.4×10^{11}	201	0.46×10^{-7}	4.4×10^9
Air	2.5×10^{11}	176	0.61×10^{-7}	2.8×10^9
Air	4.2×10^{11}	200	0.48×10^{-7}	4.1×10^9

See Note for meaning of \star

2.5 Covariance and Correlation Between Random Variables

It is common to measure more than one random variable in a given experiment. The variables are often related to one another, and it is therefore necessary to define a measure of how one variable affects the measurement of the others. Consider the measurement of both the length of one side of a square and its area; it is clear that the two quantities are related in a way that the change of one quantity affects the other in the same manner, i.e., a positive change of the length of the side results in a positive change of the area. In this case, the length and the area will be said to have a positive correlation. This section introduces key concepts to study two or more random variables simultaneously.

2.5.1 Joint Distribution and Moments of Two Random Variables

When two variables are measured at the same time, one is interested in knowing the probability of a given pair of measurements for the two variables. This information is provided by the *joint probability distribution function*, indicated as $h(x, y)$, with the meaning that $h(x, y) dx dy$ is the probability that the two variables X and Y are in a two-dimensional interval of size $dx dy$ around the value (x, y) . This two-dimensional function can be represented experimentally via its sample distribution, in the same way as one-dimensional distributions.

It is usually convenient to describe the behavior of one variable at a time, even if the experiment features more than one variable. The process of considering only one variable at a time is called *marginalization*. The marginal probability density function of X , i.e., the distribution function of X alone, is obtained from the joint probability distribution function as

$$f_X(x) = \int_{-\infty}^{+\infty} h(x, y) dy, \quad (2.16)$$

with an equivalent equation for the marginal distribution of Y . The expectation of X is defined as

$$E[X] = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} x h(x, y) dx dy = \mu_x \quad (2.17)$$

and likewise the expectation of Y is equal to

$$E[Y] = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} y h(x, y) dx dy = \mu_y.$$

The combination of these two equations leads to the result

$$E[X + Y] = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (x + y) h(x, y) dx dy = E[X] + E[Y],$$

which means that the expectation of the sum of two variables being equal to the sum of the expectations. This is a convenient result that applies to all random variables, and it can be generalized to more than two variables and it is usually referred to as the *linear property of the expectation*. In particular, it is useful to point out that the linear property applies regardless of the statistical independence among the variables, which is introduced later in this section.

The variance is similarly defined as

$$\text{Var}(X) = E[(X - \mu_x)^2] = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (x - \mu_x)^2 h(x, y) dx dy = \sigma_x^2. \quad (2.18)$$

These equations recognize the fact that the other variable, in this case Y , is indeed part of the experiment, but is considered *uninteresting* for the calculation at hand. Therefore, the uninteresting variable is integrated (or summed) over, weighted by its probability distribution function.

The *covariance* of two random variables is defined as

$$\text{Cov}(X, Y) = E[(X - \mu_x)(Y - \mu_y)] = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (x - \mu_x)(y - \mu_y) h(x, y) dx dy \quad (2.19)$$

and it is usually represented as $\text{Cov}(X, Y) = \sigma_{xy}^2$. The covariance is the expectation of the product of the deviations of the two variables, and the order of the two variables is irrelevant, so that $\text{Cov}(X, Y) = \text{Cov}(Y, X)$. The covariance is positive if, on average, a positive deviation of X is accompanied by a positive deviation of Y , or if two negative deviations are likely to occur simultaneously, so that the integrand is a positive quantity. If, on the other hand, the two variables tend to have deviations of opposite sign, the covariance will be negative. The covariance, like the mean and variance, is a parent quantity that can be calculated from the theoretical distribution of the random variables. The covariance can also be calculated as

$$\text{Cov}(X, Y) = E[XY] - E[X] \cdot E[Y]$$

and it has the linear property

$$\text{Cov}(aX + b, cY + d) = ac \cdot \text{Cov}(X, Y),$$

which can be seen immediately from the definition of covariance and the linear property of the expectation. Another useful property is

$$\text{Cov}(X + Z, Y) = \text{Cov}(X, Y) + \text{Cov}(Z, Y),$$

which can be proven by expanding the expectations, and it can be generalized for more than two variables.

The *sample covariance* for a collection of N pairs of measurements is calculated as

$$s_{xy}^2 = \frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y}), \quad (2.20)$$

using a similar equation to the sample variance. It is possible to show that the denominator ($N - 1$) is required to ensure that the sample covariance is an unbiased estimator of the parent covariance. This derivation is provided in Sect. 2.6.

The *correlation* is simply a normalized version of the covariance,

$$\text{Cor}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}} = \frac{\sigma_{xy}^2}{\sigma_x \sigma_y}, \quad (2.21)$$

and it can be indicated as $\text{Cor}(X, Y) = \rho$. The *Cauchy–Schwartz* inequality is often stated as

$$\mathbb{E}[|XY|] \leq \sqrt{\mathbb{E}[X^2]\mathbb{E}[Y^2]}, \quad (2.22)$$

and it can be used to show that $-1 \leq \rho \leq 1$. When the correlation coefficient is zero, the two variables are said to be *uncorrelated*. The *sample correlation coefficient* r

$$r = \frac{s_{xy}^2}{s_x s_y} = \frac{\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^N (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^N (y_i - \bar{y})^2}} \quad (2.23)$$

in which s_{xy}^2 is the sample covariance and s_x^2 and s_y^2 are the sample variances of the two variables, and the denominators ($N - 1$) of both the covariance and variances have canceled out. The Cauchy–Schwartz inequality can be used again to show that $-1 \leq r \leq 1$, same as for the parent correlation.

Example 2.2 (*Correlation between W/Q and I in Thomson's experiment*) Consider Thomson's experiment to measure the ratio between the mass and the charge of the electron. The data included a set of measurements for the pair of quantities W/Q and I , which were used for measuring the velocity and the mass-to-charge ratio. The sample covariances between $x = W/Q$ (the numbers need to be multiplied by 10^{11}) and $y = I$ for the 11 measurements Tube 1, and the 11 measurements in air of Tube 2, are

Tube 1: $\bar{x} = 13.275$, $s_x = 8.456$; $\bar{y} = 312.92$, $s_y = 93.36$, $s_{xy}^2 = 759.12$ ($r = 0.962$)

Tube 2: $\bar{x} = 2.918$, $s_x = 0.817$; $\bar{y} = 174.27$, $s_y = 16.92$, $s_{xy}^2 = 13.15$ ($r = 0.952$)

The large positive value of the correlation factor r (close to the maximum value of 1) indicates a very strong degree of positive correlation between the measurements of W/Q and I , for both tubes, making it clear that the two quantities are not independent of one another. This dependence could be inherent to the nature of the quantities (W is the kinetic energy, Q is the amount of electricity and I is proportional to the radius of curvature of the path in a magnetic field) or to the method of measurement, or a combination of these factors. ◇

2.5.2 Statistical Independence of Random Variables

The independence between events was described and quantified in Chap. 1, where it was shown that two events are independent only when the probability of both events occurring (or their intersection) is the product of the individual probabilities. The

concept is extended here to random variables by defining two random variables as *independent* if and only if the joint probability distribution function can be factored in the following form:

$$h(x, y) = f(x) \cdot g(y), \quad (2.24)$$

where the two functions $f(x)$ and $g(y)$ become the probability distribution functions of the two random variables. When two variables are independent, the marginal probability distribution function of each variable, obtained by marginalization over the other variable (see Eq. 2.16), is

$$f_X(x) = \int_{-\infty}^{+\infty} f(x)g(y)dy = f(x), \quad (2.25)$$

showing that $f(x)$ is indeed the marginal distribution function of X . An identical result is also applicable to $g(y)$.

It is important to remark that independence between random variables and uncorrelation are not equivalent properties. Independence, which is a property of the distribution functions (2.24), is a much stronger property than uncorrelation, which is based on a statement that involves only moments. It can be proven that independence implies uncorrelation, but not vice versa.

The fact that independence implies uncorrelation is shown by calculating the covariance of two independent random variables, with joint distribution function $h(x, y)$. The covariance is

$$\begin{aligned} \sigma_{xy}^2 &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (x - \mu_x)(y - \mu_y)h(x, y)dxdy = \\ &\int_{-\infty}^{+\infty} (x - \mu_x)f(x)dx \int_{-\infty}^{+\infty} (y - \mu_y)g(y)dy = 0, \end{aligned}$$

where each integral vanishes as the expectation of the deviation of a random variable. \square

As a counter-example of the fact that dependent variables can have non-zero correlation factor, consider the case of a random variable X with a distribution $f(x)$ that is symmetric around the origin, and another variable $Y = X^2$. They cannot be independent since they are functionally related, but it will be shown that their covariance is zero. Symmetry about zero implies $\mu_x = 0$. The mean of Y is $E[Y] = E[X^2] = \sigma_x^2$ since the mean of X is null. From this, the covariance is given by

$$\text{Cov}(X, Y) = E[X(Y - \sigma_X^2)] = E[X^3 - X\sigma_X^2] = E[X^3] = 0$$

due to the symmetry of $f(x)$. Therefore, the two variables X and X^2 are uncorrelated, yet they are not independent.

The variance of two sum of two or more independent variables has a very convenient property that is useful to present at this point. Consider the sum of N independent variables, $X = X_1 + \dots + X_N$. The variance of X has the following additive property:

$$\text{Var}(X) = \sum_{i=1}^N \text{Var}(X_i).$$

The joint distribution function of N independent variables can be factored out according to a simple generalization of Eq. 2.24,

$$h(x_1, \dots, x_n) = f_1(x_1) \dots f_N(x_N)$$

where each $f_i(x_i)$ represents the marginal distribution function of the variable X_i , with parent mean μ_i . The variance of the sum can be written as

$$\text{Var}(X) = E[(X - \mu_X)^2] = E[(X_1 + \dots + X_N - (\mu_1 + \dots + \mu_N))^2],$$

where μ_X is the mean of the sum of variables, which is always equal to the sum of the means. The square can be expanded as

$$\text{Var}(X) = \sum_{i=1}^N E[(X_i - \mu_i)^2] + 2 \sum_{i>j} E[(X_i - \mu_i)(X_j - \mu_j)],$$

where the expectations of all cross-product terms in the second sum are the covariance between variables, which are all null because of their independence. This leads to the linear property of the variance, which applies only in the case of independent variables.

2.6 The Expectation of the Sample Variance and Sample Covariance

The sample variance and sample covariance are examples of random variables that are calculated from several independent measurements of one or two random variables. As remarked earlier in the case of the sample mean, which is another random variable that is function of several independent measurements, the sample variance and sample covariance could be indicated with uppercase letters, to signify that they are not just numbers, but random variables. For the sake of keeping the notation as simple as possible, the lowercase notation will be retained, with the understanding that these quantities are random variables. This means that the sample mean, sample variance, and sample covariance all have a distribution with their accompanying moments, such as the mean (or expectation) and the variance. In the case of sample mean,

its expectation was directly calculated as $E[\bar{x}] = \mu$, where μ is the parent mean of the variable X . The agreement between the expectation of the sample mean and the parent mean indicates that the sample mean is an unbiased estimator of the mean.

Similar considerations need to be applied to the sample variance and covariance, and ascertain whether their expectations agree with the parent quantities they were designed to approximate. The measurements that contribute to these sample variables are of the type x_i , or pairs (x_i, y_i) , for $i = 1, \dots, N$. The N measurements are independent samples from the variable X or the pair of variables X, Y . This property of independence among samples is key to understand the expectation of the sample variables under consideration. The measurements x_i are said to be independent and identically distributed variables (often abbreviated as *iid*), that is, independent samples from the same parent distribution $f(x)$ with parent mean μ . The same applies to the two-dimensional case of pairs (x_i, y_i) , which are independent samples from a two-dimensional parent distribution $h(x, y)$.

The following derivation shows that the expectation of the sample variance as defined in Eq. 2.14 is equal to the parent variance of the variable X , and that the expectation of the sample covariance as defined in Eq. 2.20 is equal to the parent covariance. These properties shows that the sample variance and sample covariance are unbiased estimators of the corresponding parent quantities. The factor of $N - 1$ at the denominator ultimately derives from the fact that the sample means were estimated from the data, introducing an additional source of uncertainty. One may think of this situation as follows: the N measurements, either of a single variable or of a pair of variables, are used first to estimate the sample mean(s), thereby reducing the effective or independent number measurements to $N - 1$.

The expectation of the sample variance is obtained from

$$\begin{aligned} E[s^2] &= E\left[\frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})^2\right] = \frac{1}{N-1} E\left[\sum_{i=1}^N (x_i - \mu + \mu - \bar{x})^2\right] \\ &= \frac{1}{N-1} E\left[\sum_{i=1}^N (x_i - \mu)^2 + \sum_{i=1}^N (\mu - \bar{x})^2 + 2(\mu - \bar{x}) \sum_{i=1}^N (x_i - \mu)\right] \end{aligned}$$

The term $E[\sum_{i=1}^N (\mu - \bar{x})^2]$ is N times the variance of the sample mean. This variance can be calculated using the observation that the N measurements x_i are in fact independent random variables with the same distribution and same mean μ . The linear property of the variance, which only applies to independent variables, leads to

$$\text{Var}(\bar{x}) = \frac{\text{Var}(x_1 + \dots + x_n)}{N^2} = \frac{\sigma^2}{N},$$

where σ^2 is the parent variance of each of the X_i variables. The last term in the equation is $\sum_{i=1}^N (x_i - \mu) = N(\bar{x} - \mu)$, therefore:

$$\begin{aligned} E[s^2] &= \frac{1}{N-1} \left(E \left[\sum_{i=1}^N (x_i - \mu)^2 \right] + N E[(\mu - \bar{x})^2] + 2N E[(\mu - \bar{x})(\bar{x} - \mu)] \right) \\ &= \frac{1}{N-1} (N\sigma^2 + N\sigma^2/N - 2N E[(\mu - \bar{x})^2]). \end{aligned}$$

Since $E[(\mu - \bar{x})^2]$ is again the variance of the sample mean, the expectation of the sample variance is finally calculated as

$$E[s^2] = \frac{1}{N-1} (N\sigma^2 + N\sigma^2/N - 2N\sigma^2/N) = \sigma^2.$$

Prior to evaluating the expectation of the sample covariance, it is necessary to calculate the covariance between the two sample means \bar{x} and \bar{y} . According to the properties of the covariance,

$$\text{Cov}(\bar{x}, \bar{y}) = \frac{1}{N^2} \left(\sum_{i=1}^N \text{Cov}(x_i, y_i) + 2 \sum_{i \neq j} \text{Cov}(x_i, y_j) \right).$$

The independence between pairs of measurements means that $\text{Cov}(x_i, y_j) = 0$ when $i \neq j$. Moreover, since $\text{Cov}(x_i, y_i) = \sigma_{xy}^2$ is the parent covariance, one obtains the result

$$\text{Cov}(\bar{x}, \bar{y}) = \frac{\sigma_{xy}^2}{N}.$$

The sample covariance can be written as

$$s_{xy}^2 = \frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{N-1} \left(\sum_{i=1}^N x_i y_i - N \bar{x} \bar{y} \right).$$

Since

$$E[x_i y_i] = \sigma_{xy}^2 + \mu_x \mu_y$$

and

$$E[\bar{x} \bar{y}] = \text{Cov}(\bar{x}, \bar{y}) + E[\bar{x}] E[\bar{y}],$$

the expectation of the sample covariance is finally evaluated as

$$E[s_{xy}^2] = \frac{N(\sigma_{xy}^2 + \mu_x \mu_y) - N(\sigma_{xy}^2/N - \mu_x \mu_y)}{N-1} = \sigma_{xy}^2.$$

This shows that the sample covariance is an unbiased estimator of the parent covariance.

2.7 A Classic Experiment: Pearson’s Collection of Data on Biometric Characteristics

In 1903 K. Pearson published the analysis of a collection of biometric data on more than one thousand families in the United Kingdom, with the goal of establishing how certain characters, such as height, are correlated and inherited [78]. Pearson is also the inventor of the χ^2 test and a central figure in the development of the modern science of statistics.

Pearson asked a number of families, composed of at least the father, mother, and one son or daughter, to perform measurements of height, span of arms, and length of left forearm. This collection of data resulted in a number of tables, including some for which Pearson provides the distribution of two measurements at a time. One such table is that reporting the mother’s height versus the father’s height, Table 2.3.

The data reported in Table 2.3 represent the joint probability distribution of the two physical characters, binned in 1-inch intervals. When a non-integer count is reported (e.g., a value of 0.25, 0.5, or 0.75), it is interpreted as meaning that the original measurement fell exactly at the boundary between two cells, although Pearson does not provide an explanation for non-integer values.

Every column and row also has the sum of all counts. The bottom row in the table is therefore the distribution of the father’s height, irrespective of the mother’s height. Likewise, the rightmost column is the distribution of the mother’s height, regardless of the father’s height. The process of obtaining a one-dimensional distribution from a multidimensional illustrates the *marginalization* over variables that are not of interest. In the case of the bottom column, the marginalization of the distribution was done over the mother’s height, to obtain the distribution of father’s height. The sample marginal distributions obtained from Pearson’s data are shown in Fig. 2.3.

Basic quantities such as the sample mean, variance, and covariance require a complete list of all measurements. Pearson did not report these *raw* (i.e., unprocessed) data, and chose instead to summarize the 1,079 measurements in a table where several measurements were combined or *binned* into cells of size 1×1 square inches (Table 2.3). The number in each cell, when normalized by the total number of cells, represents the probability that the heights fall into that range. For example, the measurements indicate that there is a probability of 28/1079 that the father’s and mother’s heights fall, respectively, in the range 67–68 inches and 62–63 inches; or a probability of 140.5/1079 that the father’s height is in the 67–68 inches range, regardless of the mother’s height. The sample distributions provided by Table 2.3 can therefore still be used to estimate sample moments. For example, the sample mean of the father’s height can be evaluated as

$$\bar{x} = \sum_{i=1}^{18} x_i f_F(x_i) = 67.7$$

where $f_F(x_i)$ is the marginal sample distribution of the father’s heights, with 18 distinct intervals represented by the bottom row of Table 2.3. The sample mean of the mother’s

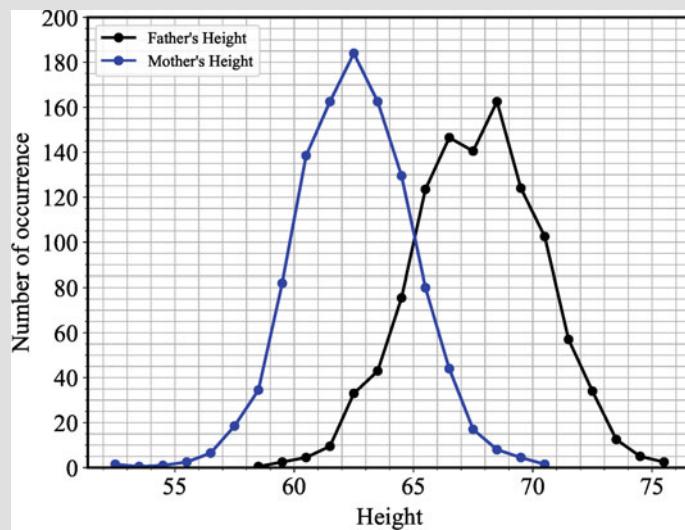


Fig. 2.3 Marginal distributions of the father's and mother's heights. These sample distributions correspond to the totals in the bottom row and in the rightmost column of Table 2.3. They can be normalized to obtain the sample probability distribution functions (i.e., $f_F(x)$ for father's height and $f_M(x)$ for mother's height) simply by dividing each number by the total number of entries ($N = 1079$), since the width of each height interval has unit size

height can be calculated with an equivalent formula,

$$\bar{y} = \sum_{i=1}^{19} y_i f_M(y_i),$$

where $f_M(y_i)$ is the marginal sample distribution of the mother's heights, with 19 distinct intervals represented by the right column of Table 2.3 (see Problem 2.7).

Table 2.3 Joint distribution of father's height (columns) and mother's height (rows) from Pearson's experiment, measured in inches

Father's height		58–	59–	60–	61–	62–	63–	64–	65–	66–	67–	68–	69–	70–	71–	72–	73–	74–	75–
Mother's height		0	0	0	0	0	0	0	1	0.5	0	0	0	0	0	0	0	0	1.5
52–53	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.5
53–54	0	0	0	0	0	0	0	0	0.5	0	0	0	0	0	0	0	0	0	0.5
54–55	0	0	0	0	0	0	0	0	0.25	0.25	0	0.5	0	0	0	0	0	0	1
55–56	0	0	0	0.5	1	0	0	0	0.25	0.25	0	0.5	0	0	0	0	0	0	2.5
56–57	0	0	0	0	0.75	1.25	0	1	1.75	1.75	0	0	0	0	0	0	0	0	6.5
57–58	0	0	0	0.25	1	1.25	1.5	4	3.25	2.5	3	1.25	0.5	0	0	0	0	0	18.5
58–59	0	0.25	0.75	1.25	1.25	2.75	4	7	5.75	4.5	3.75	1.25	2	0	0	0	0	0	34.5
59–60	0	1.25	1.25	1	4	4.5	7.75	10	15	16.75	9	5.5	3.25	1.25	1	0.5	0	0	82
60–61	0.25	0.25	0.5	2	4.25	4.5	18	16	24	14.75	23.25	12.75	7.25	5.75	4.25	0.75	0	0	138.5
61–62	0.25	0.25	0	0	8	8.25	15	17.25	25	20.75	24	14.25	14.25	10	4	0.75	0.5	0	162.5
62–63	0	0.5	0.5	1.25	4.75	7.75	10	26	21.25	28	28	23	14.25	10.75	4.5	2	1	0.5	184
63–64	0	0	0.25	2	3.5	4.5	9	21	15.75	20.75	19.5	24	22.5	10.75	4	2.25	2.25	0.5	162.5
64–65	0	0	1.25	0.75	2	6	6.5	9.75	16	18.25	23	16.75	13.75	6.75	4.75	2.25	0.25	1.5	129.5
65–66	0	0	0	0.25	1.5	1.5	3.25	5.5	9.75	7	15.5	12.75	10.5	6.25	4.25	1.75	0.25	0	80
66–67	0	0	0	0.25	1	0.75	0.5	3.5	5	3	7.25	7.75	7	3.5	2.75	1.5	0.25	0	44
67–68	0	0	0	0	0	0	0	1	2.5	1.5	2.75	3.25	2.75	1.5	1	0.5	0.25	0	17
68–69	0	0	0	0	0	0	0	0	0	1	2.5	1.25	0.5	1	0.25	0.25	0	0	8
69–70	0	0	0	0	0	0	0	0	0	0	0.25	2.25	0	2	0	0	0	0.5	4.5
70–71	0	0	0	0	0	0	0	0	0	0	0	1	0	0.5	0	0	0	0	1.5
	0.5	2.5	4.5	9.5	33	43	75.5	123.5	146.5	140.5	162.5	124	102.5	57	34	12.5	5	2.5	1079

Summary of Key Concepts for this Chapter

Random variable: A quantity that is not known exactly and is described by a probability distribution function $f(x)$.

Moments of a distribution: Expectations for the random variable or functions of the random variable, such as the mean $\mu = E[X]$ and the variance $\text{Var}(X) = \sigma^2 = E[(X - \mu)^2]$.

Sample mean and sample variance: Random variables calculated from independent measurements that are intended to approximate the corresponding parent quantities.

Law of Large Numbers: The sum of a large number of random variables with mean μ tends to a constant number equal to μ .

Joint distribution function: The distribution of probabilities for a pair or more variables.

Covariance: A measure of the tendency of two variables to follow one another, $\text{Cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$.

Correlation coefficient: A normalized version of the covariance that takes values between -1 (perfect anti-correlation) and $+1$ (perfect correlation).

Statistically independent variables: Two (or more) variables whose joint probability distribution function can be factored as the product of individual distributions.

Unbiased estimator: A random variable constructed from measurements and whose expectation is equal to the expectation of the parent variable.

Problems

2.1 Consider the exponential distribution

$$f(x) = \lambda e^{-\lambda x},$$

where $\lambda > 0$ and $x \geq 0$. Show that the distribution is properly normalized, and calculate the mean, variance, and cumulative distribution $F(x)$.

2.2 Consider the sample mean as a random variable defined by

$$\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i, \quad (2.26)$$

where x_i are identical independent random variables with mean μ and variance σ^2 . Show that the variance of \bar{x} is equal to σ^2/N .

2.3 ■ J.J. Thomson's experiment aimed at the measurement of the ratio between the mass and charge of the electron is presented on Sect. 2.4. Using the datasets for Tube 1 and Tube 2 separately, calculate the sample means and variances of the random variables W/Q and I , and the covariance and correlation coefficients between W/Q and I .

2.4 ■ Using J.J. Thomson's experiment (Sect. 2.4), verify the statement that

It will be seen from these tables that the value of m/e is independent of the nature of the gas.

You may do so by calculating the sample mean and standard deviation for the measurements in each gas (air, hydrogen, and carbonic acid), then testing whether the three measurements agree with each other within their standard deviations.

2.5 Calculate the sample covariance and correlation coefficient for the following set of data: $(0, 2), (2, 5), (1, 4), (3, 1)$.

2.6 For a random variable X with mean μ , prove that the following relationship holds:

$$\text{Var}(X) = E[X^2] - \mu^2.$$

2.7 ■ Using the Pearson experiment data in Table 2.3, calculate the sample mean of the mother's height from the $N = 1079$ measurements.

Chapter 3

Three Fundamental Distributions: Binomial, Gaussian, and Poisson



Abstract There are three distributions that play a fundamental role in statistics. The binomial distribution is used to describe binary experiments, and it is in many respects the mother distribution from which the other two distributions can be obtained. The Gaussian or *normal* distribution can be considered as a special case of the binomial when the number of tries of the experiment is sufficiently large. Its central role in probability and statistics is due to the central limit theorem, which is presented in Chap. 4. The Poisson distribution applies to integer-valued counting experiments, and it can be obtained as a limit of the binomial distribution when the probability of success is small. Although there are several more distributions that are of common use, these three distributions are central to many of the methods presented in this textbook, and they are therefore presented in more detail.

3.1 The Binomial Distribution

A *binary* experiment can only have two possible outcomes which can be interpreted as *success* or *failure*, such as the toss of a coin. Even complex experiments with a larger number of possible outcomes can be described as binary, when one is simply interested about the occurrence of a specific event A , or its non-occurrence, \bar{A} . For example, the roll of a die can be interpreted as a binary experiment if one is, say, interested in the occurrence of the number six, or the occurrence of any other number. It is of fundamental importance in statistics to determine the properties of binary experiments, and the distribution of the number of successes when the experiment is repeated a number of times under the same experimental conditions. A binary random variable is usually referred to as a *Bernoulli* variable.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-981-19-0365-6_3.

3.1.1 Derivation of the Binomial Distribution

Consider a binary experiment characterized by a probability of success p and therefore a probability of failure $q = 1 - p$. The probabilities p and q are determined according to the theory of probability and are assumed to be known for the experiment being considered. When the experiment is repeated N times under the same experimental condition, it is interesting to calculate the probability of having n successes in N tries. The order in which the n successes take place is not interesting; for example, consider tossing a coin four times, and being interested in the probability of exactly any two of these tosses showing heads up. The calculation of the binomial probability follows these steps:

(a) *Probability of an ordered sequence.* The probability of having n successes and therefore $N - n$ failures occurring in a specific order is given by

$$P(\text{specific sequence of } n \text{ successes}) = p^n \times q^{N-n}. \quad (3.1)$$

This result can be seen by using the property of independence among the N events so that the individual probabilities can simply be multiplied.

(b) *Number of ordered sequences, or permutations.* Start by counting how many ordered sequences exist that have n successes out of N tries. To begin, each of the N tries can yield the “first” success, and therefore there are N possibilities for which try is the first success. Continuing on to the “second” success, there are only $N - 1$ possibilities left for what trial will be the second success, and so on. This method of counting sequences labels each success as first, second, etc., and leads to the number of *permutations* of n successes out of N tries,

$$\text{Perm}(n, N) = N \cdot (N - 1) \cdot (N - n + 1) = \frac{N!}{(N - n)!}. \quad (3.2)$$

Example 3.1 Consider the case of $n = 2$ successes out of $N = 4$ trials. According to (3.2), the number of permutations is $4!/2! = 12$. The 12 ordered sequences that give rise to 2 successes out of 4 tries are listed in Table 3.1. Symbol H_1 denotes the “first success” and H_2 the “second success.” Consider, for example, lines 5 and 8: both represent the same situation in which tries 2 and 3 result in success. In reality, they are not different sequences, but simply a result of the method of counting time-ordered sequences. ◇

(c) *Number of unordered sequences, or combinations.* As it is clear from the previous example, the number of permutations is not quite the number sought, since it is of no consequence which success is labeled as first, second, etc. According to (3.2) for the case of $n = N$, there are $n!$ ways of ordering n successes among themselves. Therefore, the number of (time-ordered) permutations needs to be divided by $n!$ in order to avoid the double counting of equivalent sequences. It is therefore clear that the number of *combinations* of n successes out of N trials is

Table 3.1 Illustration of permutations (ordered sequences) of 2 successes out of 4 tries

Sequence	Number of try				Sequence	Number of try			
	1	2	3	4		1	2	3	4
1	H_1	H_2	–	–	7	H_2	–	H_1	–
2	H_1	–	H_2	–	8	–	H_2	H_1	–
3	H_1	–	–	H_2	9	–	–	H_1	H_2
4	H_2	H_1	–	–	10	H_2	–	–	H_1
5	–	H_1	H_2	–	11	–	H_2	–	H_1
6	–	H_1	–	H_2	12	–	–	H_2	H_1

$$C(n, N) = \frac{\text{Perm}(n, N)}{n!} = \frac{N!}{(N-n)!n!} \equiv \binom{N}{n}. \quad (3.3)$$

The number of combinations is the number of possible sequences of n successes in N tries. This number, also called the *binomial coefficient*, is indicated by the symbol in parentheses which usually reads “ N –choose– n .” The binomial coefficient is used in the binomial expansion

$$(p + q)^N = \sum_{n=0}^N \binom{N}{n} p^n q^{N-n}. \quad (3.4)$$

Example 3.2 Continue to consider the case of 2 successes out of 4 trials. There are $2! = 2$ ways to order the 2 successes among themselves (either one or the other is the first success). Therefore, the number of combinations of 2 successes out of 4 trials is 6, and not 12. As indicated above, in fact, each sequence had its “twin” sequence listed separately, and (3.3) correctly counts only different sequences. \diamond

According to the results obtained above, what remains to be done is to use the probability of each sequence (3.1) and multiply it by the number of combinations in (3.3) to obtain the overall probability of having n successes in N trials. This leads to the probability mass function

$$P_N(n) = \binom{N}{n} p^n q^{N-n} \quad n = 0, \dots, N, \quad (3.5)$$

known as the *binomial distribution*. This distribution describes the probability of n successes in N tries of a binary experiment with probability of success p . Equation (3.4) shows that the binomial distribution is properly normalized. Examples of the binomial distribution are shown in Fig. 3.1.

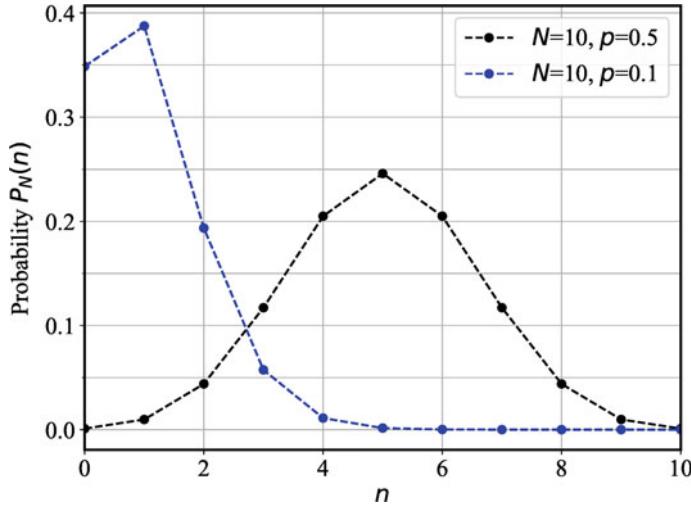


Fig. 3.1 Sample binomial distributions for $N = 10$ and $p = 0.5, p = 0.1$. The function is defined only for non-negative integers $0 \leq n \leq N$

3.1.2 Moments of the Binomial Distribution

The moments of first and second orders of a binomial distributed variable X are given by

$$\begin{cases} E[X] = \mu = pN \\ E[X^2] = \mu^2 + pqN. \end{cases} \quad (3.6)$$

Start with the mean,

$$E[X] = \sum_{n=0}^N n P_N(n) = \sum_{n=0}^N \binom{N}{n} np^n q^{N-n} = \sum_{n=0}^N \binom{N}{n} \left[p \frac{\partial}{\partial p} \right] p^n q^{N-n};$$

the linear operator $p \frac{\partial}{\partial p}$ can be applied to the entire sum, leading to

$$E[X] = p \frac{\partial}{\partial p} \left[\sum_{n=0}^N \binom{N}{n} p^n q^{N-n} \right] = p \frac{\partial}{\partial p} (p+q)^N = pN(p+q)^{N-1} = pN.$$

The derivation for the moment $E[X^2]$ is similar:

$$\begin{aligned} E[X^2] &= \sum_{n=0}^N n^2 P_N(n) = \sum_{n=0}^N \binom{N}{n} n^2 p^n q^{N-n} = \sum_{n=0}^N \binom{N}{n} \left[p \frac{\partial}{\partial p} \right]^2 q^{N-n} \\ &= \left[p \frac{\partial}{\partial p} \right]^2 (p+q)^N = p \frac{\partial}{\partial p} [pN(p+q)^{N-1}] = \\ &p [N(p+q)^{N-1} + pN(N-1)(p+q)^{N-2}] = \\ &pN + p^2 N(N-1) = pN + (pN)^2 - p^2 N = (pN)^2 + pqN. \end{aligned}$$

It follows that the variance of the binomial distribution is given by

$$\sigma^2 = E[(X - E[X])^2] = pqN. \quad (3.7)$$

Equations (3.6) and (3.7) describe the most important features of the binomial distribution, shown in Fig. 3.1 for the case of $N = 10$. The mean is naturally given by the product of the number of tries N and the probability of success p in each of the tries.

Example 3.3 (*Probability of Overbooking*) An airline knows that 5% of the persons making reservations will not show up at the gate. On a given flight that can seat 50 people, 52 tickets have been sold. The binomial distribution can be used to calculate the probability that there will be a seat available for every passenger that will arrive at the gate. This is a binary experiment in which $p = 0.95$ is the probability that a ticketed passenger will show. For a flight with $N = 52$ ticketed passengers, the probability that there is a seat available for each passenger is $P = 1 - P_N(52) - P_N(51)$, which is evaluated as

$$P = 1 - \binom{52}{52} p^{52} \cdot 1 - \binom{52}{51} p^{51} \cdot q = 1 - (0.95)^{52} - 52 \cdot (0.95)^{51} \cdot 0.05 = 0.741.$$

The airline is therefore willing to take a 25.9% chance of having an overbooked flight. \diamond

3.2 The Gaussian Distribution

The Gaussian distribution, often referred to as the *normal* distribution, plays a special role in statistics. A continuous random variable X is said to have a Gaussian distribution if its probability density function is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}. \quad (3.8)$$

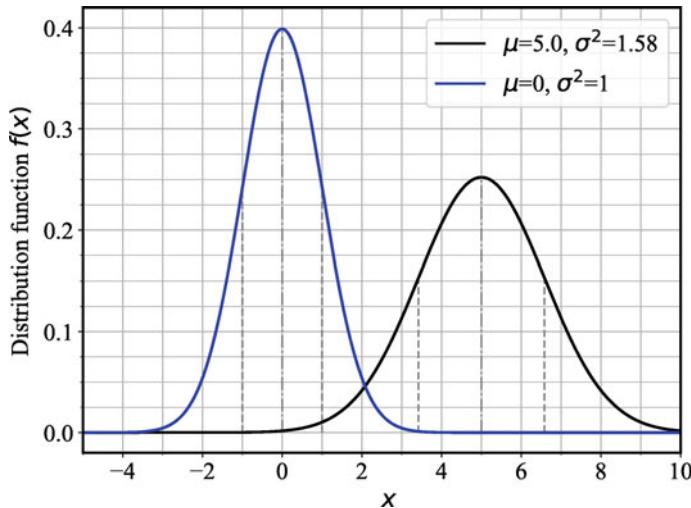


Fig. 3.2 Gaussian distributions for selected values of the mean and variance. The dashed lines around the mean mark the $\pm 1\sigma$ range, which includes 68.3% of the probability. The Gaussian curve in black has the same mean and variance as a binomial with $N = 10$ and $p = 0.5$ (shown in Fig. 3.1)

The distribution function is defined for all real values, and its shape is determined by the two parameters μ and σ^2 , which also represent the mean and variance of the random variable (see Fig. 3.2). A Gaussian of parameters μ and σ^2 is often referred to as $N(\mu, \sigma)$. It is useful to show that the Gaussian distribution can be considered as a special case of the binomial distribution, in the case of a large number of experiments performed.

3.2.1 Derivation of the Gaussian Distribution from the Binomial Distribution

The binomial distribution of (3.5) acquires a simpler form when N is large. An alternative analytic expression to the binomial distribution is a great advantage, given the numerical difficulties associated with the evaluation of the factorial of large numbers. As was evident from Fig. 3.1, the binomial distribution has a maximum at value $n = Np$. This section shows that the binomial distribution can be approximated as

$$P_N(n) \simeq \frac{1}{\sqrt{2\pi Npq}} e^{-\frac{(n-Np)^2}{2Npq}} \quad (3.9)$$

when $N \gg 1$, and for values of the variable that are close to the peak of the distribution.

Start by expanding the logarithm of the binomial probability as a Taylor series in the neighborhood of the peak value \tilde{n} ,

$$\ln P_N(n) = \ln P_N(\tilde{n}) + \sum_{k=1}^{\infty} \frac{B_k}{k!} \Delta n^k,$$

where $\Delta n = n - \tilde{n}$ is the deviation from the peak value and

$$B_k = \left. \frac{\partial \ln P_N(n)^k}{\partial^k n} \right|_{n=\tilde{n}}.$$

Since, by assumption, \tilde{n} is a point of maximum, $B_1 = 0$. Neglecting terms above the second order, one obtains the approximation

$$\ln P_N(n) \simeq \ln P_N(\tilde{n}) + \frac{1}{2} B_2 \Delta n^2,$$

where B_2 is negative, since $n = \tilde{n}$ is a point of maximum. It follows that

$$P_N(n) \simeq P_N(\tilde{n}) e^{-\frac{|B_2|\Delta n^2}{2}}.$$

Neglecting higher-order terms in Δn means that the approximation will be particularly accurate in regions where Δn is small, i.e., near the peak of the distribution. Away from the peak, the approximation will not hold with the same precision. To evaluate $|B_2|$, start with

$$\ln P_N(n) = \ln N! - \ln n! - \ln(N-n)! + n \ln p + (N-n) \ln q$$

and treat n as a continuous variable. This approximation is reasonable when n are large numbers, in particular when the mean $Np \gg 1$ and for values near the mean. The derivative of the logarithm can be approximated as a difference,

$$\frac{\partial \ln n!}{\partial n} = (\ln(n+1)! - \ln n!)/1 = \ln(n+1) \simeq \ln n.$$

From this, it follows that the first derivative of the probability function, as expected, is zero at the peak value,

$$\left. \frac{\partial \ln P_N(n)}{\partial n} \right|_{n=\tilde{n}} = \ln \frac{N-n}{n} + \ln \frac{p}{q} \Big|_{n=\tilde{n}} = \ln \left(\frac{N-n}{n} \frac{p}{q} \right) \Big|_{n=\tilde{n}} = 0$$

so that the familiar result of $\tilde{n} = N p$ is obtained. This leads to the calculation of the second derivative,

$$B_2 = \frac{\partial^2 \ln P_N(n)}{\partial n^2} \Big|_{n=\tilde{n}} = \frac{\partial}{\partial n} \ln \left(\frac{N-n}{n} \frac{p}{q} \right) \Big|_{n=\tilde{n}} = -\frac{1}{N pq}.$$

Finally, the normalization constant $P(\tilde{n})$ can be calculated making use of the integral

$$\int_{-\infty}^{\infty} e^{-ax^2} dx = \sqrt{\pi/a}$$

by enforcing the normalization condition of the probability distribution function,

$$\int_{-\infty}^{\infty} P_N(\tilde{n}) e^{-\frac{|B_2|\Delta n^2}{2}} d\Delta n = P_N(\tilde{n}) \sqrt{\frac{2\pi}{|B_2|}} = 1,$$

leading to

$$P_N(\tilde{n}) = \frac{1}{\sqrt{2\pi Npq}}.$$

The approximation of the binomial distribution for large values of n is therefore:

$$P_N(n) \simeq \frac{1}{\sqrt{2\pi Npq}} e^{-\frac{(n-Np)^2}{2Npq}}.$$

The mean of a binomial distribution is $\mu = Np$ and the variance is $\sigma^2 = Npq$. The approximation of the binomial for large values of n is therefore

$$P_N(n) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(n-\mu)^2}{2\sigma^2}}, \quad (3.10)$$

which is the standard form of the Gaussian distribution when n is a continuous variable.

3.2.2 Moments and Properties of the Gaussian Distribution

The parameters μ and σ^2 are, respectively, the mean and variance of the Gaussian distribution. These results follow from the derivation of the Gaussian distribution from the binomial and can be confirmed by direct calculation of expectations from (3.8). It can also be proven that central moments of odd order are zero since the Gaussian is symmetric with respect to the mean. Given its wide use in statistics, it is important to quantify the “effective width” of the Gaussian distribution around its mean. The probability that a Gaussian variable has values in a range of $\pm z\sigma$ is

Table 3.2 Probability associated with characteristic intervals of a Gaussian distribution

Interval around mean	Integrated probability (%)
$\pm 1\sigma$	68.27
$\pm 2\sigma$	95.45
$\pm 3\sigma$	99.73
$\pm 4\sigma$	99.99
$\pm 5\sigma$	≥ 99.9999
FWHM (or $\pm 1.18\sigma$)	76.10

$$A(z) = \int_{\mu-z\sigma}^{\mu+z\sigma} f(x)dx = \frac{1}{\sqrt{2\pi}} \int_{-z}^z e^{-\frac{t^2}{2}} dt, \quad (3.11)$$

where $f(x)$ is the probability density function of a Gaussian with mean μ and variance σ^2 . The probability that the variable is within $\pm 1\sigma$ of the mean is $A(1) = 0.683$, or 68.3%. This range of the variable is also referred to as a *confidence interval* with 68.3% probability (confidence intervals will be studied in detail in Chap. 7). The correspondence between the $\pm 1\sigma$ interval and the range that encompasses 68.3% of the probability applies strictly only to the Gaussian distribution, for which the value of σ is defined via the distribution function. It is common practice, however, to calculate to the 68.3% interval (sometimes shortened to 68%) even for those random variables that do not strictly follow a Gaussian distribution, and refer to it as the 1σ interval. The probability associated with characteristic intervals of a Gaussian variable are reported in Table 3.2.

All Gaussian distributions can be obtained from the standard $N(0, 1)$ via a simple change of variable. If X is a random variable distributed like $N(\mu, \sigma)$ and Z a standard Gaussian $N(0, 1)$, then the relationship between Z and X is given by

$$Z = \frac{X - \mu}{\sigma} \quad (\text{or } X = \mu + \sigma \cdot Z). \quad (3.12)$$

The variable Z is also referred to as the *z-score* associated with the variable X . This equation also means that samples from a standard normal can be used to generate samples from any other Gaussian distribution.

The notation $X \sim N(\mu, \sigma^2)$ indicates that the random variable X is distributed as a Gaussian variable with mean μ and variance σ^2 . The result that $Z \sim N(0, 1)$ can be obtained by a simple change of variables $z = (x - \mu)/\sigma$ with $dx = \sigma dz$, which yields

$$f(x)dx = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = g(z)dz,$$

where $f(x)$ is the probability distribution function of an $N(\mu, \sigma^2)$ variable, and

$$g(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

is the distribution for the standard normal $N(0, 1)$. This simple observation is sufficient to show that the z -score is Gaussian-distributed, and that its mean is null and its variance one. Alternatively, the mean and variance of the z -score could be calculated from the properties of X , using the basic properties of expectations:

$$\mathbb{E}[Z] = \mathbb{E}\left[\frac{X}{\sigma}\right] - \frac{\mu}{\sigma} = 0,$$

and

$$\text{Var}(Z) = \frac{1}{\sigma^2} \text{Var}(X) = 1.$$

The cumulative distribution of a normal variable $N(0, 1)$ is defined by the following integral:

$$B(z) = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt = \frac{1}{2} + \frac{A(z)}{2} \quad (3.13)$$

and it describes the integrated probability up to the value of z , with $A(z)$ according to (3.11) the integrated probability between $\pm z$.

The *half width at half maximum*, or HWHM, is defined as the distance between the point of maximum of the probability distribution function (at $x = \mu$) and the point where the distribution reaches a value of one half the peak. It can be easily shown that the HWHM is given by

$$\text{HWHM} = \sqrt{2 \ln 2} \sigma \simeq 1.18\sigma,$$

meaning that the half-maximum point is just past one standard deviation of the mean, on either side of the mean. By the same token, the *full width at half maximum*, or FWHM, is defined as the full range between the two points of half maximum, and it is $\text{FWHM} \simeq 2.36\sigma$. Tables of the Gaussian distributions are provided in Appendix A.1.

3.3 The Poisson Distribution

The Poisson distribution describes the probability of occurrence of events in counting experiments when the possible outcome is an integer number. The distribution is therefore discrete and can be derived as a limiting case of the binomial distribution.

3.3.1 Derivation of the Poisson Distribution

The binomial distribution has another useful approximation when the probability of success is small, $p \ll 1$. In this case, the number of positive outcomes is much smaller than the number of tries, $n \ll N$, and the factorial function can be approximated as

$$N! = N(N - 1) \cdots (N - n + 1) \cdot (N - n)! \simeq N^n(N - n)!.$$

The term q^{N-n} can be approximated using

$$\ln q^{N-n} = \ln(1 - p)^{N-n} = (N - n) \ln(1 - p) \simeq -p(N - n) \simeq -pN,$$

leading to

$$q^{N-n} \simeq e^{-pN}.$$

These two approximations can be used into (3.5) to give

$$P_N(n) \simeq \frac{N^n(N - n)!}{n!(N - n)!} p^n e^{-pN} = \frac{(pN)^n}{n!} e^{-pN}. \quad (3.14)$$

Since pN is the mean of the distribution, the approximation becomes approximation as

$$P(n) = \frac{\mu^n}{n!} e^{-\mu}, \quad (3.15)$$

known as the Poisson distribution. This function describes the probability of obtaining n positive outcomes, or counts, when the expected number of outcomes is μ . It can be immediately seen that the distribution is properly normalized, since

$$\sum_{n=0}^{\infty} \frac{\mu^n}{n!} = e^{\mu}.$$

A fundamental feature of this distribution is that it is described by only one parameter, the mean μ , as opposed to the Gaussian distribution that had two parameters. This clearly does not mean that the Poisson distribution has no variance—in that case, it would not be a random variable—but that the variance can be written as a function of the mean, as will be shown in the following.

3.3.2 Moments and Properties of the Poisson Distribution

Equation (3.15) has lost its reference to the binomial experiment, and only the mean $\mu = Np$ is left as a parameter. Using the definition of mean and variance, it is easy

to prove that a random variable X that follows a Poisson distribution has an expected value $E[X] = \mu$ and a variance $\text{Var}(X) = \mu$. The fact that the mean equals the variance can be seen using the values for the binomial, $\mu = Np$ and $\sigma^2 = Npq$; since $p \ll 1, q \simeq 1$, and $\mu \simeq \sigma^2$. As a result, the Poisson distribution has only one parameter which equals both the mean and the variance.

The mean and variance of a Poisson distribution can also be evaluated directly from the probability mass function. The mean can be evaluated as follows:

$$E[X] = e^{-\mu} \sum_{n=0}^{\infty} n \frac{\mu^n}{n!} = e^{-\mu} \left[\mu \frac{d}{d\mu} \right] \left(\sum_{n=0}^{\infty} \frac{\mu^n}{n!} \right) = \mu.$$

The moment of second order is

$$E[X^2] = e^{-\mu} \sum_{n=0}^{\infty} n^2 \frac{\mu^n}{n!} = e^{-\mu} \left[\mu \frac{d}{d\mu} \right] \left(\sum_{n=0}^{\infty} n \frac{\mu^n}{n!} \right) = e^{-\mu} \mu \frac{d}{d\mu} (\mu e^{\mu}) = \mu + \mu^2,$$

where the equation for the mean was used in the derivation. The variance is therefore

$$\text{Var}(X) = E[X^2] - E[X]^2 = \mu.$$

The Poisson distribution is interpreted as the probability of occurrence of n counts when the parent mean of the counts is μ . This makes the Poisson distribution the primary statistical tool for all *counting* experiments. Unlike the binomial distribution, which is limited to integer values $n \leq N$, the Poisson distribution is defined for any non-negative integer. The reference to the total number of possible events (N) and the probability of occurrence of each event (p) was lost, and only the mean μ remains to describe the primary property of the counting experiment.

As can be seen in Fig. 3.3, the Poisson distribution is not symmetric with respect to the mean, and the distribution becomes more symmetric for larger values of the mean. As for all discrete distributions, it is only meaningful to calculate the probability at a specific point or for a set of points, and not for an interval of points as in the case of continuous distributions. Moreover, the mean of the distribution itself can be a non-integer number, and still the outcome of the experiment described by the Poisson distribution can only take integer values.

Example 3.4 Consider an astronomical source known to produce photons, which are usually detected by a given detector in the amount of $\mu = 2.5$ in a given time interval. The probability of detecting $n = 4$ photons in a given time interval is therefore

$$P(4) = \frac{2.5^4}{4!} e^{-2.5} = 0.134.$$

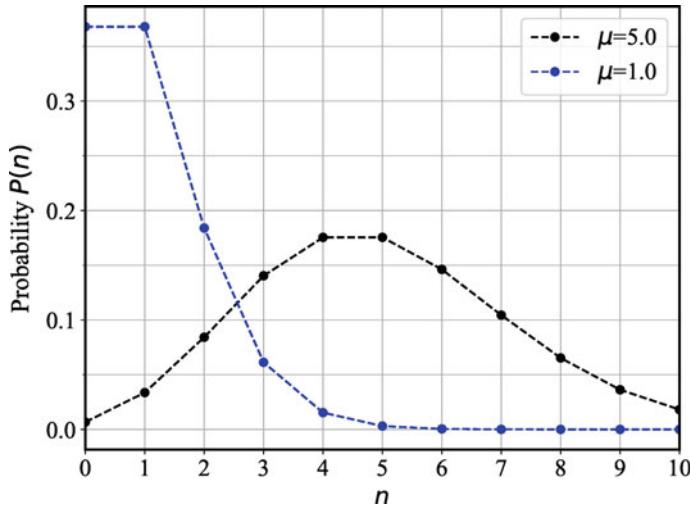


Fig. 3.3 Sample Poisson distributions for selected values of the mean. The Poisson probability mass function is defined for all values $n \geq 0$

The reason for such an apparently large probability of obtaining a measurement that differs from the expected mean is simply due to the statistical nature of the detection process. \diamond

3.3.3 The Poisson Distribution and the Poisson Process

A more formal justification for the interpretation of the Poisson distribution as the distribution of counting experiments comes from the Poisson process. Although a complete treatment of this subject is beyond the scope of this book, a short description of stochastic processes will serve to strengthen the interpretation of (3.15), which is one of the foundations of statistics. More details on stochastic processes can be found, for example, in the textbook by Ross [86].

A *stochastic counting process* $\{N(t), t > 0\}$ is a sequence of random variables $N(t)$, in which t indicates time, and $N(t)$ is a random variable that indicates the number of events that have occurred up to time t . The stochastic process can be thought of as repeating the experiment of “counting the occurrence of a given event” at various times t and $N(t)$ is the result of the experiment. The *Poisson process with rate λ* is a particular type of stochastic process, with the following properties:

1. $N(0) = 0$, meaning that at time 0 there are no counts detected.
2. The process has *independent increments*, meaning that $N(s+t) - N(s)$ is independent of $N(s)$. This property means that events occurring after time s are not being influenced by those occurring prior to it.
3. The process has *stationary increments*, i.e., the distribution of the number of events in an interval of time s depends only on the length of the time interval itself.
4. $P(N(h) = 1) = \lambda h + \mathcal{O}(h)$, in which $\mathcal{O}(h)$ is a function with the property that

$$\lim_{h \rightarrow 0} \frac{\mathcal{O}(h)}{h} = 0.$$

5. $P(N(h) \geq 2) = \mathcal{O}(h)$. The latter two properties mean that the probability of obtaining one count depends on the finite value λ , while it is unlikely that two or more events occur in a short time interval.

It can be shown that under these hypotheses, the number of events $N(t)$ recorded in any interval of length t is Poisson distributed,

$$P\{N(s+t) - N(s) = n\} = \frac{(\lambda t)^n}{n!} e^{-\lambda t}. \quad (3.16)$$

This shows that the Poisson distribution is to be interpreted as the distribution of occurrence of n events during a time interval t , under the hypothesis that the rate of occurrence of events is λ . This interpretation is identical to the one provided above, given that $\mu = \lambda t$ is the mean of the counts in that time interval.

3.4 Comparison of the Binomial, Gaussian, and Poisson Distributions

A comparison between the binomial, Gaussian, and Poisson distributions with the same mean is illustrated in Fig. 3.4. The mean and variance of a binomial distribution are determined by the choice of the number of tries N and the probability of success p of the binary experiment, with $\mu = Np$ and $\sigma^2 = Npq$. In the case of the Gaussian distribution, the mean and variance can be chosen independently as the two parameters of the probability distribution function. The Poisson distribution, on the other hand, has the special feature of having the same value of the mean and variance, in the amount of the only parameter μ in the distribution. The two sets of curves in Fig. 3.4 share the same mean, but it is not possible to simultaneously enforce also the same variance. For example, the Poisson distribution with $\mu = 5$ has a larger variance than the binomial with $N = 10$ and $p = 0.5$, which has a variance of $\sigma^2 = 2.5$.

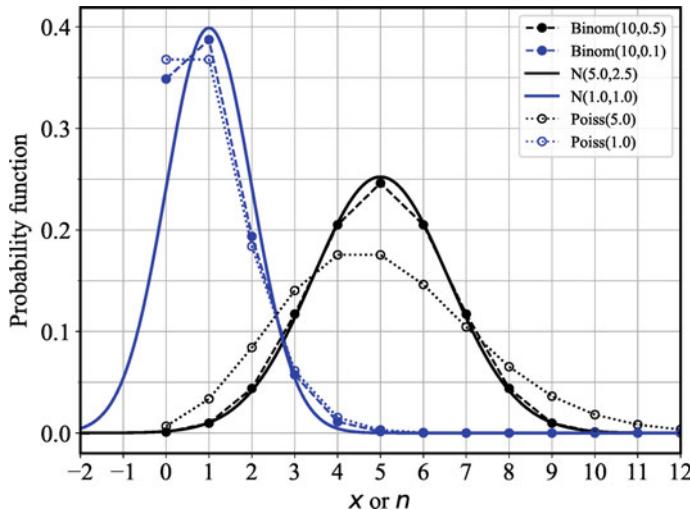


Fig. 3.4 Comparison of binomial, Gaussian, and Poisson probability distributions. The curves on the left are for random variables with a parent mean of $\mu = 1$ and to the right for $\mu = 5$

It is of practical importance to discuss the approximation of the Poisson distribution with a normal distribution of same mean and variance. For small values of the mean, the Poisson distribution is asymmetric around the mean, as illustrated in Fig. 3.4. As the mean increases, the distribution becomes progressively more symmetrical and the confidence intervals around the mean approach those of a Gaussian distribution, as illustrated in Fig. 3.5. Confidence intervals are calculated as intervals around the mean that encompass a given probability. For the Gaussian distribution, the confidence intervals are always symmetrical around the mean. For small values of the Poisson mean, the confidence intervals are shifted towards larger values of the random variable, compared to the Gaussian, and they become progressively closer to the Gaussian intervals as the mean increases. Given that the Poisson distribution is integer-valued, the jagged shape of the Poisson confidence interval curves (in red in Fig. 3.5) is explained by the requirement that the intervals are also bounded by integers. For example, the range of a Poisson variable with $\mu = 1$ that encompasses 90% of the probability is 0 to 3, while for a Gaussian distribution of same mean (and variance equal to the mean) is -0.645 to $+2.645$, or 1 ± 1.645 , see Table A.4.

The approach towards a normal distribution is also shown by the solid curve, indicating the fractional difference of the Poisson and normal probability distributions at the peak value of $x = \mu$, as an indication of the relative shape of the two curves. As a result, the approximation of a Poisson distribution with a Gaussian of same mean and variance becomes increasingly accurate as the mean becomes larger.

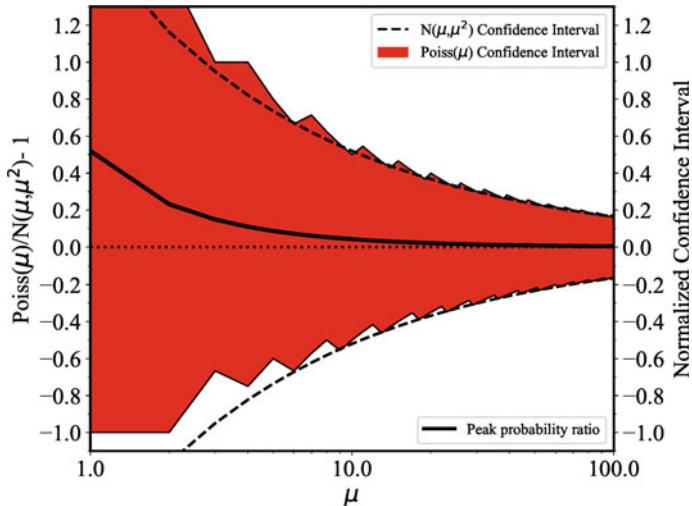


Fig. 3.5 Left vertical axis: fractional difference between the Poisson probability $P(\mu)$ and normal probability $f(\mu)$ at the peak of the distribution, $x = \mu$, shown as a solid curve. Right vertical axis: 90% confidence intervals around the mean μ for the two distributions, normalized by the value of μ . The x axis has a logarithmic scale to emphasize the behavior at low values of the parent mean

For $\mu \geq 10$, the differences in peak probability and confidence intervals are of order 10% or less, and for $\mu \geq 20$, they become of order 1%. For most applications, it is therefore reasonable to make the approximation between the two distributions when the mean exceeds a value in the range $\mu = 10 - 20$.

A quantitative measure of the symmetry of a distribution around the mean is provided by the *skewness*, which is the central moment of the third order,

$$skew(X) = E\left[\left(\frac{X - \mu}{\sigma}\right)^3\right].$$

For distributions that are symmetrical around the mean, such as the Gaussian, it is immediate to see that the skewness must be null. In fact, basic properties of the expectation and symmetry around the mean imply that $E[(X - \mu)^3] = E[(\mu - X)^3] = -E[(X - \mu)^3]$, thus this expectation must be null. In general, the skewness can be written as

$$skew(X) = \frac{E[X^3 - 3X^2\mu - 3X\mu^2 - \mu^3]}{\sigma^2} = \frac{E[X^3] - 3\mu\sigma^2 - \mu^3}{\sigma^3},$$

where σ^2 is the variance of the distribution. For a Poisson distribution, the moment of the third order can be calculated as

$$\mathbb{E}[X^3] = \mu^3 + 3\mu^2 + \mu,$$

following similar calculations to those in Sect. 3.3. It follows that the skewness of the Poisson distribution is

$$skew(X) = \frac{1}{\sqrt{\mu}},$$

and as the mean of the distribution increases, the Poisson distribution becomes increasingly symmetrical.

An additional consideration for the Poisson distribution is the presence of the factorial function, which becomes very rapidly a large number as the function of the integer n . For large values of n , one can use *Stirling's approximation* to the factorial function, which retains only the first term of the following expansion:

$$n! = \sqrt{2\pi n} n^n e^{-n} \left(1 + \frac{1}{12n} + \dots\right). \quad (3.17)$$

The Stirling approximation has a precision of better than 1% for $n \geq 10$.

Summary of Key Concepts for this Chapter

Binomial distribution: It describes the probability of occurrence of n successes in N tries of a binary event,

$$P_N(n) = \binom{N}{n} p^n q^{N-n}$$

(mean pN and variance pqN).

Gaussian or normal distribution: It is an approximation of the binomial distribution when N is large,

$$f(x)dx = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

(mean μ and variance σ^2).

Poisson distribution: It is an approximation of the binomial distribution when $p \ll 1$ that describes the probability of counting experiments,

$$P(n) = \frac{\mu^n}{n!} e^{-\mu}$$

(mean and variance have a value of μ).

Problems

3.1 Consider a random variable X that follows the Gaussian distribution

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

Calculate the mean and variance of X and show that all odd moments $E[(X - \mu)^n]$ of order $n \geq 3$ are zero.

3.2 Following the derivation provided in Sect. 3.2.1, show that the term B_2 is given by

$$B_2 = -\frac{1}{Npq}.$$

3.3 Assume that the results of an I.Q. test follow a Gaussian distribution, and that the scores are standardized in such a way that the mean is $\mu = 100$, and the standard deviation is $\sigma = 15$. Calculate:

- (a) The probability that an I.Q. score is greater or equal to 145.
- (b) The probability that the *mean* I.Q. score of a sample of 100 persons, chosen at random, is equal to or larger than 105.

3.4 A coin is tossed ten times. Find:

- (a) The probability of obtaining 5 heads up and 5 tails up;
- (b) The probability of having the first 5 tosses show heads up, and the final 5 tosses show tails up;
- (c) The probability to have at least 7 heads up.

3.5 In a given course, it is known that 7.3% of students fail.

- (a) What is the expected number of failures in a class of 32 students?
- (b) What is the probability that 5 or more students fail?

3.6 The frequency of twins in the European population is about 12 in every 1000 maternities. Calculate the probability that there are no twins in 200 births, using (a) the binomial distribution and (b) the Poisson distribution.

3.7 For a discrete random variable X that follows the Poisson distribution,

$$P(n) = \frac{\mu^n}{n!} e^{-\mu},$$

show that both the mean and the variance are given by the parameter μ .

3.8 ■ Consider the data from Mendel's experiment of Table 1.1, and refer to the "Long versus short stem" measurements.

- Use theoretical arguments to identify the parent distribution for the number of dominants X , i.e., the distribution $P(n)$ that describes the probability that $X = n$ plants display the dominant character and the remainder the recessive character.
- Calculate the uncertainty in the measurement of the number of plants that display the dominant character, based on the distribution from part (a).
- Determine the difference between the number of measured plants with the dominant character and the expected number, the latter based on the distribution of X from part (a). Normalize the difference by the standard deviation of the number of dominants,

$$z = \frac{x - E[X]}{\sigma}$$

to show that this number has an absolute value of less than one.

3.9 ■ For Mendel's experimental data in Table 1.1, consider the overall fraction of plants that display the dominant character, obtained as the combination of all seven experiments together.

- Determine the parent distribution of the overall fraction Y of plants with dominant character and its expected value.
- Calculate the measurement y of the fraction Y ;
- Using the parent variance σ^2 of the variable Y , determine the value

$$z = \frac{y - E[Y]}{\sigma}$$

which is the standardized difference between the measurement and the mean. Assuming that the binomial distribution can be approximated by a Gaussian of same mean and variance, calculate the probability of having a value of z equal to or smaller (in absolute value) than the measured value.

Chapter 4

The Distribution of Functions of Random Variables



Abstract Experiments do not always measure directly all quantities of interest to the analyst. Instead, it is sometimes necessary to infer properties of interesting variables based on the variables that have been measured directly. This chapter explains how to determine the probability distribution function of a variable that is the function of other variables of known distribution. The central limit theorem, one of the statistic's key tools, also establishes that the sum of a large number of independent variables is asymptotically distributed like a Gaussian distribution.

4.1 Functions of Random Variables

Experiments are typically designed to target one or more quantities that can be measured effectively with the available equipment. An example is the measurement of high-energy particles emitted by a radioactive source using a *Geiger* counter that detects all high-energy particles that strike the counter. Knowledge of both the instrument and the source of radiation can be used to assume that the number of counts detected in a given time interval by the counter follows a specific distribution, for example, a Poisson with an unknown mean. The Geiger counter, however, does not measure directly the number of particles from the source, but rather a combination of source-emitted particles and other particles that happen to reach the equipment, usually referred to as background. Given this limitation, another experiment might be performed to measure just the rate of counts from the background, for example, by removing the source of radiation from the field of view of the detector. The analyst now has two random variables, namely the combined source plus background counts and the background counts alone, each following a specific probability distribution.

Inferences on the quantity of interest, in this example, the rate arrival of source particles alone must be obtained from the measured variables. In some experiments, it is possible to know the probability distribution of the quantity of interest, based on the probability distribution of the other measured quantities: this is an ideal situation

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-981-19-0365-6_4.

that makes the estimation of the parameters of the distribution easier and more accurate. In the Geiger counter experiment, it can be shown that the difference between two Poisson-distributed variables is not distributed like a Poisson distribution, but if the distributions were normal instead, their difference would retain the normal distribution. The study of the distribution of functions of random variables is a complex topic that is covered exhaustively in textbooks on probability theory such as [86]. This chapter covers selected topics and methods that are applicable to typical situations encountered by the data analyst. There are also cases when it is not possible or practical to seek the probability distribution of a random variable of interest. In these cases, it is still possible to pursue approximate methods to study its mean and variance. Some of these approximate methods are introduced in this chapter and then developed in full in the following chapter.

4.2 Linear Combination of Random Variables

Experimental variables are often related by a simple linear relationship. Prior to studying the distribution of more complex functions of random variables, it is useful to understand the behavior of the linear combination of variables. The linear combination of N random variables X_i is a variable Y defined by

$$Y = \sum_{i=1}^N a_i X_i, \quad (4.1)$$

where a_i are constant coefficients. A typical example is the signal detected by an instrument, which can be thought of as the sum of the intrinsic signal from the source plus the background. The distributions of the background and the source signals will influence the properties of the total signal detected, and it is therefore important to understand the statistical properties of this relationship in order to characterize the signal from the source.

4.2.1 Mean and Variance Formulas

The mean of the linear combination of random variables can be easily calculated using the properties of the expectation. The expectation of a variable Y defined according to (4.1) is

$$\mathbb{E}[Y] = \sum_{i=1}^N a_i \mu_i, \quad (4.2)$$

where μ_i is the mean or expectation of X_i . This property follows from the linearity of the expectation operator, and it is equivalent to a weighted mean where the weights are given by the coefficients a_i . This linear property of the mean applies regardless of whether the random variables X_i are independent of one another, as shown in Sect. 2.5.1.

In the case of the variance, the situation is more complex. The variance of Y is given by the following equation:

$$\begin{aligned}\text{Var}(Y) &= \mathbb{E} \left[\left(\sum_{i=1}^N a_i X_i - \sum_{i=1}^N a_i \mu_i \right)^2 \right] = \sum_{i=1}^N a_i^2 \mathbb{E}[(X_i - \mu_i)^2] \\ &\quad + 2 \sum_{i=1}^N \sum_{j>i}^N a_i a_j \mathbb{E}[(X_i - \mu_i)(X_j - \mu_j)],\end{aligned}$$

where the cross-product terms are proportional to the covariance between the variables. The general formula for the variance of the linear sum of variables is therefore

$$\text{Var}(Y) = \sum_{i=1}^N a_i^2 \text{Var}(X_i) + 2 \sum_{i=1}^N \sum_{j=i+1}^N a_i a_j \text{Cov}(X_i, X_j). \quad (4.3)$$

Equation (4.3) shows that variances add linearly *only* for variables that are mutually uncorrelated, or $\sigma_{ij}^2 = 0$, but not in general. The following example illustrates the importance of a non-zero covariance between two variables, and its effect on the variance of the sum.

Example 4.1 (*Sum of variables with perfect correlation*) Consider two random variables X and Y and their sum $Z = X + Y$. If X and Y are perfectly anti-correlated, $\text{Cor}(X, Y) = -1$ means that $\sigma_{xy}^2 = -\sigma_x \sigma_y$. The mean of Z is simply the sum of the two means, $\mu_z = \mu_x + \mu_y$, irrespective of the value of the correlation. The variance, however, is

$$\sigma_z^2 = \sigma_x^2 + \sigma_y^2 + 2 \text{Cov}(X, Y) = (\sigma_x - \sigma_y)^2 = 0,$$

meaning that the sum of the two random variables that are perfectly uncorrelated is actually not a random variable anymore, but it is always equal to its mean. This situation is unlikely to occur in experiment, but it shows how a negative correlation is capable of reducing the variance of the sum of two variables.

An opposite situation occurs for the sum of two variables that have a perfect positive correlation. In this case, the expectations continue to add linearly, but the variance becomes

$$\sigma_z^2 = (\sigma_x + \sigma_y)^2,$$

which is always larger than $\sigma_x^2 + \sigma_y^2$. For example, if $X = Y$, then the variance becomes four times the variance of X , which is simply understood with the property $\text{Var}(aX) = a^2 \text{Var}(X)$. \diamond

4.2.2 Independent Measurements and the $1/\sqrt{N}$ Factor

Independent and therefore uncorrelated variables play a special role in probability and statistics. Many experiments are designed to provide N independent measurements of a variable X . The resulting measurements are X_i variables that are said to be independent and identically distributed (*iid*). With N independent measurements X_i of the same variable X , all of equal mean μ and variance σ^2 , one is often interested in calculating the variance of the sample mean. Using basic properties of the expectations and the properties of uncorrelated variables, the variance of the sample mean is calculated as

$$\text{Var}(s) = \frac{1}{N^2} \sum_{i=1}^N \text{Var}(X) = \frac{\sigma^2}{N},$$

showing a reduction of the variance of the sample mean by a factor of N , compared to the parent variance. An equivalent result can be obtained by defining the *relative uncertainty* of a variable as the ratio of the standard deviation to the mean. For the sample mean of N measurements, the relative uncertainty is therefore

$$\frac{\sigma_s}{\mu} = \frac{\sigma}{\mu} \times \frac{1}{\sqrt{N}}.$$

The interpretation of these two equations is simple: one expects a smaller variance between measurements of the sample mean of N measurements than between individual measurements, since the statistical fluctuations of individual measurements have averaged down with increasing sample size of the sample mean. This factor of $1/\sqrt{N}$ is essential to understand the need for repeating experiments in order to achieve a desired goal in the relative uncertainty of a variable of interest. It is important to emphasize that these results only apply to independent measurements.

4.3 The Moment Generating Function

The mean and the variance provide only partial information on the random variable. The moment generating function is a convenient mathematical tool to determine the distribution function of random variables and its moments, and it is also instrumental in proving the central limit theorem, one of the key results of statistics. The *moment generating function* of a random variable X is defined as

$$M(t) = \text{E}[e^{tX}], \quad (4.4)$$

and it has the property that all moments can be derived from it, provided they exist and are finite. The moment generating function introduces a deterministic variable t that is used for the calculation of the moments. Assuming a continuous random

variable of probability distribution function $f(x)$, the moment generating function of X can be written as

$$\begin{aligned} M(t) &= \int_{-\infty}^{+\infty} e^{tx} f(x) dx = \\ &\int_{-\infty}^{+\infty} \left(1 + \frac{tx}{1} + \frac{(tx)^2}{2!} + \dots \right) f(x) dx = 1 + t\mu_1 + \frac{t^2}{2!}\mu_2 + \dots, \end{aligned}$$

where μ_n represent the moment of the n -th order of X . The moments can therefore be obtained as partial derivatives of the moment generating function evaluated at $t = 0$,

$$\mu_r = \left. \frac{\partial^r M(t)}{\partial t^r} \right|_{t=0}. \quad (4.5)$$

4.3.1 Properties of the Moment Generating Function

The most important property of the moment generating function is that it has a one-to-one correspondence with the probability distribution function, i.e., the moment generating function is a sufficient description of the random variable. Some distributions do not have a moment-generating function, since some of their moments may be infinite, so in principle, this method cannot be used for all distributions. Equation 4.5 also applies to discrete distributions; a more complete treatment of the mathematical properties of the moment-generating function can be found in textbooks on theory of probability, such as [86]. Two properties of the moment-generating function will be useful in the determination of the distribution function of random variables:

(a) If $Y = a + bX$, where a, b are constants, the moment-generating function of Y is

$$M_y(t) = e^{at} M_x(bt). \quad (4.6)$$

This relationship can be proved by the use of the expectation operator, according to the definition of the moment-generating function:

$$E[e^{tY}] = E[e^{t(a+bX)}] = E[e^{at} e^{btX}] = e^{at} M_x(bt).$$

(b) If X and Y are independent random variables with $M_x(t)$ and $M_y(t)$ as their moment-generating functions, then the moment-generating function of $Z = X + Y$ is

$$M_z(t) = M_x(t)M_y(t). \quad (4.7)$$

The relationship is derived immediately by

$$E[e^{tZ}] = E[e^{t(X+Y)}] = M_x(t)M_y(t).$$

4.3.2 Moment-Generating Functions of Selected Distributions

Moment-Generating function of a Gaussian distribution: The moment-generating function of a Gaussian with mean μ and variance σ^2 is given by

$$M(t) = e^{\mu t + \frac{1}{2}\sigma^2 t^2}. \quad (4.8)$$

The mean and the variance appear as a linear combination in the exponent of the Gaussian moment-generating function. This property makes it such the sum of any number of independent Gaussian variables will also be a Gaussian variable with mean equal to the sum of the means, and variance equal to the sum of the variances. The addition of means and variances was already guaranteed by the independence of the variables; the fact that the distribution function remains a Gaussian is a new and stronger property that was uncovered via the analysis of the moment-generating function.

This result can be proven as follows. Start with

$$M(t) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{+\infty} e^{tx} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx.$$

The exponent can be written as

$$\begin{aligned} tx - \frac{1}{2} \frac{x^2 + \mu^2 - 2x\mu}{\sigma^2} &= \frac{2\sigma^2 tx - x^2 - \mu^2 + 2x\mu}{2\sigma^2} \\ &= -\frac{(x - \mu - \sigma^2 t)^2}{2\sigma^2} + \frac{2\mu\sigma^2 t}{2\sigma^2} + \frac{\sigma^2 t^2}{2\sigma^2} \sigma^2 t. \end{aligned}$$

It follows that

$$\begin{aligned} M(t) &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{+\infty} e^{\mu t} e^{\frac{\sigma^2 t^2}{2}} e^{-\frac{(x-\mu-\sigma^2 t)^2}{2\sigma^2}} dx \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{\mu t} e^{\frac{\sigma^2 t^2}{2}} \sqrt{2\pi\sigma^2} = e^{\mu t + \frac{\sigma^2 t^2}{2}}. \end{aligned}$$

Moment-Generating function of a Poisson distribution: The moment-generating function of the Poisson distribution is given by

$$M(t) = e^{-\mu} e^{\mu e^t}. \quad (4.9)$$

This result is obtained immediately from the definition of the expectation,

$$M(t) = \mathbb{E}[e^{tX}] = \sum_{n=0}^{\infty} e^{nt} \frac{\mu^n}{n!} e^{-\mu} = e^{-\mu} \sum_{n=0}^{\infty} \frac{(\mu e^t)^n}{n!} = e^{-\mu} e^{\mu e^t}.$$

Similar to the case of a Gaussian distribution, the parameter μ of the Poisson distribution appears linearly in the exponent of the moment-generating function. This property ensures that the sum of independent Poisson random variables is also a Poisson random variable with a mean equal to the sum of the means.

Moment-generating function of a standard uniform distribution: It is convenient to calculate the moment-generating function of a uniform random variable between 0 and 1, since this distribution is commonly used in statistics. The distribution of a continuous uniform variable X in the interval 0–1 is known as the *standard uniform distribution*. The probability distribution function is $f(x) = 1$ and the cumulative distribution function is $F(x) = x$, both defined only for $0 \leq x \leq 1$ and null otherwise. The mean and variance are $\mu = 1/2$ and $\sigma^2 = 1/12$. The value of the mean is derived immediately from the probability distribution function and the variance can be also calculated directly from (2.12), with $\mathbb{E}[X^2] = 1/3$. The moment-generating function is also obtained immediately from the expectation

$$M(t) = \mathbb{E}[e^{tX}] = \int_0^1 e^{tx} dx = \frac{1}{t}(e^t - 1). \quad (4.10)$$

4.4 The Central Limit Theorem

The *central limit theorem* is one of statistics' most important results. It establishes that the sum of a large number of independent variables has a Gaussian distribution and it provides a simple way to find its mean and variance. This theorem is what gives the Gaussian distribution the nickname of *normal* distribution. This result can be stated as:

Theorem 4.1 (Central Limit Theorem) *The sum of a large number of independent random variables is approximately distributed as a Gaussian. The mean of the distribution is the sum of the means of the variables, and the variance is the sum of the individual variances. This result holds regardless of the distribution of each individual variable, subject however to certain restrictions on the existence of moments.*

To prove this theorem, consider the variable Y as the sum of N variables X_i of mean μ_i and variance σ_i^2

$$Y = \sum_{i=1}^N X_i, \quad (4.11)$$

Since the random variables are independent, it follows immediately that both means and variances add linearly, and therefore, this portion of the central limit theorem is a simple application of the property of independence among the variables. To establish the actual shape of the distribution function of the sum, it is necessary to calculate the moment generating function of the variable Z defined by

$$Z = \frac{Y - \mu}{\sigma} = \frac{1}{\sigma} \sum_{i=1}^N (X_i - \mu_i),$$

where μ and σ^2 are, respectively, the mean and the variance of the sum Y . It is therefore necessary that these two moments exist for the theorem to apply. The variable Z is a scaled and standardized version of Y so that it has a mean of zero and unit variance. The goal is to show that Z can be approximated by a standard Gaussian. Define $M_i(t)$ as the moment generating function of the random variable $(X_i - \mu_i)$. The moment-generating function of Z is therefore

$$M_z(t) = \prod_{i=1}^N M_i(t/\sigma),$$

following the independence among X_i and property (4.6). The function $M_i(t/\sigma)$ represents the moment-generating function of $(X_i - \mu_i)/\sigma$, and it is given by

$$M_i(t/\sigma) = 1 + \mu_{x_i - \mu_i} \frac{t}{\sigma} + \frac{\sigma_i^2}{2} \left(\frac{t}{\sigma} \right)^2 + \frac{\mu_{x_i - \mu_i, 3}}{3!} \left(\frac{t}{\sigma} \right)^3 + \dots,$$

where $\mu_{x_i - \mu_i} = 0$ is the mean of $X_i - \mu_i$, σ_i^2 is the variance (or central moment of the second order) of X_i , and $\mu_{x_i - \mu_i, 3}$ is the central moment of third order of X_i . If there is a large number of random variables, the variance σ^2 of the sum is large and terms of order σ^{-3} can be ignored. It is therefore possible to make the approximation

$$\ln M_z(t) \simeq \sum \ln \left(1 + \frac{\sigma_i^2}{2} \left(\frac{t}{\sigma} \right)^2 \right) \simeq \sum \frac{\sigma_i^2}{2} \left(\frac{t}{\sigma} \right)^2 = \frac{1}{2} t^2.$$

The second approximation applies for small values of the variable t , which is appropriate since moments are derived as partial derivatives of the moment-generating function evaluated at $t = 0$. This results in the approximation of the moment-generating function of Z as

$$M_z(t) \simeq e^{\frac{t^2}{2}},$$

proving that Z is approximately distributed as a standard normal distribution, according to (4.8). A simple change of variable also shows that Y is distributed as a Gaussian with mean μ and variance σ^2 . This derivation follows closely the one provided in [16].

The central limit theorem establishes that the limiting distribution of the sum of a large number of random variables is Gaussian, no matter their original distributions but provided that they have certain finite moments. Fortunately, most distributions follow these requirements, and therefore the theorem applies to such distributions as the binomial, Poisson, exponential, normal, uniform, and many other distributions of common occurrence in statistics. The variables being added don't have to all have the same distribution, so it is possible to apply the central limit theorem to variables with different distributions. Two common cases are treated explicitly in the following, starting with the distribution of the sample mean of Gaussian measurements. Another illustrative case is the sum of uniform distributions. Although each uniform distribution is not centrally peaked, the sum rapidly approaches a Gaussian distribution when an increasing number of variables are added.

4.4.1 *The Distribution of the Sample Mean of Gaussian Measurements*

The sample mean \bar{x} of N normally-distributed measurements x_i is defined according to the usual Eq. (2.8). Since the sample mean is a linear combination of Gaussian variables, the properties of the moment-generating function described in Sect. 4.3.2 already guarantee that the sample mean is normally-distributed according to

$$\bar{x} \sim N(\mu, \sigma^2/N), \quad (4.12)$$

where μ is the parent mean of each measurement, and σ^2 its variance. The distribution (4.12) applies to any number of measurements N , even for a small value of N . When the measurements are Gaussian, it is therefore not necessary to invoke the central limit theorem, which remains applicable when N is large and leads to the same conclusion. The distribution of the sample mean of Gaussian measurement, therefore, features the $1/\sqrt{N}$ reduction in the standard error that was already highlighted, in a more general context, in Sect. 4.2.2.

4.4.2 *The Distribution of the Sum of Standard Uniform Random Variables*

According to the central limit theorem and the properties of a standard uniform distribution, the sum of N independent standard uniform variables is expected to approach a Gaussian distribution with a mean $\mu = N/2$ and a variance $\sigma^2 = N/12$. This can be shown by proving that the moment-generating function of the sum is asymptotically equal to

$$\lim_{N \rightarrow \infty} M(t) = e^{\frac{N}{2}t + \frac{t^2}{2} \cdot \frac{N}{12}}$$

Given that the N variables are independent, the moment-generating function of the sum of N standard uniform distributions can be written as the following series:

$$M(t) = M_i(t)^N = \left(1 + \frac{t}{2!} + \frac{t^2}{3!} + \dots\right)^N,$$

where (4.10) was used. Neglecting terms of order $O(t^3)$ and higher, the logarithm of the moment-generating function can be approximated as

$$\ln(M(t)^N) \simeq N \ln \left(1 + \frac{t}{2} + \frac{t^2}{6}\right).$$

The Taylor series expansion $\ln(1 + x) \simeq (x - x^2/2 + \dots)$ can now be used to obtain

$$\ln(M(t)) \simeq N \left(\frac{t}{2} + \frac{t^2}{6} - \frac{1}{2} \left(\frac{t}{2} + \frac{t^2}{6}\right)^2\right) \simeq N \left(\frac{t}{2} + \frac{t^2}{24}\right),$$

in which terms of order $O(t^3)$ and higher were again neglected, leading to the desired result. A similar derivation can also be provided for uniform distributions between different limits and not just for the standard uniform distribution.

Figure 4.1 shows simulations of the sum Y of $N = 100$ uniform and independent variables X_i between 0 and 1, with 1,000 and 100,000 samples drawn. The sample distribution of the sum approximates well the Gaussian distribution with $\mu = N/2$ and $\sigma^2 = N/12$, expected from the central limit theorem and the direct calculation of the moment-generating function of the sum. The approximation improves with when a larger number of samples are drawn from the $N = 100$ standard uniform distributions.

Example 4.2 (*Sum of Two Uniform Distributions*) An analytic exercise to develop a practical sense of how the sum of non-Gaussian distributions progressively develops a normal shape can be illustrated with the sum of just two uniform distributions. The moment generating function of a uniform distribution can be used to prove that the sum of two such variables will have a *triangular distribution*, given by the analytical form:

$$f(x) = \begin{cases} \frac{1}{2} + \frac{x}{4} & \text{if } -2 \leq x \leq 0 \\ \frac{1}{2} - \frac{x}{4} & \text{if } 0 \leq x \leq 2. \end{cases}$$

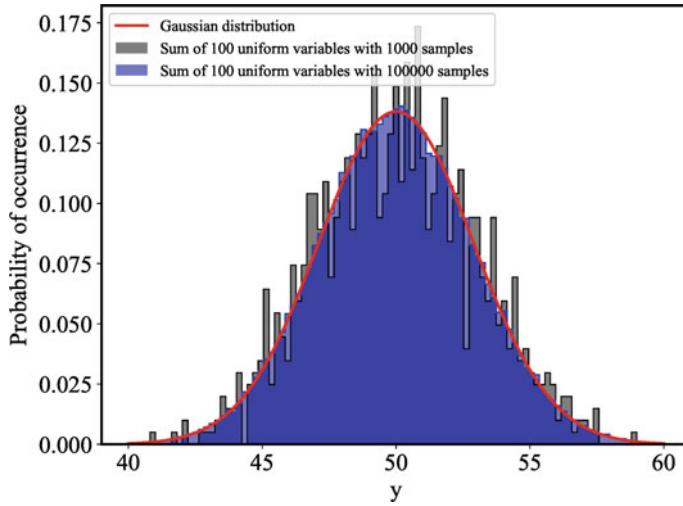


Fig. 4.1 Sample distribution functions of the sum of $N = 100$ independent uniform variables between 0 and 1, constructed from 1,000 and 100,000 simulated measurements. The solid curve is the expected limiting Gaussian distribution $N(N/2, N/12)$

This is an intuitive result that can be proven by showing that the moment-generating function of the triangular distribution is equal to the square of the moment-generating function of a standard uniform distribution (see Problem 3.3). The triangular distribution is the first step in the development of a peaked, Gaussian-like distribution. This example also shows that the sum of two or more uniform distributions in *not* a uniform distribution. \diamond

4.4.3 Certain Limitations of the Central Limit Theorem

Although the central limit theorem has wide applicability, the theorem does not hold for certain variables that do not have finite moments. One such example are the variables that follow the *Cauchy* probability distribution function

$$f(x) = \frac{1}{\pi(1+x^2)}, \quad (4.13)$$

which is defined for all real values. This seemingly simple distribution has the peculiar property that its expected value does not exist, and that certain higher-order moments are infinite, and therefore the central limit theorem does not apply to variables with this distribution.

The expectation of a variable X with a Cauchy distribution is defined by the usual integral:

$$E[X] = \int_{-\infty}^{\infty} x f(x) dx.$$

For this integral to exist, even as an extended real number, it should be possible to find a number a such that either the integral of $x f(x)$ between $-\infty, a$, or the one between a, ∞ , is a real number. But neither integral is finite, since the indefinite integral is

$$\int \frac{x}{\pi(1+x^2)} dx = \frac{\ln(1+x^2)}{2\pi}$$

and its evaluation at $\pm\infty$ causes a divergence in both integrals.

4.5 The Distribution of Functions of Random Variables

The general case of a variable that is a more complex function of other variables can be studied analytically when certain conditions are met. There are two convenient methods to obtain the distribution function of the new variable: the method of change of variables, which can be applied to monotonic transformations, and a method based on the cumulative distribution function. Additional methods can be found in probability textbooks such as [86].

4.5.1 The Method of Change of Variables

A simple method for obtaining the probability distribution function of the dependent variable $Y = Y(X)$ is by using the method of *change of variables*, which applies only if the function $Y(x)$ is a strictly increasing function of its argument. In this case, the probability distribution of $g(y)$ of the dependent variable is related to the distribution $f(x)$ of the independent variable via

$$g(y) = f(x) \frac{dx}{dy}. \quad (4.14)$$

For a decreasing function, the same method can be applied but the term dx/dy must be replaced with the absolute value $|dx/dy|$. Equation 4.14 therefore can be generalized to both increasing and decreasing functions, simply by adding an absolute value sign to the derivative on the right-hand side. This result can be interpreted as follows: the probability $f(x)dx$ that variable X is in an interval of size dx is equal to the probability $g(y)dy$ that the new variable Y is in an interval of size dy .

Example 4.3 (*Change of variables for a uniform distribution*) Consider a variable X with a uniform distribution between 0 and 1, and the variable $Y = X^2$. In the range of interest, this is a strictly increasing function of x . The method of change of variables can be used to show that the variable Y has a probability distribution function

$$g(y) = \frac{1}{2\sqrt{y}}.$$

This distribution is properly normalized in the $0 \leq y \leq 1$ domain.

The same change of variables can be applied to a uniform variable X in the negative range $-1 \leq x \leq 0$, where $dx/dy = -1/(2\sqrt{y})$, with $0 \leq y \leq 1$. The same distribution function $g(y)$ therefore applies also in this case. The method of change of variables does not apply to a uniform distribution in the range $-1 \leq x \leq 1$, since the transformation $Y = X^2$ is not monotonic in this range of the variable X . \diamond

The method of change of variables can be extended to the joint distribution of several random variables. Consider, for example, a pair of variables X, Y and the functions $U = u(X, Y)$ and $V = v(X, Y)$ that transform them to the pair of variables U, V . The two-variable version of (4.14) is

$$g(u, v) = h(x, y) |J|, \quad (4.15)$$

in which $|J|$ indicates the determinant of the matrix J , and

$$J = \begin{bmatrix} d(x, y) \\ d(u, v) \end{bmatrix} = \begin{bmatrix} \frac{dx}{du} & \frac{dx}{dv} \\ \frac{dy}{du} & \frac{dy}{dv} \end{bmatrix}$$

is the *Jacobian* of the transformation, in this case, a 2 by 2 matrix. The function $h(x, y)$ is the joint probability distribution of the independent variables X, Y .

Example 4.4 (*Transformation of Cartesian to Polar Coordinates*) Consider two random variables X and Y distributed as standard Gaussians, and independent of one another. Their joint probability distribution function is therefore

$$h(x, y) = \frac{1}{2\pi} e^{-\frac{x^2+y^2}{2}}.$$

Consider a transformation of variables from Cartesian coordinates x, y to polar coordinates r, θ , described by

$$\begin{cases} x = r \cdot \cos(\theta) \\ y = r \cdot \sin(\theta). \end{cases}$$

The Jacobian of the transformation is

$$J = \begin{pmatrix} \cos \theta & -r \sin \theta \\ \sin \theta & r \cos \theta \end{pmatrix}$$

and its determinant is $|J| = r$. Notice that to apply the method described by (4.15) one only needs to know the inverse transformation of (x, y) as the function of (r, θ) . It follows that the distribution of (r, θ) is given by

$$g(r, \theta) = \frac{1}{2\pi} r e^{-\frac{r^2}{2}}$$

for $r \geq 0, 0 \leq \theta \leq 2\pi$. The joint distribution function $g(r, \theta)$ can be written as the product of two functions: (a) the distribution

$$g_1(r) = r e^{-\frac{r^2}{2}},$$

which is known as the *Rayleigh distribution* and it is properly normalized for $r \geq 0$; and (b) $g_2(\theta) = 1/2\pi$, which can be interpreted as a uniform distribution for the angle θ . One concludes that, since $g(r, \theta)$ can be factored as the product of two functions that contain separately the two variables r and θ , the two new variables r, θ are also independent. \diamond

4.5.2 Direct Method Using the Distribution Function

The distribution of a variable Y that is a function of one or more other variables of known distribution can often be obtained directly by making use of the properties of the distribution function. This case can be illustrated by the function $Y = X^2$ when X has a uniform distribution in the range $-1 \leq x \leq 1$. The cumulative distribution $G(y)$ is obtained from

$$P(Y \geq y) = P(X \geq \sqrt{y}) + P(X \leq -\sqrt{y}) = 1 - P(X \leq \sqrt{y}) + P(X \leq -\sqrt{y}),$$

where two intervals are needed because the transformation is not monotonic. The cumulative distribution of X is $F(x) = (x - 1)/2$, and therefore the cumulative distribution of Y is given by

$$G(y) = 1 - P(Y \geq y) = \sqrt{y}, \quad 0 \leq y \leq 1.$$

The probability distribution function is then obtained via a simple derivative of the cumulative distribution as

$$g(y) = \frac{dG}{dy} = \frac{1}{2\sqrt{y}}.$$

It is often possible to arrive at the distribution of a function of random variable using this type of direct calculations also for multi-variable transformations. A common case of interest in statistics is when a variable Z is a function of two random variables X and Y , for example, $Z = X + Y$, or $Z = X/Y$. The methodology to follow in this case is illustrated with the function $Z = X + Y$, assuming that X and Y are independent. The calculation starts with the cumulative distribution function of the random variable of interest,

$$F_Z(a) = P(Z \leq a) = \iint_{x+y \leq a} f(x)g(y)dxdy,$$

in which $f(x)$ and $g(y)$ are, respectively, the probability distribution functions of X and Y , and the limits of integration must be chosen so that the sum of the two variables is less or equal than a . The portion of parameter space such that $x + y \leq a$ includes all values $x \leq a - y$, for any given value of y , or

$$F_Z(a) = \int_{-\infty}^{+\infty} dy \int_{-\infty}^{a-y} f(x)g(y)dx = \int_{-\infty}^{+\infty} g(y)F_x(a-y)dy,$$

where F_x is the cumulative distribution for the variable X . It is often more convenient to express the relationship in terms of the probability distribution function, which is related to the cumulative distribution function via a derivative,

$$f_Z(a) = \frac{d}{da} F_Z(a) = \int_{-\infty}^{\infty} f(a-y)g(y)dy. \quad (4.16)$$

This relationship is called the *convolution* of the distributions $f(x)$ and $g(y)$, and it yields the distribution function of the Z variable.

Example 4.5 (Sum of Two Independent Uniform Variables) The convolution integral can be used to calculate the probability distribution function of the sum of two independent uniform random variables between -1 and $+1$. The probability distribution function of the uniform variable is $f(x) = 1/2$, for $-1 \leq x \leq 1$, and the convolution formula gives the following integral:

$$f_Z(a) = \int_{-1}^{+1} \frac{1}{2} f(a-y) dy.$$

The distribution function of the sum Z is defined for $-2 \leq a \leq 2$, while the distribution function $f(a-y)$ is equal to $\frac{1}{2}$ only between -1 and $+1$, and it is null otherwise. The integral must therefore be divided into two parts, to ensure that $-1 \leq a-y \leq 1$. When $a < 0$, it is always true that $a-y \leq 1$ for the allowed range of y , and the remaining condition is $y \leq a+1$; the upper limit of integration must therefore be set at $a+1$. For $a > 0$, the situation is reversed, and the condition that $a-y \leq 1$ sets the lower limit of integration at $y=a-1$. The probability distribution function of Z is therefore given by

$$f_Z(a) = \frac{1}{4} \times \begin{cases} \int_{-1}^{a+1} dy = a+2 & \text{if } -2 \leq a \leq 0 \\ \int_{a-1}^1 dy = 2-a & \text{if } 0 \leq a \leq 2. \end{cases}$$

which is a *triangular distribution* between -2 and $+2$. ◊

Another useful application is for the case of $Z = X/Y$, where X and Y are again independent variables. the cumulative distribution of Z is related to the distributions of X and Y according to

$$F_Z(z) = P(Z < z) = P(X/Y < z) = P(X < zY).$$

For a given value y of the random variable Y , this probability equals $F_X(zy)$. Since Y has a probability $f_Y(y)dy$ to be in the range between y and $y+dy$, the cumulative distribution function of Z is

$$F_Z(z) = \int_{-\infty}^{\infty} F_X(zy) f_Y(y) dy.$$

The probability distribution function is again obtained via the derivative of $F_Z(z)$ with respect to z ,

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(zy) y f_Y(y) dy. \quad (4.17)$$

This is the integral than must be solved to obtain the distribution of X/Y .

Summary of Key Concepts for this Chapter

Linear combination of variables: The formulas for the mean and variance of the linear combination of N variables are

$$\begin{cases} \mu = \sum_{i=1}^N a_i \mu_i \\ \sigma^2 = \sum_{i=1}^N a_i^2 \sigma_i^2 + 2 \sum_{i=1}^N \sum_{j=i+1}^N a_i a_j \sigma_{ij}^2 \end{cases}$$

Variance of uncorrelated variables: When variables are uncorrelated the variances add linearly. The variance of the mean of N independent measurements is $\sigma_Y^2 = \sigma^2/N$.

Moment-generating function: It is a mathematical function that enables the calculation of moments of a distribution, $M(t) = E[e^{tX}]$.

Central Limit theorem: The sum of a large number of independent variables is distributed like a Gaussian of mean equal to the sum of the means and variance equal to the sum of the variances.

Method of change of variables: A method to obtain the distribution function of a variable Y that is a function of another variable X ,

$$g(y) = f(x) dx/dy$$

Problems

4.1 Consider two independent uniform variables X, Y in the range -1 to 1 .

- (a) Determine the distribution function, mean and variance, and the moment-generating function of the variables.
- (b) Speculate that the sum of the two random variables is distributed like a *triangular distribution* between the range -2 and 2 , with distribution function

$$f(x) = \begin{cases} \frac{1}{2} + \frac{x}{4} & \text{if } -2 \leq x \leq 0 \\ \frac{1}{2} - \frac{x}{4} & \text{if } 0 \leq x \leq 2 \end{cases}$$

Using the moment generation function, prove that the variable $Z = X + Y$ is in fact distributed like the triangular distribution above.

4.2 According to the central limit theorem, the distribution of the sum of $N = 100$ independent uniform variables in the range 0 to 1 is approximately a Gaussian with mean equal to the mean of the N uniform variables, and variance equal to the sum of the variances. Using a computer language of your choice, simulate the sum of the $N = 100$ uniform variables, using (a) 1,000 and (b) 100,000 samples for each variable, and compare the simulated distributions with the expected distribution based on the central limit theorem.

4.3 Consider two independent random variables X and Y , and their difference $Z = X - Y$.

- (a) Assuming that both X and Y are normally-distributed, determine whether Z also follows a normal distribution;
- (b) Assuming that both X and Y are Poisson-distributed, determine whether Z retains the Poisson distribution.

For this problem, it may be convenient to use the moment-generating function of the two distributions, and their properties.

4.4 Using the method of change of variables, determine the probability distribution of the variable $Y = \ln X$, where X is a uniform distribution between 1 and 2 , and its expectation $E[Y]$.

4.5 Calculate the mean, variance, and moment-generating function $M(t)$ for a uniform random variable in the range -1 to 1 .

Chapter 5

Error Propagation and Simulation of Random Variables



Abstract For a function of random variables, it is often necessary or convenient to develop approximate methods to estimate the mean and the variance without having full knowledge of its probability distribution function. The approximate methods to calculate the variance are of common use in data analysis and they are referred to as *error propagation* methods. Methods for the simulation of samples from a random variable using a standard uniform random variable are also presented.

5.1 The Mean of Functions of Random Variables

The general situation of interest is a random variable Y that is a function of one or more random variables, $Y = g(X)$ or $Y = g(X_1, \dots, X_n)$. It is necessary to start with the case of random variables X of known distribution function. For continuous random variables, the expectation is defined by

$$E[Y] = \int g(x_1, \dots, x_n)h(x_1, \dots, x_n)dx_1 \dots dx_n, \quad (5.1)$$

where $h(x_1, \dots, x_n)$ is the joint probability distribution function of the random variables X_i . For functions of a single variable $Y = g(X)$, the expectation is

$$E[Y] = \int g(x)f(x)dx, \quad (5.2)$$

where $f(x)$ is the distribution function of X . When the probability distribution of the X_i variables is available, the integration leads to the expectation of Y .

Example 5.1 (*Mean of the Square of a Uniform Variable*) Consider a standard uniform variable U in the range 0 to 1, with mean $1/2$. The parent mean of $Y = U^2$ is given by

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-981-19-0365-6_5.

$$\mu = \int_0^1 u^2 du = 1/3.$$

It is important to observe that the mean of U^2 is not just the square of the mean of U . This situation can also be illustrated with a practical example by using five “fair” samples from a uniform distribution, say 0.1, 0.3, 0.5, 0.7, and 0.9. Clearly their mean is $1/2$, but the mean of their squares is $1/3$ and not $1/4$, in agreement with the theoretical calculation of the parent mean. \diamond

As illustrated in the prior example, it is generally true that the mean of the function differs from the function of the mean, $E[g(X)] \neq g(E[X])$. It is useful to report a few general properties of the mean of random variables that can be of use in the analysis of data.

- (a) When X and Y are independent variables, then $E[X \cdot Y] = E[X] \cdot E[Y]$. Moreover, it is also true that $E[g(X) \cdot h(Y)] = E[g(X)] \cdot E[h(Y)]$ for any function g and h . This property follows immediately from the fact that the joint distribution function of two independent variables can be factored as the product of the individual distributions.
- (b) *Jensen’s Inequalities* state that $E[g(X)] \geq g(E[X])$ when g is a convex function (i.e., a function with non-negative second derivative), and $E[g(X)] \leq g(E[X])$ when g is a concave function. This situation was already encountered in the example of the U^2 function, which is a convex function that has $E[U^2] \geq E[U]^2$.
- (c) The *Cauchy–Schwartz* inequality, see (2.22), states that $E[|X \cdot Y|] \leq \sqrt{E[X^2] \cdot E[Y^2]}$. This property is used to show that the correlation coefficient does not exceed unity.

The equations above are only the definition of the (theoretical) expectation of the function $g(X)$, when the parameters of the probability distribution function of X are known. Obviously, the task of the data analyst is that of estimating such expectation via N independent measurements of the variable X or, more generally, of a set of random variables (X_1, \dots, X_n) , without the knowledge of the parent parameters of the distribution. Fortunately, the expectation of Y can be estimated immediately using the law of large numbers (2.10), i.e.,

$$E[Y] \simeq \frac{g(x_1) + \dots + g(x_j) + \dots g(x_N)}{N}, \quad (5.3)$$

where x_j represents one measurement (of one or more variables) needed to evaluate the function g . The law of large numbers ensures that the right-hand side of (5.3) is asymptotically an unbiased estimator of the mean, as the symbol already indicates. Notice the simplicity of this result: there is no need at all know the probability distribution of Y to estimate its mean. The task of estimating the mean of the function of a random variable is therefore immediately accomplished by (5.3).

It is important to remark that (5.3) explicitly says that one must have access to the N individual measurements of the variable X , in order to make inferences on the mean of Y . If an experimenter only reports the sample mean of the N measurements of the X variable, (5.3) cannot be later used to estimate the mean of a secondary variable $Y = g(X)$. In this case, it is simply not possible to estimate accurately $E[Y]$, since, in general, it is not true that the mean of the function g is equal to the function of the means, as illustrated earlier in the section with the example of the uniform variable (see Example 5.1). This application of the law of large numbers to the estimation of the mean of the function of random variables is therefore also a reminder of the importance of reporting the raw data of an experiment, in this case, in the form of all N measurements of the primary variable X .

5.2 The Variance of Functions of Random Variables and Error Propagation Formulas

A random variable Y that is a function of one or more variables can have its variance estimated directly if the measurements of the independent variables are available. With the mean estimated from (5.3), the variance can accordingly be estimated as

$$s_y^2 = \frac{(g(x_1) - E[Y])^2 + \dots + (g(x_N) - E[Y])^2}{N - 1}, \quad (5.4)$$

as one would normally do, treating the numbers $g(x_1), \dots, g(x_N)$ as samples from the dependent variable. When the measurements of the independent variables are available, this method provides a direct way to estimate the variance of the function of random variables.

Example 5.2 (*Sample variance from Thomson's experiment*) Using the Thomson experiment described on Sect. 2.4, consider the data collected for Tube 1, consisting of 11 measurements of two variables $X_1 = W/Q$ and $X_2 = I$, from which the variable of interest $Y = v$ is calculated as

$$Y = 2 \frac{X_1}{X_2}.$$

From the reported data, one obtains 11 measurements of Y , resulting in a mean and standard deviation of $\bar{v} = 7.9 \times 10^9$ and $s_v = 2.8 \times 10^9$. ◇

There are instances in which one does not have access to the original measurements of the independent variables required for an estimate of the variance according to (5.4). In this case, an approximate method to estimate the variance must be used instead. This method takes the name of *error propagation*. Consider a random variable Y that is a function of a number of variables, $Y = g(X_1, \dots, X_n)$. A method to approximate the variance of Y in terms of the variance of the X_j variables starts

by expanding g in a Taylor series about the means of the independent variables, to obtain

$$\begin{aligned} g(x_1, x_2, \dots) &= g(\mu_1, \mu_2, \dots) + (x_1 - \mu_1) \left. \frac{\partial g}{\partial x_1} \right|_{\mu_1} + (x_2 - \mu_2) \left. \frac{\partial g}{\partial x_2} \right|_{\mu_2} \\ &\quad + \dots + \mathcal{O}(x_1 - \mu_1)^2 + \mathcal{O}(x_2 - \mu_2)^2 + \dots \end{aligned}$$

Neglecting terms of the second order, the expectation of Y would be approximated as the function of the means, $E[Y] = g(\mu_1, \mu_2, \dots)$. This is true only if the function is linear but not, in general, as discussed in Sect. 5.1. The expansion can however be used to approximate the variance as

$$\begin{aligned} E[(Y - E[Y])^2] &\simeq E[(g(x_1, x_2, \dots) - g(\mu_1, \mu_2, \dots))^2] \simeq \\ &E \left[\left((x_1 - \mu_1) \left. \frac{\partial g}{\partial x_1} \right|_{\mu_1} \right)^2 \left((x_2 - \mu_2) \left. \frac{\partial g}{\partial x_2} \right|_{\mu_2} \right)^2 \right. \\ &\quad \left. + 2(x_1 - \mu_1) \left. \frac{\partial g}{\partial x_1} \right|_{\mu_1} \cdot (x_2 - \mu_2) \left. \frac{\partial g}{\partial x_2} \right|_{\mu_2} + \dots \right], \end{aligned}$$

where only terms relevant to two variables were shown. This formula can be rewritten as

$$\sigma_y^2 \simeq \sigma_1^2 \left. \frac{\partial g}{\partial x_1} \right|_{\mu_1}^2 + \sigma_2^2 \left. \frac{\partial g}{\partial x_2} \right|_{\mu_2}^2 + 2 \cdot \sigma_{12}^2 \left. \frac{\partial g}{\partial x_1} \right|_{\mu_1} \left. \frac{\partial g}{\partial x_2} \right|_{\mu_2} + \dots \quad (5.5)$$

which is usually referred to as the *error propagation formula*. Equation 5.5 refers to a function of two variables, but a similar formula can be obtained for any number of independent variables. This result makes it possible to estimate the variance of a function of variable, knowing simply the mean and variance of each of the independent variables, and their covariances.

The notation in these error propagation formulas uses Greek letters to refer to means, variances, and covariances. These quantities, however, are generally not parent quantities, since the distribution function of the relevant variables is usually unknown, but they should be regarded as sample quantities as estimated from the N available measurements. The Greek-letter notation is retained for ease of interpretation.

The formula is especially useful for all cases in which the measured variables are independent, and all that is known is their mean and standard deviation (but not the individual measurements used to determine the mean and variance). This method must be considered as an approximation when there is only incomplete information about the measurements. Neglecting terms of the second order in the Taylor expansion can in fact lead to significant errors, especially when the function has strong non-

linearities. Specific formulas for functions that are of common use are provided in the following.

5.2.1 Sum and Product of a Constant

Consider adding a constant a to a variable X ,

$$Y = X + a,$$

where a is a constant that can have either sign. It is clear that $\partial g/\partial a = 0$, $\partial g/\partial x = 1$, and therefore the addition of a constant has no effect on the uncertainty of X ,

$$\sigma_y^2 = \sigma_x^2. \quad (5.6)$$

The addition or subtraction of a constant only changes the mean of the variable by the same amount, but leaves its variance unchanged. If the variable is multiplied by a constant b ,

$$Y = bX$$

the error propagation formulas lead to

$$\sigma_y^2 = b^2 \sigma_x^2. \quad (5.7)$$

These properties applies in general and are not due to approximations of the error propagation method. In fact, it is true that for any variable $\text{Var}(a + bX) = b^2\text{Var}(X)$.

5.2.2 Weighted Sum of Two Variables

The variance of the weighted sum of two variables,

$$Y = aX_1 + bX_2,$$

where a, b are constants of either sign, can be calculated using $\partial g/\partial x_1 = a$, $\partial g/\partial x_2 = b$. The error propagation formula yields

$$\sigma_y^2 = a^2\sigma_1^2 + b^2\sigma_2^2 + 2ab\sigma_{12}^2. \quad (5.8)$$

The special case of two uncorrelated variables X_1 and X_2 leads to the result that the variance of the weighted mean is the weighted sum of the variances, with weights equal to the square of the weights in the function Y . This result for uncorrelated variables applies in general from basic properties of the variance. On the other hand,

the full Eq. (5.8) in the presence of the covariance is a useful approximation of the error propagation method.

Example 5.3 (*Error propagation for background subtraction*) Consider a decaying radioactive source from which $N_1 = 50$ counts and $N_2 = 35$ counts are recorded by a Geiger counter in two time intervals of same duration. During an equal interval of time, the same instrument records $B = 20$ background counts. The goal of the experiment is to estimate the number of background—subtracted source counts in the two time intervals, and their variance. Each random variable N_1 , N_2 , and B is expected to follow the Poisson distribution, since this is a counting experiment. Therefore, it is reasonable to assume that the measured counts approximate the respective parent means. Accordingly, the following means and variances can be estimated from the single measurement,

$$\begin{cases} \mu_1 = \sigma_1^2 \simeq 50 \\ \mu_2 = \sigma_2^2 \simeq 35 \\ \mu_B = \sigma_B^2 \simeq 20. \end{cases}$$

The source counts are given by $S_1 = N_1 - B$ and $S_2 = N_2 - B$. The approximate variance formulas can now be used *assuming* that the variables are uncorrelated, leading to $\sigma_{S_1}^2 = 50 + 20$ and $\sigma_{S_2}^2 = 35 + 20$. Notice the variances add even in the case of background subtraction, and therefore estimates of the source counts would be reported as $S_1 = 30 \pm 8.4$ and $S_2 = 15 \pm 7.4$.

The available data do not provide information of the correlation between N_1 , N_2 , and B , and therefore the assumption of no correlation is simply one of convenience in this example. If the example was repeated several times, and all measurements of the three quantities reported for each measurement, then it would be possible to estimate means and variances without relying on the assumption of a Poisson distribution. Moreover, it would also be possible to evaluate the sample covariance and use that estimate in the error propagation formula. In this one-measurement experiment, the variances were estimated from the expected Poisson distribution, for which the parent mean equals its variance. \diamond

5.2.3 Product and Division of Two Random Variables

Consider the product of two random variables X_1 , X_2 , with a constant factor a of either sign,

$$Y = a X_1 X_2. \quad (5.9)$$

The partial derivatives are $\partial g / \partial x_1 = ax_2$, $\partial g / \partial x_2 = ax_1$, leading to the approximate variance of

$$\sigma_y^2 = a^2 x_2^2 \sigma_1^2 + a^2 x_1^2 \sigma_2^2 + 2ax_1x_2\sigma_{12}^2.$$

According to the error propagation formula (5.5), the derivatives are to be evaluated at the mean values of the random variables. This means that, in the previous equation, x_1 represents the estimated mean value of the variable, and the same for x_2 . (Same considerations apply for other functions in this section.) The previous equation can be more conveniently rewritten as

$$\frac{\sigma_y^2}{y^2} = \frac{\sigma_1^2}{x_1^2} + \frac{\sigma_2^2}{x_2^2} + 2 \frac{\sigma_{12}^2}{x_1 x_2}. \quad (5.10)$$

Similarly, the division between two random variables

$$Y = a \frac{X_1}{X_2} \quad (5.11)$$

leads to

$$\frac{\sigma_y^2}{y^2} = \frac{\sigma_1^2}{x_1^2} + \frac{\sigma_2^2}{x_2^2} - 2 \frac{\sigma_{12}^2}{x_1 x_2}. \quad (5.12)$$

Notice how the error propagation formulas for product and division differ by just one sign, with a positive covariance between the variables leading to a reduction in the standard deviation for the division, and an increase in the standard deviation for the product.

Example 5.4 (*Error propagation with Thomson's experiment*) The Thomson data of Sect. 2.4 were used in Example 5.2 to calculate the mean and variance of v from the 11 individual measurements of Tube 1 ($\bar{v} = 7.9 \times 10^9$ and $s_v = 2.8 \times 10^9$). The two variables measured directly in the experiment were also calculated as $W/Q = 13.3 \pm 8.5 \times 10^{11}$ and $I = 312.9 \pm 93.4$ with a sample covariance between them of $s_{12}^2 = 759.1$ (see Example 2.2). The relationship between the measured variables and the variable of interest is $v = 2(W/Q)/I$. Given the availability of the individual measurements, it was possible to provide an estimate of v that accounts for the covariance between the two measured variables.

If Thomson had reported only the mean and standard deviation of W/Q and I , instead of the entire raw data, one would have been forced to estimate the mean and variance from the approximate error propagation formulas. The mean of v would have been estimated as the function of the means, $v \simeq 8.5 \times 10^9$, close to the value of 7.9×10^9 obtained from the individual measurements. The estimate of the variance requires also a knowledge of the covariance between the two variables W/Q and I . In the absence of that information, one would have to proceed with the assumption that the two variables are uncorrelated and use the error propagation formula to obtain

$$\sigma_v \simeq 2 \times \frac{13.3 \times 10^{11}}{312.9} \times \left(\left(\frac{8.5}{13.3} \right)^2 + \left(\frac{93.4}{312.9} \right)^2 \right)^{1/2} = 6 \times 10^9,$$

which is a factor of 2 larger than what is obtained directly from the data. The discrepancy is to be attributed to the neglect of the covariance between the measurement, which was in fact found to be positive, and therefore would reduce the variance of v according to (5.12). \diamond

5.2.4 Power of a Random Variable

A random variable may be raised to a constant power,

$$Y = aX^b, \quad (5.13)$$

where a and b are constants of either sign. In this case, $\partial g/\partial x = abx^{b-1}$ and the error propagation results in

$$\frac{\sigma_y}{y} = |b| \frac{\sigma_x}{x}. \quad (5.14)$$

This result states that the relative error in the variable Y is b times the relative error in the original variable X .

5.2.5 Exponential of a Random Variable

Consider the exponential of a random variable,

$$Y = ae^{bx}, \quad (5.15)$$

where a and b are constants of either sign. The partial derivative is $\partial g/\partial x = ab e^{bx}$, leading to the following error propagation formula:

$$\frac{\sigma_y}{y} = |b| \sigma_x. \quad (5.16)$$

5.2.6 Logarithm of a Random Variable

For the natural logarithm of a variable

$$Y = a \ln(bX), \quad (5.17)$$

where a is a constant of either sign, and $b > 0$, the partial derivative is $\partial g/\partial x = a/x$, leading to the error propagation formula

Table 5.1 Common error propagation formulas

Function	Error propagation formula	Notes
$Y = X + a$	$\sigma_y^2 = \sigma_x^2$	a is a constant
$Y = aX_1 + bX_2$	$\sigma_y^2 = a^2\sigma_1^2 + b^2\sigma_2^2 + 2ab\sigma_{12}^2$	a, b are constants
$Y = aX_1 X_2$	$\frac{\sigma_y^2}{y^2} = \frac{\sigma_1^2}{x_1^2} + \frac{\sigma_2^2}{x_2^2} + 2\frac{\sigma_{12}^2}{x_1 x_2}$	a is a constant
$Y = a \frac{X_1}{X_2}$	$\frac{\sigma_y^2}{y^2} = \frac{\sigma_1^2}{x_1^2} + \frac{\sigma_2^2}{x_2^2} - 2\frac{\sigma_{12}^2}{x_1 x_2}$	a is a constant
$Y = aX^b$	$\frac{\sigma_y}{y} = b \frac{\sigma_x}{x}$	a, b are constants
$Y = ae^{bX}$	$\frac{\sigma_y}{y} = b \sigma_x$	a, b are constants
$Y = a \ln(bX)$	$\sigma_y = a \frac{\sigma_x}{x}$	a, b are constants, $b > 0$
$Y = a \log(bX)$	$\sigma_y = a \frac{\sigma_x}{x \ln 10}$	a, b are constants, $b > 0$

$$\sigma_y = |a| \frac{\sigma_x}{x}. \quad (5.18)$$

For a base-10 logarithm,

$$Y = a \log(bX), \quad (5.19)$$

where a is a constant of either sign and $b > 0$. The partial derivative is $\partial g / \partial x = a / (x \ln 10)$, leading to the error propagation formula

$$\sigma_y = |a| \frac{\sigma_x}{x \ln 10}. \quad (5.20)$$

Similar error propagation formulas can be obtained for virtually any analytic function for which derivatives can be calculated. Some common formulas are reported for convenience in Table 5.1, where the terms Y , X or X_1 and X_2 refer to the random variables evaluated at their estimated mean value.

5.3 The Quantile Function and Simulation of Random Variables

Certain data analysis tasks require the simulation of random variables, that is, drawing random samples from a parent distribution. The simplest simulation is the generation of random numbers from a uniform distribution between two limits. For more

complex distribution, it is useful to learn how to generate random samples based on random samples from a uniform variable. For this purpose, it is first necessary to introduce the *quantile* of a distribution function. Given a variable X with a probability distribution function $f(x)$ and a cumulative distribution function $F(x)$, the p -quantile is the x -value of the random variable that has a cumulative probability $F(x) = p$; this means that there is a probability of p that the random variable has values less or equal to x . Sometimes the quantile is expressed as a *percentile* of the distribution, i.e., by expanding the $0 \leq p \leq 1$ range of the quantile to one hundred. For a real-valued random variable X , the *quantile function* $F^{-1}(p)$ is defined as

$$F^{-1}(p) = \min\{x \in \mathbb{R}, p \leq F(x)\} \quad (5.21)$$

with the meaning that x is the minimum value of the variable where the cumulative distribution function reaches the value p . The operation of minimum is necessary only for those distributions that have discontinuities in their cumulative distribution, but in the more common case of a strictly increasing cumulative distribution, the quantile function is simply defined by the relationship $p = F(x)$. This equation can be solved for x , to obtain the quantile function $x = F^{-1}(p)$. The basic property of the quantile function can be stated mathematically as the equivalence between the following two statements,

$$p \leq F(x) \Leftrightarrow x \leq F^{-1}(p). \quad (5.22)$$

Example 5.5 (*Quantile Function of a Uniform Distribution*) For a standard uniform variable in the range 0 to 1, the quantile function has a particularly simple form. In fact, the cumulative distribution is $F(x) = x$, and the quantile function defined by the equation $p = F(x)$ yields $x = p$, and therefore

$$x = F^{-1}(p) = p. \quad (5.23)$$

Therefore, the analytical form of both the cumulative distribution and the quantile function is identical for the standard uniform variable, meaning that, e.g., the value $x = 0.75$ of the random variable is the $p = 0.75$ quantile or 75% percentile of the distribution. \diamond

Example 5.6 (*Quantile Function of an Exponential Distribution*) An exponential random variable has probability distribution function

$$f(x) = \lambda e^{-\lambda x},$$

with $x \geq 0$, and a cumulative distribution function

$$F(x) = 1 - e^{-\lambda x}.$$

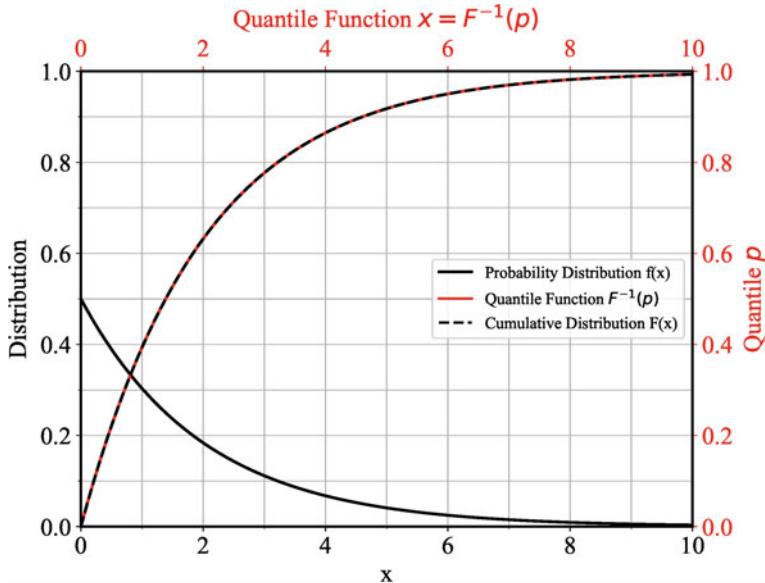


Fig. 5.1 Distribution function $f(x)$, cumulative distribution $F(x)$, and quantile function $F^{-1}(p)$ of an exponential variable with $\lambda = 1/2$. The overlap between the cumulative distribution and the quantile function (which are a function of variables along different axes) highlights the interplay between the two quantities

Starting from $p = F(x)$, one obtains $x = \ln(1 - p)/(-\lambda)$, and therefore the quantile function is

$$x = F^{-1}(p) = -\frac{\ln(1 - p)}{\lambda}.$$

Figure 5.1 shows the cumulative distribution and the quantile function for the exponential distribution. \diamond

5.3.1 General Method to Simulate a Variable

The method to simulate a random variable is summarized in the following equation:

$$X = F^{-1}(U), \quad (5.24)$$

which states that any random variable X can be expressed in terms of the uniform variable U between 0 and 1, where F is the cumulative distribution of the variable X and F^{-1} is the quantile function. If a closed analytic form for F is available, this equation results in a simple method to simulate the random variable.

It was already shown that the uniform variable has a quantile function $F^{-1}(U) = U$, i.e., the quantile function is a uniform random variable itself. The proof therefore simply consists of showing that, assuming (5.24), the cumulative distribution of X is indeed $F(X)$, or $P(X \leq x) = F(x)$. This can be shown by writing

$$P(X \leq x) = P(F^{-1}(U) \leq x) = P(U \leq F(x)) = F(x),$$

in which the second equality follows from the definition of the quantile function, and the last equality follows from the fact that $P(U \leq u) = u$ for a standard uniform variable.

Example 5.7 (*Simulation of an Exponential Variable*) Consider a random variable distributed like an exponential, $f(x) = \lambda e^{-\lambda x}$, $x \geq 0$. Given the calculations developed in the example above, the exponential variable can be simulated by

$$X = -\frac{\ln(1 - U)}{\lambda}.$$

Notice that, although this relationship is between random variables, its practical use is to draw random samples u from U , and a random sample x from X is obtained by simply using the equation

$$x = -\frac{\ln(1 - u)}{\lambda},$$

where u is a random number between 0 and 1. Therefore, for a large sample of values u , the above equation returns a random sample of values for the exponential variable X . \diamond

Example 5.8 (*Simulation of the Square of Uniform Variable*) It can be proven that the simulation of the square of a uniform random variable $Y = U^2$ is indeed achieved by squaring samples from a uniform distribution, a very intuitive result. In fact, starting with the known distribution of Y as $g(y) = 1/2 y^{-1/2}$ and its cumulative distribution $G(y) = \sqrt{y}$, the quantile function is defined by $p = \sqrt{y}$, and therefore the quantile function for U^2 is

$$y = G^{-1}(p) = p^2, \quad 0 \leq p \leq 1.$$

This result, according to (5.24), defines U^2 , or the square of a uniform distribution, as the function that needs to be simulated to draw random samples from Y . \diamond

5.3.2 Simulation of a Gaussian Variable

This method of simulation of random variables relies on the knowledge of $F(x)$ and the fact that such a function is analytic and invertible. In the case of the Gaussian

distribution, the cumulative distribution function is a special function,

$$F(x) = \frac{1}{2\pi} \int_{-\infty}^x e^{-\frac{x^2}{2}} dx,$$

which cannot be inverted analytically, and therefore this method cannot be applied. This complication must be overcome, given the importance of Gaussian distributions in probability and statistics. Fortunately, a relatively simple method is available that permits the simulation of two Gaussian distributions from two uniform random variables. In Sect. 4.5, it was shown that the transformation from Cartesian to polar coordinates results in two random variables R, Θ that are distributed, respectively, like a Rayleigh and a uniform distribution:

$$\begin{cases} h(r) = r e^{-\frac{r^2}{2}} & r \geq 0 \\ i(\theta) = \frac{1}{2\pi} & 0 \leq \theta \leq 2\pi. \end{cases} \quad (5.25)$$

Since these two distributions have a closed analytic form for their cumulative distributions, R and Θ can be easily simulated and then use the transformation given by (4.4) to simulate a pair of independent standard Gaussians. Starting with the Rayleigh distribution with a cumulative distribution function

$$H(r) = 1 - e^{-\frac{r^2}{2}},$$

the quantile function is given by

$$p = 1 - e^{-\frac{r^2}{2}},$$

and from this the quantile function is obtained as

$$r = \sqrt{-2 \ln(1 - p)} = H^{-1}(p).$$

This result shows that $R = \sqrt{-2 \ln(1 - U)}$ simulates a Rayleigh distribution from a standard uniform variable U . For the uniform variable Θ , the cumulative distribution is

$$I(\theta) = \begin{cases} \theta/(2\pi) & 0 \leq \theta \leq 2\pi \\ 0 & \text{otherwise;} \end{cases}$$

the quantile function is $\theta = 2\pi p = I^{-1}(p)$, and therefore $\Theta = 2\pi V$ simulates a uniform distribution between 0 and 2π from a standard uniform distribution V . Therefore, the use of two independent uniform distributions U and V can be used to simulate a Rayleigh and a uniform angular distribution according to

$$\begin{cases} R = \sqrt{-2 \ln(1 - U)} \\ \Theta = 2\pi V. \end{cases} \quad (5.26)$$

Then, using the Cartesian to polar coordinate transformation, the formulas needed to simulate a pair of Gaussians X and Y are

$$\begin{cases} X = R \cos \Theta = \sqrt{-2 \ln(1 - U)} \cdot \cos(2\pi V) \\ Y = R \sin \Theta = \sqrt{-2 \ln(1 - U)} \cdot \sin(2\pi V). \end{cases} \quad (5.27)$$

Equation (5.27) can be easily implemented using two independent standard uniform variables between 0 and 1. The standard Gaussians obtained from (5.27) can be further transformed to Gaussians of any mean and variance via a simple linear rescaling using the desired parent values.

Summary of Key Concepts for this Chapter

Error propagation formulas: They are approximations for the variance of a function of random variables when the variance cannot be estimated from the data. For a function of two variables $Y = g(X_1, X_2)$,

$$\sigma_Y^2 \simeq \sigma_1^2 \left. \frac{\partial g}{\partial x_1} \right|_{\mu_1}^2 + \sigma_2^2 \left. \frac{\partial g}{\partial x_2} \right|_{\mu_2}^2 + 2 \cdot \sigma_{12}^2 \left. \frac{\partial g}{\partial x_1} \right|_{\mu_1} \left. \frac{\partial g}{\partial x_2} \right|_{\mu_2}.$$

Quantile function: It is the function $x = F^{-1}(p)$ used to find the value x of a variable that corresponds to a given quantile p .

Simulation of a Variable: The general formula for the simulation of a variable X with cumulative distribution F is

$$X = F^{-1}(U),$$

where U is the standard uniform distribution.

Simulation of Gaussians: Two standard Gaussians can be obtained from two uniform random variables U, V via

$$\begin{cases} X = \sqrt{-2 \ln(1 - U)} \cos(2\pi V) \\ Y = \sqrt{-2 \ln(1 - U)} \sin(2\pi V). \end{cases}$$

Problems

5.1 ■ This problem illustrates that the error propagation formulas may give different results than the direct measurement of the mean and variance of a variable, when the individual measurements are available. Consider the data from Thomson's experiment of Tube 1, from Sect. 2.4.

- (a) Calculate the sample mean and standard deviation of the measurements of v .
- (b) Consider that the measured variables W/Q and I are related to the variable v via the relationship

$$v = \frac{2W}{QI}.$$

Use the results from Problem 2.4, in which the sample mean and standard deviation of W/Q and I were calculated, to calculate the approximate values of mean and standard deviation of v using the relevant error propagation formula, and assuming no correlation between the two measurements.

- (c) By comparison of the two estimates of the sample variance in (a) and (b), determine if there is a positive or negative correlation between W/Q and I .
- (d) Use the measurements of W/Q and I to calculate the sample covariance between the two variables.

5.2 ■ Consider the J.J. Thomson experiment of Sect. 2.4.

- (a) Calculate the sample mean and the standard deviation of m/e for Tube 1.
- (b) Calculate the approximate mean and standard deviation of m/e from the mean and standard deviation of W/Q and I , according to the equation

$$\frac{m}{e} = \frac{I^2}{2} \frac{Q}{W},$$

and assuming that W/Q and I are uncorrelated.

- (c) Compare the estimates of the sample variance of m/e from (a) and (b) and identify the reason for the difference.

5.3 Use the data provided in Example 5.3 to calculate the following:

- (a) the probability of a positive detection of source counts S in the first time period (where there are $N_1 = 50$ total counts and $B = 20$ background counts);
- (b) the probability that the source emitted ≥ 10 source counts.

Assume that the variable S can be approximated by a Gaussian distribution.

5.4 Derive a general expression for the error propagation formula when three independent random variables are present, to generalize (5.5) that is valid for two variables.

Chapter 6

Maximum Likelihood and Other Methods to Estimate Variables



Abstract This chapter addresses the problem of estimating the distribution function of a random variable based on measurements. The method of maximum likelihood is the tool of choice when the distribution of the variable is known. The maximum-likelihood criterion is also useful in more complex applications that involve the fit of two-dimensional data and the estimation of fit parameters. Alternative means to estimate variables include the method of moments and the method of maximum entropy. In particular, the method of maximum entropy can be used when the parent distribution is not specified a priori.

6.1 Estimating Random Variables with Data

Experimental data are essential to gain information on random variables. R. A. Fisher is widely credited as the most influential person to develop key concepts in modern statistics, which he defines in the book *Statistical Methods for Research Workers* [31]

as (i) the study of populations, (ii) the study of variation, and (iii) the study of methods of the reduction of data.

The word *population* refers to an aggregate of entities that has a well-defined set of properties or random variables that vary in a prescribed way. The concept of *data reduction* is that of extracting what information the data have with regard to the variables of interest. In Fisher's words,

the reduction of data may thus conveniently be divided into three types:
(i.) Problems of Specification, which arise in the choice of the mathematical form of the population.
(ii.) Problems of Estimation, which involve the choice of method of calculating, from our sample, statistics fit to estimate the unknown parameters of the population.
(iii.) Problems of Distribution, which include the mathematical deduction of the exact nature

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-981-19-0365-6_6.

of the distribution in random samples of our estimates of the parameters, and of other statistics designed to test the validity of our specification (tests of Goodness of Fit).

The first step can therefore be viewed, in its simplest form, as specifying the probability distribution for a random variable of interest. With an analytical form for the distribution of the variable available, one then needs to estimate its parameters by defining a suitable *statistic* that can be used to measure said parameters. A statistic is broadly defined as a random variable that is a function of the data, often in the form of several independent and identically distributed measurements. A simple example is that of measurements from a normally distributed variable, and the use of the sample mean or sample variance as data-based *statistics* that are used to estimate the unknown parent values of the distribution. It is important to realize that these statistics themselves follow a distribution, sometimes referred to as their sampling distribution. For example, the sample mean from N measurements is a statistic that has an expectation, a variance, and, in general, a probability distribution function. In the case of normally distributed measurements, the sample mean was shown to be distributed as a Gaussian (see Sect. 4.4.1).

One of the key issues of this process of gaining knowledge of the population is how to determine a statistic that is suitable for the task of estimating the parameters. So far, the sample mean and the sample variance were defined *ad hoc* as a reasonable data-based statistic, and *post facto* it was ensured that the statistics were in fact unbiased. One of Fisher's key accomplishments is that of establishing the *method of maximum likelihood* as the tool of choice to determine statistics:

The researches of the author have led him to the conclusion that an efficient statistic can in all cases be found by the Method of Maximum Likelihood; that is, by choosing statistics so that the estimated population should be that for which the likelihood is greatest.

The method of maximum likelihood uses the likelihood of the data with a specified model to infer the best-fit model parameters using standard analytical tools. It is necessary to remark that this method is borne out of the need to relate a data-based statistic with a parent model. Logical and successful as the method is, there are alternative methods that can also provide accurate results. Of particular interest are two alternative methods to estimate variables that are also presented in this chapter. One is the *methods of moments*, based on a traditional application of the law of large numbers. Another is the *method of maximum entropy*, based on the introduction of a measure of uncertainty or entropy and developed primarily by C.E. Shannon and E.T. Jaynes [54, 90].

6.2 The Maximum-Likelihood Method

The method of maximum likelihood is based on the postulate that the values of the unknown parameters of a distribution are those that yield the maximum probability of observing the measured data. The method therefore requires the specification of a distribution for the variable of interest, and data that contain its measurements.

6.2.1 Maximum-Likelihood Methods for a Gaussian Variable

Consider a random variable X with a Gaussian distribution of unknown mean and variance and N independent measurements x_i . The probability of obtaining a measurement between x_i and $x_i + dx$ from the parent distribution is

$$f(x_i)dx = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i - \mu)^2}{2\sigma^2}} dx,$$

where μ and σ^2 are the unknown parameters of the distribution. It is convenient to forgo the use of the differential dx and simply say that the probability of making the measurement is $P(x_i) = f(x_i)$, without specifying the interval of the random variable used in the calculation of the probability. Using the property of independence among measurements, the quantity

$$\mathcal{L} = \prod_{i=1}^N P(x_i) = \prod_{i=1}^N \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i - \mu)^2}{2\sigma^2}} \quad (6.1)$$

is the combined probability of occurrence of the N independent measurements and referred to as the *likelihood* of the data with the given model. The method of maximum likelihood consists, therefore, of finding the parameters of the distribution that maximize this likelihood. This is simply achieved by finding the point where the first derivative of the likelihood with respect to the relevant parameter of interest vanishes, to find the extremum of the function. It can be easily proven that the second derivative with respect to the two parameters is negative at the point of extremum, and therefore this is a point of maximum for the likelihood function.

To find the maximum-likelihood estimate of the mean and variance of the Gaussian distribution, it is convenient to proceed with the calculation of the derivatives of the logarithm of the likelihood, given that the logarithm is a monotonic function. The logarithm has the advantage of ease of computation of the derivatives with respect to the parameters μ and σ^2 ,

$$\begin{cases} \frac{\partial}{\partial\mu} \sum_{i=1}^N \frac{(x_i - \mu)^2}{2\sigma^2} = 0 \\ \frac{\partial}{\partial\sigma^2} \left(N \ln \frac{1}{\sqrt{2\pi\sigma^2}} \right) + \frac{\partial}{\partial\sigma^2} \left(\sum_{i=1}^N -\frac{1}{2} \frac{(x_i - \mu)^2}{\sigma^2} \right) = 0. \end{cases}$$

The solutions of the two equations lead to the two maximum-likelihood estimates of the parameters,

$$\begin{cases} \hat{\mu}_{ML} = \frac{1}{N} \sum_{i=1}^N x_i = \bar{x} \\ \hat{\sigma}_{ML}^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2. \end{cases} \quad (6.2)$$

These results are somewhat familiar. The maximum-likelihood estimates of the mean are simply the sample mean of the measurements, and therefore the method yields a familiar statistic that is known to be unbiased. The maximum-likelihood estimate of the variance, however, requires a careful look. In fact, in its derivation it was assumed that the true parent mean μ was known, and that is the reason why the second of Eq. (6.2) contains the parent mean. There may be, in principle, cases when the parent mean μ is known, and one desires to estimate the variance of the Gaussian distribution using the maximum-likelihood method: in those cases, the second of (6.2) is indeed the correct and unbiased estimator. However, in the more common situation of a simultaneous estimation of both mean and variance, one needs to replace the parent mean with its maximum-likelihood estimate, and this is consistent with regarding the two equations of the maximum-likelihood method as a coupled set of equations in the two unknowns μ and σ^2 . In this case, the second equation needs to be replaced with

$$\hat{\sigma}_{ML}^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2,$$

which is, however, a *biased* estimate of the parent mean. In fact, it was shown in Sect. 2.6 that the sample variance is an unbiased estimator of the parent variance, and therefore the maximum-likelihood estimator $\hat{\sigma}_{ML}^2$ would require a denominator of $N - 1$, and not N , to be unbiased. Fortunately, this problem is of little consequence, since one can always adjust the result by using the denominator that yields an unbiased estimator. This discussion, however, should prompt the analyst to address critically the results of the method, and when possible to ensure that estimators are unbiased.

6.2.2 Maximum-Likelihood Estimate of the Gaussian Mean for Non-uniform Uncertainties

It is sometimes the case that the data obtained to estimate a variable contain observations that are made with non-uniform uncertainties. This situation may occur in experimental situations in which one cannot repeat the experiment under the same conditions. An example is the measurement of the number of particles detected by a Geiger counter in different time intervals. Clearly one cannot directly compare the number of counts detected during different time intervals, and even the rates of arrival will have different uncertainties because a larger number of particles in a longer time intervals are subject to smaller fluctuations.

Example 6.1 (*Measurement of variables with non-uniform uncertainties*) A Geiger counter is used to measure the rate of arrival of a certain species of particles. One measurement consists of 100 counts in 10 s, another of 180 particles in 20 s, and one of 33 particles in 3 s. The measured count rates would be reported as, respectively, 10.0, 9.0, and 11.0 counts per second. Given that this is a counting experiment, the Poisson distribution applies to each of the measurements. Moreover, since the number of counts is sufficiently large, it is reasonable to approximate the Poisson distribution with a Gaussian, with variance equal to the mean. Therefore, the variance of the counts is 100, 180, and 33, and the variance of the count rate can be calculated by the property that $\text{Var}[X/t] = \text{Var}[X]/t^2$, where t is the known time of each measurement. It follows that the standard deviation σ of the count rates are, respectively, 1.0, 0.67, and 1.91 for the three measurements. The three measurements would be reported as 10.0 ± 1.0 , 9.0 ± 0.67 , and 11.0 ± 1.91 , with the last measurement being especially of lower precision because of the shorter period of observation. ◇

The situation of non-uniform measurement uncertainties, such as the one in the previous example, can be modeled by assuming that all the measurements follow a distribution function with same mean but different variance. This situation is especially relevant to normally distributed variables that feature the variance as a parameter. Following this line of reasoning, each of the N measurements can be considered as drawn from a variable $X_i \sim N(\mu, \sigma_i^2)$, with all the measurements sharing the same unknown parent mean. Assuming that the parent variances are known and equal to their measured values, the likelihood of the data can be written as

$$\mathcal{L} = \prod_{i=1}^N \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{(x_i - \mu)^2}{2\sigma_i^2}}. \quad (6.3)$$

The method of maximum likelihood can now be used to seek an estimate of the common parent mean. Following the same procedure as in Sect. 6.2.1, the maximum-likelihood estimator is

$$\hat{\mu}_{ML} = \frac{\sum_{i=1}^N \frac{x_i}{\sigma_i^2}}{\sum_{i=1}^N \frac{1}{\sigma_i^2}}, \quad (6.4)$$

which is the *weighted sample mean* of the measurements, with weights equal to the inverse of the variance of each measurement. In (6.4), it was necessary to assume that the parent variances σ_i^2 are equal to the estimates from the measurements themselves. This approximation is one of necessity, since it is not possible to estimate the mean and the different variances with the available data. This approximation can be justified if the actual precision of each measurement is known beforehand by some other means, for example, because the apparatus used for the experiment has been calibrated by prior measurements.

It is immediate to see that the weighted sample mean is an unbiased estimator of the parent mean, since $E[\hat{\mu}_{ML}] = \mu$. Moreover, its variance can be calculated as

$$\text{Var}(\hat{\mu}_{ML}) = \frac{\sum_{i=1}^N \frac{\text{Var}(x_i)}{\sigma_i^4}}{\left(\sum_{i=1}^N \frac{1}{\sigma_i^2}\right)^2} = \frac{\sum_{i=1}^N \frac{1}{\sigma_i^2}}{\left(\sum_{i=1}^N \frac{1}{\sigma_i^2}\right)^2},$$

which leads to the simple result

$$\text{Var}(\hat{\mu}_{ML}) = \frac{1}{\sum_{i=1}^N \frac{1}{\sigma_i^2}}. \quad (6.5)$$

It is also clear that the weighted sample mean from Gaussian measurements has a normal distribution, since it is a linear combination of Gaussian variables (see Sects. 4.3 and 4.4). The variance of the weighted mean on (6.5) becomes the usual σ^2/N if all variances σ_i^2 are identical.

6.3 The Maximum-Likelihood Method for the Poisson and Other Distributions

The method of maximum likelihood can be applied to any distribution. A common case in the analysis of data is that of N measurements n_i , $i = 1, \dots, N$, from a Poisson variable of parameter μ , applicable to all situations in which the measurements are derived from a counting experiment. In this case, the maximum-likelihood method can be used to estimate μ , which is the mean and variance of the random variable, and the only parameter of the Poisson distribution. Given the discrete nature of the Poisson distribution, the likelihood of making N independent measurements is simply given by

$$\mathcal{L} = \prod_{i=1}^N \frac{\mu^{n_i}}{n_i!} e^{-\mu}.$$

Working with logarithms,

$$\ln \mathcal{L} = -N \mu - \sum_{i=1}^N \ln n_i! + \ln \mu \sum_{i=1}^N n_i,$$

and requiring that the derivative of the likelihood with respect to μ is null leads to

$$\frac{1}{\mu} \sum_{i=1}^N n_i - N = 0.$$

The maximum-likelihood estimator of the μ parameter of the Poisson distribution is therefore

$$\hat{\mu}_{ML} = \frac{1}{N} \sum_{i=1}^N n_i. \quad (6.6)$$

This result was to be expected, since μ is the mean of the Poisson distribution, and the linear average of N measurements is an unbiased estimate of the mean of a random variable, according to the law of large numbers.

Example 6.2 Continuing with Example 6.1 and the assumption that the three count rates are normally distributed, Eqs. 6.4 and 6.5 yield a weighted count rate of 9.44 ± 0.53 . These data have a more direct means to obtain their count rate by using the Poisson distribution with an unknown count rate. By combining the counts for a total of 313 counts in 33 s, the count rate is 9.48 ± 0.54 , where the uncertainty in the count rate is given by the square root of the number of counts. This estimate is justified by the maximum-likelihood estimator of the Poisson parameter according to (6.6) for the single combined observation. This method to estimate the count rate should be preferred to the previous one based on the normal approximation, although the two results are virtually identical. Notice that the different times of observations make it such one cannot use (6.6) for the three individual observations in this example, since for different times of observation the parent Poisson means are different.

◇

The maximum-likelihood method can, in general, be used for any type of distribution, following the usual procedure of evaluating the likelihood of the data with the parent distribution and then taking derivatives with respect to the unknown parameters. An example with the exponential distribution is provided in Exercise 6.5.

6.4 Method of Moments

The method of moments takes a more practical approach to the estimate of the parameters of a distribution function. The basic idea is to obtain as many equations as there are free parameters in the distribution of the random variable X of interest, and then solve for the parameters of the distribution. The equations are obtained with the use of data in the usual form of N observations of the random variables. The functions $a_j(x)$ to be used for this purpose are in principle arbitrary, but they should make the distribution function integrable,

$$E[a_j(X)] = \int_{-\infty}^{\infty} a_j(x) f(x) dx = g_j(\theta), \quad (6.7)$$

so that $g_j(\theta)$ is an analytic function of the parameters θ of the distribution. Although (6.7) assumes that the random variable is continuous, the method can also be applied to discrete distributions. For example, $\theta_1 = \mu$ and $\theta_2 = \sigma^2$ for a Gaussian distribution or $\theta = \mu$ for a Poisson distribution are the parameters of the distribution that need to be estimated. Equation 6.7 is the crux of the method of moments. In fact, according to the law of large numbers, the left-hand side can be approximated by the sample mean of the function of the N measurements, and it is therefore readily calculated from the available measurements. When the integral on the right-hand side yields an analytical solution in the form of a function $g_j(\theta)$, the equation yields equations of the type

$$\frac{1}{N}(a_j(x_1) + \dots + a_j(x_N)) = g_j(\theta) \quad (6.8)$$

which can be solved for the parameters of the distribution function. Notice that one set of observations of the random variable, x_1, \dots, x_N , can be used for multiple functions a_j , typically as many as there are free parameters that need to be estimated. The reason for the name *method of moments* is that often the functions are chosen as a power of the random variable, so that the integral in (6.7) is a moment of the distribution.

As an illustration of the method, consider a Gaussian random variable with the two parameters of its distribution to estimate. First one needs to choose two functions $a_1(x)$ and $a_2(x)$. A simple and logical choice is to use $a_1(x) = x$ and $a_2(x) = x^2$, so that the right-hand side of (6.7) is, respectively, the first- and second-order moments, both with simple analytical forms. The method of moments therefore yields two equations

$$\begin{cases} E[a_1(x)] = \frac{1}{N}(x_1 + \dots + x_N) = \mu \\ E[a_2(x)] = \frac{1}{N}(x_1^2 + \dots + x_N^2) = \sigma^2 + \mu^2, \end{cases} \quad (6.9)$$

where the quantities on the right-hand sides are the unknown parameters of the Gaussian distribution, and the left-hand side quantities are measured from the data. The estimators for mean and variance are therefore

$$\begin{cases} \hat{\mu}_{MM} = \frac{1}{N}(x_1 + \dots + x_N) \\ \hat{\sigma}_{MM}^2 = \frac{1}{N}(x_1^2 + \dots + x_N^2) - \left(\frac{1}{N}(x_1 + \dots + x_N)\right)^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \hat{\mu}_{MM})^2. \end{cases} \quad (6.10)$$

For the Gaussian distribution, the method of moment therefore yields the same estimate as the maximum-likelihood method.

The method of moments for a Poisson distribution is naturally applied using $a(x) = x$, so that

$$E[x] = \frac{1}{N}(x_1 + \dots + x_N) = \mu,$$

where μ is the only parameter of the Poisson distribution. The method therefore leads to the usual estimate of the parameter as the sample mean of the N measurements, same as the maximum-likelihood method. This method is often easier computationally than the method of maximum likelihood, since it does not require the maximization of a function, but just a careful choice of the integrating functions $a_j(x)$. An application to the exponential function is provided in Exercise 6.5.

Example 6.3 Consider the five measurements presented in Example 5.1, namely, 0.1, 0.3, 0.5, 0.7, and 0.9, and assume that they are known to be drawn from a uniform distribution between 0 and a . The method of moments can be used to estimate the parameter a of the distribution from the measurements. The probability distribution function is $f(x) = 1/a$ between 0 and a , and null otherwise. Using the integrating function $a_1(x) = x$, the method of moments proceeds with the calculation of the first moment of the distribution,

$$\mathbb{E}[X] = \int_0^a x f(x) dx = \frac{a}{2}.$$

Using (6.9), the parameter a of the uniform distribution function is estimated as

$$\hat{a} = 2 \times \frac{1}{N} \sum_{i=1}^N x_i = 1,$$

where $N = 5$, for the result of $a = 1$. This estimate confirms that the five measurements are compatible with a parent mean of $1/2$. \diamond

6.5 Method of Maximum Entropy

Another method of estimation makes use of the principle that the distribution of a random variable should maximize the degree of “uncertainty,” given the constraints provided by the available data. This idea was developed primarily by E.T. Jaynes [54] and it makes use of a definition of the amount of uncertainty or *entropy* of a discrete distribution function in the form of

$$H = -K \sum_i p_i \ln p_i, \tag{6.11}$$

where p_i represents the probability mass function of the discrete variable. This quantity was derived by C.E. Shannon [90] with the intent to describe, in his own words, the “amount of uncertainty represented by this distribution function.” Given that uncertainty is at the root of the fields of probability and statistics, it appears quite reasonable that a distribution should retain as much uncertainty as possible. The entropy defined by Shannon is identical to the quantity of the same name that describes the

second law of thermodynamics. It is therefore quite suggestive that as the natural direction of a physical process tends to maximize entropy according to the second law of thermodynamics, so does the probability distribution of a random variable.¹

The method of maximum entropy is usually applied to a discrete random variable X with n possible values and a probability distribution defined by unknown probabilities p_i . The data are in the form of the expectation of a given function $f(x)$,

$$\mathbb{E}[f(x)] = \sum_{i=1}^n p_i f(x_i), \quad (6.12)$$

with the usual normalization condition for the distribution,

$$\sum_{i=1}^n p_i = 1. \quad (6.13)$$

Unlike the maximum-likelihood method, the maximum entropy method does not require to specify the functional form for the distribution. The method aims to find a set of probabilities p_i that are consistent with the given constraints while maximizing the entropy. The general form for the solution is given by

$$p_i = e^{-\lambda - \mu f(x_i)}, \quad (6.14)$$

where the unknowns λ and μ can be obtained by first defining a *partition function*

$$Z(\mu) = \sum_{i=1}^n e^{-\mu f(x_i)},$$

which is then used to derive the unknowns via the following two equations:

$$\begin{cases} \mathbb{E}[f(x)] = -\frac{d}{d\mu} \ln Z(\mu), \\ \lambda = \ln Z(\mu). \end{cases} \quad (6.15)$$

In light of the second of (6.15), it is convenient to re-write the general solution for the probability distribution as

$$p_i = \frac{e^{-\mu f(x_i)}}{Z} \quad (6.16)$$

¹ A derivation of the entropy according to (6.11) can be found in [54, 90]. A derivation based on the Boltzmann definition of entropy can be found in most textbook on statistical physics and thermodynamics, for example, pp. 61–62 of [67].

which clarifies the role of the partition function as a normalization of the probability distribution function. When using the form (6.16), the only additional condition that needs to be satisfied is the observational constraint provided by (6.12). It is possible to generalize this distribution when there is more than one constraint in the form of (6.12), as described in [54].

A proof of these equations is obtained with the method of Lagrange's multipliers, whereby the function

$$F = H + \mu \left(\sum_{i=1}^n p_i f(x_i) \right) + \lambda \left(\sum_{i=1}^n p_i \right)$$

is maximized, subject to the two constraints (6.12) and (6.13). The quantities μ and λ appearing in the function F are the two so-called Lagrange's multipliers, or unknown constants that can be specified based on the values of the two constraints. The solution is obtained by setting the derivative of F with respect to p_i to zero, which leads to

$$-K \ln p_i - K + \lambda + \mu f(x_i) = 0.$$

Since λ and μ are arbitrary numbers, they can be redefined so that the unknown probabilities have the form of (6.14). The two constraints are used to determine the values of the unknown constants. For this purpose, it is convenient to define the quantity

$$Z(\mu) = \sum_{i=1}^n e^{-\mu f(x_i)}$$

as a sum over all possible values of the random variable. This is the same function that, in statistical mechanics, is referred to as the partition function of a system in thermal equilibrium. Starting with (6.12), it is immediate to see that the average value of $f(x)$ is given by the negative of the logarithmic derivative of the partition function, as in the top equation of (6.15). (In statistical mechanics, this formula is used to calculate the average energy of the system, whereby μ is substituted by the thermal parameter β .) Use of (6.13) also immediately leads to the bottom equation of (6.15).

Example 6.4 (*Probabilities of a two-level variable with known expectation*) Consider a binary random variable X with probabilities p_1 and $p_2 = 1 - p_1$, with two possible events $x_1 = 0$ and $x_2 = 1$. The observational constraint on this variable is that the expectation of X is known to be $E[X] = \alpha$, where α is a given number. This constraint could be derived by multiple observations of the experiment or by other theoretical information on the variable. The data can be used to derive the values of the probabilities by using (6.12) as

$$E[X] = \frac{0 + e^{-\mu}}{1 + e^{-\mu}} = \alpha$$

leading to $e^{-\mu} = \alpha/(1 - \alpha)$. First, notice that the solution requires that $0 \leq \alpha \leq 1$, consistent with the possible values of the random variable. This leads to the general solution of

$$p_1 = 1 - \alpha$$

which is equal to $p_1 = 1/2$ when the expectation is $\alpha = 1/2$, corresponding to a uniform distribution, or $p_1 = 1$ when $\alpha = 0$, which corresponds to the situation of a quantity that is always equal to $x = 0$.

With regard to the uniform distribution obtained when the expectation is known to be $1/2$, it is useful to notice that a similar result would be obtained using the method of maximum likelihood using data with the same sample mean. In that case, however, one would first have to *assume* that the two-level variable has a uniform distribution. The method of maximum entropy, on the other hand, makes no such assumption, but rather infers that the distribution of maximum entropy is the uniform distribution. ◇

The method of maximum entropy has the attractive feature that it constrains the random variables while allowing as much uncertainty as possible. Jaynes' words are especially apt to describe the method:

The principle of maximum entropy may be regarded as an extension of the principle of insufficient reason (to which it reduces in case no information is given except enumeration of the possibilities x_i), with the following essential difference. The maximum-entropy distribution may be asserted for the positive reason that it is uniquely determined as the one which is maximally noncommittal with regard to missing information, instead of the negative one that there was no reason to think otherwise. Thus the concept of entropy supplies the missing criterion of choice which Laplace needed to remove the apparent arbitrariness of the principle of insufficient reason, and in addition it shows precisely how this principle is to be modified in case there are reasons for "thinking otherwise". Mathematically, the maximum-entropy distribution has the important property that no possibility is ignored; it assigns positive weight to every situation that is not absolutely excluded by the given information.

It is immediate to see that the method of maximum entropy naturally explains the uniform distribution when no measurements are available. In fact, in the absence of constraints in the form of (6.12), the maximization of the entropy with the only remaining constraint (6.13) leads to the equation $-K \ln p_i - K + \lambda = 0$, which has the simple solution of constant p_i values, i.e., a uniform distribution. The method of maximum entropy is therefore a well-motivated alternative to the more common maximum-likelihood method that can be used when there is no prior assumption on the shape of the distribution. The method is applicable to discrete distributions, although a generalization for continuous distributions is also available (see, e.g., [55]).

Summary of Key Concepts for this Chapter

Maximum-likelihood method: A method to estimate parameters of a distribution under the assumption that the parameters maximize the likelihood of the measurements.

Method of moments: A method to estimate parameters of a distribution function using expectations of known functions of the variable.

Method of maximum entropy: A method to estimate the distribution of discrete variables that maximizes the uncertainty or entropy of the variable, consistently with available constraints.

Problems

6.1 Consider the weighted sample mean of N independent and identically distributed Gaussian variables, as defined in (6.4).

- (a) Show that it is normally distributed.
- (b) Derive its variance and show that it is given by (6.5).

6.2 ■ Consider the data from Mendel's experiment in Table 1.1.

- (a) Calculate the standard deviation in the measurement of each of the seven fractions of dominants. For this purpose, it is necessary to use the parent distribution for the number of dominants.
- (b) Calculate the weighted mean and standard deviation for the seven fractions, with the weights given by the inverse of the variances as in (6.4).
- (c) Compare your result from a direct calculation of the overall fraction of dominants, obtained by grouping all dominants from the seven experiments together, and your answer from the weighted mean of part (b).

6.3 The Mendel experiment of Table 1.1 can be described as the measurement of n_i , the number of plants that displays the dominant character, out of a total of N_i plants, for each of the seven characters. The experiment is described by a binomial distribution with a probability $p = 0.75$ that the plant displays the dominant character. Using the properties of the binomial distribution, show analytically that the weighted sample mean of the measurements of the fraction $f_i = n_i/N_i$, with weights equal to the inverse of the *binomial* variance, is equal to the value calculated directly as

$$\mu = \frac{\sum_{i=1}^n n_i}{\sum_{i=1}^n N_i},$$

where $n = 7$ is the number of measurements available.

6.4 Consider a decaying radioactive source observed in a time interval of duration $T = 15$ s, where N is the number of total counts and B is the number of background counts in the same time interval (assumed to be measured independently of the total counts), measured as

$$\begin{cases} N = 19 & \text{counts} \\ B = 14 & \text{counts.} \end{cases}$$

The goal is to determine the probability of a positive detection of source counts $S = N - B$ in the time interval T .

- (a) Calculate the probability of a detection of source counts directly via:

$$\text{Prob(detection)} = \text{Prob}(S > 0 / \text{data})$$

in which S is treated as a Gaussian random variable with mean and variance calculated according to error propagation formulas. Justify why the Gaussian approximation *may* be appropriate for the variable S .

- (b) Use the same method as in (a), but assuming that the background B is known without error. This would be the situation if the background was observed for a much long time interval and then its count rate rescaled to a 15 s interval, with a negligible uncertainty.
- (c) Assume now that the background is a variable with a parent mean of 14 counts in a 15 s interval, and that it can be observed for an interval of time $T \gg 15$ s. After the observation of B for a time T , the number of background counts in a 15 s interval is B_{15} , with an estimated uncertainty of σ_{B15} . Find what interval of time T makes the error σ_{B15} of the average background over a time interval of 15 s have a relative uncertainty of 1% (i.e., $\sigma_{B15}/B_{15} = 0.01$), so that the background error is negligible compared to the uncertainty in the total counts.

6.5 Consider a random variable X with an exponential distribution with rate parameter λ , and N independent measurements x_i .

- (a) Show that the maximum-likelihood method gives the following estimate for the rate parameter:

$$\hat{\lambda} = \frac{N}{\sum_{i=1}^N x_i}.$$

- (b) Show that the method of moments provides the same estimate for the rate parameter of the distribution.

Chapter 7

Methods of Inference and Confidence Intervals of Random Variables



Abstract Confidence intervals describe a range of the variable that is estimated to have a given probability of occurrence. When the distribution is known, a confidence interval is immediately calculated from the mathematical properties of the probability distribution. A more complex task is the estimation of confidence intervals for parameters of a distribution based on available data. The method of fiducial inference makes it possible to estimate confidence intervals of unknown parameters based on the distribution of a statistic used to estimate the parameter. Bayesian methods of inference use prior distributions to determine the posterior distribution of a parameter based on the information provided by the data.

7.1 Quantiles and Confidence Intervals

The distribution function of a random variable can be used to determine the range of values that include a given probability. This range, called *confidence interval*, can be found using the quantiles of the distribution function (see Sect. 5.3). Figure 7.1 illustrates a $p = 0.90$ or 90% confidence interval for an exponential distribution, obtained as the range between the 0.05 quantile and the 0.95 quantile. The choice of probability included in a confidence intervals depends on the application. In certain applications, it is common to use 68.3% confidence intervals because this is the probability between $\pm\sigma$ of the mean for a Gaussian variable (see Sect. 7.3). Normally a confidence interval or limit at a significance lower than 68% is not considered interesting, since there is a significant probability that the random variable will be outside of this range. Commonly used confidence intervals include a 90, 95, or 99% probability (or $p = 0.90, 0.95$, and 0.99), although any choice is possible, depending on the application.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-981-19-0365-6_7.

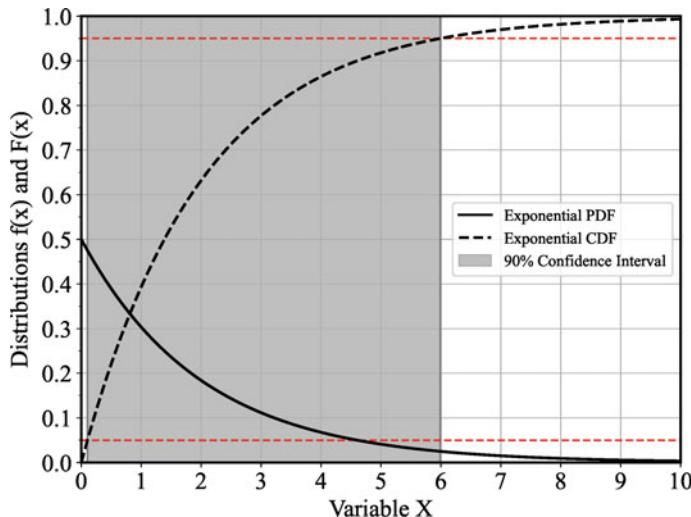


Fig. 7.1 Probability distribution function $f(x)$ (PDF) of an exponential variable with its central 90% confidence interval. The confidence interval is obtained by the intersections of the horizontal lines with the cumulative distribution function $F(x)$ (CDF)

Confidence intervals of the type illustrated in Fig. 7.1 are often referred to as *central confidence intervals* because they use symmetric quantiles (e.g., the 0.05 and the 0.95 quantiles). For a symmetric distribution such as a Gaussian, central confidence intervals are centered on the mean of the distribution. This type of confidence intervals is the most commonly used, but they are not the only intervals at a given level of probability p . For example, another $p = 0.90$ or 90% confidence interval spans from the 0.025 to the 0.925 quantile, and another spans the 0.075 to 0.975 quantiles. Central confidence intervals are usually smaller in size than non-central intervals and are the most commonly used. The largest value of a one-sided confidence interval that extend down to the 0 quantile, i.e., the lowest value allowed for that random variable, is referred to as an *upper limit*. For example, the value of $x = 6$ in Fig. 7.1 is an upper limit at the $p = 0.95$ level of probability, with the meaning that 95% of random values from that distribution are expected to fall below this value. Likewise, an interval that extend to the 1 quantile, or to the highest value allowed, defines a *lower limit*. Upper and lower limits are useful to describe how large or how small a random variable is expected to be, at a given level of probability.

7.2 Fiducial Inference

It is useful at this point to remind the distinction between a variable with known distribution from a statistic that is a function of the data. Both are random variables; but in the case of a statistic such as the sample mean, the parent values of its distri-

bution are generally unknown. The data are used to estimate the parameters of the distribution of the statistic, and therefore it becomes meaningful to also study the distribution and confidence intervals of the *parameters* themselves. This is the topic treated in this section.

The data analyst is often required to use measurements of a random variable to make inferences on the parameters of a distribution. Two examples are N measurements from a Gaussian distribution to estimate the parameters μ and σ^2 , or a single measurement of n_{obs} counts from a Poisson variable to estimate the parameter μ of the distribution. These estimation problems were already introduced, for example, using the maximum-likelihood method of Sect. 6.2.2 to provide a so-called *point estimate* for the mean of a Gaussian distribution from N measurements. The reason for such phrase is that no attempt was made to constrain the range of possible values for that parameter. It is now useful to also investigate confidence intervals at a given level of probability or confidence. Applications to Gaussian and Poisson distributions are explicitly worked out in this section, although similar methods can be used for estimating parameters of other distribution.

R.A. Fisher was instrumental in establishing a framework for the determining confidence intervals on model parameters. The method is numerically identical to the determination of confidence intervals for variables with known distributions, where one simply determines a range of possible values that are expected to occur with a given probability, as illustrated in the previous section. On the other hand, the statistical inference of possible values of an unknown parent quantity is confronted with the somewhat logical issue that a parent quantity is assumed to be fixed, and as such it is not meaningful to determine its probability distribution. Yet, being unknown, it is meaningful to ask the question of where or within which interval the parent quantity may be found. R.A. Fisher discusses this topic in a 1935 paper “*The fiducial argument in statistical inference*” [33], from which the following key passage that describes the argument of fiducial inference is excerpted:

This form of argument leads in certain cases to rigorous probability statements about the unknown parameters of the population from which the observational data are a random sample, without the assumption of any knowledge respecting their probability distributions *a priori*. For such deductions we need to know the exact sampling distributions of statistical estimates, calculable from the observations only, of the unknown parameters, and these distributions must be continuous.

The basic idea is to start with the use of a statistic that can be estimated from the data, with the property that it is an unbiased estimator of a parent quantity, for example, using the sample mean as an unbiased estimator of the μ parameter of a Gaussian distribution. Fisher’s fiducial inference argument then proposes, e.g., to turn a confidence interval of the sample mean around the unknown parent mean into an interval of the unknown parent mean around the measured sample mean. The method of fiducial inference of model parameters is described in quantitative terms below, with applications to Gaussian and Poisson variables.

7.3 Confidence Intervals for a Gaussian Variable

Central confidence intervals for a Gaussian variable are obtained according to (3.11), which is repeated here for convenience, with the notation of p as the enclosed probability,

$$p = P(|X - \mu| \leq z\sigma) = \int_{\mu-z\sigma}^{\mu+z\sigma} f(x)dx,$$

where $f(x)$ is the probability distribution function of a Gaussian with mean μ and variance σ^2 . The number z represents the number of standard deviations allowed by the interval in each direction (positive and negative relative to the mean), and it does not need to be an integer. Common central confidence intervals for a Gaussian distribution are reported in Table 7.1. When the distribution is known, confidence intervals describe the probability of occurrence of the variable. For example, for a mean μ and variance σ^2 , the interval from $\mu - \sigma$ to $\mu + \sigma$ (also indicated as $\mu \pm \sigma$) is a 68.3% confidence interval and the interval from $\mu \pm 1.65\sigma$ is a 90% confidence interval.

When the parameters of the distribution are not known, a common method to estimate them is to use data in the usual form of N independent measurements. The first step is to determine what data-based statistics to use in the estimate of the Gaussian parameters. In this case, one uses the general result that the sample mean and the sample variance are unbiased estimators of the parent parameters μ and σ^2 . The estimated quantities $\hat{\mu} = \bar{x}$ and $\hat{\sigma}^2 = s^2$ are unbiased point estimates of the unknown parent values, and it is therefore common practice to use these two statistics to determine confidence intervals for the random variable X . Approximate confidence intervals for the random variable are therefore obtained according to Table 7.1. The measurements of the X variable are therefore commonly reported as the $p = 0.9$ confidence interval

$$X = \bar{x} \pm 1.65s, \quad (7.1)$$

Table 7.1 Common confidence intervals for a Gaussian distribution

Interval	Range	Enclosed probability p
50% confidence interval	$\mu \pm 0.68\sigma$	50%
1- σ interval	$\mu \pm \sigma$	68.27%
90% confidence interval	$\mu \pm 1.65\sigma$	90%
2- σ interval	$\mu \pm 2\sigma$	95.45%
3- σ interval	$\mu \pm 3\sigma$	99.73%
4- σ interval	$\mu \pm 4\sigma$	99.99%
5- σ interval	$\mu \pm 5\sigma$	$\geq 99.9999\%$

or $X = \bar{x} \pm s$ for a $p = 0.68$ confidence interval. The meaning of this confidence interval at confidence level p is that one expects a measurement within $\pm 1.65 \hat{\sigma}$ of the estimated mean $\hat{\mu}$ to occur approximately 90% of the time, in an experiment conducted in the same way as the one that produced the data at hand.

It is also useful to ask the related but different question of what constraints the data place on the parent mean of X , since this quantity is often the one of primary interest to the analyst. This question is at the core of statistical inference, since it aims to use data to constrain parent parameters, and not just to give a probability of observation of values from a known variable. To answer this question, consider that the parent mean is estimated via the sample mean, and therefore the distribution of the sample mean needs to be used. The sample mean has a distribution that is itself normal, as a linear combination of independent normal variables, and with $E[\bar{x}] = \mu$ and $\text{Var}(\bar{x}) = \sigma^2/N$ (see Sect. 4.4.1). The expectations of X and \bar{x} are the same, but the variance of the latter is a factor of $1/N$ smaller than that of the former, and therefore the sample mean is constrained to an interval that is smaller by a factor of $1/\sqrt{N}$, compared to that of a single measurement. Now, constraining the sample mean around an unknown value of the parent mean is not useful from a practical point of view. What would be useful is to constrain the *parent* mean. Using the distribution of the sample mean to constrain the parent mean, and similar inferences on parent quantities, is subject of debate among statisticians. Fisher's concept of fiducial probability for a parent quantity is used for this task.

When using the sample mean \bar{x} to estimate the unknown parent mean μ , the foregoing discussion illustrated the known result that

$$\bar{x} \sim N(\mu, \sigma^2/N),$$

where N is the number of measurements of the variable X . One therefore expects that measurements of the sample mean follow that distribution, i.e., that 90% of measurements of the sample mean fall within $\pm 1.65(\sigma/\sqrt{N})$ of the unknown parent mean. This means that, e.g., for a probability $p = 0.9$,

$$P\left(-1.65 \leq \frac{\bar{x} - \mu}{\sigma/\sqrt{N}} \leq 1.65\right) = p. \quad (7.2)$$

So far, this expression reflects a traditional interpretation of the confidence interval of a random variable. There are two unknowns in this expression, since both mean and variance are parent quantities. The presence of the additional unknown σ^2 further complicates matters. For the time being, it is convenient to simply assume that the parent variance is accurately estimated by the sample variance, and use $\hat{\sigma}^2/N = s^2/N$ as the variance of the sample mean. The additional uncertainty associated with the variability of the sample variance will be discussed in Sect. 9.6.2. The fiducial inference argument turns the previous interval into one for the parent mean, since the previous interval can also be written as an interval for the parent mean,

$$P(\bar{x} - 1.65(\hat{\sigma}/\sqrt{N}) \leq \mu \leq \bar{x} + 1.65(\hat{\sigma}/\sqrt{N})) = p, \quad (7.3)$$

lending support to the argument that there is a 90% probability that the parent mean is in this range of values. This is Fisher's fiducial probability argument for a parent parameter: converting the traditional confidence interval for a random variable of (7.2) into a fiducial probability for the parent (but unknown) parameter according to (7.3). The approximation of the unknown parent variance with the sample variance introduces an additional source of uncertainty that will be later accounted by using the student t distribution. Notice how the use of the fiducial argument leads to a confidence interval similar to (7.1), except that it is now interpreted as a range of possible values of the parent mean. Other considerations on the fiducial inference can also be found in the textbook by M. Bulmer [16].

Example 7.1 (*Confidence intervals for m/e in Thomson's experiment*) Using the data for the Thomson experiment on the electron's mass-to-charge ratio $x = m/e$, the sample mean and standard deviation for Tube 1 can be reported as $\bar{x} = \hat{\mu} = 0.415$ and $s = \hat{\sigma} = 0.071$, for short $m/e = 0.415 \pm 0.071$, and for Tube 2 as $m/e = 0.524 \pm 0.056$. Assuming that the variable m/e is normally distributed, 90% confidence intervals can be obtained as the range within ± 1.65 estimated standard deviations of the estimated mean, respectively, $(0.297, 0.533)$ for Tube 1 and $(0.432, 0.616)$ for Tube 2. These two confidence intervals simply state that *a given measurement* has a 90% probability of being in the given interval, according to the available data.

A different estimate can be made for the parent mean of the distribution of m/e . In this case, the point estimate of the parent mean is the sample mean, which has already been calculated for both Tubes. Assuming a normal distribution for the measurements, the sample mean has itself a normal distribution with the same mean and with a variance that is a factor of $1/N$ smaller than the parent variance of the distribution, as shown in Sect. 4.2.2. Using the fiducial probability argument, the intervals

$$0.415 \pm 0.071/\sqrt{N_1} = 0.415 \pm 0.021 \text{ (Tube 1)}$$

with $N_1 = 12$ measurements for Tube 1 and

$$0.524 \pm 0.056/\sqrt{N_2} = 0.524 \pm 0.017 \text{ (Tube 2)},$$

with $N_2 = 11$ measurements for Tube 2 represent the intervals where the parent means ought to be found, with 90% probability. \diamond

7.4 Upper and Lower Limits for a Gaussian Variable

Upper and lower limits for a Gaussian distribution can be defined according to the following relationships as a means to identify semi-infinite intervals that contain a specified probability p ,

Table 7.2 Common upper and lower limits for a Gaussian distribution

Upper limit	Range	Enclosed probability (%)	Lower limit	Range	Enclosed probability (%)
50% confidence	$\leq \mu$	50	50% confidence	$\geq \mu$	50
90% confidence	$\leq \mu + 1.28\sigma$	90	90% confidence	$\geq \mu - 1.28\sigma$	90
95% confidence	$\leq \mu + 1.65\sigma$	95	95% confidence	$\geq \mu - 1.65\sigma$	95
99% confidence	$\leq \mu + 2.33\sigma$	99	99% confidence	$\geq \mu - 2.33\sigma$	99
1- σ	$\leq \mu + \sigma$	84.1	1- σ	$\geq \mu - \sigma$	84.1
2- σ	$\leq \mu + 2\sigma$	97.7	2- σ	$\geq \mu - 2\sigma$	97.7
3- σ	$\leq \mu + 3\sigma$	99.9	3- σ	$\geq \mu - 3\sigma$	99.9

$$p = P(x \leq x_{up}) = \int_{-\infty}^{x_{up}} f(x)dx = F(x_{up}) \quad (\text{upper limit } x_{up}) \quad (7.4)$$

$$p = P(x \geq x_{lo}) = \int_{x_{lo}}^{\infty} f(x)dx = 1 - F(x_{lo}) \quad (\text{lower limit } x_{lo}).$$

The quantities $F(x_{up})$ and $F(x_{lo})$ are the values of the cumulative distribution of the Gaussian, showing that x_{up} is the p -quantile and x_{lo} is the $(1-p)$ -quantile of the distribution. Common upper and lower limits for the Gaussian distribution are reported in Table 7.2. Upper limits are typically of interest when the measurements result in a low value of the variable and the analyst wants to know how high the variable can be and still be consistent with the measurement, at a given level of probability. For example, in the case of the measurement of Tube 1 for the Thomson experiment, the variable m/e was measured to be 0.415 ± 0.071 . In this case, it is interesting to ask the question of how high can m/e be and still be consistent with the data.

Example 7.2 (*Upper and lower limits for m/e in Thomson's experiment*) The Thomson data of Tables 2.1 and 2.2 can be used to calculate the upper and lower limits for the variable $x = m/e$ for Tube 1 and Tube 2. The $p = 0.90$ or 90% upper limit $x_{up} = \mu + 1.28\sigma$ means that the random variable x is expected to have values such that

$$P(x \leq \mu + 1.28\sigma) = p,$$

and, when considering the mean of N measurements,

$$P(\bar{x} \leq \mu + 1.28\sigma/\sqrt{N}) = p.$$

Using Fisher's fiducial probability argument, with the same probability it is approximately true that

$$P(\mu \geq \bar{x} - 1.28 \hat{\sigma}/\sqrt{N}) = p,$$

where

$$\mu_{lo} = \bar{x} - 1.28 \hat{\sigma}/\sqrt{N}$$

can now be viewed as the 90% fiducial lower limit on the parent mean, where the variance has been approximated by the sample variance, $\hat{\sigma}^2 = s^2$. The same argument can be made for the lower limit, therefore resulting in upper and lower limits on the parent mean that are formally identical to those in Table 7.2. Using the mean and variances calculated in Example 7.1, the fiducial 90% lower limits on the mean of the parent distribution for variable m/e from Table 7.1 can be reported as

$$\mu_{lo} = \overline{m/e} - 1.28 \hat{\sigma}/\sqrt{N_1} = 0.389.$$

Similar calculations can be made for the upper limit and for the values in Table 7.2. ◇

A common application of upper limits is when an experiment has failed to detect the variable of interest, usually referred to as a *non-detection*. In this case, the analyst still wants to place limits on possible values of the variable, based on the measurements made. This problem is typically addressed by considering the parent distribution of the variable that was not detected. This distribution usually has a mean of zero, and a variance based on the properties of the measurements. For example, a counting experiment that resulted in no counts from a source of interest can be modeled as a Poisson distribution with a parent mean given by the expected counts due to all sources of background. The following example provides an illustration of how to set upper limits from the non-detection of a quantity of interest.

Example 7.3 (*Gaussian upper limit to a non-detection with background*) This example is adapted from the analysis of an astronomical source that resulted in a non-detection of photons from a galaxy [13]. In a given time interval T , a detector recorded $n = 8$ counts when pointed toward a source of unknown intensity. The instrument used for the measurement has a background level with a mean of $\mu_B = 9.8 \pm 0.4$ counts for the given time interval, as estimated from an independent experiment of longer duration. Given that the measurement is below the expected background level, it is evident that there is no positive detection of the source, and therefore this is a case of a non-detection.

The hypothesis that the source has no emission can be modeled by assuming that the random variable X that describes the total counts recorded in a time T has a mean of approximately 9.8 counts (ignoring the uncertainty in the background level). Moreover, since this is a counting experiment, the probability distribution should be Poisson, and therefore it is reasonable to posit that $X \sim \text{Poisson}(\mu_B)$. For simplicity, the distribution is approximated with a Gaussian of same mean and variance equal to the mean, or $\sigma \simeq \sqrt{\mu_B} = 3.1$.

A 99% upper limit to the number of counts that can be recorded in the absence of genuine source counts, according to the parent distribution of X , is

$$X \leq \mu_B + 2.33\sigma \simeq 17.$$

The direct interpretation of this model is that there is a 99% probability of observing a total number of counts less than about 17, when the source does not contribute any counts; $n = 8$ total counts detected are clearly consistent with this hypothesis. Any other experiment detecting a total of $n \leq 17$ counts should be regarded as a non-detection. The variable Y describing the source counts is related to the total measured counts X and the background B as

$$Y = X - B.$$

After subtraction of the background, which was assumed constant, the analyst can conclude that the 99% upper limit to the source's true emission level in the given time interval is

$$Y \leq 2.33\sigma \simeq 8.2 \text{ counts}.$$

This upper limit depends entirely on the properties of the background, which is the only factor at play in the measurement if the source does not have any emission. A complementary way to interpret this number is that the analyst can be 99% confident that the measurement *cannot* be due to just the background *if* there is a detection of > 17 total counts. In this case, the analyst can make the positive conclusion that emission from the source was detected, at this level of confidence.

A more accurate analysis should include the possibility that the Gaussian distribution has a slightly higher mean, since the level of the background is not known exactly, and conservatively assume that perhaps 18 counts are required to establish that the source does have a positive level of emission. Alternatively, the uncertainty in the background can be accounted via error propagation, so that the variance of X becomes larger than what is assumed in this example. \diamond

7.5 Confidence Intervals for the Mean of a Poisson Variable

This simplest and most fundamental case to consider for the study of a Poisson variable X of unknown mean μ is when a measurement of n_{obs} counts from this distribution is recorded. This single datum can be used to estimate and constrain the parent mean of the distribution. The analytical form of the Poisson distribution and the discrete nature result in a different method to estimate confidence intervals on the parent mean, compared to a normal distribution. It is reasonable to assume that, given the only measurement available, the estimate of the source mean is

$$\hat{\mu} = n_{obs},$$

for example, by using the maximum-likelihood method for $N = 1$ observation. This estimate is the starting point to determine a confidence interval for the parent mean. Consider a value of the Poisson mean μ_{lo} that is sufficiently small to result in the observation of $X < n_{obs}$ with a high probability p , i.e.,

$$p = P(X < n_{obs}) = \sum_{n=0}^{n_{obs}-1} \frac{\mu_{lo}^n}{n!} e^{-\mu_{lo}}. \quad (7.5)$$

The mean μ_{lo} corresponds to the situation shown in Fig. 7.2: assuming that the parent mean is as low as μ_{lo} , there is only a small probability $\alpha = 1 - p$ (e.g., $\alpha = 0.05$ if $p = 0.95$) to make a measurement equal or greater to what was actually measured. This means that the cumulative distribution has a value of p for $x = n_{obs} - 1$. Following the fiducial probability argument, one can turn this probability statement into the fiducial probability for finding the parent mean, and say that there is only a small chance $\alpha = 1 - p$ that the actual mean could be even lower than this value, or

$$P(\mu \leq \mu_{lo}) = 1 - p. \quad (7.6)$$

The quantity μ_{lo} is therefore defined as the *lower limit* with a confidence level p . It is important to notice that the lower limit μ_{lo} is defined solely by (7.5), with the meaning implied by that equation. The use of this lower limit into (7.6) results from the use of Fisher's fiducial inference argument, which associates a probability or confidence level p to an unknown parent parameter, in this case the mean μ of the Poisson distribution.

By the same logic, the *upper limit* μ_{up} is defined as the parent mean that results in the observation of $n > n_{obs}$ with a probability p , or

$$p = P(X > n_{obs}) = \sum_{n=n_{obs}+1}^{\infty} \frac{\mu_{up}^n}{n!} e^{-\mu_{up}}. \quad (7.7)$$

This upper limit is also illustrated in Fig. 7.2, whereby the corresponding cumulative distribution has a value of $1 - p$ at $x = n_{obs}$. Assuming that the mean is as high as μ_{up} , there is only a small probability of $1 - p$ to make a measurement equal or lower than the actual measurement. Using again the fiducial argument, this upper limit can be used to state that

$$P(\mu \geq \mu_{up}) = 1 - p.$$

The lower and upper limits can be combined to provide a confidence interval for the parent mean. According to the foregoing considerations, the fiducial probability that the parent mean is between μ_{lo} and μ_{up} is

$$P(\mu_{lo} \leq \mu \leq \mu_{up}) = P(\mu \geq \mu_{lo}) - P(\mu \geq \mu_{up}) = p - (1 - p) = 2p - 1,$$

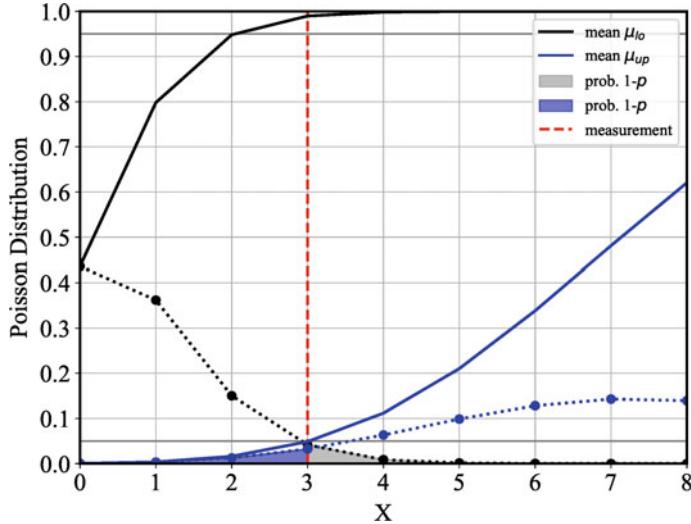


Fig. 7.2 Illustration of the upper and lower limits to the Poisson mean, assuming a measurement of $n_{obs} = 3$. The lower limit to the parent mean corresponds to a distribution (in black) that has a probability p to measure $n < n_{obs}$, as shown by the cumulative distribution (solid line). The distribution with the upper limit to the mean (in blue) has a probability p to measure $n > n_{obs}$.

i.e., for a confidence level $p = 0.95$ for each of the two limits, the range between them has a fiducial probability of 0.90 of containing the true Poisson mean.

7.6 The Gehrels Approximation for Poisson Upper and Lower Limits

A useful approximation for the solution of (7.5) and (7.7) to find the upper and lower limits to the Poisson mean was provided by N. Gehrels [38]. The solution is obtained by specifying the number of observed counts n_{obs} and the level of confidence or probability p associated with the upper and lower limits. This level of confidence is described in an equivalent way by a *Poisson parameter* S which is the number of Gaussian standard deviations that correspond to the confidence level chosen (for example, $p = 0.84$ corresponds to $S = 1$). The approximations are given by

$$\begin{cases} \mu_{up} = n_{obs} + \frac{S^2 + 3}{4} + S\sqrt{n_{obs} + \frac{3}{4}} \\ \mu_{lo} = n_{obs} \left(1 - \frac{1}{9n_{obs}} - \frac{S}{3\sqrt{n_{obs}}}\right)^3. \end{cases} \quad (7.8)$$

Table 7.3 Poisson parameters S and corresponding probabilities

Upper or lower limit	Range	Probability p (%)	Poisson S parameter
90% confidence	$\leq \mu + 1.28\sigma$	90	1.28
95% confidence	$\leq \mu + 1.65\sigma$	95	1.65
99% confidence	$\leq \mu + 2.33\sigma$	99	2.33
$1-\sigma$	$\leq \mu + \sigma$	84.1	1.0
$2-\sigma$	$\leq \mu + 2\sigma$	97.7	2.0
$3-\sigma$	$\leq \mu + 3\sigma$	99.9	3.0

The S parameter is also the p -quantile of a standard Gaussian distribution, enclosing the probabilities reported in Table 7.3. The reason for the use of this S parameter in place of the probability p is the ease of comparison between confidence intervals for the Poisson and normal distributions.

The upper and lower limits defined by (7.5) and (7.7) can be approximated analytically using a relationship that relates the Poisson sum with an analytic distribution function:

$$\sum_{n=0}^{n_{obs}-1} \frac{e^{-\mu} \mu^n}{n!} = 1 - P_{\chi^2}(\chi^2, \nu), \quad (7.9)$$

where $P_{\chi^2}(\chi^2, \nu)$ is the cumulative distribution of the χ^2 probability distribution function defined in Sect. 9.3, with parameters $\chi^2 = 2\mu$ and $\nu = 2 n_{obs}$,

$$P_{\chi^2}(\chi^2, \nu) = \int_{-\infty}^{\chi^2} f_{\chi^2}(x, \nu) dx.$$

Use of (7.9) into (7.5) and (7.7) gives a relationship between the function P_{χ^2} and the probability p ,

$$\begin{cases} P_{\chi^2}(2\mu_{lo}, 2 n_{obs}) = 1 - p \\ P_{\chi^2}(2\mu_{up}, 2 n_{obs} + 2) = p. \end{cases} \quad (7.10)$$

The simplest choice is to use the approximation for the function P_{χ^2} that provides limits within 10% of the true values. The approximation makes use of the following definitions: for every value of p , y_p is the p -quantile of a standard normal distribution,

$$F(y_p) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{y_p} e^{-t^2/2} dt = p.$$

If $P_{\chi^2}(\chi_p^2, \nu) = p$, with χ_p^2 the p -quantile of the χ^2 distribution, then the simplest approximation between χ_p^2 and y_p given by [38] is

$$\chi_p^2 \simeq \frac{1}{2} \left(y_p + \sqrt{2\nu - 1} \right)^2. \quad (7.11)$$

Consider now the upper limit in (7.10). One can solve for μ_{up} by using (7.11) with $2\mu_{up} = \chi_p^2$, $\nu = 2n_{obs} + 2$, and $S = y_p$, since y_p is the p -quantile of a standard normal distribution, thus equivalent to S . It follows that

$$2\mu_{up} = \frac{1}{2} \left(S + \sqrt{4n_{obs} + 3} \right)^2$$

and from this the top part of (7.8) follows after a simple algebraic manipulation. A similar result applies to the lower limit. In this case, Gehrels suggests the following approximation between χ_p^2 and y_p :

$$\chi_p^2 \simeq \nu \left(1 - \frac{2}{9\nu} - y_p \sqrt{\frac{2}{9\nu}} \right)^3. \quad (7.12)$$

For the lower limit, this approximation replaces (7.11). The lower limit μ_{lo} defined in (7.10) means that (7.12) can be used with $\chi_{1-p}^2 = 2\mu_{lo}$, $y_{1-p} = -S$ (due to the symmetry of the standard Gaussian) and $\nu = 2n_{obs}$, immediately leading to the second approximation in (7.8).

Upper and lower limits according to (7.8) are tabulated in Tables A.5 and A.6 for several interesting values of n_{obs} and S . A few cases of common use are also shown in Table 7.4. An useful and interesting situation that has an analytical solution is the case of a complete *non-detection* of a source, $n_{obs} = 0$. In this case, it is only meaningful to solve (7.7) in search for an upper limit with a given confidence p . In this case of $n_{obs} = 0$, the equation simplifies to

$$1 - p = e^{-\mu_{up}}$$

which leads to the exact solution of

$$\mu_{up} = -\ln(1 - p). \quad (7.13)$$

For $p = 0.84$ or $S = 1$, this corresponds to an upper limit of $\mu_{up} = 1.83$. This number is usually referred to as the *1- σ upper limit to a non-detection*, and it is particularly useful to describe the approximate uncertainty of a Poisson measurement that resulted in no detection of counts. This case can also be used to test the accuracy of the Gehrels approximation which, for $S = 1$ and $n_{obs} = 0$, yields an estimate of $\mu_{up} \simeq 1.87$ according to (7.8),

Table 7.4 Selected Upper and Lower limits for a Poisson variable using the Gehrels approximation (see Tables A.5 and A.6 for a complete list of values)

n_{obs}	Poisson parameter S or confidence level		
	$S = 1$	$S = 2$	$S = 3$
	(1- σ , or 84.1%)	(2- σ , or 97.7%)	(3- σ , or 99.9%)
<i>Upper limits</i>			
0	1.87	3.48	5.60
1	3.32	5.40	7.97
2	4.66	7.07	9.97
3	5.94	8.62	11.81
...			
<i>Lower limits</i>			
...			
3	1.37	0.58	0.17
4	2.09	1.04	0.42
...			
9	6.06	4.04	2.52
10	6.90	4.71	3.04
...			

$$\mu_{up} = 1 + \sqrt{\frac{3}{4}} = 1.87.$$

This estimate is just 2% higher than the exact result according to (7.13). An example of upper limits in the presence of a non-zero background is presented in Problem 7.4.

7.7 Bayesian Methods of Inference

Bayesian methods of inference consist of determining the posterior probability of a random variable using both the likelihood and prior probabilities, according to Bayes' theorem (1.11). The likelihood is a well-understood concept. For example, assuming a Poisson-distributed variable X , the likelihood of one measurement with $x = n_{obs}$ is

$$P(n_{obs}/\mu) = \frac{\mu^{n_{obs}}}{n_{obs}!} e^{-\mu}, \quad (7.14)$$

where the notation $P(n/\mu)$ highlights the fact that the parent mean μ was assumed as a given quantity. The likelihood is thus just the probability of occurrence of data, given a model. The analyst, however, is interested in making statistical inference on possible values of the parent model. For example, in the case of a Poisson variable,

it is interesting to know what values of the parent mean μ are possible. According to Bayes' theorem, this posterior distribution of μ , given the data, is proportional to the product of the likelihood and a prior probability distribution,

$$P(\mu/n_{obs}) \propto P(n_{obs}/\mu) \times P(\mu). \quad (7.15)$$

The use of a prior distribution is what constitutes the Bayesian approach. The prior distribution is intended to reflect the analyst's knowledge of the unknown quantity before the data are collected. It is inevitable that there is an element of subjectivity in the choice of a prior, since the understanding of an unknown quantity cannot possibly be complete and unequivocal prior to making a measurement. A commonly used prior is the uniform distribution over a fixed range of the unknown quantity, which aims to simply exclude unreasonable values from further consideration. Significant research on the use of priors was conducted by H. Jeffreys in his textbook "*Theory of Probability*" [56]. A quote from this textbook is useful to highlight the spirit of a Bayesian analysis:

If we have no information relevant to the actual value of a parameter, the probability must be chosen so as to express the fact that we have none. It must say nothing about the value of the parameter, except the bare fact that it may be possibly, by its very nature, be restricted to lie within certain definite limits.

A prior probability distribution may differ from a uniform distribution, also for the purpose to avoid normalization problems over infinite intervals. Jeffreys discusses this subject at length while establishing a framework for what are usually referred to as *non-informative priors*.

7.7.1 Bayesian Expectation of the Poisson Mean

It is of particular interest to apply Bayesian methods to the estimation of the parent mean of a Poisson distribution, in the presence of one observation with n_{obs} counts. The posterior distribution of the Poisson mean μ depends on the likelihood (7.14) and on a choice of the prior distribution according to (7.15),

$$P(\mu/n_{obs}) = \frac{P(n_{obs}/\mu)P(\mu)}{\int_0^\infty P(n_{obs}/\mu)P(\mu)d\mu},$$

where it is recognized that $\mu \geq 0$ is the entire range allowed to the parent mean. The infinite range for the possible values of the parent mean poses a challenge when choosing the prior distribution, since a simple uniform distribution cannot be normalized over an infinite range. This difficulty, also pointed out by Jeffreys [see Sect. 3.1 of [56]], can however be overcome by choosing, for example, a finite range between 0 and M and then consider the asymptotic results. In this case, the expectation of the random variable becomes

$$E[\mu/obs] = \frac{\int_0^M e^{-\mu} \mu^{n_{obs}+1} d\mu}{\int_0^M e^{-\mu} \mu^{n_{obs}} d\mu}, \quad (7.16)$$

where the constant values of the uniform prior have canceled out. It is possible to show that the asymptotic limit of this expectation is $E[\mu/obs] = n_{obs} + 1$, when $M \rightarrow \infty$.

The *Gamma or factorial function* is defined as

$$\Gamma(z) = \int_0^\infty x^{z-1} e^{-x} dx \quad (7.17)$$

with the property that, for integer arguments, $\Gamma(n) = (n - 1)!$. When the limit of integration extends to a finite value instead, the integral is referred to as the *incomplete Gamma function*,

$$\gamma(z, M) = \int_0^M x^{z-1} e^{-x} dx. \quad (7.18)$$

For integral values of z , the incomplete Gamma function can be written as

$$\gamma(n+1, M) = n! \left(1 - e^{-M} \sum_{i=0}^n \frac{M^i}{i!} \right). \quad (7.19)$$

The expectation (7.16) is therefore

$$E[\mu/n_{obs}] = \frac{\gamma(n_{obs} + 2, M)}{\gamma(n_{obs} + 1, M)}$$

which, in the asymptotic limit of a large M , yields the result of $n_{obs} + 1$, since the asymptotic limit of the ratio of the terms in parenthesis in the incomplete gamma function is unity. The same value of the posterior expectation is obtained by-passing entirely the incomplete gamma function and simply setting the expectation as a ratio of the Gamma functions, which leads to the same result. This method, however, implies the cancellation of the uniform priors, which are not properly normalized over an infinite range.

The interesting result is therefore that a measurement of n_{obs} counts implies a Bayesian expectation of $E[\mu] = n_{obs} + 1$, i.e., one count more than the observation, when using a uniform prior for the parent mean. Therefore, even a non-detection results in an expectation for the mean of the parent distribution of 1, and not 0. This somewhat surprising result can be understood by considering that even a parent mean

of 1 results in a likelihood of $1/e$ (i.e., a relatively large number) of obtaining zero counts as a result of a random fluctuations, and that the Poisson distribution is skewed. This calculation of the Poisson expectation is due to A.G. Emslie and A.M. Massone [26]. A uniform prior distribution is not the only choice, and a different expectation is obtained for other choices of the prior (see Prob. 7.6).

7.7.2 Bayesian Confidence Intervals for a Poisson Variable

The Bayesian posterior distribution of the parent mean can also be used to determine confidence intervals. Using a uniform prior, the upper limit to the source mean with probability p is given by

$$p = \frac{\int_0^{\mu_{up}} P(n_{obs}/\mu) d\mu}{\int_0^\infty P(n_{obs}/\mu) d\mu} = \frac{\int_0^{\mu_{up}} \mu^{n_{obs}} e^{-\mu} d\mu}{\int_0^\infty \mu^{n_{obs}} e^{-\mu} d\mu}, \quad (7.20)$$

where it was assumed that a uniform prior distribution over an infinite range is applicable. With the use of Bayes' theorem and a uniform prior over an infinite range, the upper limit becomes a ratio of integrals of the likelihood. The numerator is the incomplete Gamma function $\gamma(n_{obs} + 1, \mu_{up})$, and the normalization constant at the denominator is given by the Gamma function $\Gamma(n_{obs} + 1) = n_{obs}!$, so that the upper limit can be calculated according to

$$p = \frac{\gamma(n_{obs} + 1, \mu_{up})}{n_{obs}!}.$$

It is immediate to see that, according to the definition of the incomplete Gamma function (7.19), the Bayesian upper limit obtained with the choice of a uniform distribution over an infinite range and the “fiducial” upper limit of (7.7) are identical. Similarly, the Bayesian lower limit with a uniform prior and a probability p can be calculated according to

$$1 - p = \frac{\int_0^{\mu_{lo}} \mu^{n_{obs}} e^{-\mu} d\mu}{\int_0^\infty \mu^{n_{obs}} e^{-\mu} d\mu} = \frac{\gamma(n_{obs} + 1, \mu_{lo})}{n_{obs}!}, \quad (7.21)$$

which is again the same as the lower limit of (7.5). The range $\mu_{lo} \leq \mu \leq \mu_{up}$ therefore contains a posterior probability of $2p - 1$, e.g., a 90% probability if the two limits have a 95% probability.

The difference in the method of calculating the fiducial upper limits described by (7.5) and (7.7) and the Bayesian limits of (7.20) and (7.21) is summarized by the different variable of integration or summation in the relevant equations. The classical limits use the Poisson probability to make n_{obs} measurements for the unknown parent mean μ . The limits of the confidence interval are then calculated as the values of the mean that give a probability of p to observe fewer or more counts than those recorded. In this case, the probability is evaluated as a sum over the number of counts, for a fixed value of the parent mean. In the case of the Bayesian limits, on the other hand, the posterior distribution of μ is evaluated with the choice of a uniform distribution over an infinite range, and this distribution is then integrated in the usual manner to obtain confidence intervals. In general, the two methods will give different results, with the Bayesian method being sensitive to the choice of a prior. The specific choice of a uniform prior over an infinite range makes it such that the fiducial limits and the Bayesian limits are the same.

Summary of Key Concepts for this Chapter

Confidence Intervals: The range that contains a probability of occurrence p for a random variable.

Fiducial Inference: A method to derive confidence intervals for parent parameters from a sample statistic.

Upper and Lower limits: Values of the variable that contain a probability of occurrence p in a semi-infinite range ending at that value.

Bayesian Methods of Inference: Methods of inference on model parameters that make use of Bayes' theorem and use a prior probability distribution.

Problems

7.1 ■ Consider the Thomson experiments of Table 2.1 (Tube 1) and Table 2.2 (Tube 2). Calculate:

- The 90% central confidence intervals for the variable v .
- The 90% upper and lower limits for the variable v , assuming that the variable follows a normal distribution.

7.2 Consider a Poisson variable X of mean μ for which a measurement of $n_{obs} = 1$ count is recorded. The goal of this problem is to set an upper limit for the parent Poisson mean.

- Following the classical approach, find the equation that determines the $p = 0.9$ or 90% upper limit to the mean μ_{up} . Recall that the classical upper limit with probability p is defined as the value of the Poisson mean that yields a probability $P(X > n_{obs}) = p$.
- Using the Bayesian approach, which consists of defining the upper limit via

$$p = \frac{\int_0^{\mu_{up}} P(n_{obs}/\mu) d\mu}{\int_0^{\infty} P(n_{obs}/\mu) d\mu},$$

find the equation that determines the 90% upper limit to the mean μ_{up} , and show that it is identical to that from part (a).

- Show that a numerical solution for μ_{up} from parts (a) and (b) is consistent with the Gehrels approximation in (7.8).

7.3 ■ The data provided in Table 2.3 from Pearson's experiment on biometric data describes the cumulative distribution function of heights from a sample of 1,079 couples. Assuming that the Poisson statistic applies to the number of couples, calculate the 2σ upper limit to the fraction of couples in which both mother and father are taller than 68 in.

7.4 The data presented in Example 7.3 illustrate the non-detection of a source in the presence of background, where $n = 8$ counts were detected in the presence of a background that is described by a Poisson distribution with mean $\mu_B = 9.8$.

- Calculate the 99% confidence Poisson upper limit to the number of source counts *above* the background level.
- Calculate the upper limit to the non-detection of Poisson counts assuming a zero background level and compare to the upper limit from part (a).

7.5 ■ Consider the Thomson experiment analyzed in Example 7.2. Determine the lower limit to the parent mean of the m/e variable for Tube 2 at the 90% confidence level.

7.6 A Poisson variable X results in a measurement of n_{obs} counts. Show that, using a gamma distribution with parameters α and r (see Appendix A.3) as a prior on the Poisson mean, the expectation is

$$E[\mu] = \frac{n_{obs} + r}{1 + \alpha}.$$

Notice how this expectation can differ significantly from the number of observed counts (see [26]).

Chapter 8

Average Values of Random Variables



Abstract The data analyst often faces the question of what is the “best” value to report from the measurement of a random variable. This chapter investigates the use of the sample mean, the weighted sample mean, the median, and a weighted logarithmic average that may be useful when a variable has errors that are proportional to their measurements, avoiding the inherent bias arising in the weighted sample mean from measurements with small values and small errors. A relative-error weighted average is also introduced as an approximation for the weighted logarithmic average.

8.1 Point Estimates and Average Values

A *point estimate* is usually defined as a single number that is intended to be as representative as possible of the quantity of interest. For example, the sample mean (2.8) or the weighted sample mean (6.4) can be considered as point estimates of a random variable. A point estimate is obtained from a statistic or *estimator*, i.e., the sample mean is an estimator and its measured value is the point estimate. It is usually possible to go beyond a single point estimate of a variable, since the sampling distribution of the statistic, such as the distribution of the sample mean, provides a measure of the uncertainty associated with the measurement. Therefore, whenever possible, a point estimate should be accompanied by an estimate of the uncertainty associated with its value, according to its sampling distribution.

The term *average value* is loosely defined as a representative value of a random variable and does not indicate a specific mathematical operation. The sample mean and the weighted sample mean can be regarded as average values of a variable, each to be preferred according to the nature of the variable. This chapter introduces additional average values that are relevant under certain conditions.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-981-19-0365-6_8.

8.2 Linear and Weighted Averages

The method of maximum likelihood indicates that the weighted sample mean is the most likely value of the mean of a normal random variable when measurements have different variance (see Sect. 6.2.2). Therefore, the weighted sample mean is a commonly accepted quantity to report as the best estimate for the value of a measured quantity. If the measurements have same variance, then the weighted mean becomes the usual sample mean. It is quite common to use the generic term *average* in place of *mean*, and therefore refer to the sample mean as the linear average and to the weighted sample mean as the weighted average.

The difference between the linear and weighted average can be illustrated with the measurements shown in Table 8.1, which reports $N = 25$ measurements of various quantities [11]. This dataset is illustrative of a common situation of quantities with different measurement errors. Notice moreover that the “errors” are reported in asymmetrical intervals around the reported values. For the present analysis, it is sufficient to assume that a reported value $0.86 \pm^{0.22}_{0.18}$ means that 0.86 is the measurement and 0.20 (the average of the two errors) is its standard deviation, and that the measurements are normally distributed. The weighted sample mean of the N measurements of the variable ratio can be calculated as 0.90, while the sample mean is 1.01 (see Problem 8.1). The difference between these two averages can be traced to a few measurements with a low value of the ratio that carry higher weight because of the small measurement error (for example, source 15).

Which of the two values is more representative? This question can be addressed by making the following observations. The measurement errors reported in the table reflect the presence of such sources of uncertainty as Poisson fluctuations in the detection of photons from the source. This type of uncertainty is usually referred to as *statistical error*, with the broad meaning that this is a type of uncertainty that is unavoidable, given the nature of the random variable being measured. Experiments and measurements are typically also subject to other sources of uncertainty that may not be explicitly reported in the data. For example, the measurement of events recorded by a detector is affected by the calibration of the detector, and a systematic offset in the calibration would affect the numbers recorded. In the case of the data of Table 8.1, the uncertainty due to the calibration of the detector is likely to be affected by the same amount of all measurements, regardless of the precision indicated by the statistical error. This type of uncertainty is typically referred to as *systematic error*, and the inclusion of such additional source of uncertainty would modify the value of the weighted average, but not the linear average. As an example of this effect, if an error of ± 0.1 is added to all the reported errors, the weighted mean becomes 0.95 (see Problem 8.2). It is clear that the addition of a constant error for each measurement causes a de-weighting of datapoints with small statistical errors, and in the limit of a large systematic error the weighted mean becomes the linear average. Therefore, the linear average can be used when the data analyst wants to weigh equally all datapoints, regardless of the precision indicated by the statistical errors. Systematic errors are discussed in more detail in Chap. 17. This example highlights the need

Table 8.1 Dataset with measurement of energy for $N = 25$ different sources and their ratio, from [11]

Source	Radius	Energy		
		Method #1	Method #2	Ratio
1	$221.1 \pm^{11.0}_{12.3}$	$8.30 \pm^{0.76}_{0.88}$	$9.67 \pm^{1.14}_{1.12}$	$0.86 \pm^{0.08}_{0.07}$
2	$268.5 \pm^{22.1}_{20.7}$	$4.92 \pm^{0.77}_{0.70}$	$4.19 \pm^{0.82}_{0.70}$	$1.17 \pm^{0.16}_{0.15}$
3	$138.4 \pm^{12.7}_{11.9}$	$3.03 \pm^{0.53}_{0.49}$	$2.61 \pm^{0.59}_{0.49}$	$1.16 \pm^{0.20}_{0.18}$
4	$714.3 \pm^{23.5}_{34.5}$	$49.61 \pm^{3.15}_{3.19}$	$60.62 \pm^{4.84}_{6.13}$	$0.82 \pm^{0.06}_{0.05}$
5	$182.3 \pm^{18.5}_{15.1}$	$2.75 \pm^{0.49}_{0.43}$	$3.30 \pm^{0.81}_{0.61}$	$0.83 \pm^{0.14}_{0.14}$
6	$72.1 \pm^{5.5}_{5.7}$	$1.01 \pm^{0.23}_{0.20}$	$0.86 \pm^{0.14}_{0.13}$	$1.17 \pm^{0.24}_{0.21}$
7	$120.3 \pm^{8.6}_{7.5}$	$5.04 \pm^{0.66}_{0.57}$	$3.80 \pm^{0.72}_{0.57}$	$1.33 \pm^{0.16}_{0.15}$
8	$196.2 \pm^{15.1}_{15.5}$	$5.18 \pm^{0.73}_{0.70}$	$6.00 \pm^{1.17}_{1.11}$	$0.86 \pm^{0.14}_{0.11}$
9	$265.7 \pm^{8.7}_{8.6}$	$12.17 \pm^{1.22}_{1.17}$	$10.56 \pm^{0.93}_{0.95}$	$1.14 \pm^{0.13}_{0.10}$
10	$200.0 \pm^{9.6}_{10.7}$	$7.74 \pm^{0.57}_{0.58}$	$6.26 \pm^{0.78}_{0.83}$	$1.24 \pm^{0.14}_{0.11}$
11	$78.8 \pm^{5.6}_{5.1}$	$1.08 \pm^{0.16}_{0.15}$	$0.73 \pm^{0.11}_{0.10}$	$1.49 \pm^{0.26}_{0.24}$
12	$454.4 \pm^{20.3}_{20.3}$	$17.10 \pm^{2.64}_{2.03}$	$23.12 \pm^{2.36}_{2.32}$	$0.75 \pm^{0.07}_{0.06}$
13	$109.4 \pm^{8.3}_{8.3}$	$3.31 \pm^{0.34}_{0.34}$	$3.06 \pm^{0.54}_{0.52}$	$1.09 \pm^{0.18}_{0.15}$
14	$156.5 \pm^{11.5}_{10.2}$	$2.36 \pm^{0.61}_{0.58}$	$2.31 \pm^{0.36}_{0.31}$	$1.02 \pm^{0.26}_{0.23}$
15	$218.0 \pm^{6.6}_{5.9}$	$14.02 \pm^{0.75}_{0.75}$	$21.59 \pm^{1.82}_{1.82}$	$0.65 \pm^{0.04}_{0.04}$
16	$370.7 \pm^{7.6}_{8.0}$	$31.41 \pm^{1.56}_{1.56}$	$29.67 \pm^{1.56}_{1.57}$	$1.06 \pm^{0.06}_{0.06}$
17	$189.1 \pm^{16.4}_{15.4}$	$2.15 \pm^{0.45}_{0.39}$	$2.52 \pm^{0.57}_{0.51}$	$0.86 \pm^{0.22}_{0.18}$
18	$150.5 \pm^{4.2}_{4.6}$	$3.39 \pm^{0.57}_{0.50}$	$4.75 \pm^{0.44}_{0.46}$	$0.72 \pm^{0.11}_{0.11}$
19	$326.7 \pm^{12.1}_{9.9}$	$15.73 \pm^{1.43}_{1.30}$	$18.03 \pm^{1.54}_{1.26}$	$0.87 \pm^{0.06}_{0.06}$
20	$189.1 \pm^{9.9}_{9.1}$	$5.04 \pm^{0.65}_{0.55}$	$4.61 \pm^{0.61}_{0.50}$	$1.09 \pm^{0.12}_{0.12}$
21	$147.7 \pm^{8.0}_{11.1}$	$2.53 \pm^{0.29}_{0.30}$	$2.76 \pm^{0.37}_{0.48}$	$0.93 \pm^{0.12}_{0.10}$
22	$504.6 \pm^{12.5}_{11.2}$	$44.97 \pm^{2.99}_{2.74}$	$43.93 \pm^{3.08}_{2.59}$	$1.02 \pm^{0.05}_{0.05}$
23	$170.5 \pm^{8.6}_{8.1}$	$3.89 \pm^{0.30}_{0.29}$	$3.93 \pm^{0.49}_{0.42}$	$0.98 \pm^{0.10}_{0.09}$
24	$297.6 \pm^{13.1}_{13.6}$	$10.78 \pm^{1.04}_{1.02}$	$10.48 \pm^{1.34}_{1.22}$	$1.04 \pm^{0.10}_{0.11}$
25	$256.2 \pm^{13.4}_{14.4}$	$7.27 \pm^{0.81}_{0.77}$	$7.37 \pm^{0.97}_{0.95}$	$0.99 \pm^{0.09}_{0.09}$

to further investigate possible average values of a variable, according to the type of variables and measurements at hand.

8.3 The Median

Another quantity that can be calculated from the N measurements of a variable X is the *sample median* \tilde{x} , which is defined as the midpoint of the ordered measurements if N is odd, or the mean of the two ordered midpoints if it is even. The basic idea is to find a value of the variable that leaves approximately 50% of the measurements both

to its left and to its right. In the case of the measurement of the ratios in Table 8.1, this is simply obtained by ordering the 25 measurements in ascending order, and using the 13-*th* measurement as the sample median. The value of the sample median obtained in this case is 1.02, quite close to the value of the linear average, since both statistics do not take into account the measurement errors. One useful feature of the median is that it is not very sensitive to “outliers” in the distribution. For example, if one of the measurements was erroneously reported as 0.07 ± 0.01 (instead of 0.72 ± 0.11 , e.g., source 18 in the table), both linear and weighted averages would be affected by the error, but the median would not. The median may therefore be an appropriate value to report in cases where the analyst suspects the presence of outliers in the dataset.

The *median* $\tilde{\mu}$ of a continuous random variable with probability density function f and cumulative distribution F is defined by

$$F(\tilde{\mu}) = \frac{1}{2},$$

and the sample mean defined above is the corresponding sample statistic. The distribution of the sample median is described by the *Sample Median Theorem*, which bears certain similarities with the Central Limit Theorem:

Theorem 8.1 (Sample Median Theorem) *The sample median \tilde{x} of a large sample of independent measurements from a continuous distribution with probability density f of size $N = 2m + 1$ is approximately distributed as a Gaussian. The mean of the distribution is equal to the parent median $\tilde{\mu}$, and the variance of the distribution is equal to*

$$\text{Var}(\tilde{x}) = \frac{1}{8 f(\tilde{\mu})^2 m}. \quad (8.1)$$

The median is one of the *order statistics* that can be defined from a sample of measurements, and the general form of the probability distribution function of order statistics can be found in [5, 22] or other textbooks on probability theory. The Sample Median Theorem is a useful approximation that applies to most distribution functions, and it can be proven with the help of the general distribution for order statistics.

The practical importance of this asymptotic theorem is that it provides the means to estimate an uncertainty for the sample median, when the sample size is sufficiently large. Consider the example of N measurements from a normal distribution, for which the mean coincides with the median. Since $f(\tilde{\mu})^2 = 1/(2\pi\sigma^2)$ and $m \simeq N/2$, the variance of the sample median is given by

$$\text{Var}(\tilde{x}) = \frac{\sigma^2}{N} \frac{\pi}{2}$$

which is a factor $\pi/2 \simeq 1.57$ larger than the variance of the sample mean. The fact that the variance of the sample median is somewhat larger than that of the sample mean indicates that the median has more variability than the mean. The measurements of the average values of Ratio in Table 8.1 would then be reported as follows: (a) the sample mean 1.01 ± 0.04 ; (b) the weighted sample mean 0.90 ± 0.02 ; and (c) the median 1.02 ± 0.05 . The uncertainties of these point estimates are the sample standard deviations of the estimators.

8.4 The Logarithmic Average and Fractional Errors

The quantity Ratio in Table 8.1 can be used to illustrate a type of variables that may require special attention when calculating their average. Consider a variable whose errors are proportional to their measured values. In this case, a weighted average will be skewed toward *lower* values because of the smaller errors in those measurements. This is clearly what happens with the ratio measurements, whereby a few measurements with low values also have a small error (e.g., source 15). The question is whether a weighted average is appropriate for these measurements, or whether one should use a different approach to account for their measurement errors.

To illustrate this situation, consider a fictional dataset with two measurements such as $x_1 = 1.2 \pm 0.24$ and $x_2 = 0.80 \pm 0.16$. Both measurements have a relative error of 20%, their linear average is 1.00 and the weighted average is 0.923. The natural logarithms of these measurements are $\ln x_1 = 0.1823$ and $\ln x_2 = -0.2231$. Using the error propagation method (Sect. 5.2.6), the error in the logarithm is the fractional error according to

$$\sigma_{\ln x} = \frac{\sigma_x}{x}, \quad (8.2)$$

leading to the result the error of the logarithm of the two measurements has the same value of $\sigma_{\ln x} = 0.2$. The weighted average of these logarithms is therefore equal to their linear average, for a value of $\log x = -0.0204$. When transformed back to linear numbers, this average corresponds to a value of 0.980, which is much closer to the linear average of 1.00 than to the weighted average of the measurements. The base-10 logarithm can be used instead, leading to the same results.

Errors that are exactly proportional to the measurement, or

$$\sigma_x = x \cdot \sigma_r \quad (8.3)$$

may be defined as *fractional errors*, and the constant σ_r is the relative or fractional error. In most cases, including that of Table 8.1, the relative errors vary among the measurements, and therefore (8.3) applies only as an approximation. The following is an investigation of when it is in fact advisable to use the logarithm of the measurements, instead of the measurements themselves, to obtain a more accurate determination of the average value of a variable that has fractional errors.

8.4.1 The Log-Normal Distribution

A random variable X is said to have a log-normal distribution with parameters μ and σ if its natural logarithm $Y = \ln X$ has a normal distribution with the same parameters. Equivalently, when Y is a normal variable, $X = e^Y$ is said to be a log-normal variable. Since the transformation of a variable to its logarithm is monotonic, the method of transformation of variables can be used to find the log-normal distribution function according to $g(y)dy = f(x)dx$, where $f(x)$ is the sought-after log-normal probability distribution function and $g(y)$ the usual normal distribution. Since $dy/dx = 1/x$, the log-normal distribution is

$$f(x)dx = \frac{1}{x\sqrt{2\pi\sigma^2}}e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}} dx. \quad (8.4)$$

Equation 8.4 is the probability distribution function of a log-normal variable with parameters μ and σ^2 , and it is defined for $x > 0$. The parameters μ and σ in (8.4) no longer represent the mean and variance of X , like they do for the normally distributed Y . The mean and variance of a log-normal distribution can be shown to be

$$\begin{cases} E[X] = e^{\mu + \frac{\sigma^2}{2}} \\ \text{Var}[X] = e^{2(\mu + \sigma^2)} - e^{2\mu + \sigma^2}. \end{cases} \quad (8.5)$$

The parameter σ is the *shape parameter* and e^μ is the *scale parameter* of the log-normal distribution. The *standard log-normal* distribution is obtained when Y is a standard normal, and it has a distribution

$$f(x) = \frac{1}{x\sqrt{2\pi}}e^{-\frac{\ln^2 x}{2}}. \quad (8.6)$$

Useful properties of a log-normal distribution are as follows:

- (1) If X is $\text{lognormal}(\mu, \sigma^2)$, cX is $\text{lognormal}(\mu + \ln c, \sigma^2)$. This property shows that multiplying a log-normal variable by a constant leads to a change of the scale parameter.
- (2) If X is $\text{lognormal}(\mu, \sigma^2)$, X^a is $\text{lognormal}(a\mu, a^2\sigma^2)$. This property shows that the exponential of a log-normal variable remains log-normal, with a change to both the shape and scale parameters.
- (3) If X is $\text{lognormal}(\mu, \sigma^2)$, $1/X$ is $\text{lognormal}(-\mu, \sigma)$.
- (4) The *product* of independent log-normal variables is a log-normal variable with μ and σ^2 parameters equal to the sum of the individual parameters. This result is equivalent to a similar result for the *sum* of normal variables, which retains the normal shape with the additive property for both mean and variance.

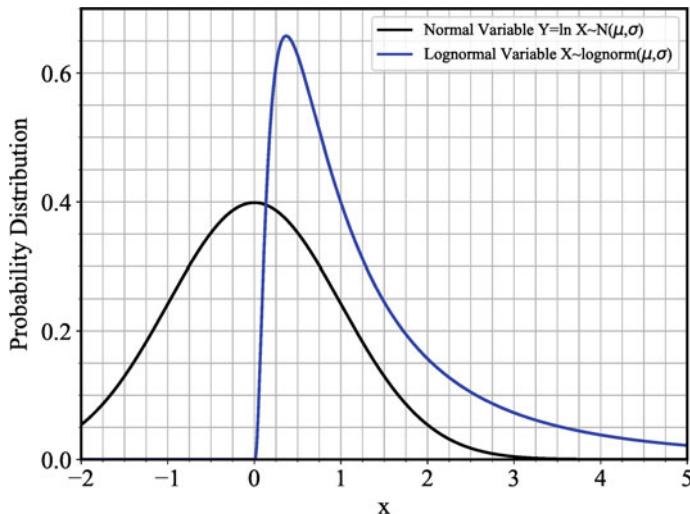


Fig. 8.1 Log-normal and normal distributions with parameters $\mu = 0$ and $\sigma = 1.0$. Positive values and a heavier right-hand tail are key feature of the log-normal distribution, compared with a normal distribution

Proofs of these properties can be found in textbooks on probability theory. Property (4) can be proved by writing $X_i = e^{Y_i}$ where $Y_i \sim N(\mu_i, \sigma_i^2)$, so that

$$X = \prod_{i=1}^N X_i = e^{\sum_{i=1}^N Y_i}.$$

Since $\sum_{i=1}^N Y_i$ has a normal distribution with mean $\sum_{i=1}^N \mu_i$ and variance $\sum_{i=1}^N \sigma_i^2$, then the product X is a log-normal variable with same parameters.

Also notice that it is possible to use a base-10 logarithm to define $Y = \log X = \ln X / \ln 10$. In this case, Y is a log-normal distribution with the same form as (8.4) but with parameters $\mu / \ln 10$ and $(\sigma / \ln 10)^2$. Therefore, all considerations for a log-normal variable apply to logarithms in any base.

Figure 8.1 illustrates the main differences between the normal and the log-normal distributions. The heavier tail at large values (i.e., a positive skewness) of the random variable is the key feature of the log-normal distribution. In practice, when a sample of measurements from a variable of unknown distribution has both properties of being positive-definite and with a positive skewness, it is a possible indication that the variable may be log-normal distributed.

8.4.2 The Weighted Logarithmic Average

The relationship between a normal variable $Y = \ln X$ and the log-normal variable $X = e^Y$ can be used for the practical purpose of estimating the parent parameters μ and σ^2 . The parameters represent the mean and variance of Y , and they are adjustable parameters for the log-normal variable X , related to its mean and variance via (8.5). Consider a log-normal variable X and data consisting of N independent measurements x_i , such that $y_i = \ln x_i$ are samples from Y . The method of maximum likelihood can be applied to these data using either the normal distribution for the measurements y_i , or the log-normal distribution for their logarithms. Following the same methods of Sect. 6.2.2 and allowing the σ_i^2 parameters to be different among the measurements, it is immediate to see that both cases return the usual maximum-likelihood estimator

$$\hat{\mu} = \frac{\sum_{i=1}^N \ln x_i / \sigma_i^2}{\sum_{i=1}^N 1 / \sigma_i^2}. \quad (8.7)$$

This is the estimator for the common parent mean of the y_i measurements, where σ_i^2 is assumed to be known and represents both the parameter of the log-normal distribution and the variance of the normally distributed $y_i = \ln x_i$ variable. The interpretation of σ_i^2 as the variance of the normally distributed variable y_i leads to a useful approximation according to the error propagation formula (8.2),

$$\sigma_i^2 \simeq \frac{\sigma_{x_i}^2}{x_i^2},$$

where $\sigma_{x_i}^2$ represents the variance of the log-normal variable that was measured. This approximate method to estimate the mean of a log-normal variable X thus defines the *weighted logarithmic average*

$$\hat{\mu} = \frac{\sum_{i=1}^N \frac{\ln x_i}{(\sigma_{x_i}/x_i)^2}}{\sum_{i=1}^N \frac{1}{(\sigma_{x_i}/x_i)^2}}, \quad (8.8)$$

where σ_{x_i} are the measured errors of the log-normal data (different from the parameter σ_i^2 , which is the variance of the associated normal variable).

The meaning of this result is the following: when measurements of a log-normal variable X are available and with different precision, the estimate of the mean $\hat{\mu}$ of the associated normal distribution $Y = \ln X$ is the weighted mean of the logarithm of the measurements, with weights equal to the square of the fractional errors. This

procedure has the advantage to weigh according to the relative error of the measurements x_i , and not to the absolute error, thus overcoming the problem of smaller measurements with smaller errors having larger weight. The estimator $\hat{\mu}$ can be considered as a “best” value of the variable $\ln X$, and thus converted to an estimator for the log-normal variable X via the corresponding exponential $\hat{x} \simeq e^{\hat{\mu}}$. This method is only approximate in that it does not attempt to be an unbiased estimator of the mean of the log-normal distribution, but it can nonetheless be useful in several practical circumstances.

The uncertainty in the weighted logarithmic average is also estimated as

$$\sigma_{\hat{\mu}}^2 = \frac{1}{\sum_{i=1}^N \frac{1}{(\sigma_{x_i}/x_i)^2}}, \quad (8.9)$$

which can be obtained following the same method as in Sect. 6.2.2. The use of this logarithmic average is justified when the variable X has a log-normal distribution, i.e., when $\ln X$ has a Gaussian distribution. In this case, the weighted logarithmic average offers an alternative to the other point estimates that de-weights measurements with small values and small errors. All results in this section also apply to logarithms in any other base, such as base-10 logarithms.

It is also useful to show that, in the limit of measurements with the same fractional error and small deviations from the mean, the weighted logarithmic average is the logarithm of the sample mean of X . In fact, when the relative errors of X are constant, (8.8) becomes the sample mean of the logarithms of the measurements. Moreover, retaining only the first-order term in the series expansion of the logarithm around the mean,

$$\ln x_i \simeq \ln \bar{x} + \frac{\Delta x_i}{\bar{x}}$$

the estimator becomes $\hat{\mu} \simeq \ln \bar{x}$, so that the exponential of the sample weighted mean approximates the linear average, or sample mean, of the variable X .

Example 8.1 (*Use of the weighted logarithmic average*) The data of Table 8.1 can be used to calculate the logarithmic average of the variable ratio according to (8.8) and (8.9) as $\hat{\mu} = -0.055 \pm 0.019$. Assuming that this is a log-normal variable, the logarithmic average represents an estimate of the logarithm of the variable, and it can be converted to linear quantities to obtain a value of 0.95 ± 0.02 . Notice that the error for the linear quantity must be obtained by expanding the range of the logarithm, and this may lead to asymmetric errors. Notice how this logarithmic average has a value that is somewhat between that of the linear average 1.01 ± 0.04 and the traditional weighted average of 0.90 ± 0.02 . It should not be surprising that the logarithmic mean is not exactly equal to the linear average. In fact, the measurements

of Table 8.1 have different relative errors. Only in the case of identical relative errors for all measurements we expect that the two averages have the same value. \diamond

8.4.3 The Relative-Error Weighted Average

Although using logarithms of measurements for the weighted logarithmic average is a simple procedure, it is useful to investigate another type of average that deals directly with the measurements of a variable X . The *relative-error weighted average* is defined as

$$\overline{x_{RE}} = \frac{\sum_{i=1}^N \frac{x_i}{(\sigma_i/x_i)^2}}{\sum_{i=1}^N \frac{1}{(\sigma_i/x_i)^2}}, \quad (8.10)$$

where the measurements are weighted by the relative errors, in the same manner as the logarithms in the case of the weighted logarithmic average of (8.8). The only difference between this average and the weighted sample mean (6.4) is the use of the extra factor of x_i in the error term, so that σ_i/x_i is the relative error of each measurement. In the limit of measurements with the same relative error and small deviations from the sample mean, this average provides a value that is the same as the exponential of the weighted logarithmic mean.

The use of the relative-error weighted average should be viewed as a convenient *ad hoc* method to obtain an average value that is consistent with the logarithmic average, especially in the limit of measurements with equal relative errors. The statistical uncertainty in this relative-error weighted average can be simply assigned as the error in the traditional weighted sample mean. In fact, the statistical error should be determined by the “physical” uncertainties in the measurements, as is the case for the variance (6.5). It would be tempting to use the inverse of the denominator of (8.10) as the variance; however, the result would be biased by our somewhat arbitrary choice of weighing the measurements by the relative errors, instead of the error themselves.

Example 8.2 Continuing with the values of Ratio in Table 8.1, the error weighted average is calculated as $\overline{x_{RE}} = 0.96$. The error in the traditional weighted average was 0.02, therefore the error weighted average can be reported as 0.96 ± 0.02 . Comparison with the value 0.95 ± 0.02 from the logarithmic average shows the general agreement between these two estimates. \diamond

Summary of Key Concepts for this Chapter

Linear average: The sample mean \bar{x} of N measurements.

Median: The 50% quantile or approximately the number below and above which there are 50% of the variable's values.

Log-normal variable: A random variable whose logarithm is normally distributed.

Logarithmic average: In some cases, e.g., when errors are proportional to the measured values or for log-normal variables, it is meaningful to calculate the weighted average of the logarithm of the variable as

$$\hat{\mu} = \frac{\sum_{i=1}^N \frac{\ln x_i}{(\sigma_{x_i}/x_i)^2}}{\sum_{i=1}^N \frac{1}{(\sigma_{x_i}/x_i)^2}}$$

Relative-error weighted average: An approximation of the logarithmic average that does not require logarithms,

$$\overline{x_{RE}} = \frac{\sum_{i=1}^N \frac{x_i}{(\sigma_i/x_i)^2}}{\sum_{i=1}^N \frac{1}{(\sigma_i/x_i)^2}}$$

Problems

8.1 ■ Table 8.1 contains the measurement of the thermal energy of certain sources using two independent methods labeled as Method 1 and Method 2, and their ratio. The error bars indicate the 68% or 1σ , confidence intervals, although the fact that most intervals are asymmetric indicate that the measurements do not follow exactly a Gaussian distribution.

- Calculate the weighted average of the ratios between the two measurements and its standard deviation, assuming that the errors are Gaussian and equal to the average of the asymmetric errors.
- Calculate the linear average of the ratios and explain why it is larger than the weighted average from (a).

8.2 ■ Consider the 25 measurements of Ratio in Table 8.1. Assume that an additional uncertainty of ± 0.1 is to be added linearly to the statistical error of each measurement reported in the table.

- Show that the addition of this source of uncertainty results in a weighted average of 0.95 ± 0.04 .
- Compare with the standard weighted average (with no additional uncertainty added) and explain the reason for the difference.

8.3 Given two measurements x_1 and x_2 with values in the neighborhood of a positive number A , show that the logarithm of the average of the measurements is approximately equal to the average of the logarithms of the measurements.

8.4 ■ For the data in Table 8.1, calculate the linear average, weighted average and median of each quantity (Radius, Energy Method 1, Energy Method 2, and Ratio). You may assume that the error of each measurement is the average of the two errors reported in the table.

8.5 ■ Calculate the weighted logarithmic average of the quantity Ratio in Table 8.1, and its uncertainty, and then convert the results back to linear scale.

Part II

**Hypothesis Testing, Regression and
Parameter Estimation**

Chapter 9

Hypothesis Testing and Fundamental Statistics



Abstract Data-based statistics are subject to the random fluctuations of the measurements, and their sampling distribution describes the probability of occurrence of the values of the statistic. The method of hypothesis testing and the associated *p*-value describe the process of using the distribution of a statistic to quantitatively test the agreement of data with a statistical hypothesis. This chapter also introduces a few fundamental statistics that play a central role in data analysis, such as the χ^2 statistic, Fisher's *F*-statistic, and *Student's t*-statistic.

9.1 Statistics and Hypothesis Testing

This book has already introduced several data-based statistics, such as the sample mean and the sample variance. These statistics are subject to random fluctuations that occur during the measurement and collection process and they are distributed according to a sampling distribution. For example, under the hypothesis that a variable X follows a Gaussian distribution of mean μ and variance σ^2 , the sample mean of N measurements is Gaussian-distributed with mean μ and variance equal to σ^2/N (see Sect. 4.2.2). This means that different samples of size N will, in general, give rise to different sample means and that ones expect a variance of order σ^2/N among the various samples. This knowledge of a statistic's sampling distribution is key to establishing whether a given sample is consistent with the parent values.

Hypothesis testing is the process that establishes whether the measurement of a given statistic, such as the sample mean, is consistent with a given theoretical distribution of interest or *hypothesis*. The process of hypothesis testing requires a considerable amount of care in the definition of the hypothesis to test, and in drawing conclusions. The method can be conveniently divided into the following four steps:

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-981-19-0365-6_9.

1. Definition of a hypothesis to test. A typical hypothesis consists of testing whether a sample of N measurements is drawn from a parent distribution for the variable X . The hypothesis therefore requires the specification of the parameters of the distribution. For example, if the analyst wants to test whether measured samples follow a standard normal distribution, the hypothesis requires the specification that X follows a standard Gaussian distribution with mean $\mu = 0$ and a variance $\sigma^2 = 1$. This initial step in the process is therefore that to identify a so-called *null hypothesis* to test with the available data. This phrase was used by R.A. Fisher in his book *The Design of Experiments* [32], with reference to the famous tea-tasting experiment, in which the null hypothesis being tested is that a person *cannot* discriminate whether the milk or the tea was added first to the cup.

2. Determination of the relevant statistic. The next step is to determine the statistic of choice to test the null hypothesis. In the example of the measurements of a variable X , the natural statistic is the sample mean. The choice of statistic means that the analyst is now in a position to use the sampling distribution for that statistic and be able to test whether the actual measurements are consistent with its expected distribution, according to the null hypothesis. There may be more than one statistic that can be used for a given hypothesis. The two main guiding principles to select a statistic are given in the following: (a) the statistic should be *unbiased* or *consistent*, meaning that the expectation of the statistic is the same as that of the parent distribution and (b) the statistic should be *efficient*, with the meaning that the statistic should have the smallest possible variance. For example, the median and the sample mean are both consistent statistics to estimate the parent mean of a normal distribution, but the sample mean has a smaller variance (see Sect. 8.3), so that in general one would choose the sample mean as the statistic of choice to determine the parent mean μ . Sometimes there are other considerations at play that may make it advisable to select a statistic that is not efficient, such as the ease of measurement of the statistic.

3. Definition of the confidence level p and the rejection region. It is necessary to quantify the degree of agreement between the statistic and its expected distribution under the null hypothesis. This *confidence level* is determined by a probability p , say $p = 0.9$ or 90 %, which defines a range of values for the statistics that are consistent with its expected distribution. For example, a standard Gaussian of zero mean and unit variance has 90 % of its values in the range between -1.65 and $+1.65$. For a confidence level of $p = 0.9$, the analyst would require that the measurement must fall within this range. The choice of probability p is somewhat arbitrary: some analysts may choose 0.9, some may require 0.99, some may even be satisfied with 0.683, which is the probability associated with the $\pm 1\sigma$ range for a Gaussian distribution.

The value of the probability p divides the range of all possible values for the statistic of choice into two regions: the *rejection region* and a complementary region with all remaining possible values of the statistic, which can be referred to as the *acceptable region* (but see the discussion below for the exact meaning of this term). Given the importance of hypothesis testing in statistics and data analysis, it is necessary to clarify the definition and meaning of these two ranges. Referring to the statistic of choice as S , the null hypothesis is usually indicated as

$$H_0 = \{\text{The statistic has values } S_1 \leq S \leq S_2\}, \quad (9.1)$$

with the boundaries of the statistic defined according to

$$p = P(S_1 \leq S \leq S_2) = \int_{S_1}^{S_2} f(s)ds, \quad (9.2)$$

where f is the probability distribution function of the statistic, under the null hypothesis; an equivalent expression holds for discrete variables. The meaning of (9.1) is that one expects values of the statistic in the range $S_1 \leq S \leq S_2$ when the null hypothesis is correct. Accordingly, values of the statistics *outside* of the range $S_1 \leq S \leq S_2$ define the rejection region. For the example of a standard Gaussian, the rejection region at $p = 0.9$ consists of values $S \geq 1.65$ and values $S \leq -1.65$, and the rejection region is said to be a *two-sided rejection region*. Other statistics may have *one-sided rejection regions*. This is often the case for positive-definite statistics, for which $S_1 = 0$, and the other limit in (9.1) $S_2 = S_{crit}$ becomes the *critical value* of the statistic. The one-sided rejection region is therefore $S \geq S_{crit}$. Clearly p and S_{crit} are related: the larger the value of the probability p , the larger the critical value. In principle, it is possible to make other choices for the rejection region, such as multiple intervals, according to the nature of the hypothesis and of the statistic. Majority of cases for the rejection region are, however, either a one-sided interval extending to infinity or a two-sided region.

4. Testing for rejection of the null hypothesis. Finally the analyst is in a position to make a quantitative statement regarding the null hypothesis. Since the range of values of the statistic has been partitioned into a rejection region and its complementary region, only two cases are possible:

(a) *Rejection of the hypothesis:* The measured value of the statistic S falls into the rejection region. This means that the distribution function of the statistic of interest, under the null hypothesis, does not allow the measured value at the confidence level p . In this case, the null hypothesis *must be rejected* at the stated confidence level p . The rejection of the null hypothesis means that the data should be tested for alternative hypotheses and the procedure can be repeated.

(b) *Non-rejection of the hypothesis:* The measured value of the statistic S is outside of the rejection region, and therefore it is said to be in the acceptable region. This means that there is a reasonable probability that the measured value of the statistic is *consistent* with the null hypothesis. In this case, the null hypothesis *cannot be rejected*, meaning that there is reasonable probability that the data are consistent with the null hypothesis. This is, however, not the same as stating that the null hypothesis is true or that it should be accepted. In fact, there could be other hypotheses that could also be acceptable, and one cannot be certain that the null hypothesis tested represents the parent model for the data. It is necessary to keep in mind that the process of hypothesis testing can only lead to a conclusive statement regarding the rejection of a specific hypothesis, at a given level of confidence. When the hypothesis cannot be rejected, it is nonetheless possible to refer to the null hypothesis as being *accept-*

able, provided that it remains clear that this statement in no way implies that the hypothesis is accepted as the correct model of the data, but simply that the data are consistent with the statistical hypothesis. The meaning of *acceptable region* is therefore intended simply as a convenient and informal short hand for the complementary of the rejection region.

It may be useful to conclude this section with a quote from R.A. Fisher's *The Design of Experiments*, who uses these words with regard to the tea-tasting experiment:

... it should be noted that the null hypothesis is never proved or established, but is possibly disproved, in the course of experimentation. Every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis.

Example 9.1 (*Method of application of hypothesis testing*) Consider $N = 5$ independent measurements of a random variable X , namely, $x_i = (10, 12, 15, 11, 13)$. The method of hypothesis testing follows these steps:

1. The hypothesis to test is that the measurements are drawn from a Gaussian random variable with $\mu = 13$ and $\sigma^2 = 2$.
2. It is now necessary to determine the test statistic to use. Since the data are in the form of N independent measurements of the same variable, a possible statistic is the *sum* of all measurements,

$$Y = \sum_{i=1}^5 X_i,$$

which is distributed like a Gaussian $N(N\mu, N\sigma^2) = N(65, 10)$. The sample mean could have been chosen instead, with a parent distribution $N(\mu, \sigma^2/N)$. It can be proven that the results of the hypothesis testing are equivalent for the two statistics. In fact, both statistics are unbiased, and they have the same relative variance.

3. The next step requires the choice of a confidence level for our hypothesis. This choice is somewhat arbitrary, and it will be illustrated with two possible values of $p = 0.95$, corresponding to a $\pm 1.96\sigma$ interval around the parent mean $\mu = 65$, and $p = 0.683$, corresponding to a $\pm 1\sigma$ interval. The corresponding rejection regions are shown by the hatched and cross-hatched regions in Fig. 9.1.
4. Finally, the measured value of the statistic $Y = 61$ must be compared with the rejection region. If the analysts chooses a 95% confidence level, the measured value *does not* fall within the region of rejection, and the analyst can conclude that the data are consistent with the hypothesis that the measurements are drawn from the parent Gaussian at the 95 % probability level (or 1.96σ level). If the analyst is satisfied with a $p = 68.3$ % probability, instead of 95 the measured value of the test statistic Y now falls in the rejection region. In this case, the analyst can conclude that the hypothesis must be rejected at the 68 % probability level (or at the 1σ level). ◇

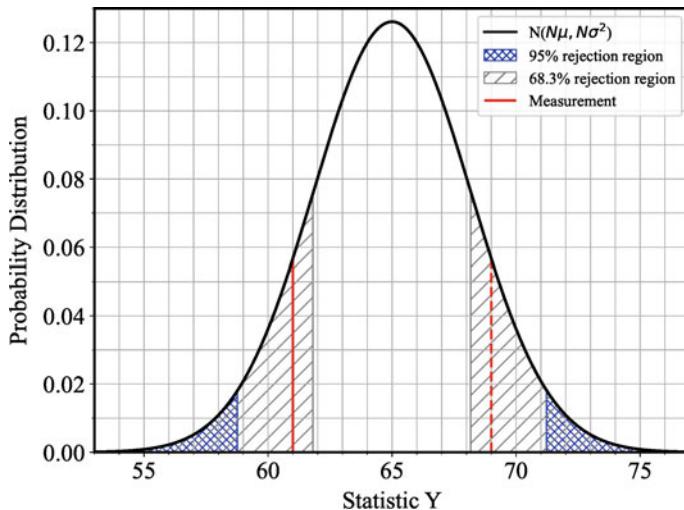


Fig. 9.1 Two-sided rejection regions for the test of a Gaussian origin for the five measurements of Example 9.1

9.2 The *P*-Value of a Statistical Analysis

The example provided in the previous section illustrates the importance of the choice of the confidence level p for hypothesis testing. In fact, the same null hypothesis must be rejected, or not, depending on the choice of a value for the probability p , which is necessarily arbitrary. To avoid this ambiguity, analysts often prefer to take a *post-facto* approach to the choice of p . Continuing with the previous example, the measured value of the sum or the sample mean of the measurements corresponds to a -1.26σ deviation from the parent mean. Such deviation marks one of the two boundaries of a 79 % central confidence interval around the parent mean, i.e., there is a 21% probability that the measurement exceeds $\pm 1.26\sigma$ from the mean. In fact, prior to the measurement, there was no reason to expect a negative deviation, so the analyst must entertain the possibility of both negative and positive deviations, so that the measurement with an absolute deviation larger than 1.26σ is indicated as more *extreme* than the measured value. It is therefore possible to report this result with the statement that the data are consistent with the parent model at the p confidence level, which in this example is the 79% confidence level.

Following this line of reasoning, it is customary to define the *p-value* of a statistical hypothesis as the probability that the statistic is as extreme or more extreme than the measured value, under the null hypothesis. The distribution of the statistic provides guidance on how to determine the *p*-value. For example, for a two-sided rejection region such as the one in Fig. 9.1, and for a measurement S_{data} of the statistic that can be on either side of the parent mean μ_S , the quantity $S_{1/2} = |S_{\text{data}} - \mu_S|$ represents one half of the interval between the measurement and the mean (the range between

the red line and the mean in Fig. 9.1). The p -value is therefore calculated as the probability

$$P(|S - \mu_S| > S_{1/2}) = 1 - \int_{\mu_S - S_{1/2}}^{\mu_S + S_{1/2}} f(s) ds = 1 - p, \quad (9.3)$$

where the integral represents the probability under the curve between the two red lines in Fig . 9.1. Equation 9.3 is a special case of the general Equation (9.2) for the choice of a symmetric two-sided rejection region. There are cases when, on the other hand, it is meaningful to reject a null hypothesis only when there are measurements of the associated statistic that are either sufficiently smaller or larger than a critical value, in a specific direction that depends on the statistic of choice for the null hypothesis, and the statistical problem at hand. In those cases, it is possible to find the p -value in a similar way, starting with (9.2). A common situation is that of a one-sided rejection where for large values of the test statistic are unacceptable, whereby the p -value is calculated as

$$P(S > S_{\text{data}}) = 1 - \int_{-\infty}^{S_{\text{data}}} f(s) ds = 1 - p. \quad (9.4)$$

For a positive-definite statistic, the lower limit of integration will naturally be replaced with zero in (9.4). Notice that the p -value is the value of the residual probability $(1 - p)$, which represents the probability that the statistic has a value that is at least as extreme as the measured value. This choice of notation is somewhat non-standard, in that other authors often prefer to refer to the p -value by the letter p itself, which is used in this book as the enclosed probability or confidence level. The choice is ultimately immaterial, since a smaller p -value always must represent a small probability of occurrence of extreme values of the test statistic, and this is reflected by (9.3) and (9.4).

Example 9.2 (*Calculation of p -values*) Continuing with the data of Example 9.1, the test statistic of choice was the sum Y of the five numbers, measured as $Y_{\text{data}} = 61$ and therefore with a half-interval of $S_{1/2} = 4$. The distribution of Y under the null hypothesis is a Gaussian with $\mu = 65$ and variance $\sigma^2 = 10$, or standard deviation $\sigma = 3.16$, meaning that the measured statistic deviates approximately -1.3σ from the parent mean. Choosing a two-sided rejection region around the mean $\mu = 65$ results in a p -value of

$$P(|Y - \mu| > 4) \simeq 1 - 0.80 = 0.20,$$

so that the statistical hypothesis would be reported as having a p -value of approximately 20%, meaning that there is a 20% probability of obtaining such a value of the statistic, or a more extreme value, under the null hypothesis that the data follow the parent model. This is clearly a sufficiently large probability that most analysts would conclude that the null hypothesis should not be rejected. \diamond

The interpretation of *p*-values is often a source of debate among statisticians and data analysts, and it is therefore useful to understand certain issues and limitations. The *p*-value is based on a null hypothesis or statistical model *and* on the data at the same time. Since the null hypothesis is abstract in nature, it is usually considered appropriate to phrase conclusions based on *p*-values in terms of the likelihood of the data being collected, rather than the probability that a null hypothesis is true. In 2016, the *American Statistical Association* (ASA) issued the *ASA Statement on Statistical Significance and P-Values* as a means to clarify *p*-values and the method of hypothesis testing [101]. Key excerpts from the statement include the definition of the *p*-value:

Informally, a *p*-value is the probability under a specified statistical model that a statistical summary of the data (e.g., the sample mean difference between two compared groups) would be equal to or more extreme than its observed value.

and the principles of use of *p*-values:

1. P-values can indicate how incompatible the data are with a specified statistical model.

A *p*-value provides one approach to summarizing the incompatibility between a particular set of data and a proposed model for the data. The most common context is a model, constructed under a set of assumptions, together with a so-called ‘null hypothesis.’ Often the null hypothesis postulates the absence of an effect, such as no difference between two groups, or the absence of a relationship between a factor and an outcome. The smaller the *p*-value, the greater the statistical incompatibility of the data with the null hypothesis, if the underlying assumptions used to calculate the *p*-value hold. This incompatibility can be interpreted as casting doubt on or providing evidence against the null hypothesis or the underlying assumptions.

2. P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.

Researchers often wish to turn a *p*-value into a statement about the truth of a null hypothesis, or about the probability that random chance produced the observed data. The *p*-value is neither. It is a statement about data in relation to a specified hypothetical explanation, and is not a statement about the explanation itself.

3. Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.

Practices that reduce data analysis or scientific inference to mechanical ‘bright-line’ rules (such as ‘*p* < 0.05’) for justifying scientific claims or conclusions can lead to erroneous beliefs and poor decision making. A conclusion does not immediately become ‘true’ on one side of the divide and ‘false’ on the other. Researchers should bring many contextual factors into play to derive scientific inferences, including the design of a study, the quality of the measurements, the external evidence for the phenomenon under study, and the validity of assumptions that underlie the data analysis. Pragmatic considerations often require binary, ‘yes-no’ decisions, but this does not mean that *p*-values alone can ensure that a decision is correct or incorrect. The widespread use of ‘statistical significance’ (generally interpreted as ‘*p* ≤ 0.05’) as a license for making a claim of a scientific finding (or implied truth) leads to considerable distortion of the scientific process.

4. Proper inference requires full reporting and transparency.

P-values and related analyses should not be reported selectively. Conducting multiple analyses of the data and reporting only those with certain *p*-values (typically those passing a significance threshold) renders the reported *p*-values essentially uninterpretable. Cherry-picking promising findings, also known by such terms as data dredging, significance chasing,

significance questing, selective inference, and ‘p-hacking,’ leads to a spurious excess of statistically significant results in the published literature and should be vigorously avoided. One need not formally carry out multiple statistical tests for this problem to arise: Whenever a researcher chooses what to present based on statistical results, valid interpretation of those results is severely compromised if the reader is not informed of the choice and its basis. Researchers should disclose the number of hypotheses explored during the study, all data collection decisions, all statistical analyses conducted, and all p-values computed. Valid scientific conclusions based on p-values and related statistics cannot be drawn without at least knowing how many and which analyses were conducted, and how those analyses (including p-values) were selected for reporting.

5. A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.

Statistical significance is not equivalent to scientific, human, or economic significance. Smaller p-values do not necessarily imply the presence of larger or more important effects, and larger p-values do not imply a lack of importance or even lack of effect. Any effect, no matter how tiny, can produce a small p-value if the sample size or measurement precision is high enough, and large effects may produce unimpressive p-values if the sample size is small or measurements are imprecise. Similarly, identical estimated effects will have different p-values if the precision of the estimates differs.

6. By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

Researchers should recognize that a p-value without context or other evidence provides limited information. For example, a p-value near 0.05 taken by itself offers only weak evidence against the null hypothesis. Likewise, a relatively large p-value does not imply evidence in favor of the null hypothesis; many other hypotheses may be equally or more consistent with the observed data. For these reasons, data analysis should not end with the calculation of a p-value when other approaches are appropriate and feasible.

It is also useful to further discuss the meaning of the word “acceptable” with regard to the null hypothesis, which was introduced at the end of the previous section. The fact that the data yield a large p -value does not imply that the null hypothesis is the correct one, as pointed out in the second item of the ASA statement. In fact, there could be other hypotheses that are equally well “acceptable.” Therefore, any null hypothesis can only be conclusively rejected but *never conclusively proven* to be true, since this would imply exhausting and discarding all possible alternative hypotheses. The process of hypothesis testing is therefore slanted toward trying to disprove the null hypothesis, possibly in favor of alternative hypotheses, but with the clear understanding that it is not possible to prove that a hypothesis is true, as also highlighted by Fisher’s quote in Sect. 9.1. The rejection of a null hypothesis by means of a small p -value should also not be directly interpreted as evidence in favor of an alternative hypothesis, since only the null hypothesis was being tested.

9.3 The χ^2 Statistic

Consider N random variables X_i , each normally distributed with mean μ_i and variance σ_i^2 , and independent of one other. For each variable X_i , the associated z-score,

$$Z_i = \frac{X_i - \mu_i}{\sigma_i}$$

is therefore a standard Gaussian of zero mean and unit variance. The sum of the square of all the deviations,

$$\chi^2 = \sum_{i=1}^N Z_i^2, \quad (9.5)$$

is said to be a χ^2 -distributed variable. The same symbol, derived from the Greek letter χ (usually transliterated as “chi” and pronounced “kai”), represents both a statistic or random variable and a distribution function. Usually data-based statistics are not represented with Greek letter, but this notation is traditional and therefore it is convenient to use a single symbol to represent both the name of the variable and the name of the distribution. The interest in this variable is that the X_i can be considered as independent measurements of a normal variable, and therefore χ^2 also becomes a statistic that can be used to test the hypothesis that the measurements are drawn from the respective normal distributions. The χ^2 statistic was first introduced by K. Pearson to test for association between two variables [76, 77], and further studied and developed by R.A. Fisher [28]. It plays a fundamental role in statistics and data analysis, because of the common occurrence of normally distributed measurements.

9.3.1 The Probability Distribution Function

The probability distribution of the χ^2 statistic is a special case of a *gamma distribution*, which is defined as

$$f_\gamma(x) = \frac{\alpha(\alpha x)^{r-1} e^{-\alpha x}}{\Gamma(r)}, \quad (9.6)$$

where α, r are positive numbers, and $x \geq 0$. Its name derives from the relationship with the Gamma function (7.17), which serves as the normalization constant for the distribution. It can be shown that the random variable χ^2 is distributed like a gamma distribution with parameters $r = N/2$ and $\alpha = 1/2$, and therefore it has the following probability distribution function:

$$f_{\chi^2}(x) = \left(\frac{1}{2}\right)^{N/2} \frac{1}{\Gamma(N/2)} e^{-x/2} x^{N/2-1}. \quad (9.7)$$

The parameter N of the distribution is usually referred to as the number of *degrees of freedom* of the χ^2 distribution.

To derive the distribution function of χ^2 , it is necessary to show that the moment generating function of the square of each z-score is given by

$$M_{Z_i^2}(t) = \sqrt{\frac{1}{1-2t}}. \quad (9.8)$$

In fact,

$$M_{Z_i^2}(t) = E[e^{Z_i^2 t}] = \int_{-\infty}^{+\infty} e^{x^2 t} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-x^2(\frac{1}{2}-t)} dx.$$

Using $\int_{-\infty}^{+\infty} e^{-y^2} dy = \sqrt{\pi}$ and changing variables $y^2 = x^2(1/2 - t)$, with $2x dx(1/2 - t) = 2y dy$ leads to

$$dx = \frac{y}{x} \frac{dy}{(1/2 - t)} = \frac{\sqrt{1/2 - t}}{1/2 - t} dy = \frac{dy}{\sqrt{1/2 - t}}.$$

This results in the following moment generating function for Z_i^2 :

$$M_{Z_i^2}(t) = \int_{-\infty}^{+\infty} \frac{e^{-y^2}}{\sqrt{\pi}} \frac{dy}{\sqrt{2(1/2 - t)}} = \sqrt{\frac{1}{1-2t}}. \quad (9.9)$$

Since the variables X_i are independent of one another, so are the variables Z_i^2 . Therefore, the moment generating function of Z is given by

$$M_{\chi^2}(t) = \left(M_{Z_i^2}(t) \right)^N = \left(\sqrt{\frac{1}{1-2t}} \right)^{N/2}, \quad (9.10)$$

where the property $M_{x+y}(t) = M_x(t) \cdot M_y(t)$ for independent variables was used.

Next, it is possible to show that the moment generating function of a gamma distribution is

$$M_{\gamma}(t) = \frac{1}{\left(1 - \frac{t}{\alpha}\right)^r}. \quad (9.11)$$

In fact, with G a γ -distributed variable,

$$\begin{aligned} M_{\gamma}(t) &= E[e^{tG}] = \int_0^{\infty} e^{tz} f_{\gamma}(z) dz = \int_0^{\infty} \frac{\alpha^r}{\Gamma(r)} z^{r-1} e^{-z(\alpha-t)} dz \\ &= \frac{\alpha^r}{(\alpha-t)^r} \int_0^{\infty} \frac{(\alpha-t)^{r-1}}{\Gamma(r)} z^{r-1} e^{-z(\alpha-t)} (\alpha-t) dz. \end{aligned}$$

The change of variable $x = z(\alpha - t)$ leads to

$$M_\gamma(t) = \frac{\alpha^r}{(\alpha - t)^r} \int_0^\infty \frac{x^{r-1}}{\Gamma(r)} e^{-x} dx = \frac{\alpha^r}{(\alpha - t)^r} = \frac{1}{\left(1 - \frac{t}{\alpha}\right)^r}, \quad (9.12)$$

where the property that (9.6) integrates to one was used. Comparison of the moment generating functions for the χ^2 and γ distributions show that χ^2 is γ -distributed with parameters $\alpha = 1/2$ and $r = N/2$.

9.3.2 Moments and Other Properties

Since the mean and variance of a gamma distribution with parameters r and α are, respectively, r/α and r/α^2 (see Appendix A.3), the χ^2 statistic has the following mean and variance:

$$\begin{cases} E[\chi^2] = N \\ \text{Var}(\chi^2) = 2N. \end{cases} \quad (9.13)$$

An example of χ^2 distribution is shown in Fig. 9.2. The distribution is unimodal, although not symmetric with respect to the mean. It is common to use the *reduced χ^2 square* variable defined by

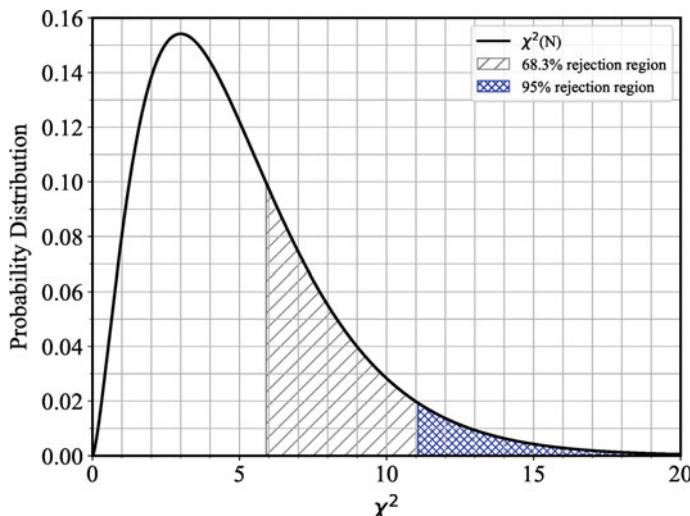


Fig. 9.2 The χ^2 distribution with $N = 5$ degrees of freedom, and the associated 95% and 68.3% one-sided rejection regions.

$$\chi_{red}^2 = \frac{\chi^2}{N}, \quad (9.14)$$

with unit expectation. The ratio between the standard deviation and the mean of χ^2 , a measure of the spread of the distribution, decreases with the number of degrees of freedom. As the number of degrees of freedom increases, the values of the reduced χ^2 are more closely distributed around 1. The moment generating function (9.10) leads to the property that the sum of two independent χ^2 variables with N and M degrees of freedom is a χ^2 variable with $N + M$ degrees of freedom. In fact, the moment generating function of the sum of two independent variables is the product of the two moment generating functions, and the exponents in (9.10) will add to give the desired result.

9.3.3 Hypothesis Testing

The null hypothesis used to obtain the probability distribution of χ^2 (9.7) is that all variables X_i are consistent with parent Gaussians with specified means and variances. If the N measurements are consistent with their parent distributions, one expects a value of approximately $\chi^2 \simeq N$, with each measurement contributing approximately a value of one, on average, to the statistic. Large values of χ^2 indicate that some of the measurements differ by several standard deviations from the expected mean, either in defect or in excess, in both cases leading to large and always positive contributions to χ^2 . Likewise, values of $\chi^2 \ll N$ are also not expected. Consider, for example, the extreme case of N measurements all identical to the parent mean, resulting in $\chi^2 = 0$. Statistical fluctuations of the random variables make it extremely unlikely that all N measurements match the mean. Clearly such extreme cases of perfect or nearly perfect agreement between the data and the parent model are suspicious. In those cases, the data should be checked for possible errors in the collection or analysis, and the model should be investigated for possible erroneous assumptions of the mean and variance of the contributing Gaussians.

Despite the fact that very small values of χ^2 are unlikely, it is customary to test for the agreement between a measurement of χ^2 and its theoretical distribution using a one-sided rejection region. The rejection region consists of values of χ^2 exceeding a critical value with confidence level p that can be calculated via

$$P(\chi^2 \geq \chi_{crit}^2) = \int_{\chi_{crit}^2}^{\infty} f_{\chi^2}(x) dx = 1 - p. \quad (9.15)$$

Critical values for the χ^2 distribution are tabulated in Table A.7.

Example 9.3 (*Hypothesis testing with χ^2*) Consider the $N = 5$ measurements of a variable X , (10, 12, 15, 11, 13), presented in Example 9.1, with the same hypothesis that these are independent measurements of a Gaussian variable X of mean $\mu = 13$

and variance $\sigma^2 = 2$. The χ^2 statistic can be used to try and falsify the null hypothesis that the data are drawn from the given Gaussian. The procedure for a quantitative answer to this hypothesis test requires a level of probability p , then to calculate the value of the statistic as

$$\chi^2 = \sum_{i=1}^5 \frac{(x_i - \mu)^2}{\sigma^2} = 9.$$

Figure 9.2 shows the rejection regions for a probability $p = 0.68$ and $p = 0.95$, which are determined according to (9.15) with $N = 5$ degrees of freedom: $\chi^2_{crit} = 5.89$ marks the beginning of the 68.3 % rejection region, and $\chi^2_{crit} = 11.07$ that of the 95 % rejection region. The hypothesis is therefore rejected at the 68 % probability level, but cannot be rejected at the 95 % confidence level.

Moreover, the p -value is calculated according to (9.4) as

$$P(\chi^2 \geq 9) \simeq 0.11,$$

for example, by interpolation of the numbers in Table A.7. It is therefore possible to conclude that there is a $\sim 11\%$ probability of observing such value of χ^2 , or higher, under the hypothesis that the measurements were made from a Gaussian distribution of such mean and variance (see Fig. 9.2). The result of hypothesis testing using the χ^2 statistic is therefore different from that obtained with the sum (or sample mean) of the five measurements in Example 9.1, which resulted in a $\sim 20\%$ p -value. The main difference is that the present hypothesis testing with the χ^2 statistic makes use of a one-sided rejection region, while the earlier test with a Gaussian distribution used a two-sided rejection region. \diamond

9.4 The Distribution of the Sample Variance

Consider N independent measurements of a random variable X that is normally distributed with mean μ and variance σ^2 . The sample variance s^2 of the measurements can be calculated according to (2.14), without the need to specify the parent mean. The probability distribution of the sample variance, or *sampling distribution of the variance*, is useful to compare a given measurement of the sample variance with the parent variance. It is convenient to define the statistic S^2 as

$$S^2 = (N - 1)s^2 = \sum_{i=1}^N (x_i - \bar{x})^2 \quad (9.16)$$

and seek a distribution function for S^2 that enables a comparison of the measured sample variance with the parent variance.

In determining the sampling distribution of the variance it is not necessary to specify a value for the mean of the parent Gaussian. This is a common experimental situation, since usually one does not know a priori the parent mean of the distribution, but the sample mean is an easily calculated statistic. In the use of (9.16), it is therefore the case that \bar{x} is itself a random variable, and not an exactly known quantity. This fact must be taken into account when calculating the expectation of S^2 . The S^2 statistic is equal to

$$S^2 = \sum_{i=1}^N (x_i - \mu + \mu - \bar{x})^2 = \sum_{i=1}^N (x_i - \mu)^2 - N(\mu - \bar{x})^2, \quad (9.17)$$

and dividing both terms by σ^2 yields the following result:

$$\frac{\sum_{i=1}^N (x_i - \mu)^2}{\sigma^2} = \frac{S^2}{\sigma^2} + \frac{(\bar{x} - \mu)^2}{\sigma^2/N}. \quad (9.18)$$

According to the result in Sect. 9.3, the left-hand side term is distributed like a χ^2 variable with N degrees of freedom, since the parent mean μ and variance σ^2 appear in the sum of squares. For the same reason, the second term in the right-hand side is also distributed like a χ^2 variable with 1 degree of freedom, since it was already determined that the sample mean is distributed like a Gaussian with mean μ and with variance σ^2/N . Although it may not be apparent at first sight, it can be proven that the two terms on the right-hand side are two independent random variables. If the independence between these two variables can be established, then it must be true that the first variable in the right-hand side, S^2/σ^2 , is also distributed like a χ^2 variable with $N - 1$ degrees of freedom. This follows from the fact that the sum of two independent χ^2 variables is also a χ^2 variable featuring the sum of the degrees of freedom of the two variables, as shown in Sect. 9.3.

The proof of the independence between the two statistics in the right-hand side of (9.18) and the fact that both are distributed like χ^2 distributions with, respectively, $N - 1$ and 1 degree of freedom, can be obtained by making a suitable change of variables from the original N standard normal variables that appear in the left-hand side of (9.18),

$$Z_i = \frac{X_i - \mu}{\sigma},$$

to a new set of variables Y_i . The desired transformation is an *orthonormal transformation* that has the property

$$Z_1^2 + \dots + Z_N^2 = Y_1^2 + \dots + Y_N^2.$$

In matrix form, this transformation can be expressed by a transformation matrix A of dimensions $N \times N$, such that a row vector $\mathbf{z} = (Z_1, \dots, Z_N)$ is transformed into another vector \mathbf{y} by way of the product $\mathbf{y} = \mathbf{z} A$. For such a transformation, the dot product between two vectors is expressed as $\mathbf{y}\mathbf{y}^T = \mathbf{z} A A^T \mathbf{z}^T$. Since for an orthonormal transformation the relationship $A A^T = I$ holds, where I is the $N \times N$ identity matrix, then the dot product remains constant upon this transformation. An orthonormal transformation, expressed in extended form as

$$\begin{cases} Y_1 = a_1 Z_1 + \dots + a_N Z_N \\ Y_2 = b_1 Z_1 + \dots + b_N Z_N \\ \dots \end{cases}$$

is obtained when, for each row vector, $\sum a_i^2 = 1$, and for any pair of row vectors, $\sum a_i b_i = 0$, so that the Y_i 's are independent of one another.

Any such orthonormal transformation, when applied to N independent variables that are standard Gaussians, $Z_i \sim N(0, 1)$, as is the case in this application, is such that the transformed variables Y_i are also independent standard Gaussians. In fact, the joint probability distribution function of the Z_i 's can be written as

$$f(\mathbf{z}) = \frac{1}{(2\pi)^{N/2}} e^{-\frac{z_1^2 + \dots + z_N^2}{2}},$$

and, since the transformed variables have the same dot product, $z_1^2 + \dots + z_N^2 = y_1^2 + \dots + y_N^2$, the N variables Y_i have the same joint distribution function, proving that they are also independent standard Gaussians.

These general properties of orthonormal transformations can be used to find a transformation that will enable a proof of the independence between S^2/σ^2 and $(\bar{x} - \mu)^2/\sigma_\mu^2$. The first variable is defined by the following linear combination:

$$Y_1 = \frac{Z_1}{\sqrt{N}} + \dots + \frac{Z_N}{\sqrt{N}}$$

in such a way that the following relationships hold:

$$\begin{cases} Y_1^2 = \frac{(\bar{X} - \mu)^2}{\sigma^2/N} \\ \sum_{i=1}^N Z_i^2 = \frac{1}{\sigma^2} \sum_{i=1}^N (X_i - \mu)^2, \text{ or} \\ \sum_{i=1}^N Z_i^2 = Y_1^2 + \sum_{i=2}^N Y_i^2. \end{cases}$$

The other $N - 1$ variables Y_2, \dots, Y_N can be chosen arbitrarily, provided they satisfy the requirements of orthonormality. Since

$$\sum_{i=1}^N Z_i^2 - Y_1^2 = \sum_{i=2}^N Y_i^2 = \frac{S^2}{\sigma^2}, \quad (9.19)$$

it is immediate to see that the statistic S^2/σ^2 is distributed like a χ^2 distribution with $N - 1$ degrees of freedom, as the sum of squares of $N - 1$ independent standard Gaussians. This variable is also independent of the sampling distribution of the mean, Y_1^2 , since the variables Y_i are independent of each other. This proof is due to M.G. Bulmer [16], who used a derivation done earlier by F.R. Helmert [50].

It is therefore possible to conclude that the ratio S^2/σ^2 is distributed like a χ^2 variable with $N - 1$ degrees of freedom,

$$\frac{S^2}{\sigma^2} \sim \chi^2(N - 1). \quad (9.20)$$

The difference between the χ^2 distribution (9.7) and the distribution of the sample variance (9.20) is that in the latter case the mean of the parent distribution is not assumed to be known, but it is calculated from the data. This is in fact the more common situation, and therefore when N measurements are obtained, the quantity

$$\frac{S^2}{\sigma^2} = \sum_{i=1}^N \left(\frac{x_i - \bar{x}}{\sigma} \right)^2$$

is distributed like a χ^2 distribution with just $N - 1$ degrees of freedom, not N . This reduction in the number of degrees of freedom can be expressed by saying that 1 degree of freedom is being used to estimate the mean.

Example 9.4 (*Hypothesis testing for the sample variance*) Consider $N = 10$ measurements of a random variable, $(10, 12, 15, 11, 13, 16, 12, 10, 18, 13)$. The hypothesis to test is whether these measurements are consistent with being drawn from a Gaussian distribution of unspecified mean and variance $\sigma^2 = 2$. If the measurements are derived from the *same* variable, then the actual measurement of the sample variance should be consistent with its theoretical distribution (9.20). For these measurements, the sample mean is $\bar{x} = 13$ and $S^2 = 62$. Therefore, the measurement $S^2/\sigma^2 = 62/2 = 36$ must be compared with the χ^2 distribution with $N - 1 = 9$ degrees of freedom. The measurement is equivalent to a reduced χ^2 value of 4, which is inconsistent with a χ^2 distribution with 9 degrees of freedom at more than the 99 % confidence level, for a p -value of ~ 0.0002 . It is therefore necessary to conclude that the hypothesis must be rejected with high level of confidence. It is useful to point out that this calculation assumes that the parent variance is known, which is uncommon in practical applications. The following section introduces another test that can be used to compare two measurements of the variance that does not require knowledge of the parent variance. ◇

9.5 The F -Statistic

The distribution of the sample variance of normally distributed measurements (9.20) requires the parent variance to test the agreement between the data and the model. Alternatively, one can compare two different measurements of the variance and ask the associated question of whether the ratio between the two measurements is statistically equivalent. The *F -statistic* is defined as

$$F = \frac{\chi_1^2/f_1}{\chi_2^2/f_2}, \quad (9.21)$$

where χ_1^2 and χ_2^2 are two χ^2 -distributed variables with f_1 and f_2 degrees of freedom. When this statistic is applied to the ratio of two sample variances, with S^2/σ^2 distributed as a χ^2 variable, the parent variance σ^2 cancels out and one can compare two measurements of the sample variance under the hypothesis of normally distributed measurements, without any reference to a specific value of the parent mean or variance. The distribution of the F -statistic is known as the *F -distribution* with parameters f_1 and f_2 . It is named after R.A. Fisher [29], who was the first to study a distribution for the ratio of sample variances, although in a different form. It is also sometimes referred to as the *Snedecor's F -distribution* after G.W. Snedecor, who reported it in his book *Statistical Methods* [95].

9.5.1 The Probability Distribution Function

The probability distribution function of the statistic F is given by

$$f_F(x) = \frac{f_1/f_2}{B(f_1/2, f_2/2)} \frac{\left(x \frac{f_1}{f_2}\right)^{f_1/2-1}}{\left(1 + x \frac{f_1}{f_2}\right)^{f_1/2+f_2/2}}, \quad (9.22)$$

where f_1 and f_2 are the number of degrees of freedom of the χ^2 variable at the numerator and at the denominator, respectively, and

$$B(f_1/2, f_2/2) = \frac{\Gamma(f_1/2)\Gamma(f_2/2)}{\Gamma(f_1/2 + f_2/2)} \quad (9.23)$$

is the *Beta function*, with Γ the usual Gamma function defined in (7.17).

The proof of the F -distribution makes use of the methods described in Sects. 4.5.1 and 4.5.2. First the distributions of the numerator and denominator of (9.21) are found, and then the distribution function for the ratio of two variables is calculated using (4.17).

Given that $\chi_1^2 \sim \chi^2(f_1)$ and $\chi_2^2 \sim \chi^2(f_2)$, the distribution functions of $X' = \chi_1^2/f_1$ and $Y' = \chi_2^2/f_2$ are found using a straightforward change of variables. The distribution of X' is therefore

$$f_{X'}(x') = f(z) \frac{dz}{dx'} = f(z) f_1,$$

where $f(z)$ is the distribution of χ_1^2 and $x' = z/f_1$. This results in

$$f_{X'}(x') = \frac{z^{f_1/2-1} e^{-z/2}}{\Gamma(f_1/2) 2^{f_1/2}} f_1 = \frac{(x' f_1)^{f_1/2-1} e^{-(x' f_1)/2}}{\Gamma(f_1/2) 2^{f_1/2}} f_1,$$

with an identical transformation also for Y' . The two distribution functions for the numerator and denominator can now be used in (4.17) to give the probability distribution of F ,

$$\begin{aligned} f_F(z) &= \int_0^\infty f_{X'}(z\zeta) \zeta f_{Y'}(\zeta) d\zeta \\ &= \int_0^\infty \frac{(z\zeta f_1)^{f_1/2-1} e^{-(z\zeta f_1)/2}}{\Gamma(f_1/2) 2^{f_1/2}} f_1 \zeta \frac{(\zeta f_2)^{f_2/2-1} e^{-(\zeta f_2)/2}}{\Gamma(f_2/2) 2^{f_2/2}} d\zeta \\ &= \frac{z^{f_1/2-1} f_1^{f_1/2} f_2^{f_2/2}}{\Gamma(f_1/2) \Gamma(f_2/2) 2^{(f_1+f_2)/2}} \int_0^\infty \zeta^{(f_1+f_2)/2-1} e^{-1/2\zeta(z f_1 + f_2)} d\zeta. \end{aligned}$$

After another change of variables, $t = \zeta(z f_1 + f_2)/2$, $dt = d\zeta(z f_1 + f_2)/2$, the integral becomes

$$\begin{aligned} &\int_0^\infty \left(\frac{2t}{z f_1 + f_2} \right)^{(f_1+f_2)/2-1} e^{-t} \frac{dt}{\left(\frac{z f_1 + f_2}{2} \right)} \\ &= \frac{2^{(f_1+f_2)/2}}{(z f_1 + f_2)^{1+(f_1+f_2)/2-1}} \int_0^\infty t^{(f_1+f_2)/2-1} e^{-t} dt \\ &= \frac{2^{(f_1+f_2)/2}}{(z f_1 + f_2)^{(f_1+f_2)/2}} \Gamma\left(\frac{f_1 + f_2}{2}\right). \end{aligned}$$

Finally, the distribution of F is given by

$$\begin{aligned} f_F(z) &= \frac{z^{f_1/2-1} f_1^{f_1/2} f_2^{f_2/2}}{\Gamma(f_1/2)\Gamma(f_2/2) 2^{(f_1+f_2)/2}} \frac{2^{(f_1+f_2)/2}\Gamma\left(\frac{f_1+f_2}{2}\right)}{(z f_1 + f_2)^{(f_1+f_2)/2}} \\ &= \left(\frac{f_1}{f_2}\right)^{f_1/2} \frac{\Gamma\left(\frac{f_1+f_2}{2}\right)}{\Gamma\left(\frac{f_1}{2}\right)\Gamma\left(\frac{f_2}{2}\right)} \frac{z^{f_1/2-1}}{\left(1+z\frac{f_1}{f_2}\right)^{(f_1+f_2)/2}}. \end{aligned}$$

9.5.2 Moments and Other Properties

The mean and higher-order moments of the F -distribution can be calculated making use of the Beta function and the properties of the Gamma function, to find that

$$\begin{cases} E[F] = \frac{f_2}{f_2 - 2} & \text{(for } f_2 > 2\text{)} \\ \text{Var}(F) = \frac{2 f_2^2 (f_1 + f_2 - 2)}{f_1 (f_2 - 2)^2 (f_2 - 4)} & \text{(for } f_2 > 4\text{).} \end{cases} \quad (9.24)$$

The mean of the F -distribution diverges for $0 < f_2 \leq 2$, and likewise the variance diverges for $2 < f_2 \leq 4$, and it is undefined for $0 < f_2 \leq 2$. The above formulas can therefore be used only for values of the degrees of freedom where they are defined.

The mean is approximately 1, provided that f_2 is not too small. It is possible to find an approximation to the F -distribution when either f_1 or f_2 is a large number:

$$\begin{cases} \lim_{f_2 \rightarrow \infty} f_F(z; f_1, f_2) = f_{\chi^2}(x; f_1), & \text{where } x = f_1 z \\ \lim_{f_1 \rightarrow \infty} f_F(z; f_1, f_2) = f_{\chi^2}(x; f_2), & \text{where } x = f_2/z. \end{cases} \quad (9.25)$$

This approximation (see, e.g., [1]) is very convenient, since it overcomes the problems with the evaluation of the Gamma function for large numbers.

9.5.3 Hypothesis Testing

In general, the F -statistic is a ratio between two independent χ^2 measurements of, respectively, f_1 and f_2 degrees of freedom. The underlying null hypothesis is

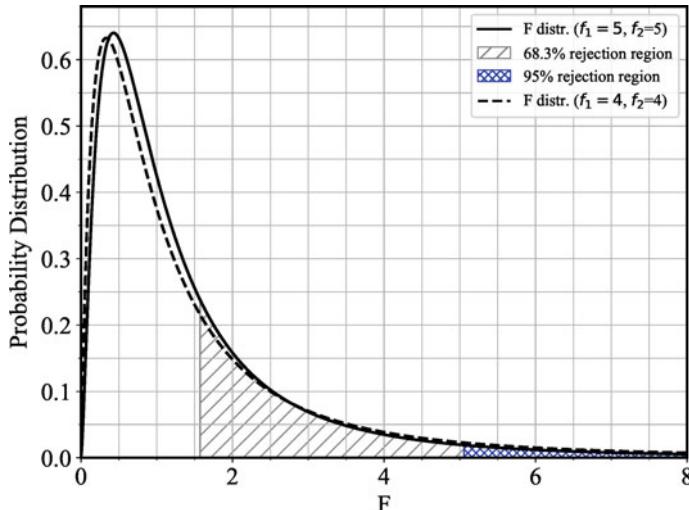


Fig. 9.3 F -distribution with $f_1 = 5$, $f_2 = 5$ degrees of freedom, and corresponding one-sided rejection regions. For comparison, the F -distribution with $f_1 = 4$, $f_2 = 4$ degrees of freedom is shown as the dashed curve, and its rejection regions are shifted to *higher* values, relative to the F -distribution with $f_1 = 5$, $f_2 = 5$ degrees of freedom, because of its heavier tail.

that both sets of measurements follow the respective Gaussian distribution, so that two χ^2 statistics results. When these hypotheses are satisfied, the measured ratio of the statistics, normalized by the number of degrees of freedom, will follow the F -distribution. There is a potential ambiguity when calculating the F -statistic, and that is given by the choice of which of the two statistics is at the numerator, and which at the denominator. It is customary to take the ratio of the two statistics so that the value of F is greater than one. In this case, extreme values of the statistic are contained in a one-sided interval that exceeds a critical value calculated according to

$$P(F > F_{crit}) = \int_{F_{crit}}^{\infty} f_F(x) dx = 1 - p. \quad (9.26)$$

Critical values are illustrated in Fig. 9.3, which shows that the F -distribution has a heavy tail, especially for small values of the number of degrees of freedom. When the lower limit of the integral (9.26) is the measured value of the statistic, the resulting value of $1 - p$ leads to the p -value of the measurement. Critical values are a function of the probability p and the two numbers of degrees of freedom, and they are tabulated in Tables A.8 through A.15. The value of F_{crit} calculated from (9.26) indicates how large the F -statistic can be and still be consistent with the hypothesis that the two quantities at the numerator and denominator are χ^2 -distributed variables.

The approximations for the F -distribution in (9.25) can be used to calculate critical values when one of the degrees of freedom is very large. For example, the critical value of F at 90 % confidence, $p = 0.90$, for $f_1 = 100$ and $f_2 \rightarrow \infty$ (e.g.,

Table A.13) is calculated from Table A.7 as $\bar{F} = 1.185$. Note that Table A.7 reports the value of the reduced χ^2 .

Example 9.5 (*Hypothesis testing with the F -statistic—Part I*) Consider the same ten measurements as in Example 9.4,

$$(10, 12, 15, 11, 13, 16, 12, 10, 18, 13).$$

Under the null hypothesis that the measurements follow a Gaussian distribution of mean of $\mu = 13$ and variance σ^2 , the F -distribution can be used to compare the χ^2 of the first five measurements with that of last five, to address whether both subsets are equally likely to be described by the same Gaussian. The data yield $\chi_1^2 = 18/\sigma^2$ and $\chi_2^2 = 44/\sigma^2$, respectively, for the first and the second set of five measurements. Both variables, under the null hypothesis that the measurements follow the reference Gaussian, are distributed like χ^2 with 5 degrees of freedom (since both mean and variance are assumed to be known). The corresponding F -statistic of $F = 44/18 = 2.44$. The initial five measurements appear at the denominator, so that $F > 1$. In the process of calculating the F -statistic, the variances σ^2 have been canceled, and therefore the null hypothesis is that of a normally distributed variable with mean of $\mu = 13$ and same variance for both sets, regardless of its value. Figure 9.3 plots the F -distribution for $f_1 = 5$ and $f_2 = 5$ as the solid line, and two representative rejection region marked by the respective critical values. The measurement of the F -statistic is therefore consistent with the null hypothesis at the 95 % confidence level, since the critical value is $F_{crit} = 5.05$. Clearly the first set of five numbers follows the parent Gaussian more closely than the second set, yet there is a reasonable chance ($> 5\%$) that both sets follow the Gaussian. For these data, the p -value associated with the F -test is 0.174, meaning that there is a probability of 17.4% to observe such value of F or larger under the null hypothesis. This number is sufficiently large that the null hypothesis should not be rejected, as already concluded based on the 95% confidence interval.

The parent variance, if specified, could have been used to test both subsets *independently* for the hypothesis that they follow a Gaussian of mean $\mu = 13$ and variance $\sigma^2 = 2$, using the χ^2 distribution. The two measurements are $\chi_1^2 = 9$ and $\chi_2^2 = 22$ for 5 degrees of freedom. Assuming a confidence level of $p = 0.95$, the critical value of the χ^2 distribution is $\chi_{crit}^2 = 11.07$. At this confidence level, the null hypothesis for the second measurement would be rejected at the 95% confidence level, while the first set of measurements is consistent with the parent model. Given the result of the χ^2 test, an analyst who is willing to follow the 95% confidence level would not even attempt to calculate the F -statistic. This example can be seen as an illustration of the *sensitivity* of two different tests of the same hypothesis, when the parent variance is specified. The χ^2 statistic is able to detect a disagreement between the data and the model, while the F -statistic is not, at the same confidence level, primarily because it does not make use of the available information on the parent variance. The F -test, though, remains of great use for a majority of applications where the parent variance

of normally distributed variables is not known, and therefore the χ^2 statistic cannot be used. \diamond

Perhaps the main application of the F -statistic is to compare the ratio between two measurements of the sample variance. For two independent sets of data with N_1 and N_2 measurements, the sample variances s_1^2 and s_2^2 are related to the parent variances σ_1^2 and σ_2^2 of the Gaussian models and to the F -statistic via

$$F = \frac{\chi_1^2/f_1}{\chi_2^2/f_2} = \frac{\frac{s_1^2}{\sigma_1^2 f_1}}{\frac{s_2^2}{\sigma_2^2 f_2}}, \quad (9.27)$$

where $f_1 = N_1 - 1$ and $f_2 = N_2 - 1$ and

$$\begin{cases} S_1^2 = (N_1 - 1)s_1^2 = \sum_{i=1}^{N_1} (x_i - \bar{x})^2 \\ S_2^2 = (N_2 - 1)s_2^2 = \sum_{j=1}^{N_2} (y_j - \bar{y})^2. \end{cases} \quad (9.28)$$

The quantities $\chi_1^2 = S_1^2/\sigma_1^2$ and $\chi_2^2 = S_2^2/\sigma_2^2$ are χ^2 -distributed variables with, respectively, f_1 and f_2 degrees of freedom. The F -statistic can be used to test whether both measurements of the variance are equally likely to have come from the respective models. Notice that, in general, one can test a hypothesis where the two parent variances are different, but usually the interesting case is when the two variances are equal, $\sigma_1^2 = \sigma_2^2$, so that the value of the variance drops out of the equation and the F -statistic becomes

$$F = \frac{S_1^2/f_1}{S_2^2/f_2} \text{ (case of same parent variance).} \quad (9.29)$$

In this case, the null hypothesis is that the two samples are Gaussian-distributed with the same variance, regardless of the values of the means. The F -statistic therefore measures if the variance or degree of variability of the data in the two measurements is consistent with one another. If the value of F exceeds the critical value, then the null hypothesis must be rejected and the conclusion is that the measurement with the largest value of the sample variance per degree of freedom, which is placed at the numerator, is unlikely to have the same parent variance as the other set, and therefore likely to have larger variance.

Example 9.6 (*Hypothesis testing with the F -statistic—Part II*) Using the same 10-measurement data as in the previous example, the sample variance in each of the two 5-measurement subsets can be calculated using the respective sample means of $\bar{x}_1 = 12.2$ and $\bar{x}_2 = 13.8$, for values of $S_1^2 = 14.8$ and $S_2^2 = 40.8$. The corresponding

F-statistic is therefore $F = 2.76$. Given that the sample mean was estimated from the data, the null hypothesis is that both sets are drawn from the same Gaussian distribution, without specification of the value of either variance or mean. Each measurement of S^2/σ^2 is distributed now like a χ^2 variable with just four degrees of freedom (and not five, as in the case of the previous example where μ was specified). The value of the *F*-statistic must therefore be compared with an *F*-distribution with $f_1 = 4$ and $f_2 = 4$ degrees of freedom, reported in Fig. 9.3 as a dashed line, with a 95 % critical value of $F_{crit} = 6.39$, which is significantly larger than the critical value for the *F*-distribution with $f_1 = f_2 = 5$ of the previous example. In fact, fewer degrees of freedom result in a significantly heavier right tail, and therefore the ability to allow larger values for the *F*-statistic. The measurements of the sample variance of the two subsets are therefore consistent at the 95 % confidence level. The *p*-value associated with this *F*-test is 0.175, virtually identical to the one calculated in the previous example.

9.6 The Sampling Distribution of the Mean and *Student's t*-Statistic

Comparing the sample mean to a parent mean and comparing two sample means among themselves are some of statistic's most common situations. When the measurements are from a normal distribution with given mean and variance, the problem is easily addressed. Complications arise when the parent variance is not known, but must itself be estimated from the data. This section addresses some of the methods used for the sample mean.

9.6.1 *Student's t*-Statistic for the Sample Mean

When N measurements of a Gaussian variable X of mean μ and variance σ^2 are available, the sample mean is distributed as a Gaussian of mean μ and variance σ^2/N . Therefore, if both the mean and the variance of the parent distribution are known, the sample mean can be used to calculate the *z*-score (3.12) as

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{N}} \sim N(0, 1).$$

Hypothesis testing for the *z*-score of the sample mean is immediately accomplished using the standard normal as the null hypothesis probability distribution (see Sect. 7.3). This simple case—when also σ^2 is known—therefore does not require the development of any new statistic to compare the sample mean to a parent mean.

Example 9.7 (*Comparison of sample mean with parent mean—Part I*) The same five measurements of a random variable of Example 9.1, (10, 12, 15, 11, 13), are assumed to be drawn from a Gaussian with $\mu = 13$ and $\sigma^2 = 2$. Assuming knowledge of the parent mean and variance, the z -score of the sample mean is

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{N}} = \frac{12.2 - 13}{\sqrt{2/5}} = -1.27.$$

The p -value of this z -score is 0.21, for a probability of about 21 % to exceed the absolute value of this measurement according to the parent distribution $N(0, 1)$. Therefore, the null hypothesis that the measurements are distributed like a Gaussian of $\mu = 13$ and $\sigma^2 = 2$ cannot be rejected at the 90 % confidence level. Notice that this is the same result obtained by using the sum of the five measurements as the statistic of choice, instead of the sample average. This was to be expected, since the mean differs from the sum by a constant, and therefore the two statistics are equivalent. \diamond

A more common situation is when the mean μ of the parent distribution is known, but the parent variance is unknown. In this case, the parent variance can only be estimated from the data themselves via the sample variance s^2 , leading to an additional source of uncertainty when estimating the distribution of the sample mean. To this end, the statistic of choice is defined as

$$T = \frac{\bar{x} - \mu}{s/\sqrt{N}}, \quad (9.30)$$

where the parent variance is replaced with the sample variance. This statistic is usually referred to as *Student's t-statistic*. The additional uncertainty associated with estimating the parent variance with the sample variance leads to a deviation of the distribution function of the T -statistics from the Gaussian distribution for the z -score. The variable T can be written as

$$T = \frac{\bar{x} - \mu}{s/\sqrt{N}} = \frac{\left[\frac{\bar{x} - \mu}{\sigma/\sqrt{N}} \right]}{\frac{s}{\sigma}} = \frac{\left[\frac{\bar{x} - \mu}{\sigma/\sqrt{N}} \right]}{\left[\frac{S^2}{(N-1)\sigma^2} \right]^{1/2}}, \quad (9.31)$$

where S^2 is the usual sum of the squares of the deviations from the sample mean. As shown earlier, the random variables at the numerator and at the denominator follow well-known distributions,

$$\frac{\bar{x} - \mu}{\sigma/\sqrt{N}} \sim N(0, 1)$$

and

$$\frac{S^2}{\sigma^2} \sim \chi^2(N-1).$$

The *T*-statistic is therefore defined as the ratio

$$T = \frac{X}{\sqrt{Z/f}}, \quad (9.32)$$

where $X \sim N(0, 1)$ and $Z \sim \chi^2(f)$ (a χ^2 distribution with f degrees of freedom) are independent variables. It is possible to show that the probability distribution of the *T*-statistic is

$$f_T(t) = \frac{1}{\sqrt{f\pi}} \frac{\Gamma(f/2 + 1/2)}{\Gamma(f/2)} \left(1 + \frac{t^2}{f}\right)^{-\frac{f+1}{2}}. \quad (9.33)$$

This distribution is symmetric around its mean of zero and goes under the name of *Student's t-distribution*. This distribution was studied first by W.S Gosset in 1908 [45], who published the seminal paper *The probable error of a mean* under the pseudonym of "Student."

The proof of (9.33) follows the same method as for the *F*-distribution. First, the distribution of $Y = \sqrt{Z/f}$ is found using the usual method of change of variables,

$$g(y) = h(z) \frac{dz}{dy} = 2h(z)\sqrt{fz},$$

where

$$h(z) = \frac{z^{f/2-1} e^{-z/2}}{2^{f/2} \Gamma(f/2)}$$

is the distribution of a χ^2 random variable with $f = N - 1$ degrees of freedom, according to (9.7). The distribution of Y is given by substituting $z = fy^2$ into the equation,

$$g(y) = \frac{f^{(f-1)/2} y^{f-1} e^{-fy^2/2} \sqrt{f}}{2^{f/2-1} \Gamma(f/2)}. \quad (9.34)$$

The distribution function of the numerator of (9.32) is simply

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

and therefore the distribution of T is given by applying (4.17),

$$f_T(t) = \int_0^\infty \frac{1}{\sqrt{2\pi}} e^{-(ty)^2/2} y \frac{f^{(f-1)/2} y^{f-1} e^{-fy^2/2} \sqrt{f}}{2^{f/2-1} \Gamma(f/2)} dy. \quad (9.35)$$

The integral can be shown to be equal to (9.33) following a few steps of integration as in the case of the *F*-distribution.

The T -statistic defined in (9.30) is therefore said to have a Student t -distribution with $f = N - 1$ degrees of freedom. The key moments of the distribution are

$$\begin{cases} \text{E}[T] = 0 & (\text{for } f > 1) \\ \text{Var}(T) = \frac{f}{f - 2} & (\text{for } f > 2) \end{cases} \quad (9.36)$$

with an undefined mean and variance if $f \leq 1$, and an infinite variance if $1 < f \leq 2$. The t -distribution is defined also for non-integer values of f , although usually in statistics f is an integer number of degrees of freedom. The moments can be calculated from (9.32) using the independence between the numerator and the denominator.

It is interesting to illustrate the case of a t -statistic derived from $N = 2$ measurements and therefore $f = 1$ degree of freedom. In this case, it is immediate to see that the distribution function becomes

$$f(x) = \frac{1}{\pi(1+x^2)}$$

which is the Cauchy distribution (4.13). This distribution has a number of peculiar features, particularly that the mean and the variance do not exist, and that it has especially heavy tails. One way to understand the peculiar nature of this distribution is that the distribution function f decreases toward infinity as a polynomial, and not as an exponential as in the case of a normal distribution. In practice, this means that it is not possible to use the t -statistic for datasets with just two measurements. More properties on the Cauchy distribution can be found in textbooks of probability theory.

It is important to notice the difference between the sample distribution of the mean when the variance is known, which is $N(0, 1)$, and the t -distribution. In particular, the latter depends on the number of measurements, while the former does not. One expects that, in the limit of a large number of measurements, the t -distribution tends to the standard normal (see Problem 9.10). The t -distribution has in fact broader wings than the standard Gaussian, and in the limit of an infinite number of degrees of freedom, the two distributions are identical; an example of the comparison between the two distributions is shown in Fig. 9.4. The t -distribution has heavier tails than the Gaussian distribution, as a result of the additional uncertainty associated with the fact that the variance is estimated from the data and not known a priori.

9.6.2 Hypothesis Testing with the t -Statistic

Hypothesis testing with the t -distribution typically uses a two-sided rejection region, with a critical value for a confidence level p calculated via

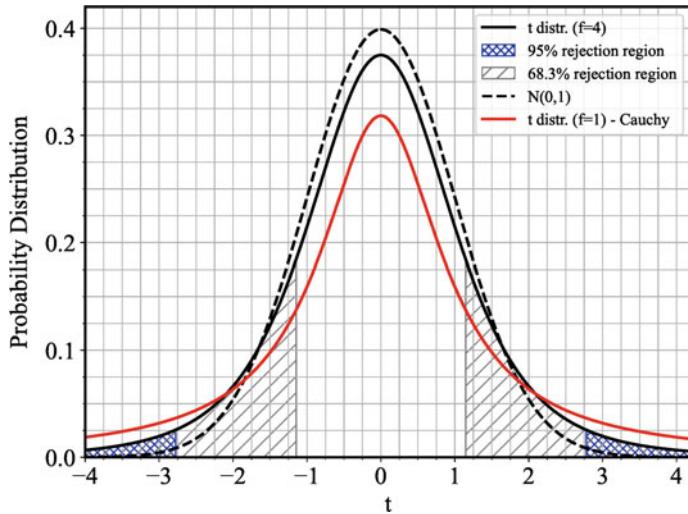


Fig. 9.4 Student's *t*-distribution with $f = 4$ degrees of freedom. The dashed curve is the $N(0, 1)$ Gaussian, to which the *t*-distribution tends for a large number of degrees of freedom. The solid red line is a *t*-distribution for $f = 1$, also known as the Cauchy distribution.

$$P(|T| \geq T_{crit}) = 1 - \int_{-T_{crit}}^{T_{crit}} f_T(t) dt = 1 - p. \quad (9.37)$$

The critical value is a function of the number of degrees of freedom of the *t*-statistic. Equation 9.37 also defines the p -value, when the critical value is replaced with the measured value of the statistic. Tables A.16, A.17, A.18, A.19, A.20, A.21, and A.22 report the value of p as function of the critical value T_{crit} for selected degrees of freedom, and Table A.23 compares the *t*-distribution with the standard Gaussian. For convenience, Table 9.1 also reports selected critical values of the Student *t*-distribution for a fixed value of the enclosed probability p , according to (9.37). The heavy tails of the Cauchy distribution can also be noted in the $f = 1$ row of the table.

Example 9.8 (*Comparison of sample mean with parent mean—Part II*) Assume now that the five measurements of Example 9.7 (10, 12, 15, 11, 13) are distributed like a Gaussian of $\mu = 13$, but without reference to a parent variance. In this case, the sample variance is

$$s^2 = \frac{1}{4} \sum (x_i - \bar{x})^2 = 3.7,$$

and the *t*-statistic is calculated as

$$t = \frac{\bar{x} - \mu}{s/\sqrt{5}} = \frac{12.2 - 13}{1.92/\sqrt{5}} = -0.93.$$

Table 9.1 Two-sided critical values T_{crit} of the t -distribution for a fixed value of the enclosed probability p , and for selected values of the number of degrees of freedom

Number of d.o.f. (f)	Enclosed probability p			
	0.683	0.90	0.95	0.99
Values of T_{crit}				
1	1.839	6.314	12.706	63.657
2	1.322	2.920	4.303	9.925
5	1.111	2.015	2.571	4.032
10	1.053	1.812	2.228	3.169
20	1.026	1.725	2.086	2.845
50	1.011	1.676	2.009	2.678
100	1.006	1.660	1.984	2.626
∞ (Gaussian)	1.000	1.645	1.960	2.576

This value of t corresponds to a p -value of 0.405, using the t -distribution with four degrees of freedom of Table A.23. It is clear that the estimation of the variance from the data has added a new source of uncertainty in the comparison of the measurement with the parent distribution. This results in a larger p -value than in the case of Example 9.7, when the sample mean was compared to parent mean while using the parent variance. \diamond

It is useful to also address the fiducial confidence intervals on the parent mean of the Gaussian distribution from the measurement of the sample mean, following the same treatment of Sect. 7.3. In that case, it was assumed that the parent variance was known, leading to the traditional confidence interval (7.2) and the fiducial confidence interval (7.3). When the parent variance is not known, the Student t -distribution needs to be used instead, as discussed in this section. That leads, for example, to a $p = 0.90$ fiducial confidence interval on the parent mean of the type

$$P(\bar{x} - T_{crit} \cdot (s/\sqrt{N}) \leq \mu \leq \bar{x} + T_{crit} \cdot (s/\sqrt{N})) = p, \quad (9.38)$$

where the critical value of the t -statistic, T_{crit} , is now also function of the number of measurements, instead of just the level of probability p . This confidence interval replaces (7.3) when the variance is not known a priori. In the limit of a large number of measurements, the two intervals become progressively closer to one another, since the critical values of the t -distribution approach those of a normal Gaussian in the large- N limit. As can be seen from Fig. 9.4 and Table 9.1, the two-sided critical values of the t -distribution, for a fixed value of the enclosed probability p , are always *larger* than that of the corresponding Gaussian distribution. This is caused by the additional uncertainty associated with the unknown parent mean, and as a result the t -based fiducial confidence level (9.38) is always large than the Gaussian-based interval (7.3).

9.6.3 Comparison of Two Sample Means and Hypothesis Testing

The same distribution function is also applicable to the comparison between two sample means \bar{x}_1 and \bar{x}_2 , derived from samples of size N_1 and N_2 , respectively. In this case, a new *t*-statistic is defined as

$$T = \frac{\bar{x}_1 - \bar{x}_2}{s\sqrt{1/N_1 + 1/N_2}}, \quad (9.39)$$

where

$$\begin{cases} S^2 = S_1^2 + S_2^2 \\ s^2 = \frac{S^2}{N_1 + N_2 - 2} \\ S_1^2 = \sum_{i=1}^{N_1} (x_i - \bar{x}_1)^2 \\ S_2^2 = \sum_{j=1}^{N_2} (x_j - \bar{x}_2)^2. \end{cases}$$

This new statistic is distributed like a *T*-distribution with $f = N_1 + N_2 - 2$ degrees of freedom, and therefore the same distribution as (9.30) can also be used for testing the agreement between two sample means, when neither the parent mean or the parent variance are specified.

Under the hypothesis that all measurements are drawn from the same parent distribution, $X \sim N(\mu, \sigma)$,

$$\begin{cases} \frac{\bar{x}_1 - \mu}{\sigma/\sqrt{N_1}} \sim N(0, 1) \\ \frac{\bar{x}_2 - \mu}{\sigma/\sqrt{N_2}} \sim N(0, 1) \end{cases}$$

and from (9.20)

$$\begin{cases} \frac{S_1^2}{\sigma^2} \sim \chi^2(N_1 - 1) \\ \frac{S_2^2}{\sigma^2} \sim \chi^2(N_2 - 1). \end{cases}$$

First, the distribution function for the difference between the sample means is needed. Assuming that the measurements are independent, then the difference of the sample means is normally distributed with zero mean and with variances added in quadrature, leading to

$$X = \frac{\bar{x}_1 - \bar{x}_2}{\sigma \sqrt{1/N_1 + 1/N_2}} \sim N(0, 1).$$

Next, since the sum of independent χ^2 variables is also distributed like a χ^2 distribution with a number of degrees of freedom equal to the sum of the individual degrees of freedom,

$$Z = \frac{S_1^2}{\sigma^2} + \frac{S_2^2}{\sigma^2} \sim \chi^2(N_1 + N_2 - 2).$$

The distribution of $\sqrt{Z/f}$ follows (9.34), with $f = N_1 + N_2 - 2$ as the number of degrees of freedom for both datasets combined. As a result, the new T -statistic can be written as

$$T = \frac{X}{\sqrt{Z/f}} \quad (9.40)$$

in a form that is identical to the original T -statistic, except for the different number of degrees of freedom. It is therefore possible to conclude that the random variable defined in (9.39) has a t -distribution with $f = N_1 + N_2 - 2$ degrees of freedom.

Example 9.9 (*Comparison between two sample means*) For the usual ten measurements (10, 12, 15, 11, 13, 16, 12, 10, 18, 13), the sample means of the first five and second five measurements are $\bar{x}_1 = 12.2$ and $\bar{x}_2 = 13.8$, the sample variances are proportional to $S_1^2 = 14.8$ and $S_2^2 = 40.8$, for

$$s^2 = \frac{S_1^2 + S_2^2}{8} = 6.95.$$

This results in a measurement of the t -distribution for the comparison between two means of

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s \sqrt{1/5 + 1/5}} = -0.96. \quad (9.41)$$

This number is to be compared with a t -distribution with 8 degrees of freedom, for a p -value of 0.365. Since there is a substantial probability that the difference between the two means can have the observed value or a more extreme one, the conclusion is that the measurements are consistent with being drawn from normal distributions with the same mean, without committing to a specific parent value. ◇

Summary of Key Concepts for this Chapter

Hypothesis Testing: A four-step process that consists of (1) defining a null hypothesis to test, (2) determine the relevant statistic (e.g., χ^2), (3) a confidence level (e.g., 90 %), and (4) whether the null hypothesis is rejected or not.

χ^2 distribution: The theoretical distribution of the sum of the squares of independent z -scores,

$$\chi^2 = \sum_{i=1}^N \left(\frac{x_i - \mu_i}{\sigma_i} \right)^2,$$

with mean N and variance $2N$.

Sampling distribution of variance: The distribution of sample variance s^2 is such that

$$\frac{s^2}{\sigma^2} \sim \chi^2(N - 1),$$

where $S^2 = (N - 1)s^2$.

F-Statistic: Distribution of the ratio of independent χ^2 variables

$$F = \frac{\chi_1^2/f_1}{\chi_2^2/f_2}$$

(mean $f_2/(f_2 - 2)$ for $f_2 > 2$).

Student's t-distribution: Distribution for the variable

$$T = \frac{\bar{x} - \mu}{s/\sqrt{N}},$$

useful to compare the sample mean to the parent mean when the variance is estimated from the data, and two sample means with each other.

Problems

9.1 Five students score 70, 75, 65, 70, and 65 on a test. Determine whether the scores are compatible with the following hypotheses:

- (a) The mean is $\mu = 75$.
- (b) The mean is $\mu = 75$ and the standard deviation is $\sigma = 5$.

Test both hypotheses at the 95 % or 68 % confidence levels, assuming that the scores are normally distributed.

9.2 Prove that the mean and variance of the F -distribution are given by the following relationships:

$$\begin{cases} \mu = \frac{f_2}{f_2 - 2} \\ \sigma^2 = \frac{2 f_2^2 (f_1 + f_2 - 2)}{f_1 (f_2 - 2)^2 (f_2 - 4)}, \end{cases}$$

where f_1 and f_2 are the degrees of freedom of the variables at the numerator and denominator, respectively.

9.3 Using the same data as Problem 9.1, test whether the sample variance is consistent with a parent variance of $\sigma^2 = 25$, at the 95 % level. For this problem, you should not specify a parent mean for the null hypothesis.

9.4 ■ Consider the J.J. Thomson experiment data of Sect. 2.5.

- (a) Measure the ratio of the sample variances of the m/e measurements in Air for Tube 1 and Tube 2.
- (b) Use the F -statistic to determine if the null hypothesis that the two sets of measurements for Tube 1 and Tube 2 are drawn from the same distribution can be rejected at the 90 % confidence level.
- (c) State the null hypothesis being tested in (b) and all assumptions required to use the F -distribution.

9.5 Consider a dataset of ten measurements (10, 12, 15, 11, 13, 16, 12, 10, 18, 13), assumed to be drawn from a normal distribution.

- (a) Calculate the ratio of the sample variance of the first two measurements with that of the last eight.
- (b) Determine at what confidence level the two subsets are consistent with the null hypothesis that the two subsets have the same parent variance.

9.6 Six measurements of the length of a wooden block gave the following measurements: 20.3, 20.4, 19.8, 20.4, 19.9, and 20.7 cm.

- (a) Estimate the mean and the standard error of the length of the block.

- (b) Assume that the block is known to be of length $\mu = 20\text{ cm}$. Establish if the measurements are consistent with the known length of the block, at the 90% probability level.

9.7 ■ Consider the “Long vs. short stem” data in Mendel’s experiment of Table 1.1 at Sect. 1.5.

- (a) Identify the distribution and an expression that describes the probability of making that measurement, assuming Mendel’s hypothesis of independent assortment. You do not need to evaluate the probability.
- (b) Using the distribution function that pertains to that measurement, determine the mean and variance of the parent distribution. Using the Gaussian approximation for the distribution, determine if the null hypothesis that the measurement is drawn from the parent distribution is compatible with the data at the 68% confidence level.

9.8 ■ Consider Mendel’s experimental data in Table 1.1 at Sect. 1.5. Using all seven measurements, calculate the level of confidence for the agreement of the combined fraction of dominant characters with the theoretical expectation of 0.75. For this purpose, you may use the t -statistic.

9.9 Starting with (9.35), complete the derivation of (9.33).

9.10 Show that the t -distribution,

$$f_T(t) = \frac{1}{\sqrt{f\pi}} \frac{\Gamma(f/2 + 1/2)}{\Gamma(f/2)} \left(1 + \frac{t^2}{f}\right)^{-\frac{f+1}{2}},$$

becomes a standard Gaussian in the limit of large f . You can make use of the asymptotic expansion of the Gamma function (A.12).

9.11 Prove that the moments of the t -distribution follow (9.36). You can use the definition (9.32) and the independence of the numerator and denominator.

Chapter 10

Contingency Tables and Diagnostic Tests



Abstract Contingency tables are a convenient form to report and analyze data that contain characteristics of two properties. In the simplest and most common form, each property has two possible outcomes, and the contingency table contains four numbers. Contingency tables are often used to test for the independence between the properties, although more complex hypothesis tests can also be performed. K. Pearson was the first to use an approximate χ^2 test for independence in contingency tables, while R.A. Fisher devised an exact method to do hypothesis testing on contingency tables based on the binomial distribution. Contingency tables are convenient for binary diagnostic testing and provide a convenient framework to study such properties as true or false positives, and the sensitivity and specificity of tests.

10.1 A Classic Experiment: The 1915 Greenwood and Yule Inoculation Statistics

In 1915, M. Greenwood and U. Yule published a collection of statistics on the effect of immunization against typhoid fever and cholera, and statistics on other diseases [46]. The data presented by the authors include several samples, with the primary goals to investigate whether various forms of vaccines were effective or not. The data are typically presented in the form of Table 10.1, which reproduces Table II of [46]. Another example is in Table 10.2, which reproduces their Table XV, and Table 10.3, which reproduces their Table VIII. These data are all in the form of *contingency tables*, which report the outcomes of two binary properties, in this case, inoculation status and “attack” or disease status.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-981-19-0365-6_10.

Table 10.1 Reproduction of Table II from Greenwood and Yule [46]. The authors reported the associated statistics of $\chi^2 = 56.23$ and $P = \text{less than } 0.0001$

	Not attacked	Attacked	Total
Inoculated	6,759	56	6,815
Not inoculated	11,395	272	11,668
Total	18,155	328	18,483

Table 10.2 Reproduction of Table XV from Greenwood and Yule [46]. Associated statistics were reported as $\chi^2 = 5.61$ and $P = 0.1351$

	Not attacked	Attacked	Total
Inoculated	54	5	59
Non-inoculated	46	15	61
Total	100	20	120

Table 10.3 Reproduction of Table VIII from Greenwood and Yule [46]. Associated statistics were reported as $\chi^2 = 3.18$ and $P = 0.3682$

	Not attacked	Attacked	Total
Inoculated	105	5	110
Non-inoculated	88	11	99
Total	193	16	209

From data in the form of contingency table, it is possible to determine whether there is an *association* between the two properties, with the goal to determine quantitatively whether inoculated people were less susceptible to being attacked by the disease, and likewise non-inoculated people were more susceptible to the disease. Methods of analysis for contingency tables, with applications to these data, are presented throughout the chapter.

10.2 2×2 Contingency Tables

Contingency tables describe the occurrence of two properties, say properties A and B , in a tabular form of the type illustrated in Table 10.4. For example, property A may be whether a person has been vaccinated, and property B whether the person contracts the disease. When the two properties have a binary outcome, then the data are in the form of a 2×2 contingency table. A compact form for a 2×2 contingency table is $(a; b; c; d)$. Quantities $n_1 = a + b$, $n_2 = c + d$, $m_1 = a + c$ and $m_2 = b + d$ are called *margins* or marginal values, and $N = n_1 + n_2 = m_1 + m_2$ is the total number of entries in the table. The table can always be arranged in such a way that, for convenience, $m_1 \leq m_2$, and $n_1 \leq n_2$. The following probabilities are defined:

Table 10.4 Notation for 2×2 contingency tables, following Yates [107]. A_1 and A_2 are the two possible outcomes of property A , and B_1 and B_2 the two possible outcomes of property B

	B_1	B_2	Total	
A_1	a	b	n_1	$p_1 = a/n_1$
A_2	c	d	n_2	$p_2 = c/n_2$
Total	m_1	m_2	N	$p = m_1/N$

$$\begin{cases} p_1 = a/n_1 : \text{probability of occurrence of } B_1, \text{ given } A_1, P(B_1/A_1); \\ p_2 = c/n_2 : \text{probability of occurrence of } B_1, \text{ given } A_2, P(B_1/A_2); \\ p = m_1/N : \text{probability of occurrence of } B_1, P(B_1). \end{cases} \quad (10.1)$$

A typical goal of contingency tables is to test whether there is a dependence, or *association*, between the two events. For example, if A is the binary event that describes if a person is vaccinated, and the binary event B describes whether the person becomes ill, the null hypothesis being tested is that there is no dependence between the two events, i.e., the inoculation has no effect.

To test whether there is an association between the properties, it is necessary to develop a model for the data that follows the null hypothesis of no association between the two properties. If the two properties A and B are independent of one another, conditioning on the occurrence of A_1 or A_2 is irrelevant. It is convenient to make the assumption that the margins are fixed, meaning that the total number of occurrence of property A_1 (e.g., total number n_1 of inoculated persons) and of property A_2 (e.g., total number n_2 of persons that were not inoculated) are fixed, the number of occurrence of property B_1 (number m_1 of persons who are healthy) and B_2 (number m_2 of person who become ill) are also fixed, in addition to the grand total N . With this assumption, only one of the four numbers in the table is free to vary to meet the constraints enforced by the margins.

A model for the independence between the two properties is obtained with the use of the ratios p_1 , p_2 and p . With fixed margins, the parent probabilities π_1 and π_2 for the ratios p_1 and p_2 are derived from a binomial distributions. Under the hypothesis of independence between A and B , the two parent probabilities are the same, $\pi_1 = \pi_2 \equiv \pi$, and this parent value applies also to the probability p . Assuming that the number N is fixed, this probability is obtained from the data as $\pi \simeq m_1/N$. It follows that the expectation of a , assuming independence between the events, is

$$E[a] = n_1\pi = \frac{m_1n_1}{N}. \quad (10.2)$$

This is the expectation for a assuming independence between the events. The expectation of the remaining three numbers is calculated accordingly since all the margins are assumed to be fixed. These expectations are used for testing the hypothesis of independence.

10.2.1 The χ^2 Test

K. Pearson was the first to introduce the χ^2 test for association in contingency tables in 1900 [76, 77]. Under the null hypothesis of independence, the expectations of the four numbers $O_i = (a, b, c, d)$, hereafter referred to as E_i , can be used to construct the following χ^2 statistic

$$\chi^2 = \sum_{i=1}^4 \frac{(O_i - E_i)^2}{E_i}. \quad (10.3)$$

This statistic appears to differ, at first sight, from a χ^2 -distributed variable, defined in Sect. 9.3 as the sum of the squares of the standardized deviations of f independent Gaussian variables. In fact, the sum here extends over four terms that are correlated by the binomial character of the variables. It is possible to show that Eq. 10.3 is *approximately* distributed like a χ^2 variable, with a number of degrees of freedom that depends on how many parameters of the contingency table are constrained by the data. In the case under consideration of a 2×2 contingency table with fixed margins, the parent distribution under the hypothesis of independence is $\chi^2(1)$.

The four data points O_i can be considered as drawn from independent Poisson-distributed variables, plus the additional condition that their sum is constrained to be N . The Poisson distribution has the desirable property that $E_i = N P_i$, where the Poisson probabilities P_i are given by

$$\begin{cases} P_1 = \frac{m_1}{N} \frac{n_1}{N} \\ P_2 = \frac{N - m_1}{N} \frac{n_1}{N} \\ P_3 = \frac{m_2}{N} \frac{n_2}{N} \\ P_4 = \frac{N - m_2}{N} \frac{n_2}{N}, \end{cases} \quad (10.4)$$

satisfying the condition that $P_1 + P_2 + P_3 + P_4 = 1$. The expectations E_i are the same as for the binomial case, but the variance is now equal to the expectation. As a result, the denominator in Eq. 10.3 provides the proper standardization, provided that the count rates are not too small, so that a Poisson distribution is reasonably approximated by a normal distribution.

The additional and more subtle complication comes from estimating certain proportions (or parameters) of the table from the data, instead of being specified *a priori* as part of the null hypothesis. In addition to using the value of N as observed from the data, the probabilities P_i used the ratios m_1/M and m_2/N , estimated from the measured marginal values m_1 and m_2 . This brings the total of three parameters (N, m_1 and m_2) being estimated from the data.

It is worth noticing that this reduction of the number of degrees of freedom was not initially appreciated by Pearson, who suggested that the fit statistic was distributed like a $\chi^2(3)$, i.e., with $N - 1 = 3$ degrees of freedom. R. A. Fisher was the first to realize that when certain parameters of a model are estimated from the data, and not known exactly *a priori*, there must be a change in the parent distribution of the χ^2 fit statistic. In particular, Fisher examined the case of contingency tables in which a number m of proportions were fixed to the measured value, and not assumed *a priori*. In this case, Fisher showed that the test statistic is distributed like $\chi^2(N - m - 1)$, and not $\chi^2(N - 1)$ as previously assumed. This modification helped ease concerns in the interpretation of a dataset published by Greenwood and Yule [46] and reported in Sect. 10.1. The result proposed by R. A. Fisher in a 1922 paper [28] can be summarized as the following theorem:

Theorem 10.1 (Fisher's theorem on the limiting distribution of χ^2_{min} for contingency tables)

For an $r \times c$ contingency table with fixed margins, the asymptotic parent distribution for the measured χ^2 (10.3) in the large count limit is $\chi^2(f)$, with $f = (c - 1) \times (r - 1)$. A more complete discussion and proof of the theorem are provided in [28].

For 2×2 contingency tables with fixed margins, the number of degrees of freedom is therefore $f = 1$, consistent with the use of N and the margins m_1 and m_2 in the model of the data.

For 2×2 tables with fixed margins, the χ^2 of Eq. 10.3 can be written in a more convenient form as

$$\chi^2 = \frac{(ad - bc)^2 N}{m_1 m_2 n_1 n_2}. \quad (10.5)$$

This χ^2 statistic can be tested for association between the events using critical values of the parent distribution $\chi^2(1)$, provided the numbers are sufficiently large.

Example 10.1 (*Greenwood and Yule's Table II*) Using Table II of Greenwood and Yule, the expected values based on the hypothesis of independence are $E = (6694.06, 120.94; 11460.94, 207.06)$. The fit statistic is $\chi^2 = 56.234$, as reported by Greenwood and Yule, and for such a large value the null hypothesis that there is independence between the events that must be discarded. As discussed in Sect. 9.2, hypothesis testing with p values should not *by itself* be interpreted as evidence in favor of a specific hypothesis. In this example, however, given the binary nature of the statistical model (i.e., independence versus dependence), failure of the null hypothesis could be seen as qualitative evidence that there is a relationship between vaccination and the probability of becoming ill. The test itself does not tell in which direction the correlation goes, but since there is a larger fraction of ill persons in the non-vaccinated group, one can conclude that vaccination is effective. ◇

10.2.2 χ^2 Test with the Yates Continuity Correction

In the low-count regime, the approximation of a Poisson distribution with a normal is not accurate. In addition, the χ^2 distribution is continuous, whereas the Poisson or binomial distributions it aims to approximate are integer-valued and therefore discontinuous. As will be shown in the following section, it is possible to provide an exact test for 2×2 contingency tables, and that method is to be preferred to the approximate χ^2 test, especially in the low-count regime. Nonetheless, it remains convenient to use the χ^2 test even in the low-count regime. For this purpose, F. Yates [106] used true probabilities to show that the χ^2 distribution always underestimates the true probability of a given value of the Pearson χ^2 statistic, and devised a simple correction to improve its accuracy. With this correction, referred to by Yates as a *continuity correction*, the statistic becomes

$$\chi_c^2 = \frac{(|ad - bc| - N/2)^2 \cdot N}{m_1 m_2 n_1 n_2}. \quad (10.6)$$

The use of Eq. 10.6 in place of Eq. 10.5 is suggested when there are fewer than approximately five counts in some of the cells of the contingency table.

Example 10.2 (*Greenwood and Yule's Table VIII and Table XV*) Using Table VIII of Greenwood and Yule, one obtains a Pearson statistic of $\chi^2 = 3.18$, as originally reported. The use of the Yates continuity correction leads to a corrected value of $\chi_c^2 = 2.32$. It is instructive to use Greenwood and Yule's Table VIII to illustrate the effect of the Yates correction and of the correct number of degrees of freedom when interpreting contingency tables:

Statistic used	Value of statistic	p -value using reference statistic $\chi^2(3)$	p -value using reference statistic $\chi^2(1)$
χ^2	3.18	0.365	0.075
χ_c^2 (with Yates correction)	2.32	0.509	0.128

Greenwood and Yule followed Pearson in assuming that the χ^2 statistic had a parent distribution with three degrees of freedom and reported a probability of $P = 0.368$ to exceed this value, which is nearly identical to the one reported in the table above. Instead, using the appropriate distribution with just one degree of freedom (as per Fisher's recommendation), the probability to exceed this value is just $P = 0.075$, when using the uncorrected value of χ^2 . Using the Yates correction, the probability to exceed the measured value is now $P = 0.128$, using the correct $\chi^2(1)$ reference distribution. It is clear that the measured data cannot conclusively discard the null hypothesis of independence between immunization and contracting the disease, when using the Yates correction and the correct reference distribution. The vaccine under consideration in this table is therefore not proven to be effective with high confidence.

For Table XV, the following results are obtained:

Statistic used	Value of statistic	<i>p</i> -value using reference statistic	
		$\chi^2(3)$	$\chi^2(1)$
χ^2	5.61	0.132	0.018
χ^2_c (with Yates correction)	4.51	0.212	0.034

Greenwood and Yule would erroneously conclude, based on the $\chi^2(3)$ parent distribution, that the vaccination was not very effective, having inferred a *p*-value of 0.1351 that is very similar to the one reported in the table above. However, using the correct parent distribution, the data are quite inconsistent with the null hypothesis, for a *p*-value of approximately 1.8%, which becomes 3.4% using the Yates correction, suggesting that vaccination was in fact quite effective.

10.2.3 The Fisher Exact Test for 2×2 Contingency Tables

In his 1925 textbook *Statistical Methods for Research Workers* [30], Fisher provides an exact method for the treatment of 2×2 contingency tables. The method is briefly described below, leading to the following probability for the observation of a given contingency table, assuming that all margins are fixed:

$$P(a, b, c, d) = \frac{n_1! n_2! m_1! m_2!}{N!} \frac{1}{a! b! c! d!} \quad (10.7)$$

If p is the probability of occurrence of an event, say event B_1 according to Eq. 10.1, the probability that it will occur a times out of $n_1 = a + b$ times is given by the binomial probability

$$\frac{n_1!}{a! b!} p^a q^b.$$

Same is true for the second row of the table, whereby the probability of c occurrences out of $n_2 = c + d$ tries is

$$\frac{n_2!}{c! d!} p^c q^d.$$

As a result, the probability of observing a specific contingency table $(a, b; c, d)$ with fixed margins is given by the product of the two probabilities,

$$\frac{n_1! n_2!}{a! b! c! d!} p^{m_1} q^{m_2}.$$

If the probability p was given, then the probability above could be used for hypothesis testing. In most cases, however, p is not known a priori, and additional considerations are required. Continuing with the assumption that the margins are fixed, the probability of observing a specific table can be written as

$$P(a, b, c, d) \propto \frac{1}{a! b! c! d!}$$

regardless of the value of p , provided that this probability remains fixed. This distribution must be normalized by evaluating the sum of probabilities for all possible tables. This is easily accomplished because a 2×2 contingency table with N and the margins held fixed has just one degree of freedom. For example, assuming that $n_1 \leq m_1$ (this assumption can always be satisfied by rearranging the order of the numbers in the table), all possible contingency tables are of the following form:

$$(i, n_1 - i; m_1 - i, m_2 - (n_1 - i)), \text{ for } i = 0, \dots, n_1.$$

The sum of probabilities for all possible tables is therefore evaluated as

$$\sum_{i=0}^{n_1} \frac{1}{i! (n_1 - i)! (m_1 - i)! (m_2 - (n_1 - i))!} = \sum_{i=0}^{n_1} \binom{m_1}{i} \frac{1}{m_1!} \binom{m_2}{n_1 - i} \frac{1}{m_2!}.$$

Using the fact that

$$\sum_{k=0}^p \binom{n}{k} \binom{m}{p-k} = \binom{n+m}{p}$$

leads to a sum of

$$\frac{1}{m_1! m_2!} \binom{m_1 + m_2}{n_1} = \frac{N!}{m_1! m_2! n_1! n_2!}.$$

Finally, the normalized probability distribution is:

$$P(a, b, c, d) = \frac{m_1! m_2! n_1! n_2!}{a! b! c! d! N!}$$

This distribution is commonly known as the *hypergeometric distribution*, which is intended to describe the probability of occurrence of a type-I objects in a binary population of size N with a total of m_1 type-I objects and the remaining m_2 of type-II, when n_1 of the N objects are sampled at random. The usual and equivalent form for the hypergeometric distribution is

$$P(a, b, c, d) = \frac{\binom{m_1}{a} \binom{m_2}{b}}{\binom{N}{n_1}}$$

where the numerator is the product of the number of ways to sample a objects from m_1 and the number of ways to sample the remaining $b = n_1 - a$ objects from the remaining $m_2 = N - m_1$, and the denominator is the number of ways to sample the n_1 objects from the total N . When $n_1 \leq m_1$, the possible values of a range from 0 to n_1 .

The Fisher exact test makes use of the distribution in Eq. 10.7, and it consists of calculating the exact cumulative probability of obtaining a table that has a probability of occurrence, according to Eq. 10.7, that is *equal or lower* than the observed table. This means that the probabilities need to be summed over all possible datasets (consistent with the fixed margins) that have an equal or lower probability of occurrence than the observed data.

Example 10.3 (*Fisher exact test on tables with $N = 10$*) To illustrate the Fisher exact test and the associated cumulative probability, it is convenient to follow an example of a 2×2 table with $N = 10$,

$$(2, 1; 3, 4),$$

similar to the one presented by Yates in his Table 2. With N fixed, the table is specified by the margins and by the value of the remaining degree of freedom. The table has $m_1 = m_2 = 5$, and $n_1 = 3$, $n_2 = 7$, and the margins are considered fixed. Notice that the requirements that $m_1 \leq m_2$ and $n_1 \leq n_2$ are already satisfied, so the table needs not be rearranged.

With the margins fixed, the table has just one degree of freedom, taken to be the number a . Therefore, only possible values of a that are consistent with the margins are $i = 0, 1, 2, 3$, or $n_1 + 1$ values (since $n_1 < m_1$). The corresponding tables have these probabilities

Value of a	Corresponding table	Fisher's exact probability
0	(0, 3; 5, 2)	0.0833
1	(1, 2; 4, 3)	0.4167
2	(2, 1; 3, 4)	0.4167 (observed table)
3	(3, 0; 2, 5)	0.0833

The Fisher exact test concludes that there is a cumulative probability of $P = 1$, or 100%, of obtaining a distribution of counts *at least as extreme* as the one observed. Clearly, there is nothing “extreme” about the observed table, and the test reflects it.

Consider now another $N = 10$ table, $(2, 3; 1, 4)$, with $n_1 = 5$, $n_2 = 5$, $m_1 = 3$ and $m_2 = 7$. It is easy to show that there are again four possible tables with the same

fixed margins, and that the results of the Fisher exact test are identical to the ones of the previous table (see Problem 10.1).

Finally, consider the table $(0, 5; 4, 1)$. Since $n_1 > m_1$, there are $m_1 + 1$ possible values for a , namely $i = 0, \dots, 4$, corresponding to the following tables:

Value of a	Corresponding table	Fisher's exact probability
0	$(0, 5; 4, 1)$	0.0238* (this is the observed table)
1	$(1, 4; 3, 2)$	0.2381
2	$(2, 3; 2, 3)$	0.4761
3	$(3, 2; 1, 4)$	0.2380
4	$(4, 1; 0, 5)$	0.0238*

For the Fisher exact test, one would add the probabilities marked with an asterisk, corresponding to the extreme cases, leading to a cumulative probability of 0.048 or 4.8% of observing a table at least as extreme as the measured one. For this table, there is less than 5% probability that the data are consistent with the hypothesis of independence, and many analysts would conclude that it is likely that there is correlation between the events. For this table with low-count data, the χ^2 test with the Yates correction would provide a value of $\chi_c^2 = 3.75$, for a p -value of 0.053 when compared with the $\chi^2(1)$ distribution. The approximate χ^2 test with the Yates correction, therefore, provides a reasonable approximation of the Fisher exact test. Without the Yates correction, the χ^2 test would result in a significantly larger value $\chi^2 = 6.67$ with a p -value of approximately 0.01, that is significantly at odds with the exact test. \diamond

10.2.4 Exact Tests Based on the Binomial Distribution

The Fisher exact test makes two notable assumptions: that the margins are fixed, and that there is independence between the two events. There are situations when the events can be modeled with a binomial distribution of *known* probability p , instead of estimating the proportions based on the observed margins. Assuming that the events are independent, it is immediate to see that the probability of obtaining a given 2×2 contingency table is given by the product of two binomial distributions, as pointed out by G. A. Barnard [7],

$$P(a, b, c, d) = \frac{n_1!}{a! b!} p_1^a q_1^{n_1-a} \times \frac{n_2!}{c! d!} p_2^c q_2^{n_2-c}.$$

This result can be seen as two independent Bernoulli trials of, respectively, n_1 and n_2 events, from binomial distributions with probability of success of respectively p_1 and p_2 . If, moreover, the hypothesis to test predicts that $p_1 = p_2 = p$, then one obtains the following probability:

$$P(a, b, c, d) = \frac{n_1! n_2!}{a! b! c! d!} p^{m_1} q^{m_2}. \quad (10.8)$$

Equation 10.8 differs from the hypergeometric probability of the Fisher exact test (Eq. 10.7) by a factor of

$$\frac{N!}{m_1!m_2!} p^{m_1} q^{m_2}$$

that corresponds to the binomial probability of choosing at random m_1 of N events with probability p , instead of holding the numbers fixed to their measured values. Given that this factor is clearly smaller than one, the probability of Eq. 10.8 is always smaller than that of Fisher's exact test.

Hypothesis testing of 2×2 contingency tables with this probability distribution can be more complicated than for the Fisher exact test, depending on the constraints imposed on the table. The Fisher exact test assumed fixed margins, thus leaving only one degree of freedom, i.e., the value of the number a . When there are no constraints or only one constraint to take into account, all possible tables that are consistent with the constraints need to be enumerated accordingly. An example of data with one additional constraint is reported in the following example, which reproduces the treatment presented by Yates [107] in his Table 2.

Example 10.4 (*Binomial test of 2×2 tables with N and n_1 fixed*) Consider a table with $N = 10$, constrained to a fixed margin n_1 , with $p_1 = p_2 = 1/2$. It is easy to see that tables consistent with these constraints have two degrees of freedom; for example, $a = 0, \dots, n_1$ and, for a given value of a , $c = 0, \dots, n_2$ (or $m_1 = a, \dots, a + n_2$) for a total of $n_2 + 1$ values of the number c . This results in a total of $(n_2 + 1) \times (n_1 + 1)$ tables, which corresponds to 36 tables when $n_1 = n_2 = 5$. For example, there is one table with margins $(m_1, m_2) = (10, 0)$, namely $(5, 0; 5, 0)$, and two tables with margins $(m_1, m_2) = (9, 1)$, namely $(4, 1; 5, 0)$ and $(5, 0; 4, 1)$, etc. For each one of these tables, two probabilities apply. One is the hypergeometric distribution (10.7), which assumes fixed margins. The other is the binomial distribution

$$P(a, b, c, d) = \binom{n_1}{a} \left(\frac{1}{2}\right)^N,$$

which is the probability of choosing a specific value of a , without specifying the m_1, m_2 margins. The binomial probability will therefore have to be summed over all tables with different m_1, m_2 margins to obtain a probability of one. For example:

Contingency table	Hypergeometric probability	Binomial probability
$(5, 0; 5, 0)$	1	$1 \times (1/2)^{10}$
$(4, 1; 5, 0)$	$1/2$	$5 \times (1/2)^{10}$
$(5, 0; 4, 1)$	$1/2$	$5 \times (1/2)^{10}$
	...	

These probabilities, illustrated in Table 10.5, are the basis for hypothesis testing. As an example, assume that the table $(1, 4; 5, 0)$ is measured, with margins $(m_1, m_2) = (6, 4)$ and $(a - c)/n_1 = -0.8$, corresponding to the top entry in the $(6, 4)$ column of Table 10.5. This table is the most extreme case for the given margins.

Table 10.5 Frequency of occurrence of the 36 tables obtained from $n_1 = n_2 = 5$, $N = 10$ and $p_1 = p_2 = 1/2$; the numbers must be multiplied by $(1/2)^{10}$ to obtain a probability. In parenthesis are the probabilities obtained from the hypergeometric distribution that assumes the margins m_1 , m_2 as fixed. This table is partially reproduced from [107]. Since the table is symmetrical, duplicate entries are not shown

$\frac{a - c}{n_1}$	(m_1, m_2) margins							Total
	(10,0)	(9,1)	(8,2)	(7,3)	(6,4)	(5,5)	...	
-1.0						1 (0.004)	...	1
-0.8					5 (0.024)	—	...	10
-0.6				10 (0.083)	—	25 (0.099)	...	45
-0.4			10 (0.22)	—	50 (0.238)	—	...	120
-0.2		5 (0.5)	—	50 (0.417)	—	100 (0.397)	...	210
0.0	1 (1.0)	—	25 (0.556)	—	100 (0.476)	—	...	252
0.2		5 (0.5)	—	50 (0.417)	—	100 (0.397)	...	210
0.4			10 (0.22)	—	50 (0.238)	—	...	120
0.6				10 (0.083)	—	25 (0.099)	...	45
0.8					5 (0.024)	—	...	10
1.0						1 (0.004)	...	1
Total	1	10	45	120	210	252	...	1024

The probability to obtain a table as extreme or more extreme than this regardless of margins, as measured by the $(a - c)/n_1$ metric, is given by $(1 + 10)/1024 = 0.0107$ if the hypothesis is one-sided, or twice that amount if the hypothesis is double-sided. This probability is obtained by including all tables with $(a - c)/n_1$ as extreme or more extreme than the one measured, in this example the tables in the first two rows of Table 10.5 for a one-sided hypothesis, and also the tables in the last two rows for a double-sided hypothesis.

If the experiment was constrained by the knowledge of the fixed m_1 , m_2 margins, then the probability of such an extreme or more extreme table is $5/210 = 0.0238$ for a one-sided hypothesis, or twice as large for a double-sided hypothesis; the latter corresponds to the Fisher exact test, which yields a probability of 0.0476. As expected, knowledge of the margins increases the probability of occurrence of a given table. ◇

Another contingency table that can be studied with this method is Mendel's Table 1.2 for the first-generation plants from the hybrids. In this case, the Mendel

hypothesis on recessive and dominant characters implies that binomial distributions with $p_1 = p_2 = 3/4$ should be used.

Example 10.5 (*Mendel's data for first-generation plants from the hybrids*) Mendel's data from Table 1.2 is an example of 2×2 contingency table, where one event is the color of the seed (yellow or green) and the other event is the shape (round or wrinkled). The contingency table can be reported as

$$(101, 32; 315, 108),$$

with rows inverted so that $n_1 \leq m_1$. The table has $N = 556$, with margins $n_1 = 133$ as the number of wrinkled (recessive) seeds and $m_1 = 416$ as the number of yellow (dominant) seeds. Using the simple χ^2 test for independence yields a value of $\chi^2 = 0.12$, which shows a nearly perfect match between the data and the independence model with fixed margins. The χ^2 statistic only tests the independence between the color and the shape, with fixed ratios

$$\frac{\text{wrinkled seeds}}{\text{all seeds}} = \frac{n_1}{N} = 0.239$$

and

$$\frac{\text{yellow seeds}}{\text{all seeds}} = \frac{m_1}{N} = 0.748.$$

For example, these ratios are used to calculate the expectation of a as

$$E[a] = N \times \frac{m_1}{N} \cdot \frac{n_1}{N} = 99.51$$

which is in fact very close to the measured value of $a = 101$. The Fisher exact test can also be used to test independence, with probabilities calculated according to (10.7) for fixed margins. For this test, there are a total of $n_1 + 1$ tables to consider that are consistent with the fixed margins. Of these, only two (namely (99, 34; 317, 106), with probability 0.0902, and (100, 33; 316, 107), with probability 0.0908 have a probability of occurrence that is *less* extreme than that of the observed table (with probability 0.0868). As a result, the Fisher exact test for independence shows a cumulative probability of 0.819 to observe the current table or a more extreme one, meaning that the observed data are perfectly consistent with the hypothesis of independence. These tests have not made use of the expected 3:1 ratio, but rather the estimated ratios from the fixed margins.

The Mendel hypothesis is that the dominant characters occur in a proportion of 3:1 relative to the recessive character. It is therefore expected that the parent probabilities for the measured ratios $a/n_1 = 0.759$ and $c/n_2 = 0.745$ are $\pi = \pi_1 = \pi_2 = 3/4$. The simplest way to test the Mendel hypothesis is by using a χ^2 test using a fully-specified hypothesis, whereby the model for the four numbers is

$$(\pi(1 - \pi)N, (1 - \pi)^2N, \pi^2N, \pi(1 - \pi)N) = (104.25, 34.75, 312.75, 104.25).$$

This model is a simple modification of (10.4) using the parent probability π instead of the ratios estimated from the data, and it yields $\chi^2 = 0.47$ for a p -value of 0.925, using the null hypothesis $\chi^2(3)$ distribution. In fact, the χ^2 statistic was calculated enforcing just one constraint on the number N , for a total of 3 degrees of freedom. The agreement between the data and the Mendel 3:1 model is remarkable, and the model is clearly consistent with the data. \diamond

10.3 Higher Dimension $r \times c$ Contingency Tables

Exact tests for tables of higher dimensions are also possible, but they are usually computationally more intensive (see, e.g., [42]). On the other hand, the χ^2 test of independence can be immediately extended to contingency tables of higher dimensions, by changing the number of degrees of freedom to $f = (r - 1) \times (c - 1)$, where r is the number of rows and c the number of columns, according to Fisher's theorem. In fact, only $r - 1$ of the r entries for each column are determined independently, and likewise only $c - 1$ of the c entries for each independent row, because of the constraints from assuming that N and the margins are fixed. Therefore, for tables of dimensions larger than 2×2 , the χ^2 test is the most direct way to test for independence.

Example 10.6 (*COVID-19 mortality statistics*) Consider as an example the 6×2 contingency table reported as Table 10.6, with $r = 6$ rows and $c = 2$ columns, reproduced from a study that analyzed a large cohort of patients in England ($N = 17,278,392$) to study possible risks of death from the COVID-19 disease, by E.J. Williamson and colleagues [104]. Table 10.6 shows the number of individuals in this sample, divided by age group (6 classes) and by death from the disease (or outcome, which is a binary classification). These data can be used to test the hypothesis of independence between age and outcome. The data can be rearranged, without loss of information, as number of deaths and survivals in each age group.

For each age group $i = 1, \dots, 6$, the data provide a ratio of deaths per number of individuals, $P(A_i) = p_i$. The data also provide the probability of death for all age groups combined, as the column-wise ratio $P(B_1) = m_1/N$; since there are only two columns, $P(B_2) = m_2/N$ is the complementary probability of survival, regardless of age, with $m_1 + m_2 = N$. For a given age group i and outcome ($j = 1, 2$), the hypothesis of independence between age and outcome results in the following expectations:

$$E_{ij} = N \times p_i \cdot P(B_j) = N \times \frac{n_i}{N} \cdot \frac{m_j}{N} = \frac{n_i m_j}{N}$$

which is obtained via a product of the margins, in the same way as for 2×2 tables; notice how the assumption of independence leads to the products of probabilities.

Table 10.6 Data from the OpenSAFELY COVID-19 study [104]

	Number of persons (%)	Number of deaths (%)
Age 18–39	5,914,384(34.2)	54(0.00)
Age 40–49	2,849,984(16.5)	140(0.00)
Age 50–59	3,051,110(17.7)	522(0.02)
Age 60–69	2,392,392(13.8)	1,101(0.05)
Age 70–79	1,938,842(11.2)	2,635(0.14)
Age 80+	1,131,680(6.5)	6,474(0.57)

The expectations for death outcome in each age group are reported below, under the null hypothesis of independence between age and outcome.

	Deaths	Survivals	Ratio p_i	Death Expectation
Age 18–39	54	5,914,330	9.13×10^{-6}	3,739.96
Age 40–49	140	2,849,844	4.91×10^{-5}	1,802.19
Age 50–59	522	3,050,588	1.71×10^{-4}	1,929.37
Age 60–69	1,101	2,391,291	4.60×10^{-4}	1,512.83
Age 70–79	2,635	1,936,207	1.36×10^{-3}	1,226.02
Age 80+	6,474	1,125,206	5.75×10^{-3}	715.62
	10,926	17,267,466		10,926.00

As an example, the observed ratio $p_1 = 9.13 \times 10^{-6}$ of the first row is the conditional probability that the outcome is death, given that the individual is in that age group. This is the lowest probability of the six age groups. Summed over all age groups, the probability of death becomes a much larger $m_1/N = 6.32 \times 10^{-4}$.

It is immediate to see that there is a significant mismatch between the data and the expectations based on the hypothesis of independence amongst rows. This is confirmed by the value $\chi^2 = 54294.16$, which is to be compared with a $\chi^2(5)$ distribution. The result is an overwhelming evidence of higher mortality rate for older individuals in this sample, compared to younger individuals. Notice how the different number of persons in each age group is accounted for by considering the row-wise ratios in the expectation of deaths for each age group. \diamond

10.4 Binary Diagnostic Tests

Probabilities of an event with two possible outcomes, such as the presence or absence of a medical condition, can be studied by means of contingency tables and Bayesian statistics using the general framework for 2×2 contingency tables of (10.4). Within this framework, event A is the result of a diagnostic test for the presence of a disease, and event B is the actual presence of a disease. In clinical testing, the actual presence of a disease would be assessed by means of a tried-and-true or *golden standard*

method that is assumed to be correct, while the diagnostic test is one being evaluated using a sample of size N . Each event is binary, and the four possible outcomes can be described in a contingency table such as the one in Table 10.4:

	Disease (B_1)	No Disease (B_2)	Total
Positive Test (A_1)	a	b	n_1
Negative Test (A_2)	c	d	n_2
Total	m_1	m_2	N

The four numbers in the table have the following meaning:

$$\begin{cases} a = \text{number of } \textit{true positives} \\ b = \text{number of } \textit{false positives} \\ c = \text{number of } \textit{false negatives} \\ d = \text{number of } \textit{true negatives}. \end{cases}$$

An ideal diagnostic test would have $b = c = 0$. The test provides the number of positive tests (n_1) and of negative tests (n_2), with $N = n_1 + n_2$. The m_1 margin represents the total number of people in the sample with the disease, and it is generally unknown from the diagnostic testing alone. This number determines the *prevalence* of the disease within the sample, defined as

$$PR = P(B_1) = \frac{m_1}{N}$$

and indicating the fraction of the sample with the given condition. If the sample of size N is representative of a larger population, than this fraction applies also to the entire population. The prevalence has an important role in diagnostic testing, since it is also the *prior* or *pre-test* probability of a person having the disease.

10.4.1 Sensitivity, Specificity, and Likelihood Ratios

It is often the case that a diagnostic test is calibrated with a population of known properties, to determine the accuracy of the test. In this case, all the numbers in the table are known. The following ratios have a specific meaning:

$$\begin{cases} SE = \frac{a}{m_1} = P(A_1/B_1) : \text{sensitivity of the test} \\ SP = \frac{d}{m_2} = P(A_2/B_2) : \text{specificity of the test.} \end{cases} \quad (10.9)$$

The sensitivity is the conditional probability of obtaining a positive test, given that the person has the disease, and as such it determines the test's ability to discern

the presence of the disease (true positives) in sick patients. The specificity is the conditional probability of a negative test, given that the person does not have the disease, and so it is the ability to discern the absence of the disease (true negatives) in a healthy person. These probabilities are test-specific, and independent of the prevalence of the disease, which is a property of the population.

The *positive* and *negative likelihood ratios* are defined as

$$\begin{cases} PLR = \frac{P(A_1/B_1)}{P(A_1/B_2)} = \frac{SE}{1-SP} \\ NLR = \frac{P(A_2/B_1)}{P(A_2/B_2)} = \frac{1-SE}{SP} \end{cases}$$

and indicate, respectively, the ratio of the probability of a positive test for a sick patient versus the probability of a positive test for a healthy patient; and the ratio of the probability of a negative test for a sick patient versus the probability of a negative test for a healthy patient. A likelihood ratio of 1 means that healthy and sick patients have the same probability of that test result. As the name indicates, these are ratios of probabilities *given* that the health status of the patient is known: these ratios simply forecast the ability of a test to yield positive or negative result, in an alternative but substantially equivalent way to sensitivity and specificity.

10.4.2 Posterior Probabilities: The Positive and Negative Predictive Values

The primary goal of diagnostic testing is to determine the probability of the presence, or absence, of a disease, given a specific test result. These probabilities are given by the following ratios:

$$\begin{cases} PPV = \frac{a}{n_1} = P(B_1/A_1) : \text{positive predictive value} \\ NPV = \frac{d}{n_2} = P(B_2/A_2) : \text{negative predictive value.} \end{cases}$$

The positive predictive value represents the probability that the disease is present, given a positive test. Likewise, the negative predictive value is the probability of correctly diagnosing the absence of the disease, given a negative test. These conditional probabilities are posterior probabilities describing the effect of testing in diagnosing a condition (i.e., probabilities *after* taking the diagnostic test). They quantify the change in the probability of the disease being present or absent, after a test was performed. Notice the inverted order of conditioning, relative to the sensitivity and specificity defined in (10.9).

The positive and negative predictive values are related to the sensitivity and specificity of the test, and the prevalence of the disease, via Bayes' theorem and the total probability theorem:

$$\begin{cases} PPV = \frac{SE \cdot PR}{SE \cdot PR + (1 - SP)(1 - PR)} \\ NPV = \frac{SP \cdot (1 - PR)}{SP \cdot (1 - PR) + (1 - SE) \cdot PR}. \end{cases} \quad (10.10)$$

To obtain the equation for the positive predictive value, start with Bayes' theorem:

$$PPV = P(B_1/A_1) = \frac{P(A_1/B_1)P(B_1)}{P(A_1)} = \frac{SE \cdot PR}{P(\text{'positive test'})}.$$

The probability $P(A_1)$ of a positive test can be further evaluated via the total probability theorem as

$$P(A_1) = P(A_1/B_1)P(B_1) + P(A_1/B_2)P(B_2) = SE \cdot PR + (1 - SP)(1 - PR)$$

which leads to the result of Eq. 10.10. Similar calculations lead to the equation for the negative predictive value.

After a diagnostic test is calibrated and therefore the sensitivity and the specificity become known, performing the test on a sample of size N yields the number n_1 of positive results,

$$FP = \frac{n_1}{N} \text{ (fraction of positive tests).}$$

To determine the posterior probabilities, i.e., the positive and negative predictive values, and accordingly all numbers in the corresponding contingency table, more information is needed. One option that is often used is to estimate or assume a value for the prevalence, and then use Equations 10.10 to calculate the posterior probabilities. Figure 10.1 summarizes the dependence of the positive and negative predictive values on sensitivity, specificity, and prevalence. It is useful to illustrate the behavior of the PPV and NPV as a function of prevalence with the following cases.

(a) Consider first a test with 50% sensitivity and 50% specificity. A value of the sensitivity of $SE = 0.5$ means that there is an equal number of true positives ($a = m_1/2$) and false negatives ($c = m_1/2$), and a value of the specificity of $SP = 0.5$ means that there is an equal number of true negatives ($d = m_2/2$) and false positives ($b = m_2/2$) in a sample with m_1 persons with the disease and m_2 healthy persons. Such a test will provide no improvement top the post-test probabilities of the presence or absence of the disease, and accordingly (10.10) yields $PPV = PR$ and $NPV = 1 - PR$, as shown in Fig. 10.2.

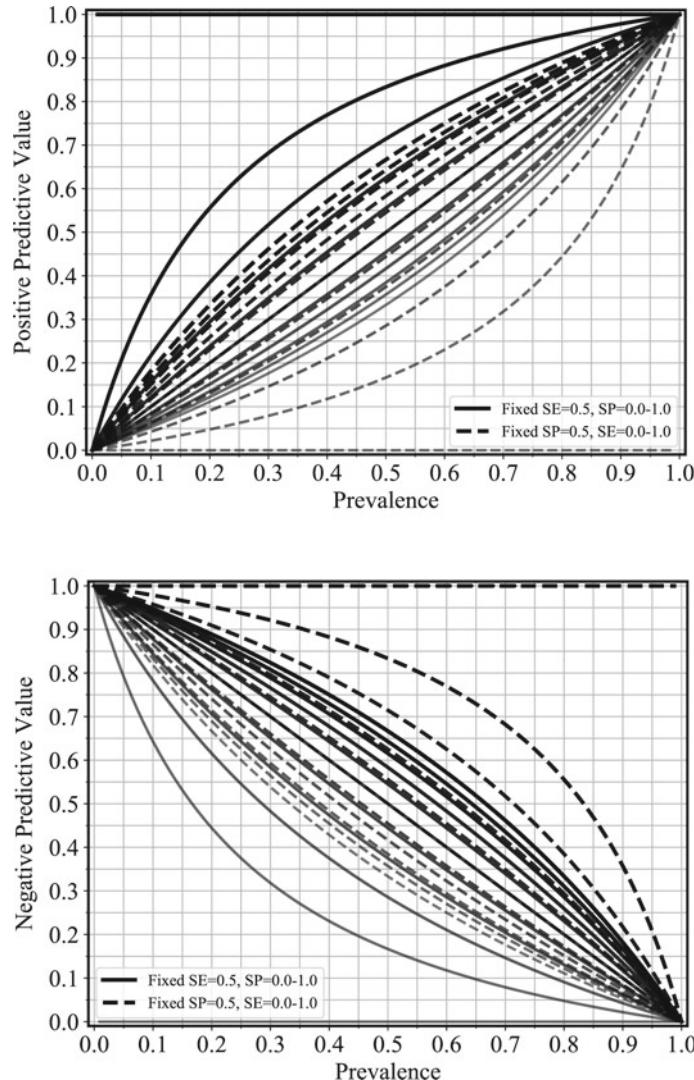


Fig. 10.1 Positive (a/n_1) and negative (d/n_2) predictive values as a function of prevalence. For the PPV , solid curves are at a fixed value of the sensitivity, with specificities increasing from bottom to top curves in steps of 0.1; dashed curves are at a fixed value of the specificity, also with sensitivity increasing from bottom to top curve in steps of 0.1. Similar patterns are also reported for the NPV . For ease of viewing, the line width also increases with increasing variable parameter (either sensitivity or specificity, depending on the curves)

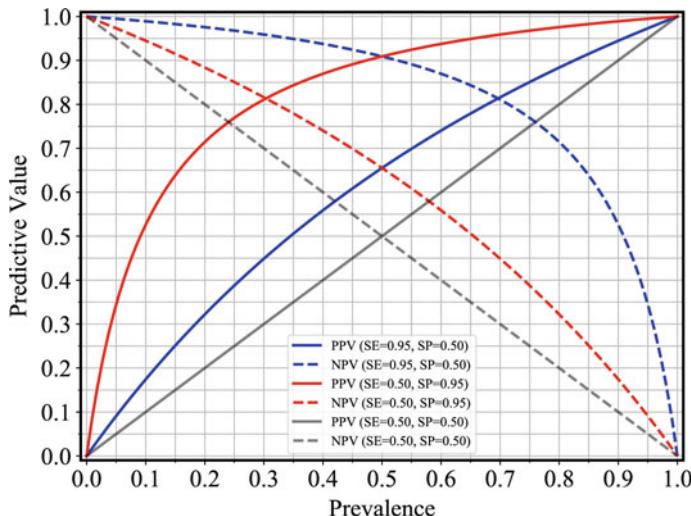


Fig. 10.2 Positive and negative predictive values for three specific cases: (a) high sensitivity and medium specificity (blue curves); (b) a high specificity and medium sensitivity (red curves); and (c) a test with 50% sensitivity and 50% specificity (grey curves)

(b) Consider next a test with high sensitivity, $SE = 0.95$, and a lower specificity of $SP = 0.5$. In this case, there is a large number of true positives and a small number of false negatives, or a good ability to detect a true positive. The low specificity, however, makes it such there is an equal number of false positives and true negatives, $b = d = m_2/2$, or a 50% ability to detect a true negative. This case corresponds to the blue curves in Fig. 10.2. At a low prevalence of the disease, $m_1 \ll N$, a highly sensitive and moderately specific test has a high NPV and a low PPV . This can be seen from

$$PPV = \frac{a}{n_1} = \frac{a}{a+b},$$

where the large number of false positives b drives the positive predictive value down. Conversely,

$$NPV = \frac{d}{n_2} = \frac{d}{c+d}$$

is a large number because of the small number c of false negatives. The result is that, in this case, a negative test is most likely an indicator of the absence of the disease, but a positive test does not provide strong indication of the disease. This is due to the small number of (mostly correctly identified) positive patients, and the large number of (randomly identified) negative patients.

Example 10.7 (*A high-sensitivity and low-specificity test*) A test with high sensitivity but low specificity results in many patients who are disease-free being told of the possibility that they have the disease (false positive) and are then subject to

further testing. For example, for a 10% prevalence of the disease, also the pre-test probability, a test with 95% sensitivity but only 50% specificity leads only to a modest increase in the post-test (or posterior) probability of $PPV \approx 17.5\%$: a positive test cannot be taken as a definitive statement of the presence of the disease. In other words, many false positives are expected with such a test. \diamond

(c) Conversely, a highly specific test with moderate sensitivity has a higher PPV and a lower NPV : a positive test provides a stronger indication of the disease, but there may be more false negatives. This situation is illustrated by the red curves in Fig. 10.2.

Example 10.8 (*A high-specificity and low-sensitivity test*) In a high-specificity and low-sensitivity test, patients who do have a disease may falsely be told they are negative (false negatives). For a 10% prevalence, a 95% specific test with only 50% sensitivity leads to a $PPV \approx 52.5\%$: a positive test provides a significantly higher post-test probability of the disease. This probability will tend to 100% as the specificity approaches 100%, as shown in Fig. 10.1, also illustrating the highly non-linear effect of the specificity on PPV . At higher values of the prevalence, say 90%, a highly specific test has now a $NPV \approx 17.5\%$, providing only a modest improvement over the 10% pre-test probability of the absence of the disease. This is caused by the large number of randomly identified positive cases. \diamond

10.4.3 Change in Posterior Probability with Repeated Testing

The sensitivity and specificity of a diagnostic test are fixed parameters that, together with the prevalence of the disease, determine the (posterior) probability that a single positive or negative test indicates the presence or absence of the disease. It is often possible to improve the diagnostic power of a test by performing multiple independent tests. The effect of repeated testing on the predictive values can be studied with Bayes' theorem. For a given value of the prevalence, the positive and negative predictive values can be written in equivalent forms to Eq. 10.10 as

$$\left\{ \begin{array}{l} PPV = \left(1 + \frac{1 - SP}{SE} \cdot \frac{1 - PR}{PR} \right)^{-1} = \left(1 + \frac{1}{PLR} \cdot \frac{1 - PR}{PR} \right)^{-1} \\ NPV = \left(1 + \frac{1 - SE}{SP} \cdot \frac{PR}{1 - PR} \right)^{-1} = \left(1 + NLR \cdot \frac{PR}{1 - PR} \right)^{-1} \end{array} \right. \quad (10.11)$$

The prevalence represents the prior probability $P(B_1)$, or the probability of the presence of the disease before the test. A positive test has the effect to *update* this probability to the posterior probability, as given by the PPV . Before a second test, it is thus reasonable to use this PPV as the prior probability, and therefore a second positive test will yield a positive predictive value of

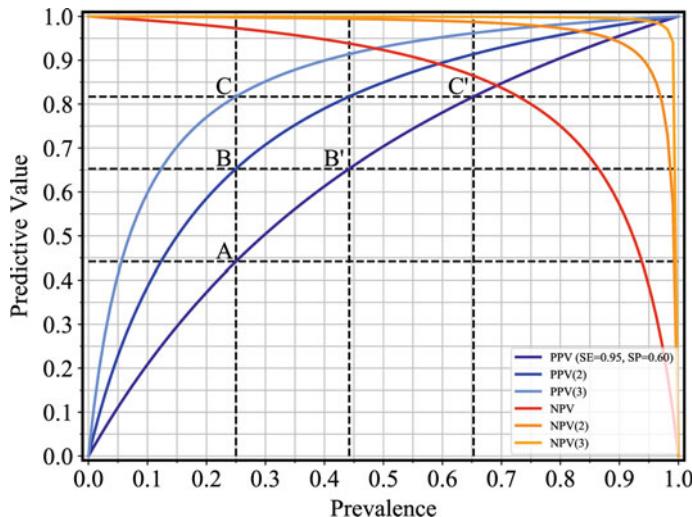


Fig. 10.3 Predictive values for a repeated high-sensitivity and low-specificity test

$$PPV(2) = \left(1 + \frac{1}{PLR} \cdot \frac{1 - PPV}{PPV} \right)^{-1},$$

with a similar expression for the negative predictive value for two consecutive negative tests. It can be shown that the generalization to n consecutive positive or negative tests leads to the following expressions for the positive and negative predictive values:

$$\begin{cases} PPV(n) = \left(1 + \left(\frac{1}{PLR} \right)^n \cdot \frac{1 - PR}{PR} \right)^{-1} \\ NPV(n) = \left(1 + NLR^n \cdot \frac{PR}{1 - PR} \right)^{-1}. \end{cases} \quad (10.12)$$

Equations 10.12 gives the posterior probabilities after n positive or negative tests, respectively. The tests are assumed to be independent, so that the result of one test has no effect on the others.

Equation 10.12 can be proven by induction. In fact,

$$\frac{1 - PPV}{PPV} = \frac{1}{PLR} \cdot \frac{1 - PR}{PR}$$

leading to the expression for $PPV(2)$. Using this same expression in $PPV(n-1)$ leads to the $PPV(n)$ of Eq. 10.12, thus completing the proof. A similar calculation can be used to prove the formula for $NPV(n)$.

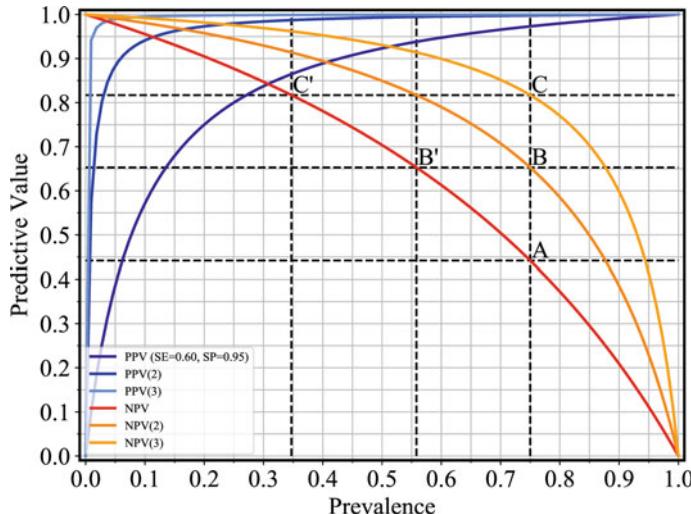


Fig. 10.4 Predictive values for a repeated high-specificity and low-sensitivity test

Example 10.9 (*Repeated high-sensitivity and low-specificity tests*) Figure 10.3 shows the positive predictive values and the negative predictive values of a test with 95% sensitivity and 60% specificity. At low values of the prevalence, say 25% (left vertical dashed line), a positive test only provides a $\sim 44\%$ posterior probability for the presence of the disease; a second positive test increases the posterior probability to $\sim 65\%$, and a third to just over 80%. These posterior probabilities can be calculated by using the $PPV(n)$ curves according to Eq. 10.12, specifically their intersections with the vertical prevalence line (points A, B, and C). Alternatively, using the intersection of the traditional PPV curve for one positive test with vertical lines with the updated prevalences. For example, the second vertical line has a value of PPV as the prevalence, intersecting the PPV curve at a point B' , with same vertical coordinate as B; the third vertical line corresponds to a prevalence of $PPV(2)$, intersecting the PPV curve at a point C' , with same vertical coordinate as C.

Example 10.10 (*Repeated high-specificity and low-sensitivity tests*) Figure 10.4 shows the positive predictive values and the negative predictive values of a test with 60% sensitivity and 95% specificity. At high values of the prevalence (75%, right dashed line), a negative test increases the negative predictive value from $1 - PR = 25\%$ to approximately $\sim 44\%$, thus lowering the prior on the prevalence to $\sim 56\%$ (second dashed vertical line). After the third negative test, there is a posterior probability of just over 80% that the patient is disease-free (intersection of the NPV curve with the $1 - NPV(2)$ vertical line, point C' ; or intersection of the $NPV(3)$ curve with the PR vertical line, point C).

Table 10.7 Contingency table for inoculation statistics

	B_1 (Healthy)	B_2 (Sick)	Total
A_1 (vaccinated)	a	b	n_1
A_2 (not vaccinated)	c	d	n_2
Total	m_1	m_2	N

Comparison of Eqs. 10.11 and 10.12 shows that n consecutive tests with the same outcome are equivalent to a test with positive and negative likelihood ratios of

$$\begin{cases} PLR(n) = PLR^n \\ NLR(n) = NLR^n. \end{cases}$$

There are several additional quantities to summarize and interpret the diagnostic ability of a test, such as the odds ratio, the *Youden's index* or the *receiver operating characteristic* curve. Specialized textbook on medical diagnostics such as [108] can be used for further reference on the subject.

10.5 Vaccine Efficacy

The methods of analysis of 2×2 contingency tables developed in Sect. 10.2 can be applied to inoculation statistics such as those Tables 10.1–10.3 to provide a quantitative assessment of the effectiveness of the inoculation. The *vaccine efficacy* is therefore defined as

$$VE = 1 - \frac{\text{risk for vaccinated individuals}}{\text{risk for unvaccinated individuals}}$$

with a high efficacy indicating that there is smaller risk of the disease for vaccinated individuals. A contingency table with inoculation statistics is of the form of Table 10.7, with b indicating the number of individuals who became sick among those who were vaccinated, and d the number of sick individuals among those who were not vaccinated (in a clinical trial, these patients would receive a placebo). The vaccine efficacy is, therefore,

$$VE = 1 - \frac{b/n_1}{d/n_2} \tag{10.13}$$

where b/n_1 is the risk of contracting the disease for vaccinated individuals, and d/n_2 is the risk for non-vaccinated individuals. The definition provided by (10.13) therefore makes it straightforward to evaluate the nominal value of the vaccine efficacy.

Uncertainties in the vaccine efficacy depend on the size of the sample of vaccinated and unvaccinated individuals. Given that the health status is binary, it is natural to use a binomial distribution to describe the probability of the number of vaccinated individuals who become sick:

$$P_{m_2}(b) = \binom{m_2}{b} \rho^b (1 - \rho)^{m_2 - b} \quad (10.14)$$

where ρ is the parent probability of a vaccinated individual to become sick. This distribution assumes that the margins are fixed, since m_2 (the number of sick individuals in the sample) is fixed at its measured value. Eq. 10.14 is the starting point to determine the range of possible values of b , which is then converted to a confidence interval on VE using (10.13). In fact, when the margins are fixed, there is only one degree of freedom in the 2×2 table, and for any value of b there corresponds only one value of d needed in (10.13) to evaluate the vaccine efficacy. There are two alternative methods to provide confidence intervals on the vaccine efficacy, based on the binomial distribution (10.14):

- (a) Although the parent probability ρ is not known a priori, it can be estimated as the measured ratio ($\rho \simeq b/m_2$) if the statistics are sufficiently large.
- (b) Alternatively, confidence intervals on the parent probability ρ and therefore on VE can be obtained using the fiducial argument (see Sect. 7.2), without the need to assume a fixed value for the binomial probability.

These two methods are illustrated in the following example with the aid of the data from the COVID-19 vaccine provided in a 2020 paper by M. Knoll and C. Wonodi [61], which summarizes the results of an extensive clinical study.

Example 10.11 (*The Oxford–AstraZeneca COVID-19 vaccine efficacy*) On November 23, 2020, the pharmaceutical company Astra–Zeneca released preliminary information of the efficacy of their COVID-19 vaccine AZD1222, followed by a complete release of their clinical trials on December 8, 2020 [61]. The company initially reported that their double-blind clinical trials reported a total of 131 COVID-19 cases combined between two dosing regimens. The company also informed that one dosing regimen with $n = 2,741$ participants showed a vaccine efficacy of 90%, while another dosing regimen with $n = 8,895$ participants yielded an efficacy of 62%, for a combined efficacy of 70%. The preliminary information was not sufficient to reconstruct exactly the contingency tables for the entire sample, or for either of the two subsamples. The subsequent release of information on December 8 contained all the necessary information, to calculate the vaccine efficacy. The data are reported in Table 10.8.

It is now possible to provide confidence intervals for the efficacy of the vaccine in either of the two sub-samples, and for the entire sample. A simple method consists of assuming that the margins m_2, m_{21} and m_{22} are fixed, which corresponds to assuming that the total number of persons who became sick in each sample is fixed. Following this assumption, the number of sick individuals in the vaccinated group is described by the binomial distribution (10.14) reported here for convenience:

Table 10.8 Immunization statistics for the Oxford–AstraZeneca COVID-19 vaccine, from [61]

	Healthy	Sick	Total	
SD/SD Trial				
Vaccinated	...	27	4,440	
Unvaccinated	...	71	4,455	
Total	...	$m_{21} = 98$	$N_1 = 8,895$	
LD/SD Trial				
Vaccinated	...	3	1,367	
Unvaccinated	...	30	1,374	
Total	...	$m_{22} = 33$	$N_2 = 2,741$	
Entire trial				
Vaccinated	...	30	$n_1 = 5,807$	
Unvaccinated	...	101	$n_2 = 5,829$	
Total	...	$m_2 = 131$	$N = 11,636$	

Table 10.9 Confidence intervals of vaccine efficacy from the immunization statistics of Table 10.8

Sample	VE	95 % Confidence interval			
		Using measured b value		Using parent β value	
		b	VE	β/m_2	VE
Entire trial	0.7108	21 – 40	0.56 – 0.81	0.17 – 0.31	0.55 – 0.79
LD/SD trial	0.8995	0 – 7	0.73 – 1.00	0.03 – 0.24	0.68 – 0.97
SD/SD trial	0.6184	19 – 36	0.42 – 0.76	0.20 – 0.38	0.40 – 0.75

$$P_{m_2}(b) = \binom{m_2}{b} \rho^b (1 - \rho)^{m_2 - b} \quad \text{for } b = 0, \dots, m_2, \quad (10.15)$$

where ρ is the unknown and sought-after parent probability of a vaccinated individual to become sick. This distribution applies also for both subsamples, where m_2 is replaced by the two margins m_{21} and m_{22} . This probability can be used to establish a confidence intervals on the number b and therefore on VE , since $d = m_2 - b$ when m_2 is fixed.

(a) Consider first the *observed* number of sick patients b in the sample and use a parent probability equal to the observed probability, $\rho = b/m_2$. With this fixed value of ρ , a confidence interval on b is obtained immediately from the probability mass function Eq. 10.15, by using a central confidence interval for the number b , around the observed value. For a confidence interval at 95% level or $p = 0.95$, the binomial distribution gives the following confidence intervals as shown in Table 10.8. This method to obtain confidence intervals is not exact, since it assumes a fixed probability for the binomial distribution.

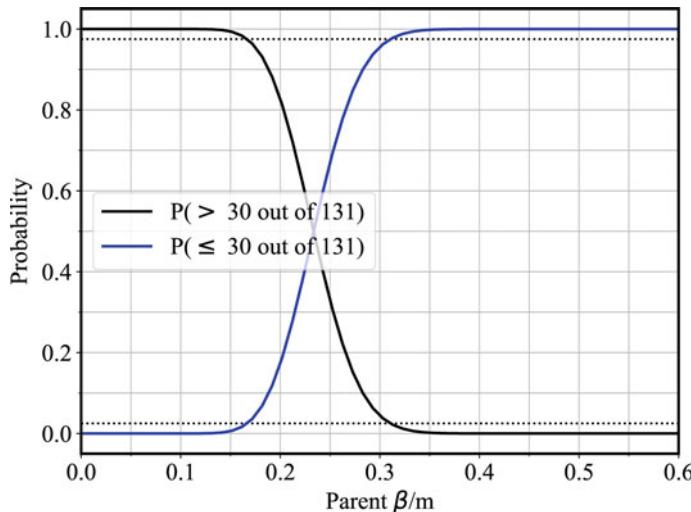


Fig. 10.5 Cumulative distribution $F(b)$ for a binomial with fixed number of counts $b = 30$, corresponding to the total number of COVID-19 cases among vaccinated people in the entire Oxford-AstraZeneca sample, and of the variable $\rho = \beta/m_2$, where $m_2 = 131$ is the total number of COVID-19 cases among all participants (vaccinated and placebo recipients). In black is the associated survival function $1 - F(b)$. The confidence interval on β/m is obtained by the two intersections of the cumulative distribution with the horizontal dashed lines, which mark the 0.025 and the 0.975 quantiles, for a 95% confidence interval

(b) A more accurate method consists of using the fiducial probability of the parent number β of sick individuals in the vaccinated group, so that the parent probability for the occurrence of a sick patient is $\rho = \beta/m_2$, where m_2 is still kept fixed at the measured value. The distribution (10.15) is still used via its cumulative distribution, but in this case as a function of the variable ρ instead of b , as illustrated in Fig. 10.5. Confidence intervals on the parent value of β/m_2 , and of the corresponding VE , are reported in Table 10.8.

The two methods of determining confidence intervals give similar results, although they use different assumptions. The vaccine efficacy reported in the study [61] is 70.4% (54.8–80.6) for the entire sample, 90.0% (67.4–97.0) for the LD/SD sample, and 62.1% (41.0–75.7) for the SD/SD sample, in good agreement with the numbers of Table 10.8, in particular those of method (b) using the parent distribution of the number b .

This analysis shows that there are significant uncertainties in the estimate of the vaccine efficacy with such a limited number of ill people in the samples. The overlap between the 95% confidence intervals provides qualitative evidence that there is no statistically significant difference in the efficacy of the two subsamples, once the uncertainties caused by the small number statistics are accounted. It is possible to quantify the degree of agreement between the two subsamples and the full sample, using a simple test that makes use of the assumption that the full sample provides

Table 10.10 Binomial tests of consistency between samples

Parameters of the binomial		Number of positives b		Test result
Number of tries	Probability (ρ)	95% CI	Observed value	
$m_{21} = 98$	0.229	15–31	27	Consistent
$m_{22} = 33$	0.229	3–12	3	Consistent
$m_{22} = 33$	0.276	4–14	3	Inconsistent

an estimate of the parent mean of the binomial distribution, with $\rho = b/m_2 = 0.229$ ($b = 30$ and $m_2 = 131$ for the full sample). It is now possible to test whether the measured values of b for the two subsamples agree with this parent distribution, continuing with the assumption that $m_{21} = 98$ and $m_{22} = 33$ are the fixed number of sick individuals in the two subsamples. The confidence intervals for the two binomial distributions are:

This analysis shows that the two subsamples are consistent with the results from the entire sample. A similar test can be done using the measured probability for the larger subsample as the parent probability of the binomial ($\rho = b/m_{21} = 0.276$), and testing whether the smaller subsample is consistent with that distribution. This test is also reported in the table above, indicating that the LD/SD subsample is not consistent with the model from the SD/SD sample, at the 95% confidence level. This test, therefore, lends support to the hypothesis that the LD/SD regimen is significantly better than the SD/SD regimen. Care must, however, be exercised when interpreting this result, since no uncertainties in the model based on the SD/SD sample was included in this test. ◇

Summary of Key Concepts for this Chapter

Contingency table: An $r \times c$ contingency table reports the r possible outcomes of one event, as a function of the c outcomes of another event. When the two events are binary, the contingency table is a 2×2 table and it is usually reported as $(a, b; c, d)$. Contingency tables are used commonly to study the relationship and dependence between two events.

χ^2 test for contingency tables: Pearson devised the first χ^2 test for 2×2 contingency tables in 1900, with the goal to test whether the two events reported in the table are independent. The χ^2 statistic has a simple analytical form when all the margins of the table are fixed at the measured value,

$$\chi^2 = \frac{(ad - bc)^2 N}{m_1 m_2 n_1 n_2}$$

where $n_1 = a + b$, $n_2 = c + d$, $m_1 = a + c$ and $m_2 = b + d$ are the margins of the table, and N is the sum of all numbers. Under the hypothesis that the two events are independent and that the numbers in the table are sufficiently large, its parent distribution is $\chi^2(1)$. When the numbers in the table are small, it is possible to apply the Yates correction and still use the χ^2 test.

Fisher exact test: The independence of events in a 2×2 contingency table can be tested with Fisher's exact test, with probability density

$$P(a, b, c, d) = \frac{n_1! n_2! m_1! m_2!}{N!} \frac{1}{a! b! c! d!}.$$

The Fisher exact test consists of summing the probability of all tables that are as extreme as, or more extreme than, the observed table.

Binary diagnostic testing: Contingency tables can be used effectively to study binary diagnostic tests, where one event is the test result, and the other is the presence of a disease. The use of Bayes' theorem makes it possible to determine a *posterior* probability that, e.g., a person who tests positive is in fact ill (positive predictive value), or the probability that a person who tests negative is healthy (negative predictive value). The same framework can also be used to study the *efficacy* of a vaccine and its uncertainties from a clinical study.

Problems

10.1 Consider the 2×2 contingency table $(2, 3; 1, 4)$ and assume that all margins are fixed.

- (a) Calculate the margins m_1, m_2, n_1 and n_2 and, if needed, rearrange the table so that $m_1 \leq m_2$ and $n_1 \leq n_2$.
- (b) Illustrate all possible tables with the same fixed margins, and the results of the Fisher exact test.

10.2 Consider the contingency table $(0, 5; 5, 0)$, assumed to be measured from independent binomial distributions with $p_1 = p_2 = 1/2$, with fixed margin $n_1 = 5$ and $N = 10$.

- (a) Calculate the probability of obtaining such an extreme table, or more extreme ones, if the m_1, m_2 margins are unknown.
- (b) Calculate the same probability assuming that the measured m_1, m_2 margins are fixed.

10.3 The contingency table $(1, 4; 5, 0)$ of Example 10.4 has fixed margins, so that the distribution of probabilities follows that of the Fisher exact test.

- (a) Calculate the Fisher exact probability.
- (b) Calculate the approximate χ^2 statistic and its null hypothesis probability.
- (c) Calculate the χ^2 statistic with the Yates correction, and its null hypothesis probability.
- (d) Determine whether the measured contingency table is consistent with the hypothesis of independence between the events, at the 99% confidence level.

10.4 Calculate the number of all possible 2×2 contingency tables that can be obtained with a fixed number N , but without constraints on any of the margins. Evaluate the number and list all possible tables for $N = 2$.

10.5 ■ Calculate the probability that the vaccination data of Table XV of Greendwood and Yule (Table 10.2) are consistent with the null hypothesis that the vaccine was not effective. For this test, you may assume that the margins are fixed.

10.6 ■ Consider the vaccination data of Table XV of Greendwood and Yule (Table 10.2), and the reported statistic stated as $\chi^2 = 5.61$ and $P = 0.1351$.

- (a) Evaluate the Pearson χ^2 statistic for the given contingency table.
- (b) Determine the associated null hypothesis probability, and determine whether the null hypothesis probability reported by Greendwood and Yule is accurate.

10.7 Use Bayes' theorem and the total probability theorem to prove the equation for the negative predictive value in (10.10).

10.8 A diagnostic test is calibrated with a sample of $N = 280$ persons, with the result of 120 true positives, 20 false positives, 40 false negatives, and 100 true negatives.

- (a) Calculate the sensitivity, specificity, positive likelihood ratio, and negative likelihood ratio of this diagnostic test.
- (b) What is the probability that a healthy person tests negative with this diagnostic test?

10.9 A diagnostic test has a sensitivity of 75% and a specificity of 83.3%.

- (a) Explain whether this test is better at diagnosing a disease in sick individuals, or ruling out a disease in healthy individuals.
- (b) The disease is assumed to have a prevalence of 10 %. Calculate the probability that an individual is sick, following a positive test.
- (c) Calculate the probability that an individual is sick, following a positive test, when the prevalence is now 1%.
- (d) The disease is assumed to have a prevalence of 10 %. Calculate the probability that an individual is healthy, following a negative test.

10.10 A diagnostic test with 70% sensitivity and 99% specificity is used to test for the presence of a disease that has a prevalence of 10%.

- (a) Determine the probability that the disease is in fact present, when a test is negative.
- (b) Determine the same probability, when the same test is repeated 3 times independently on the same individual.
- (c) Repeat the calculations for (a) and (b) if the prevalence is 50%.

10.11 A diagnostic test with the same sensitivity and specificity as those of Problem 10.10 is used to test for the presence of a disease. Determine the positive predictive value when there is a low prevalence of the disease ($PR = 0.01$), and discuss whether such test can be used effectively to determine the presence of the disease in the individual being tested, under these circumstances.

10.12 A COVID-19 vaccine was tested on 21,720 persons, 8 of whom contracted the disease. A control sample of 21,728 persons was administered a placebo, and it resulted in 162 unvaccinated persons contracting the disease. Determine the vaccine efficacy and estimate its 95% confidence interval. This problem is based on [79], which reported a 95% efficacy, with a 95% credible interval of 90.3 to 97.6.

Chapter 11

Linear and Non-linear Regression for Gaussian Data



Abstract One of the most common tasks in the analysis of scientific data is to establish a relationship between two quantities. Many experiments feature the measurement of a quantity of interest as a function of another control quantity that is varied as the experiment is performed. This chapter describes the use of the maximum-likelihood method to determine whether a certain relationship between the two quantities is consistent with the available measurements, and the best-fit parameters of the relationship. The application to a linear function, known as *linear regression*, has a simple analytic solution, but the method can also be applied to more complex analytic functions.

11.1 Measurement of Pairs of Variables and Regression

A general problem in data analysis is to establish a relationship $y = y(x)$ between two variables X and Y , given the availability of N measurements of the two variables. In general, the two variables are described by a joint distribution function that determines the probability of occurrence of a specific pair of values. The problem becomes much more easily tractable if one of the two variables can be treated as an *independent variable* having an uncertainty that is negligible compared to that of the other *dependent variable*. An example of this situation is the measurement of the flux of a source as a function of time, where time can be regarded as the independent variable, since it can be measured with much higher accuracy than the flux. This may not always be the case, and there are some instances when both uncertainties need to be considered.

The starting point for the analysis of a two-dimensional dataset is an analytic function $y(x)$ for the relationship between the two variables, with a number of adjustable

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-981-19-0365-6_11.

parameters, referred to as a_k , that are to be constrained according to the measurements. With the independent variable X assumed to be known exactly and Y the dependent variable, a two-variable dataset is usually reported in the form

$$(x_i, y_i) \quad i = 1, \dots, N.$$

For example, x_i is the time of observation, assumed to be known exactly, and y_i is the measurement of the flux of a source at that specific time. Although each y_i is referred to as a “measurement,” it may itself be obtained from a number of measurements. For example, the flux of the source at a *given* fixed time may result from a series of observations that let the analyst estimate not just the value of the flux, but also its standard error. In general, the variable Y_i can have any distribution, but a common situation is that of a normal distribution, so that the measurement is often reported in the form $y_i \pm \sigma_i$, indicating a measurement of the mean and standard deviation of a normal variable.

The estimation of the parameters of the function $y(x)$ from the data is referred to as *model fitting* or *regression*.¹ The process of finding the parameters that “best” describe the relationship between the two variables is two-fold: not only fit parameters must be determined, but it is also necessary to determine whether the data are actually well described by the model, so that the estimated parameters are meaningful. A powerful means to accomplish these goals is via the method of maximum likelihood, whereby the best-fit parameters are chosen so that the likelihood of the data with the model is maximized, while in the process defining a *fit statistic* that is used to test whether the fit is acceptable. The main assumption of the regression is that the function $y(x)$ is the correct description of the relationship between the two variables. This means that, at each fixed value x_i of the independent variable, the random variable Y_i should have an expectation of $y(x_i)$. This expectation is used in the calculation of the likelihood, which varies according to the data type. Two cases of particular interest are those of Gaussian data and of Poisson data, each resulting in a different fit statistic, but other types of data are also possible.

11.2 Regression Using Maximum Likelihood for Gaussian Data

Normally distributed measurements are very common in data analysis. In this case, the data are usually reported in the form

$$(x_i, y_i \pm \sigma_i) \quad i = 1, \dots, N.$$

¹ Sir F. Galton is usually credited for the use of the word *regression* as a measure of the strength of the relationship between a dependent variable and one or more independent variables [36].

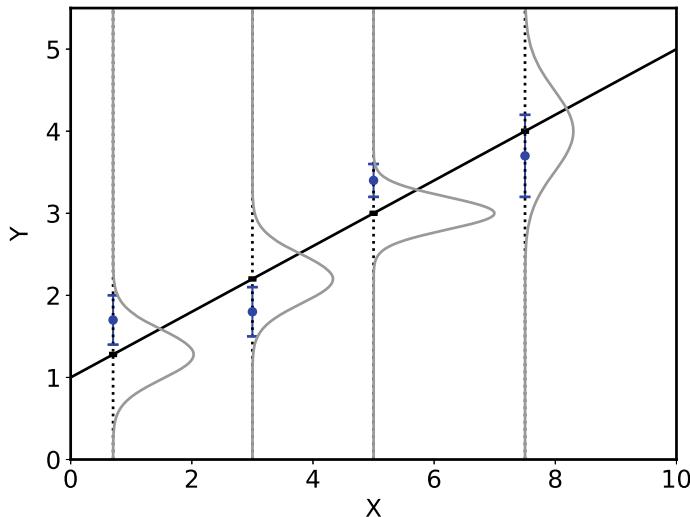


Fig. 11.1 Illustration of the method of linear regression with Gaussian data for a dataset with four measurements. Each measurement, at a fixed value of the X variable, is marked by a circle with an error bar representing the standard deviation. The parent mean of each Gaussian is determined by the model, and the σ^2 parameter is determined by the data

with the meaning that the Gaussian-distributed Y_i variable was estimated to have a mean of y_i and a variance of σ_i^2 . Each measurement at a fixed value of x_i could be derived, for example, from a set of measurements with sample mean \bar{y}_i and sample variance s_i^2 , which are known to be unbiased estimators of the parent quantities. The likelihood of the data with the model requires the specification of the two parameters of the Gaussian distribution for each datapoint, and the measurements must independent of one other. With these assumptions, the likelihood of the data with the model is given by

$$\begin{aligned} \mathcal{L} &= \prod_{i=1}^N \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{(y_i - y(x_i))^2}{2\sigma_i^2}} \\ &= \left(\prod_{i=1}^N \frac{1}{\sqrt{2\pi\sigma_i^2}} \right) e^{-\frac{1}{2} \sum_{i=1}^N \frac{(y_i - y(x_i))^2}{\sigma_i^2}}. \end{aligned} \quad (11.1)$$

Figure 11.1 illustrates the calculation of the likelihood: for each value x_i of the independent variable X , the Gaussian variable Y_i has a mean of $y(x_i)$ determined by the model, and a variance σ_i^2 that is derived from the measurements. It is reasonable that the variance of the Gaussian variable Y_i is estimated from the data, since the distribution function of Y_i must reflect the statistical precision of the measurements.

The same maximum-likelihood method used to estimate parameters of a distribution in Chap. 6 can be used also to estimate the parameters of the regression curve $y(x)$, by requiring that the unknown parameters a_k of the model $y = y(x)$ are those that maximize the likelihood of the data. The key feature of (11.1) is that the product term is independent of the parameters of the model, and therefore maximization of the likelihood is equivalent to the minimization of the statistic

$$\chi^2 = \sum_{i=1}^N \left(\frac{y_i - y(x_i)}{\sigma_i} \right)^2. \quad (11.2)$$

The reason for the name χ^2 is that, when the variable Y_i follows a Gaussian with fixed mean $y(x_i)$ and variance σ_i^2 , (11.2) is a χ^2 variable with N degrees of freedom (see Sect. 9.3). Equation (11.2) also defines the goodness-of-fit statistic χ_{\min}^2 , which is the specific value of (11.2) for the parameter values that minimize it. Estimation of the free parameters of $y(x)$ will lead χ_{\min}^2 still being χ^2 -distributed, but with a reduced number of degrees of freedom, as will be explained in detail in Chap. 12. In this chapter, the minimization of (11.2) will be used to estimate the free parameters and their uncertainties. Given the form of (11.2), the maximum-likelihood method, when applied to Gaussian distribution, is also known as the *least squares* method.

11.3 Linear Regression with Gaussian Data

When the fitting function is a simple linear function $y(x) = a + bx$, the problem of finding the values of the two adjustable parameters a and b that minimize the χ^2 statistic (11.2) can be solved analytically, and it is often referred to as *linear regression*. The conditions of minimum χ^2 are written as partial derivatives with respect to the two unknown parameters:

$$\begin{cases} \frac{\partial \chi^2}{\partial a} = -2 \sum_{i=1}^N \frac{1}{\sigma_i^2} (y_i - a - bx_i) = 0 \\ \frac{\partial \chi^2}{\partial b} = -2 \sum_{i=1}^N \frac{x_i}{\sigma_i^2} (y_i - a - bx_i) = 0 \end{cases} \quad (11.3)$$

leading to a linear system of two equations in two unknowns,

$$\begin{cases} \sum_{i=1}^N \frac{y_i}{\sigma_i^2} = a \sum_{i=1}^N \frac{1}{\sigma_i^2} + b \sum_{i=1}^N \frac{x_i}{\sigma_i^2} \\ \sum_{i=1}^N \frac{x_i y_i}{\sigma_i^2} = a \sum_{i=1}^N \frac{x_i}{\sigma_i^2} + b \sum_{i=1}^N \frac{x_i^2}{\sigma_i^2} \end{cases} \quad (11.4)$$

with solution that can be conveniently written as

$$a = \frac{1}{\Delta} \begin{vmatrix} \sum_{i=1}^N \frac{y_i}{\sigma_i^2} & \sum_{i=1}^N \frac{x_i}{\sigma_i^2} \\ \sum_{i=1}^N \frac{x_i y_i}{\sigma_i^2} & \sum_{i=1}^N \frac{x_i^2}{\sigma_i^2} \end{vmatrix} \quad (11.5)$$

$$b = \frac{1}{\Delta} \begin{vmatrix} \sum_{i=1}^N \frac{1}{\sigma_i^2} & \sum_{i=1}^N \frac{y_i}{\sigma_i^2} \\ \sum_{i=1}^N \frac{x_i}{\sigma_i^2} & \sum_{i=1}^N \frac{x_i y_i}{\sigma_i^2} \end{vmatrix}$$

$$\Delta = \sum_{i=1}^N \frac{1}{\sigma_i^2} \sum_{i=1}^N \frac{x_i^2}{\sigma_i^2} - \left(\sum_{i=1}^N \frac{x_i}{\sigma_i^2} \right)^2. \quad (11.6)$$

Equations (11.5) and (11.6) provide the solution for the best-fit parameters of the linear model, and all the sums extend over the N measurements. It is also possible to provide analytical estimates of the uncertainty associated with the best-fit parameters. For this purpose, it is convenient to rewrite the solution in matrix notation, starting with the definition of a 2×2 symmetric matrix \mathbf{A}

$$\mathbf{A} = \begin{bmatrix} \sum_{i=1}^N \frac{1}{\sigma_i^2} & \sum_{i=1}^N \frac{x_i}{\sigma_i^2} \\ \sum_{i=1}^N \frac{x_i}{\sigma_i^2} & \sum_{i=1}^N \frac{x_i^2}{\sigma_i^2} \end{bmatrix} \quad (11.7)$$

with inverse given by

$$\boldsymbol{\varepsilon} = \mathbf{A}^{-1} = \frac{1}{\Delta} \begin{bmatrix} A_{22} & -A_{12} \\ -A_{12} & A_{11} \end{bmatrix} \quad (11.8)$$

and a row vector

$$\boldsymbol{\beta} = \left(\sum_{i=1}^N \frac{y_i}{\sigma_i^2}, \sum_{i=1}^N \frac{y_i x_i}{\sigma_i^2} \right),$$

so that the best-fit parameters (11.5) can be written as

$$(a, b) = (\beta_1, \beta_2) \begin{bmatrix} \epsilon_{11} & \epsilon_{12} \\ \epsilon_{21} & \epsilon_{22} \end{bmatrix} = \boldsymbol{\beta} \boldsymbol{\varepsilon}.$$

It is possible to show that the matrix $\boldsymbol{\varepsilon} = \mathbf{A}^{-1}$ is a symmetric *error matrix* or *covariance matrix* that contains the variances of the measured parameters along the diagonal, and the covariance between the parameters in the two off-diagonal positions,

$$\boldsymbol{\varepsilon} = \begin{bmatrix} \hat{\sigma}_a^2 & \hat{\sigma}_{ab}^2 \\ \hat{\sigma}_{ab}^2 & \hat{\sigma}_b^2 \end{bmatrix}, \quad (11.9)$$

where the notation $\hat{\sigma}_a^2$ etc. indicates that these are estimated quantities, and not parent quantities known a priori. It is important to notice that the parameters a and b are *normally distributed*, since they are a linear combination of normally distributed and independent measurements, and their best-fit values and variances are obtained analytically using basic tools of linear algebra. This simple solution makes the linear regression very simple to implement and interpret. It may be useful to point out that the covariance matrix is independent of the measurements y_i , and dependent only on the values of the independent variable x_i and the variances σ_i^2 of the dependent variable.

Proof that the matrix $\boldsymbol{\varepsilon}$ contains the variances and the covariance can be obtained using standard methods of linear algebra and the error propagation formula for the sum of independent variables. For example, the best-fit value of the parameter a can be written as

$$a = \frac{1}{\Delta} \left(\sum_{i=1}^N \frac{y_i}{\sigma_i^2} \sum_{i=1}^N \frac{x_i^2}{\sigma_i^2} - \sum_{i=1}^N \frac{x_i}{\sigma_i^2} \sum_{i=1}^N \frac{x_i y_i}{\sigma_i^2} \right),$$

and it is a linear function of N independent normal measurements y_i with variance σ_i^2 . It is therefore possible to estimate the variance of the parameter a as the sum

$$\hat{\sigma}_a^2 = \sum_{i=1}^N \left(\frac{\partial a}{\partial y_i} \right)^2 \sigma_i^2$$

using the error propagation formula (see Sect. 5.2) and the independence between the measurements. Given the linearity of the parameter a on the N measurements y_i , the partial derivative is immediately obtained as

$$\frac{\partial a}{\partial y_k} = \frac{1}{\sigma_k^2} \cdot \epsilon_{11} + \frac{x_k}{\sigma_k^2} \cdot \epsilon_{12}$$

where it is convenient to indicate the sums that are independent of the measurements y_i with the corresponding term in the error matrix. The error propagation formula therefore leads to the following expression:

$$\hat{\sigma}_a^2 = \sum_{k=1}^N \left[\sigma_k^2 \left(\sum_{i=1}^2 \epsilon_{1i} \cdot \frac{f_i(x_k)}{\sigma_k^2} \right)^2 \right]$$

using the notation $f_1(x_k) = 1$ and $f_2(x_k) = x_k$. The second-power term can be expanded with the aid of a different dummy index, leading to

$$\begin{aligned}\hat{\sigma}_a^2 &= \sum_{i=1}^2 \epsilon_{1i} \sum_{j=1}^2 \epsilon_{1j} \sum_{k=1}^N \frac{f_i(x_k) f_j(x_k)}{\sigma_k^2} \\ &= \epsilon_{11} \left(\epsilon_{11} \sum_{k=1}^N \frac{1}{\sigma_k^2} + \epsilon_{12} \sum_{k=1}^N \frac{x_k}{\sigma_k^2} \right) + \epsilon_{12} \left(\epsilon_{11} \sum_{k=1}^N \frac{x_k}{\sigma_k^2} + \epsilon_{12} \sum_{k=1}^N \frac{x_k^2}{\sigma_k^2} \right).\end{aligned}$$

Finally, the two terms in parenthesis in the last expression evaluate to respectively 1 and 0, as can be seen by

$$\left(\epsilon_{11} \sum_{k=1}^N \frac{1}{\sigma_k^2} + \epsilon_{12} \sum_{k=1}^N \frac{x_k}{\sigma_k^2} \right) = \frac{1}{\Delta} (A_{22} A_{11} - A_{12}^2) = 1$$

and

$$\left(\epsilon_{11} \sum_{k=1}^N \frac{x_k}{\sigma_k^2} + \epsilon_{12} \sum_{k=1}^N \frac{x_k^2}{\sigma_k^2} \right) = \frac{1}{\Delta} (A_{22} A_{21} - A_{12} A_{22}) = 0$$

leading to the result that ϵ_{11} is the variance of the a parameter. Similar calculations can be done to show that ϵ_{22} is the variance of the b parameter, and that the ϵ_{12} term is the covariance.

The solution to the problem of linear regression with Gaussian data presented in this section is particularly convenient for the simple analytic forms for both the best-fit parameters in (11.5) and (11.6), and for their errors in (11.7) and (11.8). This solution includes the general case of measurements with different variances, since the σ_i^2 values can be different among the measurements. This situation is applicable when measurements have different precision, and therefore the variance of a measurement provides a sort of "weight" for that data point. A simplified method of linear regression is when all variances are equal, and it will be presented below in Sect. 11.5, along with a numerical application. Prior to that, it is useful to provide a generalization of the method of linear regression presented in this section, which allows a more general form for the fitting function.

11.4 Multiple Linear Regression

The linear regression to two-dimensional data of Sect. 11.3 can be generalized to a fitting function of the form

$$y(x) = \sum_{k=1}^m a_k \cdot f_k(x). \quad (11.10)$$

This function is linear in the m parameters a_k , which are the coefficients of functions $f_k(x)$ that can have any analytical form. In this case one speaks of *multiple linear regression*, or simply multiple regression. It is necessary to keep in mind that in this type of regression the *multiple* applies to the number of (fixed) fit functions, and that the *linear* applies to the linearity in the m adjustable parameters. A common use of the multiple regression is with polynomials, where the fit functions are chosen as

$$f_k(x) = x^k, \quad (11.11)$$

and the linear regression is a special case with only two such function, $f_1(x) = 1$ and $f_2(x) = x$. Minimization of χ^2 is achieved by taking partial derivatives with respect to the unknown parameters a_k , following a simple generalization of the linear regression. This yields the following m equations:

$$-2 \sum_{i=1}^N \left(\frac{y_i - \sum_{k=1}^m a_k \cdot f_k(x_i)}{\sigma_i^2} \right) f_l(x_i) = 0$$

that can be written, for $l = 1, \dots, m$ as

$$\sum_{i=1}^N \frac{f_l(x_i)}{\sigma_i^2} \left(y_i - \sum_{k=1}^m a_k \cdot f_k(x_i) \right) = 0 \quad (11.12)$$

leading to

$$\sum_{i=1}^N \frac{f_l(x_i) y_i}{\sigma_i^2} = \sum_{k=1}^m a_k \sum_{i=1}^N \frac{f_k(x_i) f_l(x_i)}{\sigma_i^2} \quad l = 1, \dots, m. \quad (11.13)$$

Equation (11.13) are m coupled equations in the parameters a_k , which can be solved using matrix algebra, as described below. Notice that the term $f_l(x_i)$ is the l -th model component (thus the index l is not summed over), and the index i runs from 1 to N , where N is the number of data points. The best-fit parameters are therefore obtained by defining a row vector $\beta = (\beta_1, \dots, \beta_m)$ with k -th element

$$\beta_k = \sum_{i=1}^N \frac{f_k(x_i) y_i}{\sigma_i^2},$$

and an $m \times m$ symmetric matrix A with (l, k) -th element equal to

$$A_{lk} = \sum_{i=1}^N \frac{f_l(x_i) f_k(x_i)}{\sigma_i^2}.$$

Defining $\mathbf{a} = (a_1, \dots, a_m)$ as the row vector with the best-fit m model parameters, the linear system of m Eq.(11.13) can be rewritten in matrix form as

$$\mathbf{a} \mathbf{A} = \boldsymbol{\beta}, \quad (11.14)$$

and therefore the task of estimating the best-fit parameters is that of inverting the matrix \mathbf{A} , which in general can be done numerically. The m best-fit parameters are therefore given by

$$\mathbf{a} = \boldsymbol{\beta} \mathbf{A}^{-1}. \quad (11.15)$$

The vector $\boldsymbol{\beta}$ and the matrix \mathbf{A} can be calculated from the data and the functions $f_k(x)$. It is also possible to show that the inverse of the matrix \mathbf{A} is the *error* or *covariance matrix*,

$$\boldsymbol{\varepsilon} = \mathbf{A}^{-1} = \begin{bmatrix} \hat{\sigma}_1^2 & \hat{\sigma}_{12}^2 & \dots \\ \hat{\sigma}_{12}^2 & \hat{\sigma}_2^2 & \dots \\ \vdots & \vdots & \ddots \\ & & \hat{\sigma}_m^2 \end{bmatrix} \quad (11.16)$$

containing the estimates of the parameter variances along the diagonal, and the covariances in the off-diagonal positions. All variances and covariances are therefore obtained simply by inversion of the $m \times m$ matrix \mathbf{A} . The parameters of the multiple linear regression are normally distributed, since they are a linear combination of normally distributed independent measurements, same as in the case of the simple linear regression.

To calculate variances and covariances in the best-fit parameters, the parameters a_k are treated as functions of the measurements, $a_k = a_k(y_i)$. The error propagation method is used to estimate variances and covariances as

$$\begin{cases} \hat{\sigma}_{a_k}^2 = \sum_{i=1}^N \left(\frac{\partial a_k}{\partial y_i} \right)^2 \sigma_i^2 \\ \hat{\sigma}_{a_l a_j}^2 = \sum_{i=1}^N \frac{\partial a_l}{\partial y_i} \frac{\partial a_j}{\partial y_i} \sigma_i^2, \end{cases} \quad (11.17)$$

using independence between the measurements. For convenience of notation, in (11.16), it is set $\hat{\sigma}_1^2 = \hat{\sigma}_{a_1}^2$ etc. The matrix equation $\mathbf{a} = \boldsymbol{\beta} \boldsymbol{\varepsilon}$ is used to write

$$a_l = \sum_{k=1}^m \beta_k \varepsilon_{kl} = \sum_{k=1}^m \varepsilon_{kl} \sum_{i=1}^N \frac{y_i f_k(x_i)}{\sigma_i^2}$$

so that

$$\frac{\partial a_l}{\partial y_i} = \sum_{k=1}^m \varepsilon_{kl} \frac{f_k(x_i)}{\sigma_i^2}.$$

The equation above can be used into (11.17) to show that

$$\hat{\sigma}_{a_l a_j}^2 = \sum_{i=1}^N \left[\sigma_i^2 \sum_{k=1}^m \left(\varepsilon_{jk} \frac{f_k(x_i)}{\sigma_i^2} \right) \times \sum_{p=1}^m \left(\varepsilon_{lp} \frac{f_p(x_i)}{\sigma_i^2} \right) \right]$$

in which the indices k and p indicate the m model parameters, and the index i is used for the sum over the N measurements. This leads to

$$\hat{\sigma}_{a_l a_j}^2 = \sum_{k=1}^m \varepsilon_{jk} \sum_{p=1}^m \varepsilon_{lp} \sum_{i=1}^N \frac{f_k(x_i) f_p(x_i)}{\sigma_i^2} = \sum_{k=1}^m \varepsilon_{jk} \sum_{p=1}^m \varepsilon_{lp} A_{pk}.$$

Now recall that the matrix A is the inverse of ε , and therefore the expression above can be simplified to

$$\sigma_{a_l a_j}^2 = \sum_k \varepsilon_{jk} \mathbf{1}_{kl} = \varepsilon_{jl}.$$

The multiple regression is therefore a straightforward extension of the simple linear regression, with solutions for the m adjustable parameters given by (11.15), and with parameter uncertainties given by the error matrix (11.16). The additional complication is that, in general, the matrix inversion does not have a simple analytical solution, and therefore the inversion must be performed numerically.

11.5 Linear Regression with Uniform Variance

It is a common experimental situation that all the measurements y_i in a dataset have the same uncertainty. Continuing with the assumption that the measurements are normally distributed, this means that the parent variance of all measurements is the same, $\sigma_i^2 = \sigma^2$. In this case, the solution to the linear regression (11.5) is simplified to

$$\begin{cases} a = \frac{1}{\Delta} \frac{1}{\sigma^4} \left(\sum_{i=1}^N y_i \sum_{i=1}^N x_i^2 - \sum_{i=1}^N x_i \sum_{i=1}^N x_i y_i \right) \\ b = \frac{1}{\Delta} \frac{1}{\sigma^4} \left(N \sum_{i=1}^N x_i y_i - \sum_{i=1}^N y_i \sum_{i=1}^N x_i \right) \\ \Delta = \frac{1}{\sigma^4} \left(N \sum_{i=1}^N x_i^2 - \left(\sum_{i=1}^N x_i \right)^2 \right) \end{cases} \quad (11.18)$$

and the variances and covariance of the parameters in (11.9) become

$$\begin{cases} \hat{\sigma}_a^2 = \frac{1}{\Delta} \frac{1}{\sigma^2} \sum_{i=1}^N x_i^2 \\ \hat{\sigma}_b^2 = \frac{1}{\Delta} \frac{N}{\sigma^2} \\ \hat{\sigma}_{ab}^2 = -\frac{1}{\Delta} \frac{1}{\sigma^2} \sum_{i=1}^N x_i. \end{cases} \quad (11.19)$$

The important feature of the solution (11.18) is that the best-fit parameters are *independent* of the value of σ . This means that, for the purpose of finding the best-fit values of the model parameters, the actual value of the parent variance of the measurements is irrelevant, and this is the case because all measurements carry the same weight. The value of the variance is still relevant for the measurement of the variances of the parameters, as can be seen by (11.19). For datasets that do not report the uncertainty of individual measurements, it is reasonable to assume that all measurements have the same variance and therefore calculate the best-fit parameters without the need to specify a value for σ^2 . However, uncertainties in the parameters are not possible without assuming a value for the parent variance.

11.5.1 Alternative form of the Solution with Sample Moments

There is an alternative and perhaps more insightful way to describe the best-fit parameters for the linear regression with uniform variance. In fact, it is easy to show that the best-fit parameters estimated by the least-squares method can be written as a function of the sample means and of the sample variances and covariance,

$$\begin{cases} b = \frac{s_{xy}^2}{s_x^2} \\ a = \bar{y} - b \bar{x} \end{cases} \quad (11.20)$$

where s_{xy}^2 is the usual sample covariance between the X and Y measurements and s_x^2 is the sample variance of the X measurements. This means that, in the absence of correlation between the two variables, the best-fit slope will be zero and the value of a is simply the linear average of the measurements. Equation 11.20 also leads to a simple way to show that the sampling distribution of the parameters a and b is normally distributed, and that the parameter estimates are unbiased. In fact, the assumption of normal measurements from a parent linear model can be written as

$$y_i = \alpha + \beta x_i + \delta_i$$

where α and β are the parent values of the linear model, and δ_i is a normally distributed variable with zero mean and variance σ^2 , representing the random error in the i -th measurement. Simple algebraic manipulations of (11.20) show that the parameter b can be written as

$$b = \beta + \frac{\sum_{i=1}^N \delta_i \cdot (x_i - \bar{x})}{\sum_{i=1}^N (x_i - \bar{x})^2}.$$

Since the terms $(x_i - \bar{x})$ are constant, b is normally distributed as the sum of N normally distributed and independent variables. The expectation is immediately evaluated as

$$\mathbb{E}[b] = \beta$$

thus showing that the linear regression provides an unbiased estimate of the parent value. The variance is also immediately evaluated as

$$\text{Var}(b) = \frac{\sigma^2}{\sum_{i=1}^N (x_i - \bar{x})^2},$$

which is the same as in (11.19). Similar calculations can be performed for the a parameter, which can be written as

$$a = \alpha + (\beta - b)\bar{x} + \bar{\delta},$$

where $\bar{\delta}$ represents the average of the N independent random errors. The statistic a is therefore normally distributed, as the linear combination of independent Gaussian variables. The expectation is therefore

$$\mathbb{E}[a] = \alpha,$$

showing that also the least-squares parameter a is an unbiased estimate of the parent parameter. The variance is also immediately calculated as

$$\text{Var}(a) = \sigma^2 \left(\frac{1}{N} + \frac{\bar{x}^2}{\sum_{i=1}^N (x_i - \bar{x})^2} \right),$$

which is consistent with (11.19). Equation 11.20 are therefore equivalent solutions for the best-fit parameters of the linear regression with uniform variance to those of (11.18).

11.5.2 Choice of Independent Variable

There are situations when it may not be clear which of the two variables is to be treated as independent. Equation 11.20 clearly shows that the best-fit parameters are dependent on the choice of which variable is considered independent. In fact, if Y is regarded as the independent variable instead of X , and the data are fit to the model

$$x = a' + b'y, \quad (11.21)$$

the least-squares method gives the best-fit slope of

$$b' = \frac{s_{xy}^2}{s_y^2}.$$

When the model is rewritten in the usual form

$$y = a_{X/Y} + b_{X/Y} x,$$

in which the notation X/Y means “ X given Y ”, the best-fit model parameters are

$$\begin{cases} b_{X/Y} = \frac{1}{b'} = \frac{s_y^2}{s_{xy}^2} \\ a_{X/Y} = \bar{y} - b_{X/Y} \bar{x}. \end{cases}$$

This new linear regression of X on Y differs from the linear regression (11.20) of Y on X , and therefore the two linear models assuming x or y as independent variable will be different from one another. It is up to the data analyst to determine which of the two variables is to be considered independent when there is a dataset with no errors reported in either variable. Normally the issue is resolved by knowing how the experiment was performed, e.g., which variable had to be assumed or calculated first in order to calculate or measure the second.

11.6 A Classic Experiment: Edwin Hubble's Discovery of the Expansion of the Universe

In the early twentieth century, astronomers were debating whether “nebulae,” now known to be external galaxies, were in fact part of our own Galaxy. At the time, there was no notion of the Big Bang and the expansion of the Universe. Edwin Hubble pioneered a revolution in the understanding of the Universe via a seemingly simple observation that a number of “nebulae” moved away from the Earth with a velocity v that is proportional

to their distance d . This relationship between distance and velocity is known as *Hubble's law*,

$$v = H_0 d. \quad (11.22)$$

The quantity H_0 is the *Hubble constant*, typically measured in units of $\text{km s}^{-1} \text{Mpc}^{-1}$, where Mpc indicates a distance of 10^6 parsec. The data used by Hubble [51] is summarized in Table 11.1. The quantity m is the apparent magnitude, related to the distance via the following logarithmic relationship,

$$\log d = \frac{m - M + 5}{5} \quad (11.23)$$

where $M = -13.8$ is the absolute magnitude, also measured by Hubble as part of the same experiment, and considered as a constant for the purpose of this dataset, and d is measured in parsecs.

The first part of Hubble's analysis consisted in fitting the (v, m) dataset to a relationship that is linear in $\log v$,

$$\log v = a + b \cdot m \quad (11.24)$$

where a and b are the adjustable parameters of the linear regression. Hubble reported two different fit results, one in which he determined also the error in a ,

$$\log v = (0.202 \pm 0.007) \cdot m + 0.472 \quad (11.25)$$

and one in which he fixed $a = 0.2$, and determined the error in b :

$$\log v = 0.2 \cdot m + 0.507 \pm 0.012. \quad (11.26)$$

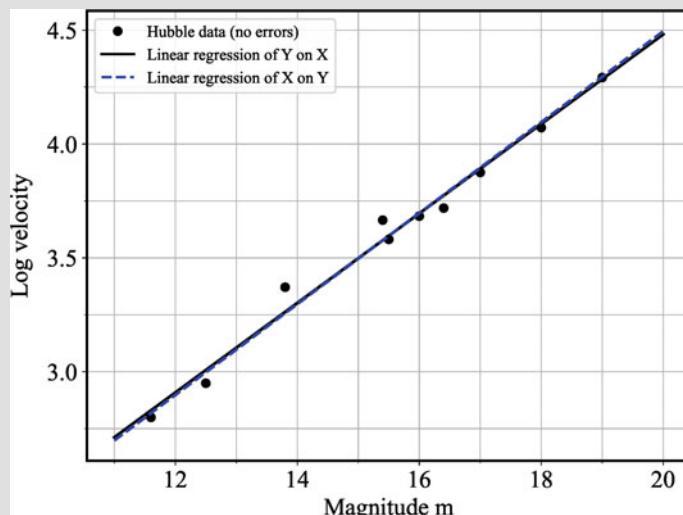
Using (11.26) into (11.23), Hubble determined the following relationship between velocity and distance,

$$\log \frac{v}{d} = 0.2M - 0.493 = -3.253 \quad (11.27)$$

and this results in the measurement of his name-sake constant, $H_0 = v/d = 10^{-3.253} = 558 \times 10^{-6} \text{ km s}^{-1} \text{ pc}^{-1}$, or $558 \text{ km s}^{-1} \text{ Mpc}^{-1}$. The value of the Hubble constant is now known to be approximately ten times smaller, but Hubble's observation of an approximately linear relationship between distance and velocity of extragalactic nebulae was a new breakthrough for cosmology, eventually leading to the discovery that the Universe is expanding.

Table 11.1 Data from E. Hubble's measurements

Name of nebula	Mean velocity km s ⁻¹	Number of velocities	Mean m
Virgo	890	7	12.5
Pegasus	3810	5	15.5
Pisces	4630	4	15.4
Cancer	4820	2	16.0
Perseus	5230	4	16.4
Coma	7500	3	17.0
Ursa Major	11,800	1	18.0
Leo	19,600	1	19.0
(No name)	2350	16	13.8
(No name)	630	21	11.6

**Fig. 11.2** Best-fit linear regression model for the data in Table 11.1. The analysis reproduces very closely Hubble's results in Fig. 4 of his 1931 paper [51]

The data from Hubble's experiment are a typical example of a dataset for which no errors were reported. A linear fit can be performed assuming that m is the independent variable. Using the equal-variance formulas (11.20), the best-fit model is

$$\log v = 0.548 + 0.197 m$$

as shown in Fig. 11.2. The best-fit parameters reported by Hubble's original analysis have a nearly identical slope as the one calculated above, although with a slightly different value of the intercept. It is useful to point out that Hubble performed quite a complex analysis leading to the measurements of Table 11.1, with certain data points measured with better accuracy than others. However, Hubble did not report the

uncertainties associated with the velocities, but just the mean velocity for each nebula. This omission prevents a linear regression with variable variances, which would weigh each measurement according to (11.5). It is also unclear which variable should be considered as independent, since both are subject to measurement errors. Assuming that $\log v$ is the independent variable instead, the linear regression leads to a best-fit model of

$$\log v = 0.503 + 0.200 m$$

that is virtually identical to the one reported by Hubble. As can be seen in Fig. 11.2, the two linear regressions (of Y on X and of X on Y) lead to very similar results. This is due to the fact that there is a strong degree of correlation between the X and Y measurements, with a sample correlation coefficient of $r = 0.99$.

11.7 Non-linear Regression

The methods described in this chapter assume that the model is linear in the fitting parameters. This requirement is, however, not necessary: the same χ^2 statistic (11.2) applies to any analytical relationship $y(x)$, provided that each variable y_i is Gaussian distributed. The main complication for non-linear functions is that an analytic solution for the best-fit values and the errors is in general no longer available. This limitation can be overcome with the use of numerical methods to minimize the χ^2 statistic. The most straightforward way to achieve a minimization of the χ^2 as function of all parameters is to construct an m dimensional grid of all possible parameter values, evaluate the χ^2 at each point, and then find the global minimum. The parameter values corresponding to this minimum can be regarded as the best estimate of the model parameters. The direct grid-search method, however, becomes rapidly unfeasible as the number of free parameters increases. Moreover, to find the parameter uncertainties using grid-search methods requires a knowledge of the expected variation of the χ^2 around the minimum. Among the methods that can be used to estimate parameters and their covariance matrix by passing the calculation of the entire grid is the Markov chain Monte Carlo technique. These methods will be covered in later chapters.

Summary of Key Concepts for this Chapter

Regression: A method to find best-fit parameters of a model $y(x)$ being fit to two-dimensional data, assuming that one variable is the independent variable, and based on the principle of maximum likelihood.

Linear regression: Fit of two-dimensional data to a simple linear model that, under the assumption of Gaussian distributions with uniform variances, has simple analytical results for the slope and intercept,

$$\begin{cases} b = \frac{s_{xy}^2}{s_x^2} \\ a = \bar{y} - b\bar{x}. \end{cases}$$

Multiple linear regression: An extension of the linear regression to models that are linear in m adjustable parameters, of the type

$$y(x) = \sum_{k=1}^m a_k \cdot f_k(x).$$

Minimum chi-squared statistic: Fit statistic that results from the maximum-likelihood method of fit for Gaussian data.

Error or covariance matrix: An $m \times m$ symmetric matrix ϵ , obtained by the inversion of a data-based matrix A and containing the variances and covariances of the m fit parameters.

Problems

11.1 ■ Consider the data from Hubble's experiment in Table 11.1. Determine the best-fit values of the fit to a linear model for $(m, \log v)$ assuming that the dependent variables have uniform variance.

11.2 ■ Consider the following two-dimensional data, in which X is the independent variable, and Y is the dependent variable assumed to be derived from a photon-counting experiment:

x_i	y_i
0.0	25
1.0	36
2.0	64
3.0	49
4.0	81

- (a) Determine the errors associated with the dependent variables Y_i .
- (b) Find the best-fit parameters a, b of the linear regression curve

$$y(x) = a + bx,$$

the errors in the best-fit parameters and the correlation coefficient.

- (c) Calculate the minimum χ^2 of the fit, and the corresponding probability to exceed this value. For this χ^2_{\min} value, determine if the fit is acceptable at the 90% confidence level.

11.3 Consider the following Gaussian dataset in which the dependent variables are assumed to have the same unknown standard deviation σ ,

x_i	y_i
0.0	0.0
1.0	1.5
2.0	1.5
3.0	2.5
4.0	4.5
5.0	5.0

The data are to be fit to a linear model.

- (a) Evaluate the sums $\sum x_i$, $\sum y_i$, $\sum x_i y_i$, $\sum x_i^2$.
- (b) Show that, according to (11.18), the best-fit values of the model parameters are $a = 0$ and $b = 1$.

11.4 Show that sufficient conditions for having best-fit parameters $a = 0$ and $b = 1$ in the linear regression with uniform variance are

$$\begin{cases} \sum_{i=1}^N y_i = \sum_{i=1}^N x_i \\ \sum_{i=1}^N x_i^2 = \sum_{i=1}^N x_i y_i. \end{cases} \quad (11.28)$$

11.5 Show that the best-fit slope parameter b of a linear fit to a Gaussian dataset is insensitive to a shift of all datapoints by the same amount Δx , or by the same amount Δy . You can show that this property applies to the case of equal errors in the dependent variable, although the same result applies also for the general case of different errors.

11.6 ■ Use the data provided in Table 8.1 to calculate the maximum-likelihood estimates of the parameters a and b for the fit to the radius vs. pressure ratio data. For the fit, you can assume that the radius is known exactly, and that the standard deviation of the pressure ratio is obtained as a linear average of the positive and negative errors.

11.7 Show that, when all measurement errors are identical, the least squares estimators of the linear parameters a and b are given by

$$\begin{cases} b = s_{xy}^2 / s_x^2 \\ a = \bar{y} - b \bar{x}, \end{cases}$$

where s_{xy}^2 is the sample covariance and s_x^2 is the sample variance of x .

Chapter 12

Goodness of Fit and Parameter Uncertainty for Gaussian Data



Abstract The minimum χ^2 statistic is the fit statistic of choice for normally distributed data. The sampling distribution of this statistic is necessary to determine whether the model is actually a correct description of the data. As the name suggests, this fit statistic is χ^2 -distributed, but the number of degrees of freedom depends on the number of parameters that were minimized. The estimation of parameter uncertainties makes use of a new $\Delta\chi^2$ statistic, which is also χ^2 -distributed with a number of degrees of freedom equal to the number of free parameters in the model, leading to a simple method to determine confidence intervals on fit parameters.

12.1 The χ^2_{\min} Goodness-of-Fit Statistic

For Gaussian-distributed data, finding the maximum of the likelihood is equivalent to the minimization of the χ^2 statistic (11.2). The minimum χ^2 statistic is evaluated with the adjustable parameters a_k that minimize it, and indicated as

$$\chi^2_{\min} = \sum_{i=1}^N \frac{(y_i - \hat{y}_i)^2}{\sigma_i^2} \quad (12.1)$$

where

$$\hat{y}_i = y(x_i)|_{\text{best-fit}}$$

is the value of the model function $y(x)$ calculated with the m best-fit parameter values indicated by \hat{a}_k . This statistic applies to the fit of any analytical model to Gaussian data. For the simple linear model, the minimization leads to an analytical solution for the best-fit parameters, while for more complex fit functions the minimization must be performed by numerical means. Regardless of the method used to obtain the best-fit parameters, it is necessary to establish if the parameters that minimize χ^2 give an

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-981-19-0365-6_12.

acceptable result, as not all models provide an accurate description of the data, even when the adjustable parameters are optimized. The quantitative method to establish if the fit was “successful” is via a hypothesis testing using the sampling distribution of the minimum χ^2 statistic. It would be tempting to say that the χ_{\min}^2 statistic is distributed like a χ^2 random variable with N degrees of freedom, since it is the sum of the square of N normally distributed random variables. If the function $y(x)$ has no free parameters, this is in fact the case, as shown in Sect. 9.3. The complication is that the fit function has m free parameters that were adjusted in such a way as to minimize the χ^2 statistic. This has two implications on the χ_{\min}^2 statistic: the free parameters will reduce the value of χ^2 with respect to the case in which no free parameters were present, and, more importantly, the fit function $y(x)$ introduces a dependence among the N random variables in the sum. Given that the χ_{\min}^2 is no longer the sum of N independent terms, one cannot immediately determine the distribution of the minimum χ^2 statistic, as was done in Sect. 9.3.

Given the importance of χ^2 in statistics and data analysis, much work has been devoted to the determination of the sampling distribution of the minimum χ^2 statistic, under as general conditions as possible. One of the most general results on this topic was provided by H. Cramér [21], who generalized a simpler case considered earlier by R.A. Fisher [28]. Cramér showed that in the fit of N independent measurements to a function with m free parameters, χ_{\min}^2 is in fact still distributed as a χ^2 variable,

$$\chi_{\min}^2 \sim \chi^2(f) \quad (12.2)$$

but with a reduction in the number of degrees of freedom f according to

$$f = N - m. \quad (12.3)$$

This result is remarkably general, since it applies to any type of function $y(x)$, provided that the m parameters follow certain regularity conditions, as is normally the case for most “meaningful” fit functions. The result may be simply stated as follows:

Theorem 12.1 (Cramér’s theorem on the limiting distribution of χ_{\min}^2) *Consider N Gaussian measurements and the fit to a function $y(x)$ with m adjustable parameters. When a number of regularity conditions on the m adjustable parameters are satisfied, the best-fit parameters estimated via minimization of χ^2 lead to a χ_{\min}^2 statistic that is asymptotically distributed like $\chi^2(N - m)$, when the number of measurements is sufficiently large.*

The general proof of this theorem is rather elaborate and can be found in Sect. 30.3 of the textbook *Mathematical Methods of Statistics* by H. Cramér [21]. It is useful to provide a proof for the specific case of a one-parameter constant model $y(x) = a$, to illustrate the reduction of degrees of freedom from N to $N - 1$ when there is just one free parameter that can be used to minimize χ^2 .

When performing a maximum likelihood fit to the function $y(x) = a$, it was found that the best-fit parameter is estimated as the sample mean of the measurements,

$$\hat{a} = \bar{x} = \frac{1}{N} \sum_{i=1}^N x_i,$$

under the assumption that all measurements are drawn from the same distribution $N(\mu, \sigma^2)$ (see Sect. 6.2). Therefore, the χ^2 statistic for the (unknown) parent model is related to the χ^2_{\min} statistic by

$$\chi^2 = \sum_{i=1}^N \frac{(x_i - \mu)^2}{\sigma^2} = \frac{(\bar{x} - \mu)^2}{\sigma^2/N} + \sum_{i=1}^N \frac{(x_i - \bar{x})^2}{\sigma^2} = \frac{(\bar{x} - \mu)^2}{\sigma^2/N} + \chi^2_{\min}.$$

This equation is identical to the relationship (9.18) used to derive the sampling distribution of the variance, leading to the conclusion that $\chi^2_{\min} \sim \chi^2(N - 1)$ and that χ^2_{\min} and χ^2 are independent random variables. This simple model shows how the estimation of the sample mean, represented by the first term on the right-hand side of the previous equation, effectively “uses” one of the N degrees of freedom of the χ^2 on the left-hand side, leaving the χ^2_{\min} statistic with $N - 1$ degrees of freedom.

Now that the sampling distribution of the fit statistic χ^2_{\min} is known, one can use the hypothesis testing methods of Sect. 9.3.3 to determine whether a value of the statistic is acceptable or not. The power of Cramér’s theorem is that it allows the use of the χ^2 distribution with a number of degrees of freedom that is immediately calculated by subtracting the number of free parameters m from the number of independent measurements N . The caveat of Cramér’s theorem is that it applies accurately to data with a large number of measurements. Although there is no fixed prescription as to what constitutes a “large” number, the analyst needs to be aware that this is only an asymptotic distribution.

Example 12.1 (*Hypothesis testing with χ^2_{\min}*) Figure 12.1 shows a linear fit using data from Table 8.1. The quantity Energy 1 is used as the independent variable, and its errors are neglected. The quantity Energy 2 is the dependent variable, and errors are calculated as the average of the positive and negative error bars. The best-fit linear model is represented as the dotted line, for a fit statistic of $\chi^2_{\min} = 60.5$ for 23 degrees of freedom. The value of the fit statistic is too large since its p -value is less than 0.0001, and the linear model must be discarded. Despite failing the χ^2_{\min} test, the best-fit model appears visually to be a reasonable match to the data. The large value of the test statistic is clearly caused by few datapoints with small error bars, but there appears to be no systematic deviation from the linear model. One reason for the poor fit statistic could be that errors in the independent variables were neglected. Chapter 18 provides an alternative fitting method that takes into account errors in both variables. Another possibility for the poor fit is that there are other

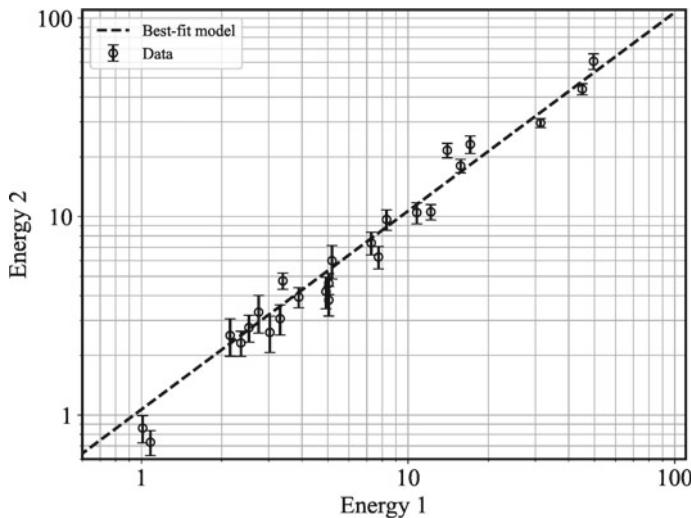


Fig. 12.1 Linear fit to the data of Table 8.1, assuming that the independent variable is Energy 1 with its errors neglected in the fit. Note the logarithmic scale for both axes

sources of error that are not accounted for. These additional errors are often referred to as systematic errors. Chapter 17 addresses the presence of systematic errors and how one can handle the presence of such errors in the fit. ◇

12.2 Data with No Errors and the Model Sample Variance

For datasets that do not report the variance of individual measurements, such as the Hubble data of Sect. 11.6, a least-squares regression is still possible, but it is not possible to evaluate the goodness of fit. In fact, the χ^2_{min} statistic does require the specification of the variances, or the value of a common variance. If neither is available, the analyst cannot perform a test on the goodness of fit. Since the absence of measurement errors is a very common occurrence, it is useful to discuss it further.

The Hubble data are an example of an experiment where the analyst could have reported the variances of the measurements. For example, the mean velocity for each nebula was likely calculated from several individual measurements, and these “raw” data would be sufficient to estimate the variance of each measurement. If the analyst chooses not to report this information, it is simply not possible to reconstruct the variance a posteriori. There may be cases where the nature of the experiment provides a simple way to estimate such variance. For example, a counting experiment with data in the form of integer numbers is likely to be derived from a Poisson variable, and it may be reasonable to assume a variance equal to the mean of the distribution. In that case, it must be kept in mind that the χ^2 statistic applies only to normal data, and

therefore the approximation of a Poisson distribution with a Gaussian must be further justified. In other cases, the variance of the measurements can be estimated from the knowledge of the apparatus used for the data collection. An example is the use of a device that was previously calibrated to a given accuracy, and that information can be used to estimate a common variance for all measurements.

When data uncertainties are simply not available, performing a regression assumes that the fitting function $y(x)$ is the correct description for the data, without the possibility of testing such assumption. In this case, it may become useful to answer the inverse question of what is a reasonable value of the variance of the measurements, if the model is an accurate description of the data. Under this assumption, one can define a *model sample variance* as the value of the variance that yields a minimum reduced χ^2 of one,

$$\hat{s}^2 = \frac{1}{N - m} \sum_{i=1}^N (y_i - \hat{y}_i)^2 \quad (12.4)$$

where \hat{y}_i is the value of the fitting function $y(x_i)$ evaluated with the best-fit parameters, as obtained by a fit assuming identical errors. This model sample variance can be used as an unbiased estimator of the unknown data variance.

The reason for the definition of this model sample variance according to (12.4) is that, according to (12.1), using this value as the common variance leads to a reduced χ^2_{\min} of exactly one, which is the expectation of the χ^2_{\min} statistic under the null hypothesis that the data are drawn from the parent model. In other words, the sum on the right-hand side of (12.4), $S_r^2 = \sum(y_i - \hat{y}_i)^2$, is the residual variance of the data after the fit is performed. Therefore, according to Cramér's theorem,

$$\frac{S_r^2}{\sigma^2} \sim \chi^2(N - m),$$

where σ^2 is the parent variance of the measurements, which is an unknown. The model sample variance is therefore an unbiased estimator of the parent variance, since

$$E\left[\frac{\hat{s}^2}{\sigma^2}\right] = \frac{1}{N - m} E\left[\frac{S_r^2}{\sigma^2}\right] = 1.$$

The estimate (12.4) can be therefore viewed as using the measured residual variance S_r^2 to estimate σ^2 by requiring that the statistic S_r^2/σ^2 is equal to its expected value. It is clear that one does not expect a reduced χ^2_{\min} of exactly one for every fit, even when the model is accurate, and therefore this value of the variance is to be considered as an approximation that is convenient for an order-of-magnitude estimate of the model uncertainties.

The underlying assumption behind the use of (12.4) is to treat each measurement y_i as drawn from a parent distribution $y(x_i)$, $i = 1, \dots, N$, e.g., assuming that the

model is the correct description for the data. In the case of a linear regression, $m = 2$, since two parameters (a and b) are estimated from the data. In turn, this value of σ^2 can now be used to estimate the error matrix, and therefore uncertainties on the fit parameters. When σ^2 is estimated via (12.4), any following analysis is based on the assumption that the model is a correct description of the data.

12.3 The $\Delta\chi^2$ Statistic

Minimization of the fit statistic χ^2 leads to the calculation of the best-fit parameters of the parent model. Assuming that the value of χ^2_{\min} does not lead to a rejection of the null hypothesis model $y(x)$, i.e., the data are consistent with the model, it is necessary to determine the uncertainties of the best-fit parameters. Clearly, there is no reason to believe that the estimated parameters \hat{a}_k are the same as the parent or true parameters (say, a_k), which remain unknown. In fact, repeating the experiment under the same conditions will in general lead to different measurements and different estimates of the adjustable parameters. It is therefore necessary to develop a method to calculate confidence intervals on model parameters. Much of the theory needed for this task was developed in Cramér's book [21], and presented in a more data analysis friendly format by M. Lampton, B. Margon and S. Bowyer [64].

Under the assumption that a model $y(x)$ with m adjustable parameters is the correct description of the data, the fit statistic χ^2 calculated with these true parent values,

$$\chi^2_{\text{true}} = \sum_{i=1}^N \left(\frac{y_i - y(x_i)|_{\text{true}}}{\sigma_i^2} \right)^2 \quad (12.5)$$

is distributed as $\chi^2(N)$. This is so because the true parameters are fixed and no minimization of the χ^2 function can be performed. The quantity χ^2_{true} is clearly only a mathematical construct, since the true values of the parameters are unknown. One does not expect that $\chi^2_{\text{true}} = 0$, meaning a perfect match between the data and the model. In fact, even if the model is correct, statistical fluctuations of the measurements result in random deviations from the parent model.

On the other hand, after minimizing the χ^2 statistic to find the best-fit parameters \hat{a}_k , the fit statistic

$$\chi^2_{\min} = \sum_{i=1}^N \left(\frac{y_i - \hat{y}_i}{\sigma_i} \right)^2 \quad (12.6)$$

is distributed as $\chi^2(N - m)$, according to Cramér's theorem. It is also clear that the values of the best-fit parameters are not the same as the true parameters, as remarked earlier. After finding χ^2_{\min} , any change in the parameters (say, from \hat{a}_k to a'_k) will yield a larger value of the test statistic, $\chi^2 > \chi^2_{\min}$. The main idea behind the method to identify a confidence interval for the model parameters is to test whether a new

set of parameters a'_k can be the true (yet unknown) values of the model parameters, e.g., whether the corresponding χ^2 can be considered to be χ_{true}^2 . For this purpose, a new statistic is constructed:

$$\Delta\chi^2 = \chi^2 - \chi_{\min}^2 \geq 0 \quad (12.7)$$

where χ^2 is obtained for a given set of model parameters and, by definition, $\Delta\chi^2$ is non-negative. The question, therefore, turns to the sampling distribution of this new statistic $\Delta\chi^2$, under the null hypothesis that $\chi^2 = \chi_{\text{true}}^2$. It is possible to show that, under the regularity conditions required by the Cramér theorem, χ_{true}^2 and χ_{\min}^2 are independent and the statistic $\Delta\chi^2 = \chi_{\text{true}}^2 - \chi_{\min}^2$ has a sampling distribution

$$\Delta\chi^2 \sim \chi^2(m). \quad (12.8)$$

This result provides a quantitative way to determine the amount by which χ^2 can increase, relative to χ_{\min}^2 , while still remaining consistent with χ_{true}^2 . The method to use the $\Delta\chi^2$ statistic to determine parameter uncertainties is described in the following section.

The independence between χ_{true}^2 and χ_{\min}^2 follows from a similar derivation as the one presented in Sect. 9.4 for the sample variance. In the case of the sample variance, it was possible to find an orthonormal transformation between the N variables Z_i , whose squares are summed to form the χ^2 statistic, and a new set of variables Y_i with the property (9.19). In the more general case considered by Cramér, it is possible to find first a non-orthogonal transformation between the original data variables Z_i (x in Cramér's notation) and new variables Y_i , followed by a new orthogonal transformation (referred to as K transformation) which diagonalizes the $\sum_{i=1}^N Y_i^2$ in terms of the original variables and shows that $\Delta\chi^2$ is independent of χ_{\min}^2 . The two transformations allow for $\Delta\chi^2$ to be written as the difference of two χ^2 -distributed variables according to (12.7), where the three statistics $\chi^2 = \chi_{\text{true}}^2$, $\Delta\chi^2$ and χ_{\min}^2 are mutually independent. A complete mathematical proof of these statements is found in Sect. 30.3 of [21] and in the appendix of [64].

12.4 Confidence Intervals of Model Parameters

Equation (12.8) provides a quantitative method to estimate the confidence interval on the m adjustable parameters. Since the $\Delta\chi^2$ statistic follows the $\chi^2(m)$ distribution, the m parameters can deviate from the best-fit estimates so long as the increase in χ^2 is consistent with the critical value of the respective $\Delta\chi^2$ distribution. For example, in the case of a model with $m = 2$ free parameters, one can expect a change $\Delta\chi^2 \leq 4.6$ for a $p = 0.9$ confidence level, or, for a model with $m = 1$ parameter, a change $\Delta\chi^2 \leq 2.7$. Table 12.1 reports selected critical values of the χ^2 statistic that are

Table 12.1 Values of χ^2_{crit} for selected degrees of freedom and confidence levels p . For the $\Delta\chi^2$ statistic, the number m is equal to the number of free parameters

Degrees of freedom	Confidence level p		
	0.683	0.90	0.99
Values of χ^2_{crit}			
1	1.00	2.71	6.63
2	2.30	4.61	9.21
3	3.53	6.25	11.34
4	4.72	7.78	13.28
5	5.89	9.24	15.09

commonly used for the $\Delta\chi^2$ statistic (see Table A.7 for a more complete table). The method to determine the confidence interval on the parameters starts with the value of χ^2_{min} . From this, one constructs an m -dimensional volume in parameter space where

$$\Delta\chi^2 \leq \chi^2_{crit},$$

bounded by the surface with $\Delta\chi^2 = \chi^2_{crit}$, where the critical value of a χ^2 distribution depends on the number m and on the confidence level p . Parameter values within this m -dimensional volume are said to be an *m -dimensional confidence interval* on all adjustable parameters.

Example 12.2 (*Confidence intervals of a two-parameter model*) Consider the case of a linear fit to the following normally distributed data:

X (Indep. variable)	Y (Dependent variable)	
	Mean Value	Standard Deviation
0	25	5
1	36	6
2	64	8
3	49	7
4	81	9

The $N = 5$ datapoints have different variances, and therefore the equations of Sect. 11.3 can be used to obtain the following best-fit estimates of the parameters: $a = 25.44 \pm 4.26$ and $b = 12.06 \pm 2.11$. For this fit to a model with $m = 2$ parameters, the parent distribution of the fit statistic, under the null hypothesis that the data are drawn from the parent model, is $\chi^2(3)$, where $f = N - m = 3$ is the number of degrees of freedom. The critical value of the $\chi^2(3)$ statistic at a confidence level $p = 0.9$ is $\chi^2_{crit} = 6.25$, and it becomes $\chi^2_{crit} = 11.34$ at a higher confidence level of $p = 0.99$. The fit statistic is $\chi^2_{min} = 7.24$ for a p -value of 0.065, and the best-fit model is shown in Fig. 12.2. The model is therefore consistent with the data at the 99% confidence level but should be discarded at the 90% confidence level. In practice, most analysts would regard this fit as *acceptable*, given that there is a significant

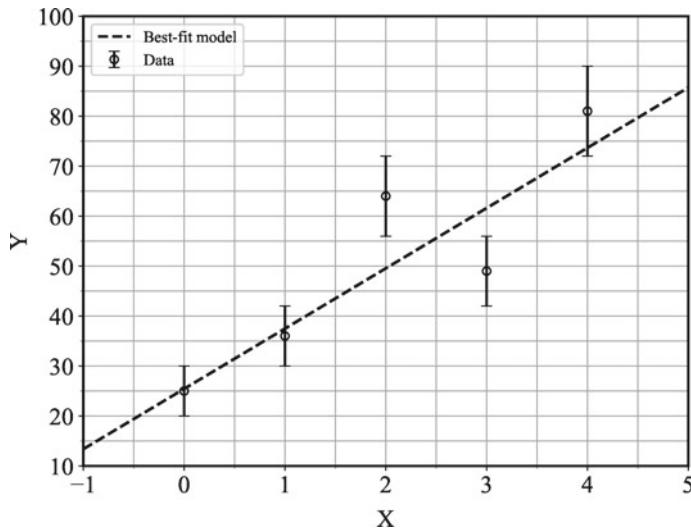


Fig. 12.2 Best-fit linear model (dashed line) to the data of Example 12.2, with a best-fit statistic $\chi^2_{\min} = 7.24$ for $f = N - m = 3$ degrees of freedom

probability (as given by the p -value, or 6.5%) that such a large value of χ^2 can be obtained by chance under the null hypothesis. Notice how the small number of degrees of freedom makes it such that a *reduced* χ^2 value of well over 2 is actually quite reasonable.

The two-dimensional parameter space must now be sampled to determine variations in the fit statistic χ^2 around the minimum value. The result is shown in Fig. 12.3, in which the contours mark the $\chi^2_{\min} + 1.0$, $\chi^2_{\min} + 2.3$ and $\chi^2_{\min} + 4.6$ boundaries. In this fit with $m = 2$ free parameters, the statistic $\Delta\chi^2$ is distributed like $\chi^2(2)$, and a value of $\Delta\chi^2 = 4.6$ or larger is expected only 10% of the time. Accordingly, the $\Delta\chi^2 = 4.6$ contour marks the boundary of the 90% 2-dimensional confidence interval (or surface): the true values of a and b are expected to be within this area 90% of the time, if the null hypothesis is correct. The smaller areas are only shown for reference, since they encompass a smaller probability of containing the true value of the parent parameters. Larger areas could be also constructed; for example, a surface bounded by $\Delta\chi^2 = 9.2$ would encompass a 99% probability, since that is the critical value at $p = 0.99$ for a $\chi^2(2)$ distribution (see Table 12.1). ◇

12.5 Confidence Intervals on a Reduced Number of Parameters

In the case of a large number m of free parameters, it is customary to report the uncertainty on each of the fitted parameters or, in general, on just a subset of $l < m$ parameters considered to be of interest. In this case, the l parameters a_1, \dots, a_l are

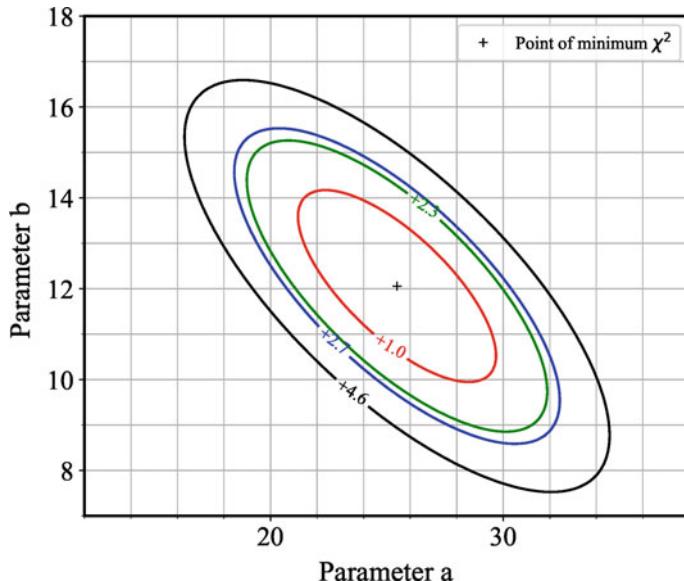


Fig. 12.3 Contours of $\Delta\chi^2 = 1.0, 2.3, 2.7$ and 4.6 (from smaller to larger areas) for the data of Example 12.2. For the two-dimensional confidence surface for both parameters, the relevant curves are $\Delta\chi^2 = 2.3$ (68% confidence level) and $\Delta\chi^2 = 4.6$ (90% confidence level)

said to be the *interesting* parameters, and the remaining $m - l$ parameters are said to be *uninteresting*. This can be thought of as reducing the number of parameters of the model from m to l , often in such a way that only one interesting parameter is investigated at a time ($l = 1$). This is a situation that is of practical importance for several reasons. First, it is not convenient to display surfaces in more than two or three dimensions. Also, sometimes there are parameters that are truly uninteresting to the interpretation of the data, although necessary for its analysis. One case of this is the presence of a measurement background, which must be taken into account for a proper analysis of the data, but it is of no interest in the interpretation of the results.

When considering only a few interesting parameters, the χ^2_{\min} is calculated in the usual way, that is, by fitting all parameters and adjusting them until the minimum χ^2 is found, so that χ^2_{\min} continues to be distributed like $\chi^2(N - m)$. New considerations must be applied to χ^2_{true} and its distribution when considering only a subset of interesting parameters. The goal is to treat the $y(x)$ model as if only l interesting parameters are present, and therefore assume that they are fixed at the true values, same as was done when considering all parameters. The question remains as to what to do with the remaining $m - l$ uninteresting parameters. If they are *marginalized* over, meaning that they are left free to adjust themselves to the values that yield the lowest value of χ^2 , then this process ensures that

$$\chi^2_{\text{true}} \propto \chi^2(N - (m - l)). \quad (12.9)$$

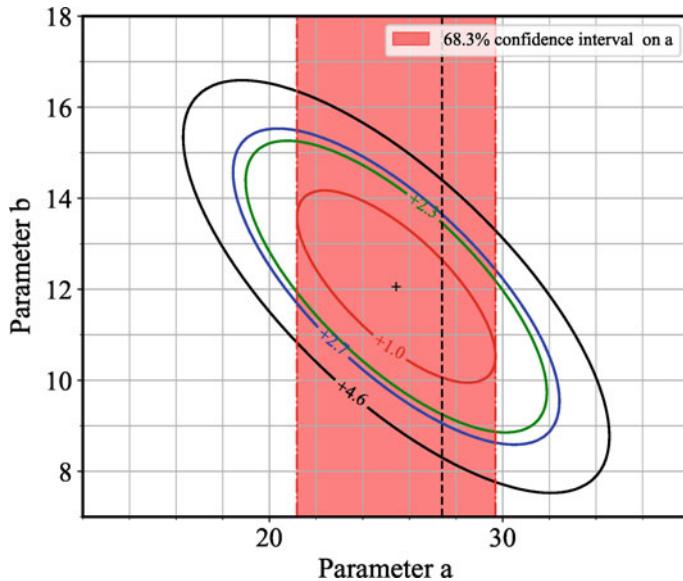


Fig. 12.4 Illustration of confidence intervals on a reduced number of parameters, using the data of Example 12.2. A 68.3% confidence interval on a single parameter is the projection of the $\Delta\chi^2 = 1$ surface, while a 68.3% confidence region for both parameters is bounded by the $\Delta\chi^2 = 2.3$ curve

Notice that the marginalization does *not* mean fixing the values of the uninteresting parameters to their best-fit values. Rather, when calculating $\Delta\chi^2$ for the interesting parameters, at each point in l -dimensional parameter space, the value of χ^2 is the one obtained by minimizing χ^2 over the uninteresting parameters. This minimization makes it such the resulting χ_{true}^2 is really a “minimum χ^2 ” for a fit with $N - (m - l)$ degrees of freedom (the datapoints minus the uninteresting parameters), so that (12.9) applies along with the independence between χ_{true}^2 and the original χ_{\min}^2 . In summary, when only l interesting parameters are considered, the l -dimensional confidence interval is obtained according to

$$\Delta\chi^2 \sim \chi^2(l). \quad (12.10)$$

The process of marginalization is quite simple to accomplish in practice, and it is illustrated in the following example.

Example 12.3 (*Confidence intervals on interesting parameters*) Consider the same data as in Example 12.2 and assume that the interesting parameter is a . This means that the confidence interval on a is an interval of the x axis in Fig. 12.4 bounded by values of a that give $\Delta\chi^2 = 1$ relative to the minimum obtained at the point indicated by the cross. For each fixed value of a , the χ^2 depends on the value of the uninteresting parameter b as it varies along a vertical line. When seeking a $1-\sigma$ or 68.3% confidence interval for a , the limiting values of a are those on either side of the best-fit value

that result in a $\Delta\chi^2 = 1$, since the critical value at $p = 0.683$ confidence level for a distribution $\chi^2(1)$ is 1. Therefore, the 68.3% confidence interval for a is found by projecting the $\chi_{\min}^2 + 1.0$ contour along the a axis. Likewise, a 90% confidence interval on a is given by the projection of the $\chi_{\min}^2 + 2.7$ contour along the a axis, since 2.7 is the critical value of a $\chi^2(1)$ distribution for $p = 0.9$.

On the other hand, the 2-dimensional 68% confidence surface on both parameters a and b is given by the $\chi_{\min}^2 + 2.3$ contour, and the 90% confidence surface by the $\chi_{\min}^2 + 4.6$. It is important not to confuse the one-dimensional and two-dimensional confidence regions (or, in general, l - and m -dimensional regions). A 90% two-dimensional surface ($\Delta\chi^2 = 4.6$) contains 90% of all possible parameter pairs, while a 90% one-dimensional interval ($\Delta\chi^2 = 2.7$) contains 90% of a values, regardless of the value of the b parameter. All considerations for a also apply to the other parameter b , so that the same $\Delta\chi^2$ contours can be projected along the other axis to give confidence intervals on the b parameter. ◇

The use of (12.10) for a subset of interesting parameters applies to any number of interesting parameters. When $l > 1$, the confidence region is obtained as a “projection” of the full m -dimensional $\Delta\chi^2$ volume, bounded by the surface with the appropriate critical value, along the l dimensions of the interesting parameters, much like the 2-dimensional surface of Fig. 12.4 is projected along each axis. Such operation becomes geometrically less easy to represent, but it must follow the same general method of letting the uninteresting parameters adjust themselves so that, at each point in l -dimensional space, the χ^2 is minimized with respect to the $m - l$ uninteresting parameters. This procedure for estimation of intervals on a reduced number of parameters was not well understood until the 1976 paper of Lampton, Margon and Bowyer [64], and it is now widely accepted as the correct method to estimate errors in a subset of the model parameters.

Summary of Key Concepts for this Chapter

The χ_{\min}^2 statistic: The χ_{\min}^2 statistic is the best-fit statistic that applies to Gaussian data, and it is distributed like a $\chi^2(N - m)$ distribution, when there are N measurements and m free parameters.

Confidence intervals on model parameters: A region of parameter space obtained from the condition that $\Delta\chi^2 \sim \chi^2(m)$, where m is the number of parameters of interest.

Interesting parameters: A subset $l < m$ of model parameters for which confidence intervals are of interest. Their confidence intervals is calculated from the condition $\Delta\chi^2 \sim \chi^2(l)$ by marginalizing over the remaining uninteresting parameters.

Marginalization: the process of varying uninteresting parameters, for fixed values of interesting parameters, so that the χ^2 is minimized.

Problems

12.1 ■ Consider the data from Hubble's experiment in Table 11.1, same as for Problem 11.1.

- (a) Using the best-fit model determined for the fit to a linear model for $(m, \log v)$ (see Problem 11.1), estimate the model sample variance from the data and the best-fit model, and then estimate the errors in the parameters a and b and the covariance between a and b .
- (b) Calculate the statistic χ^2_{\min} of the linear fit, using the common error as estimated in part (a).

12.2 ■ Evaluate the minimum χ^2 of the linear fit to the data provided in Table 8.1 and in Problem 11.6. Determine whether the maximum-likelihood fit is acceptable at the 90% confidence level.

12.3 ■ Consider the same 5-point data as in Problem 11.2 and Example 12.2.

- (a) Plot the 2-dimensional confidence contours at 68 and 90% significance, by sampling the (a,b) parameter space in a suitable interval around the best-fit values.
- (b) Using a suitable 2-dimensional confidence contour, determine the 68% confidence intervals on each parameter separately, and compare with the analytic results obtained from the linear regression method.

12.4 The background rate in a measuring apparatus is assumed to be constant with time. Assume that of an even number N of measurements taken, $N/2$ result in a value of $\bar{y} + \Delta$, and $N/2$ in a value $\bar{y} - \Delta$. Use the data and the best-fit model to estimate the sample variance of the background rate.

12.5 Consider the following 3-measurement dataset,

x	y
0	1
1	1
2	1

and assume that the y measurements are Gaussian, with variances equal to the measurements. For the one-parameter fit function $y = e^{ax}$, find the best-fit parameter a and show that the 68% confidence interval is given by the range $\Delta a = \pm\sqrt{1/5}$, under the assumption of small a .

12.6 ■ Use the data from Table 8.1 for the radius versus ratio, assuming that the radius is the independent variable with no error, same as in Problem 11.6. Draw the $\Delta\chi^2 = 1, 2.7$ and 4.6 contours on the two fit parameters a and b . Comment on whether these two-dimensional regions can be used as confidence intervals on the best-fit parameters.

Chapter 13

Multi-variable Regression



Abstract In many situations, a variable of interest depends on several other variables. Such multi-variable data are common across the sciences and in many other fields such as economics and business. Multi-variable analysis can be performed in a simple and effective way when the relationship that links the variable of interest to the other quantities is linear. This chapter presents the method of multi-variable regression and shows how it is related to the multiple regression described in Chap. 11, which applies to the traditional two-variable dataset. This chapter also presents the methods for hypothesis testing on the multi-variable regression and its parameters.

13.1 Multi-variable Datasets

The two-dimensional datasets studied so far include an independent variable X and a dependent variable Y , and the data took the form of a collection of (x_i, y_i) , where $i = 1, \dots, N$, with N indicating the total number of measurements. Chapter 11 described a method to fit such two-dimensional data, including a linear regression with a model $y(x) = a + bx$, where a and b are the two adjustable parameters.

This chapter provides an extension of those two-variable models to data that have more than two variables. Datasets that have measurements for three or more variables are referred to as *multi-variable datasets*. An example of multi-variable dataset is presented in Sect. 13.2, which reports measurements of different characteristics of irises performed by R.A. Fisher and E. Anderson in 1935–1936 [3, 34]. Each of those measurements comprises four quantities: the sepal length, sepal width, petal length, and petal width of irises. For several multi-variable datasets such as that of Fisher and Anderson, it is often unclear which variable is the dependent one. It typically depends on the question at hand: if the goal is to determine the sepal length of an iris flower based on the sepal width, petal length, and petal width, then the sepal length becomes the dependent variable and the remaining three are the independent variables.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-981-19-0365-6_13.

Using multi-variable datasets to predict or forecast the behavior of one quantity based on several other variables is a fundamental topic in data analysis. It is common throughout the sciences and especially used in such fields as economics or behavioral sciences, where a number of possible factors can be used to predict one quantity of interest. An example is to predict the score on a college-admission test based on several factors, such as the grade-point average during the sophomore and the junior years, a measure of the motivation of the student, and their economic status. Another example is to predict the price of a stock based, e.g., on the overall index of the stock exchange, a consumer's index for goods in the relevant class and the rate of treasury bonds. To address any such questions clearly requires a multi-variable dataset that has several measurements for all quantities of interest. This chapter develops a method to determine the relationship between one of the quantities of multi-dimensional datasets based on the others, assuming a linear relationship among the variables. This method will also determine whether one or more of the quantities are in fact not useful in predicting the variable of interest. For example, one may find that the treasury bond rates are irrelevant in predicting the stock value of a given corporation and therefore the analyst may focus only on those variables that are useful in predicting its stock price.

13.2 A Classic Experiment: The R.A. Fisher and E. Anderson Measurements of Iris Characteristics

R.A. Fisher is one of the fathers of modern statistics. In 1936, he published the paper *The Use of Multiple Measurements in Taxonomic Problems* reporting measurements of several characteristics of three species of the iris plant [34]. Figure 13.1 reproduces the original measurements, performed by E. Anderson [3], of the petal length and the sepal length of 150 iris plants of the species *iris setosa*, *iris versicolor*, and *iris virginica*. The measurements are in millimeters (mm). Fisher's aim was to find a linear combination of the four characteristics that would be best suited to identify one species from the others. It is already clear from the data in Fig. 13.1 that one of the quantities (e.g., the sepal length) may be used as a discriminator among the three species. R.A. Fisher used this dataset to find a linear combination of the four quantities that would improve the classification of irises.

The dataset is a classic example of a multi-variate dataset, in which several variables are measured simultaneously and independently. In addition to Fisher's original purpose, these data can also be used to determine whether one of the characteristics, e.g., the sepal length can be efficiently predicted based on any (or all) of the other characteristics. For example, one could expect that the length of the sepal (*SL*, which is part of the calyx of the flower) is related linearly to its width *SW*, or to the length of the petal *PL* or the width of the petal *PW*. Assuming a linear relationship among the variables, the relationship is

$$SL = a + b \cdot SW + c \cdot PL + d \cdot PW \quad (13.1)$$

where a , b , c , and d are coefficients that we must be estimated from the data, using the method described in Sect. 13.3.

Throughout this chapter, these data are used to study the linear regression (13.1) for the species *iris setosa*. It will be found that the most important variable needed to predict the sepal length is the sepal width, while the measurements of characteristics of petals are not very important in predicting the sepal length.

<i>Iris setosa</i>				<i>Iris versicolor</i>				<i>Iris virginica</i>			
Sepal length	Sepal width	Petal length	Petal width	Sepal length	Sepal width	Petal length	Petal width	Sepal length	Sepal width	Petal length	Petal width
5.1	3.5	1.4	0.2	7.0	3.2	4.7	1.4	6.3	3.3	6.0	2.5
4.9	3.0	1.4	0.2	6.4	3.2	4.5	1.5	5.8	2.7	5.1	1.9
4.7	3.2	1.3	0.2	6.9	3.1	4.9	1.5	7.1	3.0	5.9	2.1
4.6	3.1	1.5	0.2	5.5	2.3	4.0	1.3	6.3	2.9	5.6	1.8
5.0	3.6	1.4	0.2	6.5	2.8	4.6	1.5	6.5	3.0	5.8	2.2
5.4	3.9	1.7	0.4	5.7	2.8	4.5	1.3	7.6	3.0	6.6	2.1
4.6	3.4	1.4	0.3	6.3	3.3	4.7	1.6	4.9	2.5	4.5	1.7
5.0	3.4	1.5	0.2	4.9	2.4	3.3	1.0	7.3	2.9	6.3	1.8
4.4	2.9	1.4	0.2	6.6	2.9	4.6	1.3	6.7	2.5	5.8	1.8
4.9	3.1	1.5	0.1	5.2	2.7	3.9	1.4	7.2	3.6	6.1	2.5
5.4	3.7	1.5	0.2	5.0	2.0	3.5	1.0	6.5	3.2	5.1	2.0
4.8	3.4	1.6	0.2	5.9	3.0	4.2	1.5	6.4	2.7	5.3	1.9
4.8	3.0	1.4	0.1	6.0	2.2	4.0	1.0	6.8	3.0	5.5	2.1
4.3	3.0	1.1	0.1	6.1	2.9	4.7	1.4	5.7	2.5	5.0	2.0
5.8	4.0	1.2	0.2	5.6	2.9	3.6	1.3	5.8	2.8	5.1	2.4
5.7	4.4	1.5	0.4	6.7	3.1	4.4	1.4	6.4	3.2	5.3	2.3
5.4	3.9	1.3	0.4	5.6	3.0	4.5	1.5	6.5	3.0	5.5	1.8
5.1	3.5	1.4	0.3	5.8	2.7	4.1	1.0	7.7	3.8	6.7	2.2
5.7	3.8	1.7	0.3	6.2	2.2	4.5	1.5	7.7	2.6	6.9	2.3
5.1	3.8	1.5	0.3	5.6	2.5	3.9	1.1	6.0	2.2	5.0	1.5
5.4	3.4	1.7	0.2	5.9	3.2	4.8	1.8	6.9	3.2	5.7	2.3
5.1	3.7	1.5	0.4	6.1	2.8	4.0	1.3	5.6	2.8	4.9	2.0
4.6	3.6	1.0	0.2	6.3	2.5	4.9	1.5	7.7	2.8	6.7	2.0
5.1	3.3	1.7	0.5	6.1	2.8	4.7	1.2	6.3	2.7	4.9	1.8
4.8	3.4	1.9	0.2	6.4	2.9	4.3	1.3	6.7	3.3	5.7	2.1
5.0	3.0	1.6	0.2	6.6	3.0	4.4	1.4	7.2	3.2	6.0	1.8
5.0	3.4	1.6	0.4	6.8	2.8	4.8	1.4	6.2	2.8	4.8	1.8
5.2	3.5	1.5	0.2	6.7	3.0	5.0	1.7	6.1	3.0	4.9	1.8
5.2	3.4	1.4	0.2	6.0	2.9	4.5	1.5	6.4	2.8	5.6	2.1
4.7	3.2	1.6	0.2	5.7	2.6	3.5	1.0	7.2	3.0	5.8	1.6
4.8	3.1	1.6	0.2	5.5	2.4	3.8	1.1	7.4	3.8	6.1	1.9
5.4	3.4	1.5	0.4	5.5	2.4	3.7	1.0	7.9	3.8	6.4	2.0
5.2	4.1	1.5	0.1	5.8	2.7	3.9	1.2	6.4	2.8	5.6	2.2
5.5	4.2	1.4	0.2	6.0	2.7	5.1	1.6	6.3	2.8	5.1	1.5
4.9	3.1	1.5	0.2	5.4	3.0	4.5	1.5	6.1	2.6	5.6	1.4
5.0	3.2	1.2	0.2	6.0	3.4	4.5	1.6	7.7	3.0	6.1	2.3
5.5	3.5	1.3	0.2	6.7	3.1	4.7	1.5	6.3	3.4	5.6	2.4
4.9	3.6	1.4	0.1	6.3	2.3	4.4	1.3	6.4	3.1	5.5	1.8
4.4	3.0	1.3	0.2	5.6	3.0	4.1	1.3	6.0	3.0	4.8	1.8
5.1	3.4	1.5	0.2	5.5	2.5	4.0	1.3	6.9	3.1	5.4	2.1
5.0	3.5	1.3	0.3	5.5	2.6	4.4	1.2	6.7	3.1	5.6	2.4
4.5	2.3	1.3	0.3	6.1	3.0	4.6	1.4	6.9	3.1	5.1	2.3
4.4	3.2	1.3	0.2	5.8	2.6	4.0	1.2	5.8	2.7	5.1	1.9
5.0	3.5	1.6	0.6	5.0	2.3	3.3	1.0	6.8	3.2	5.9	2.3
5.1	3.8	1.9	0.4	5.6	2.7	4.2	1.3	6.7	3.3	5.7	2.5
4.8	3.0	1.4	0.3	5.7	3.0	4.2	1.2	6.7	3.0	5.2	2.3
5.1	3.8	1.6	0.2	5.7	2.9	4.2	1.3	6.3	2.5	5.0	1.9
4.6	3.2	1.4	0.2	6.2	2.9	4.3	1.3	6.5	3.0	5.2	2.0
5.3	3.7	1.5	0.2	5.1	2.5	3.0	1.1	6.2	3.4	5.4	2.3
5.0	3.3	1.4	0.2	5.7	2.8	4.1	1.3	5.9	3.0	5.1	1.8

Fig. 13.1 Measurements of three iris species, reproduced from the 1936 R.A. Fisher paper [34]. Measurements are in mm for all quantities

13.3 The Multi-variable Linear Regression

Consider a dataset with N measurements of $m + 1$ variables, of which one will be indicated as the dependent variable Y , and the remaining are indicated as the independent variables X_k , for $k = 1, \dots, m$. The data are therefore of the form

$$(y_i, x_{1i}, \dots, x_{mi}) \text{ for } i = 1, \dots, N.$$

The model that describes the variable Y is a linear function of the m variables X_i ,

$$y(x) = a_0 + a_1 x_1 + \dots + a_m x_m = a_0 + \sum_{k=1}^m a_k x_k. \quad (13.2)$$

The goal is to find the values for the $m + 1$ coefficients a_k that minimize the χ^2 statistic

$$\chi^2 = \sum_{i=1}^N \left(\frac{y_i - y(x_i)}{\sigma_i} \right)^2, \quad (13.3)$$

where σ_i is a measure of the variance of the dependent variable. The quantity $y(x_i)$ is the value of the model $y(x)$ calculated for the i -th set of measurements of the X_k variables. This form for the χ^2 function is similar to that used for the multiple linear regression of Sect. 11.4. The only changes are that the variables x_k take the place of the functions $f_k(x)$, and the presence of the a_0 coefficient, which is equivalent to the constant a for the two-dimensional linear regression function $y = a + bx$. The quantity σ_i is interpreted as the error in the variable Y , which is the dependent quantity in this regression. As in the case of the two-variable dataset, errors in the independent variables X_k are ignored. When the multi-variable dataset has no errors reported, as in the case of the Fisher and Anderson iris data, it is assumed that the measurements have a uniform variance, and the equations use a constant $\sigma_i^2 = \sigma^2$, same as in the multiple regression of Sect. 11.4.

The similarity in form between the χ^2 for the present multi-variable linear regression and for the one in the multiple regression means that there is already a solution available for the coefficients of the regression and their errors. What is needed is to make the following substitutions:

$$\begin{cases} f_1(x) = 1 = x_0 \\ f_{k+1}(x) = x_k, \text{ for } k = 1, \dots, m \end{cases} \quad (13.4)$$

and use the solution from Sect. 11.4 with $m + 1$ terms. The best-fit parameters a_k can be found via the matrix Eq. (11.14), which requires the usual inversion of the matrix A :

$$\mathbf{a} = \boldsymbol{\beta} \mathbf{A}^{-1}, \quad (13.5)$$

where the $(m + 1)$ -dimensional row vectors β and a , and the $(m + 1) \times (m + 1)$ symmetric matrix A are given by

$$\begin{cases} \beta = (\beta_0, \beta_1, \dots, \beta_m) \\ a = (a_0, a_1, \dots, a_m) \\ A_{lk} = \sum_{i=1}^N \frac{x_{li} x_{ki}}{\sigma_i^2} \end{cases} \quad \begin{array}{l} \text{where } \beta_k = \sum_{i=1}^N \frac{x_{ki} y_i}{\sigma_i^2} \\ \text{(model parameters)} \\ (l, k \text{ component of } A). \end{array}$$

The errors and covariances among parameters are likewise given by the usual error matrix $\varepsilon = A^{-1}$. The multi-variable linear regression can therefore be solved with the same methods as the multiple linear regression between two variables.

13.4 Multi-variable Linear Regression with Uniform Variance

The multi-variable linear regression has a simpler solution when there is a constant value for the variance σ^2 of the measurements. This is for example the case for the Fisher and Anderson iris data, for which no measurement errors are reported, and therefore a uniform variance is assumed. In this case, the matrix A and the vector β can be written in extended form as

$$A = \frac{1}{\sigma^2} \begin{bmatrix} N & \sum_{i=1}^N x_{1i} & \dots & \sum_{i=1}^N x_{mi} \\ \sum_{i=1}^N x_{1i} & \sum_{i=1}^N x_{1i}^2 & \dots & \sum_{i=1}^N x_{1i} x_{mi} \\ \vdots & & & \\ \sum_{i=1}^N x_{mi} & \sum_{i=1}^N x_{mi} x_{1i} & \dots & \sum_{i=1}^N x_{mi}^2 \end{bmatrix} \quad (13.6)$$

$$\beta = \frac{1}{\sigma^2} \left(\sum_{i=1}^N y_i, \sum_{i=1}^N x_{1i} y_i, \dots, \sum_{i=1}^N x_{mi} y_i \right) \quad (13.7)$$

where all sums are over the N measurements. A solution for the best-fit parameters a according to (13.5) is possible even when the value of the variance is not specified, but the error matrix requires a value for the variance, same as for the case of the two-variable linear regression.

An alternative notation for finding the coefficients a_k makes use of the following definitions:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_N \end{bmatrix}; \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & \dots & x_{m1} \\ 1 & x_{12} & \dots & x_{m2} \\ \dots \\ 1 & x_{1N} & \dots & x_{mN} \end{bmatrix} \text{ and } \mathbf{a}^T = \begin{bmatrix} a_0 \\ a_1 \\ \dots \\ a_m \end{bmatrix} \quad (13.8)$$

where the Y measurements and the coefficients are arranged in column vectors, and \mathbf{X} is called the *design matrix*, which is an $N \times (m + 1)$ matrix with each row containing the m measurements of the independent variables for the i -th observation. With this notation, the least-squares method gives the following solution for the coefficients [100]:

$$\mathbf{a}^T = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} \quad (13.9)$$

It is easy to show that (13.5) and (13.9) are equivalent, with $\mathbf{X}^T \mathbf{X} = \sigma^2 \mathbf{A}$ and $\mathbf{X}^T \mathbf{Y} = \sigma^2 \boldsymbol{\beta}$ (see Problem 13.3). Using this notation, the error matrix is therefore given by

$$\boldsymbol{\varepsilon} = \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1} = \mathbf{A}^{-1}. \quad (13.10)$$

The two methods to calculate the coefficients of the multiple regression and their errors are therefore equivalent. The latter form (13.8) may be convenient if the data are already tabulated according to the measurements in the form of matrix \mathbf{X} , and therefore the parameters can be found using the matrix algebra of (13.9). The drawback is that the design matrix can have a large size, $N \times (m + 1)$ where N is the number of measurements. The form of (13.5) is more compact, since the matrix \mathbf{A} is $(m + 1) \times (m + 1)$, and the summation over the N measurements must be performed beforehand to obtain the matrix \mathbf{A} .

Example 13.1 (*Multi-variable linear regression on iris setosa data*) The data of Fig. 13.1 for the *iris setosa* species are fit to the linear model of (13.1), where the sepal length is used as the Y variable and the remaining three variables are the independent variables. The measurements are first arranged in 50-long vectors, then the X measurements are combined into the 4×50 design matrix \mathbf{X} , and finally the solution for the best-fit parameters is found using (13.9). The best-fit parameters of the regression are

$$\mathbf{a} = (2.352, 0.655, 0.238, 0.252).$$

It was assumed that the data have uniform variances, and it is not necessary to specify a value for the variance to find the least-squares solution to the multiple-variable linear regression. In the absence of the data variance, it is not possible to establish the goodness of fit. \diamond

13.5 Goodness of Fit of Multi-variable Regression

The fit statistic that determines the goodness of fit is the usual χ^2_{\min} obtained from (13.3) using the best-fit model $\hat{y}(x)$. When the data variances or a common data variance is specified, the null hypothesis that the model is an accurate description of the data implies that $\chi^2_{\min} \sim \chi^2(N - m - 1)$, according to Cramér's theorem, since there are $m + 1$ adjustable parameters in the regression. The hypothesis testing for the multi-variable linear regression therefore follows the same procedure as for the standard near regression, when the data variance is known or can be somehow estimated from the data. It is a common occurrence, though, that multi-variable data do not report the value of the variance of the independent variable. This means that usually the χ^2_{\min} statistic cannot be used to test the goodness of fit for the multi-variable regression.

Example 13.2 (*Goodness of fit of iris setosa data using a fiducial data variance*) Fisher and Anderson did not report uncertainties in their measurements of the lengths associated with the *iris* data, and it is therefore difficult to determine an accurate value for the uncertainty in the measurements. It is nonetheless possible to speculate that a typical uncertainty in the measurements is of the order of 0.1 inches, since this is the precision to which the measurements were reported. It is also likely that there are other sources of uncertainty in the measurements, and that certain quantities might have been measured with better precision than others. To illustrate the test of goodness of fit for multi-variable regressions, the starting point is the sum of the square of the residuals, which is independent of the value of the unknown common variance. For the *iris setosa* data, the best fit model parameters reported in Example 13.1 lead to

$$\chi^2_{\min} = \frac{2.59}{\sigma^2},$$

where σ^2 is the unknown value of the common data variance. Using fiducial values for the data variance, the measured sum of the square of the residuals corresponds to $\chi^2_{\min} = 258.7$ if $\sigma = 0.1$ and $\chi^2_{\min} = 64.7$ if $\sigma = 0.2$. These values of the fit statistic show that, if $\sigma = 0.1$, then the null hypothesis should be discarded, since the critical value of a $\chi^2(46)$ distribution, for example at the $p = 0.99$ confidence level, is 71.2. On the other hand, if there are reasons to believe that the parent data variance is such that $\sigma = 0.2$, then the data are consistent with the null hypothesis. In summary, if the data analyst has good reasons to use a fiducial value of the data variance, a test of the goodness of fit is possible. ◇

When the data variance is not available, it is possible to approximate it with a model sample variance defined in the same manner as (12.4),

$$\hat{s}^2 = \frac{1}{N - m - 1} \sum_{i=1}^N (y_i - \hat{y}_i)^2 \quad (13.11)$$

taking into account that there are now $m + 1$ free parameters that were estimated by the least-squares method. When this estimate is used to replace the unknown σ^2 , it is necessary to remember that the analyst has made the assumption that the model is correct, and therefore the goodness of fit cannot be established. Using the model sample variance has the advantage that uncertainties in the model parameters can now be estimated according to (13.10). As already remarked in Sect. 12.2, this estimate is based on the assumption that the data variance yields a unit reduced value of the minimum χ^2 .

Example 13.3 (*Model sample variance for the iris setosa data*) Using the same *iris setosa* data as the previous example, the model sample variance is estimated as

$$\hat{s}^2 = \frac{2.59}{N - m - 1} = 0.056$$

where $N = 50$ and $m = 3$, which corresponds to $\hat{s} = 0.237$. Notice how in the previous example a fiducial value of $\sigma = 0.2$ yielded a reduced χ_{\min}^2 just above unity. The model sample variance is the estimated value of the variance that gives a reduced value of exactly one, assuming that the data are accurately described by the model. \diamond

13.6 Tests for the Significance of Multiple Regression Coefficients

The multi-variable linear regression model of (13.2) is specified by the $m + 1$ coefficients a_k . After determining their best-fit values and, if possible, the overall goodness of fit, it is also possible to establish whether there are any independent variables X_i that do not provide a significant contribution to the prediction of the Y variable. Two tests are introduced for this purpose: the first test is based on the Student t statistic and it makes use of the estimated model sample variance, while the second one is based on the F statistic and it does not require the specification of a data variance.

13.6.1 *t*-Test for the Significance of Model Components

For the multi-variable linear regression, it is often useful to test the significance of each of the $m + 1$ parameters, with the goal to determine whether the independent variables are in fact useful in predicting the dependent variable. For this purpose, it is necessary to have available the data variance σ^2 or a suitable estimate such as the model sample variance \hat{s}^2 according to (13.11), so that the error matrix (13.10) can be used to provide estimates of the parameter variances. The starting point for testing

the regression coefficients is the normal distribution of the parameters, since they are a linear combination of normal measurements as in the simple linear regression.

When the data variance σ^2 is known, the variance σ_k^2 of the model parameter a_k , obtained as a diagonal term in the error matrix (13.10),¹ can be used to test whether the parameter is consistent with a non-zero value according to

$$\frac{a_k}{\sigma_k} \sim N(0, 1).$$

In this case, the null hypothesis is that the parent value of the parameter is $\alpha_k = 0$, and hypothesis testing is therefore accomplished with a simple two-tailed test based on the standard normal distribution.

When the true data variance is unknown—which is the more common case—and the error matrix (13.10) is estimated with the model sample variance \hat{s}^2 replacing σ^2 , it is possible to show that the ratio of the parameter's best-fit value a_k to its standard deviation, now indicated as s_k , is distributed like a Student's t distribution with $N - m - 1$ degrees of freedom,

$$t_k = \frac{a_k}{s_k} \sim t(N - m - 1) \quad (13.12)$$

under the null hypothesis that the parameter's true value is zero. In this case, hypothesis testing for a non-zero model parameter is conducted with the t distribution instead of the normal distribution.

When the data variance σ^2 is replaced by the model sample variance \hat{s}^2 , the ratio of the estimated to true parameter variance is, according to (13.10),

$$\frac{s_k^2}{\sigma_k^2} = \frac{\hat{s}^2}{\sigma^2} = \frac{S_r^2}{\sigma^2} \cdot \frac{1}{N - m - 1} \sim \frac{\chi^2(N - m - 1)}{N - m - 1}.$$

The test statistic t_k can then be written as

$$t_k = \frac{(a_k - \alpha_k)/\sigma_k}{s_k/\sigma_k} \quad (13.13)$$

where $\alpha_k = 0$ is the null hypothesis that the parameter has a value of zero, and σ_k^2 is the unknown parent variance for the parameter. It is clear that, under the null hypothesis, the numerator of t_k is distributed like a standard normal distribution.

Since the denominator is

¹ The error matrix for the multi-variable linear regression has a form similar to (11.9), with the estimated parameter variances along the diagonal. For simplicity of notation, the *hat* symbol is dropped from the parameter variances in this section.

Table 13.1 Multiple regression parameters for the *iris setosa* data

Parameter	Best-fit value	Estimated error	<i>t</i> score	<i>p</i> value
a_0	2.352	0.393	5.99	<0.000001
a_1	0.655	0.092	7.08	<0.00000001
a_2	0.238	0.208	1.14	0.26
a_3	0.252	0.347	0.73	0.47

$$\frac{s_k}{\sigma_k} \sim \sqrt{\frac{\chi^2(N-m-1)}{N-m-1}}, \quad (13.14)$$

it follows that t_k is distributed like a *t* distribution,

$$t_k \sim \frac{N(0, 1)}{\sqrt{\chi^2(N-m-1)/(N-m-1)}} \sim t(N-m-1) \quad (13.15)$$

according to the definition of the *t* distribution (9.32) as the ratio of a normal distribution and the square root of a reduced χ^2 variable.

Example 13.4 (*t*-Test on multi-variable regression of *iris setosa* data) The data of Fig. 13.1 for the *iris setosa* species are fit to the linear model of (13.1), where the sepal length is used as the Y variable and the remaining three variables are the independent variables. These data are assumed to have uniform variance, which is estimated via the model sample variance \hat{s}^2 . The parameter uncertainties are therefore estimated according to (13.10), where the unknown σ^2 is replaced with \hat{s}^2 . The multi-variable linear regression leads to the results shown in Table 13.1, including the *t* scores for the four parameters of the multiple regression. For each parameter is reported the two-sided *p* value of the measured *t* statistics according to a *t* distribution with $f = 46$ degrees of freedom, where $f = N - m - 1$ with $N = 50$ measurements and $m = 3$ independent variables. The analysis of the significance of the parameters via the *t* scores leads to the conclusion that parameters a_2 and a_3 , corresponding to the petal length and width, are not significantly different from zero (because of the large probability *p* to exceed their measured values, under the null hypothesis). Accordingly, it would be meaningful to repeat the linear regression using only the sepal width as an estimator for the sepal length. ◇

13.6.2 F-Test for the Significance of the a_1, \dots, a_m Parameters

The purpose of the multi-variable linear model is to provide a fit to the data that is more accurate than a simple constant predictor, i.e., the average of the *Y* measurements.

In other words, it is necessary to establish whether *any* of the slope parameters a_1, a_2, \dots, a_m provides a significant improvement over the constant model with $a_1 = a_2 = \dots = a_m = 0$. For this purpose, it is useful to write the total variance of the data as follows:

$$\sum_{i=1}^N (y_i - \bar{y})^2 = \sum_{i=1}^N (y_i - \hat{y}_i)^2 + \sum_{i=1}^N (\hat{y}_i - \bar{y})^2 \quad (13.16)$$

where $\hat{y}_i = y(x_i)$ is evaluated for the best-fit values of the parameters a_k . This equation holds because

$$\sum_{i=1}^N (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = 0 \quad (13.17)$$

(see Problem 13.5). The parent variance σ^2 of the data is assumed to be unknown and it is not required for this test. The three terms in (13.16) are interpreted as follows. The left-hand side term is the total variance of the data and it is distributed like

$$\frac{S^2}{\sigma^2} = \sum_{i=1}^N \frac{(y_i - \bar{y})^2}{\sigma^2} \sim \chi^2(N - 1). \quad (13.18)$$

The total variance S^2 can be interpreted as the variance obtained using a model with $a_1 = \dots = a_m = 0$, i.e., a constant model with only one free parameter a_0 whose best-fit value is equal to the average of the Y measurements. The first term on the right-hand side of (13.16) is the *residual variance* after the data are fit to the linear model with $m + 1$ parameters and it follows the usual χ^2 distribution

$$\frac{S_r^2}{\sigma^2} = \sum_{i=1}^N \frac{(y_i - \hat{y}_i)^2}{\sigma^2} \sim \chi^2(N - m - 1) \quad (13.19)$$

because of the $m + 1$ free parameters used in the fit, according to Cramér's theorem. This is the usual χ^2_{\min} obtained using the full model in which at least some of the a_k parameters are not equal to zero. Finally, the second term on the right-hand side can be interpreted as the variance *explained* by the best-fit model and it is distributed like

$$\frac{S_e^2}{\sigma^2} = \sum_{i=1}^N \frac{(\hat{y}_i - \bar{y})^2}{\sigma^2} \sim \chi^2(m). \quad (13.20)$$

The distribution of this term can be explained by the independence between the two variables on the right-hand side of the equation and the distribution of the left-hand side term, following a derivation similar to that of Sect. 9.4 for the sample variance.

The assumption made in the derivation of the parent distributions for these statistics is that the parent values of the a_1, \dots, a_m coefficients are zero, as indicated after

(13.18). These m parameters were, however, used in the regression (it is immediate to see that, if they had been fixed to zero, $S^2 = S_r^2$ and $S_e^2 = 0$). If the addition of these slope coefficients is not significant, meaning that these m best-fit parameters are not significantly different from zero, then (13.20) applies, and the variance explained by the full linear model is small compared to the total variance, since typically $m \ll N$. On the other hand, if the null hypothesis that the parent coefficients are zero is false, then the S_e^2 variance term will be large compared to S^2 and S_r^2 , per unit degree of freedom. This would mean that at least some of the parent slope coefficients are non-zero and able to account for a significant fraction of the total variance of the data. The variances described above can be used to define a statistic that tests the hypothesis that the parent regression parameters are null. For this purpose,

$$F = \frac{S_e^2/m}{S_r^2/(N-m-1)} = \frac{\sum_{i=1}^N (\hat{y}_i - \bar{y})^2/m}{\sum_{i=1}^N (y_i - \hat{y}_i)^2/(N-m-1)}, \quad (13.21)$$

is distributed as an F variable with m , $N-m-1$ degrees of freedom under the null hypothesis that the parent parameters are $\alpha_1 = \dots = \alpha_m = 0$ (see Sect. 9.5 for the definition of an F -distributed variable). This F statistic is the ratio between the variance explained by the fit and the residual variance, each normalized by the respective degrees of freedom. Notice how the unknown data variance has canceled out, and therefore its value is not needed for this test. The measurement of F that results from the fit of a dataset to the multi-variable linear model can therefore be used to test the null hypothesis. If the measurement of the F statistic exceeds the critical value of the F distribution for the desired confidence level, the null hypothesis must be rejected. In practice, this is a useful outcome, since it indicates that the constant model is rejected. The rejection of the null hypothesis is usually interpreted as meaning that the linear model with non-zero slope coefficients a_1, \dots, a_m is suggested by the data. To be clear, it does not mean that the multi-variable linear model is the correct model, or even that the model is consistent with the data. In fact, the unavailability of the data variance prevents a χ^2 test of the model, as explained in Sect. 13.5. Nonetheless, a rejection of the null hypothesis via this F test does indicate that the additional slope parameters provide a better fit to the data than the constant model, thus providing evidence in its favor.

Example 13.5 (*F*-Test of *iris setosa* data) The variances for the multi-variable linear regression to the *iris setosa* data are shown in Table 13.2. The F statistic to test for the significance of the slope parameters is

$$F = \frac{S_e^2/3}{S_r^2/46} = \frac{3.50/2}{2.59/46} = 20.76. \quad (13.22)$$

The $p = 0.99$ critical value for an F distribution with 3 and 46 degrees of freedom is 4.24. Therefore, the null hypothesis that the parent values of the a_1, \dots, a_m

Table 13.2 Variances and F -test results for the *iris setosa* data

Statistic	Value	d.o.f	p value
S^2/σ^2	6.09	$N - 1 = 49$	—
S_r^2/σ^2	2.59	$N - m - 1 = 46$	—
S_e^2/σ^2	3.50	$m = 3$	—
$F = \frac{S_e^2/m}{S_r^2/(N - m - 1)}$	20.76	$m, N - m - 1$	1.2×10^{-8}

parameters are null must be rejected. In practice, this means that the linear model is warranted. The probability to exceed the measured value of 20.7 for the test statistic under the null hypothesis is 1.2×10^{-8} , i.e., exceedingly small. The result of this F test is consistent with the t test for individual coefficients, which resulted in the conclusion that the a_1 parameter is significantly different from zero. \diamond

13.6.3 The Coefficient of Determination

A commonly used measure of the ability of the linear model to describe the data is the ratio of the explained variance S_e^2 to the total variance S^2 , defined as the *coefficient of (multiple) determination*

$$R^2 = \frac{S_e^2}{S^2} = \frac{\sum_{i=1}^N (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^N (y_i - \bar{y})^2}. \quad (13.23)$$

According to its definition, it has values $0 \leq R^2 \leq 1$, with values close to 1 indicating that the model describes the data with little additional residual variance. It is possible to relate the coefficient R^2 to the F statistic defined in (13.21) via

$$F = \frac{R^2(N - m - 1)}{(1 - R^2)m}$$

and therefore test for the significance of the R^2 coefficient by testing the associated F statistic in the same way as done in Sect. 13.6.2. The advantage of reporting explicitly a value for R^2 is that it identifies in a simple way the amount of variance in the data that is explained by the multi-variable linear regression model.

Example 13.6 (R^2 Value for the *iris setosa* data) The data in Table 13.2 yield a coefficient of multiple determination $R^2 = 0.575$. This number means that 57.5% of the total data variance is explained by the best-fit regression model. \diamond

Summary of Key Concepts for this Chapter

Multi-variable dataset: Simultaneous measurements of several variables, often without reference to a specific independent variable.

Multi-variable linear regression: Extension of the simple linear regression to the case of multi-variable data, with a model

$$y(x) = a_0 + a_1 x_1 + \dots + a_m x_m.$$

Best-fit coefficients are given by the matrix equation

$$\mathbf{a} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}.$$

t-test for the significance of a coefficient: when the data variance is estimated via the sample model variance \hat{s}^2 , the statistic

$$t = \frac{a_k}{s_k}$$

is distributed like a Student's $t(N - m - 1)$ variable.

F-test for multi-variable linear regression: The statistic

$$F = \frac{S_e^2/m}{S_r^2/(N - m - 1)}$$

follows an F distribution with $m, N - m - 1$ degrees of freedom under the null hypothesis that the coefficients a_1, \dots, a_m are not required by the model.

Coefficient of determination: The ratio between the explained variance and total variance

$$R^2 = \frac{S_e^2}{S^2} \leq 1.$$

Problems

13.1 ■ Consider the *iris setosa* data of Fig. 13.1. Calculate the best-fit parameters for the multi-variable regression featuring the sepal length as the dependent Y variable and the other three quantities (sepal width, petal length and petal width) as the three independent X_k variables.

13.2 ■ Use an F test to determine whether the multi-variable regression of the *iris setosa* of Problem 13.1 data is justified or not as a predictor of the dependent variable.

13.3 Prove that the best-fit parameters for the multi-variable regression according to (13.5) and (13.9) are equivalent. Take into consideration that in (13.5) the vectors a and β are row vectors, and that you may re-write (13.5) using column vectors.

13.4 Prove that the coefficient of determination R^2 for the simple linear regression $y = a + bx$ is equivalent to the sample correlation coefficient of (2.23).

13.5 Prove (13.17).

13.6 ■ Consider the *iris versicolor* data of Fig. 13.1.

- Calculate the best-fit parameters for the multi-variable regression featuring the sepal length as the dependent Y variable and the other three quantities (sepal width, petal length and petal width) as the three independent X_k variables.
- Use the model sample variance to estimate the uncertainties in the best-fit parameters and calculate the t -test statistics for the significance of model component to identify which of the three independent variables is the best predictor of the sepal length.

13.7 ■ Consider the *iris virginica* data of Fig. 13.1.

- Calculate the best-fit parameters and uncertainties for the multi-variable regression featuring the sepal length as the dependent Y variable and the other three quantities (sepal width, petal length and petal width) as the three independent X_k variables, using the model sample variance.
- Determine whether there is sufficient evidence that the multi-variable regression provides a better prediction than a constant model.
- Determine what fraction of the data variance is explained by the multi-variable regression model.

13.8 ■ Consider the *iris setosa* data of Fig. 13.1 and Problem 13.1.

- Use the best-fit multi-variable regression of Problem 13.1 to estimate a model sample variance.
- Use this estimated variance to determine uncertainties in the best-fit model parameters.

Chapter 14

The Linear Correlation Coefficient



Abstract The linear correlation coefficient r is a simple and convenient measure of the degree of linear correlation between two random variables, and it is related to the slopes of the linear regressions of Y on X and of X on Y . The sampling distribution for the linear correlation coefficient, under the hypothesis of no correlation between two variables, also makes it possible to perform a quantitative test for the presence of correlation, as a necessary first step to establish a functional relationship between two variables. For the multiple linear regression of an independent variable Y on several regressors X_i , the coefficient of determination R^2 takes the place of r^2 , and its sampling distribution can be likewise used to test the hypothesis of correlation.

14.1 Linear Regression and Choice of the Independent Variable

Consider N independent and normally distributed measurements of two random variables X and Y . An example is the biometric data collected by Pearson (Table 2.3), where the two variables are the height of the mother and the height of the father. The data, shown in Fig. 14.1, have several datapoints with the same values, due to the method of data collection. Pearson did not report an uncertainty for the measurements, but it is reasonable to assume that all measurements have a similar error, provided that all couples had their heights measured with the same method. The Pearson measurements are a typical example of data where no precedence should be given to either variable when assigning the tag of ‘independent’ when performing a regression. Instead, it is meaningful to proceed with two separate fits: one where the father’s height (X) is considered as the independent variable, or the regression of Y on X , and the other where the mother’s height (Y) is the independent variable, or linear

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-981-19-0365-6_14.

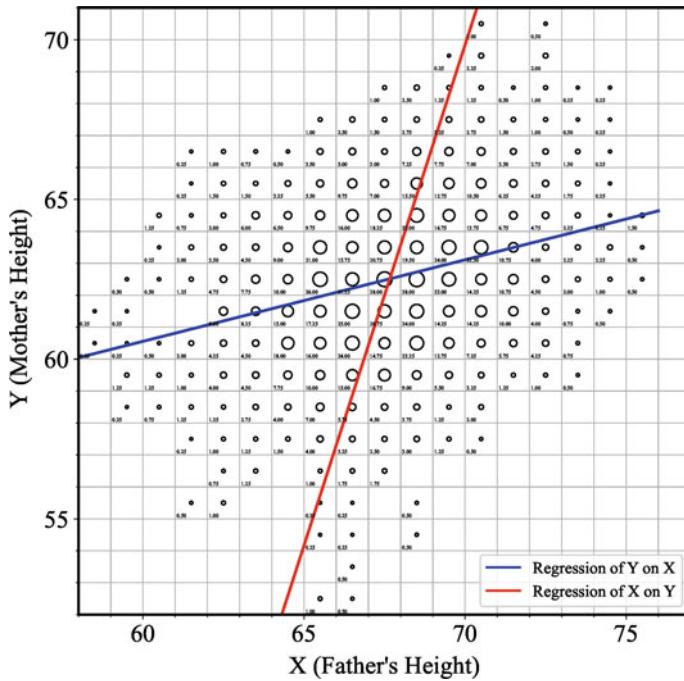


Fig. 14.1 Linear regressions based on the data collected by Pearson, Table 2.3. Larger circles indicate a higher number of occurrences. The number of measurements is reported in the bottom left corner of each bin

regression of X on Y . The two fits are reported in Fig. 14.1, obtained by the usual maximum-likelihood method assuming equal and normally distributed errors for the dependent variables.

It should not be surprising that the two regressions give different results, as shown in Sect. 11.5.2. In the case of the regression of Y on X , the father's height is assumed to be the independent variable, and the slope is given by

$$b_{Y/X} = \frac{s_{xy}^2}{s_x^2},$$

where s_{xy}^2 is the sample covariance and s_x^2 the sample variance of the X measurements. This regression is used to predict Y , given X . On the other hand, the regression of X on Y has a slope of

$$b_{X/Y} = \frac{s_y^2}{s_{xy}^2}.$$

It is clear that and the two slopes are identical only when $s_{xy}^2 = \sqrt{s_x^2 s_y^2}$, which is not true in general.

Example 14.1 (*Linear regressions on the Pearson biometric data*) The Pearson measurements of biometric data of $N = 1,079$ couples contain non-integer counts that require additional attention prior to performing the regressions according to the equations of Sect. 11.5. In fact, it is not possible to reconstruct exactly the (X, Y) pairs of measurements that Pearson collected, given his choice to split the measurements among adjacent bins. Nonetheless, it is possible to calculate sample variances and covariance from these data, by making the following observations. The data are reported in a table with $n = 18$ columns (i.e., 18 bins at fixed father's height, indicated as X) and $m = 19$ rows (at fixed mother's height, indicated as Y). Accordingly, the sample variance of X can be written as

$$s_x^2 = \frac{1}{N-1} \sum_{k=1}^N (x_k - \bar{x})^2 = \frac{1}{N-1} \sum_{i=1}^n n_i \cdot (x_i - \bar{x})^2$$

where n_i is the number of couples with the i -th value x_i of the father's height, obtained as a sum of the i -th column of the data. An equivalent formula applies to the variance of Y . As for the sample covariance,

$$s_{xy}^2 = \frac{1}{N-1} \sum_{k=1}^N (x_k - \bar{x})(y_k - \bar{y}) = \frac{1}{N-1} \sum_{i=1}^n \sum_{j=1}^m n_{ji} \cdot (x_i - \bar{x})(y_j - \bar{y})$$

where n_{ji} are the number of couples with the i -th value of the father's height x_i and the j -th value of the mother's height y_j . In other words, n_{ji} is the Pearson $m \times n$ matrix in Table 2.3 with m rows for the mother's height, and n columns for the father's height. In practice, these equations weigh each of the N coordinate pairs according to the number of reported occurrences, including the $m \times n$ non-integer n_{ji} counts, and the column- and row-summed n_i and n_j counts.

The data have the following sample statistics: $\bar{x} = 67.7$, $\bar{y} = 62.5$, $s_x^2 = 7.20$, $s_y^2 = 5.76$ and $s_{xy}^2 = 1.84$. Accordingly, the two linear regressions have the following parameters:

$$\begin{cases} b_{Y/X} = \frac{s_{xy}^2}{s_x^2} = 0.255 \\ a_{Y/X} = \bar{y} - b_{Y/X} \cdot \bar{x} = 45.24, \end{cases}$$

corresponding to the blue line, $y = a_{Y/X} + b_{Y/X} \cdot x$, or simply $y = a + bx$; and

$$\begin{cases} b_{X/Y} = \frac{s_y^2}{s_{xy}^2} = 3.14 \\ a_{X/Y} = \bar{y} - b_{X/Y} \cdot \bar{x} = -149.60, \end{cases}$$

corresponding to the red line $y = a_{X/Y} + b_{X/Y} \cdot x$. Notice that this regression can also be written as $x = a' + b'y$, where $b' = b_{X/Y}^{-1}$ is the reciprocal slope dx/dy . \diamond

14.2 The Linear Correlation Coefficient

It is now necessary to define a statistic that describes the degree of linear relationship between two random variables X and Y . This can be accomplished with the use of the slope b of the linear regression of Y on X , $y = a + bx$, and the slope b' of the linear regression of X on Y , $x = a' + b'y$. The linear correlation coefficient r is defined by the product of the slopes of the two regressions,

$$r^2 = bb' = \frac{\left(N \sum_{i=1}^N x_i y_i - \sum_{i=1}^N x_i \sum_{i=1}^N y_i \right)^2}{\left(N \sum_{i=1}^N x_i^2 - \left(\sum_{i=1}^N x_i \right)^2 \right) \left(N \sum_{i=1}^N y_i^2 - \left(\sum_{i=1}^N y_i \right)^2 \right)}, \quad (14.1)$$

using the results of (11.18). It is easy to show that this expression can be rewritten as

$$r^2 = \frac{\left(\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y}) \right)^2}{\sum_{i=1}^N (x_i - \bar{x})^2 \sum_{i=1}^N (y_i - \bar{y})^2} \quad (14.2)$$

and therefore r is also the sample correlation coefficient as defined in (2.23), i.e., the ratio of the square of the sample covariance to the product of the two sample variances. The Cauchy–Schwartz inequality (2.22) ensures that $r^2 \leq 1$, with $-1 \leq r \leq 1$.

If the two variables X and Y are uncorrelated, then the two best-fit slopes b and b' have an expectation of zero. In fact, as one variable varies through its range, the other is not expected to either decrease (negative correlation) or increase (positive correlation), resulting in null best-fit slopes for the two fits. One, therefore, expects that the sample distribution of r has zero mean, under the null hypothesis of no correlation between X and Y .

On the other hand, if there is a true linear correlation between the two variables, i.e., $y = a + bx$ is satisfied with $b \neq 0$, then it is also true that $x = a' + b'x = -a/b + (1/b)y$, and one therefore expects $bb' = r^2 = 1$. It is therefore clear that a small absolute value of the r coefficient indicates no correlation between X and Y , and a large absolute value indicates the presence of correlation.

Example 14.2 (*Correlation between mother's and father's height in Pearson's data*) Continuing with the use of the Pearson biometric data, the linear correlation coefficient for these binned data can be evaluated with the usual formula

$$r = \frac{s_{xy}^2}{\sqrt{s_x^2 s_y^2}}$$

where the sample covariance and variances are evaluated as in Example 14.1. The result is $r = 0.285$, or $r^2 = 0.081$. The coefficient can also be calculated from the slope of the two linear regressions obtained in Example 14.1, $b = b_{Y/X} = 0.255$ and $b' = 1/b_{X/Y} = 1/3.135$, whose product leads again to $r^2 = b \cdot b' = 0.081$. \diamond

14.3 Sampling Distribution of r and Hypothesis Testing

A quantitative test for the correlation between two random variables requires knowledge of the sampling distribution of r . It is possible to obtain such distribution under the null hypothesis that the two variables X and Y are uncorrelated. The distribution is given by

$$f_r(r) = \frac{1}{B(1/2, f/2)} (1 - r^2)^{f/2 - 1} \quad (14.3)$$

where $f = N - 2$ is the number of degrees of freedom and N is the number of independent data points. The quantity $B(1/2, f/2)$ is the Beta function defined in (9.23), with

$$B(1/2, f/2) = \frac{\Gamma(1/2)\Gamma(f/2)}{\Gamma(f/2 + 1/2)}$$

where $\Gamma(1/2) = \sqrt{\pi}$. A variable with probability distribution function (14.3) is said to have a *symmetric beta distribution* with parameter f , where symmetry refers to the fact that the random variable is defined in a symmetric interval around 0, namely $-1 \leq r \leq 1$. This distribution is reminiscent of the t distribution, which in fact plays a role in its derivation. This distribution is also referred to as the *r -distribution* with parameter f .

The proof starts with the determination of the probability distribution function of a suitable function of r , and then, by change of variables, the distribution of r is obtained. The best-fit parameter b is given by

$$b^2 = \frac{\left(N \sum_{i=1}^N x_i y_i - \sum_{i=1}^N x_i \sum_{i=1}^N y_i \right)^2}{\left(N \sum_{i=1}^N x_i^2 - (\sum_{i=1}^N x_i)^2 \right)^2} = \frac{\left(\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y}) \right)^2}{\left(\sum_{i=1}^N (x_i - \bar{x})^2 \right)^2},$$

and accordingly

$$r^2 = \frac{\left(\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y}) \right)^2}{\sum_{i=1}^N (x_i - \bar{x})^2 \sum_{i=1}^N (y_i - \bar{y})^2} = b^2 \times \frac{\sum_{i=1}^N (x_i - \bar{x})^2}{\sum_{i=1}^N (y_i - \bar{y})^2}. \quad (14.4)$$

Since $a = \bar{y} - b\bar{x}$, it is also true that

$$S^2 = \sum_{i=1}^N (y_i - a - bx_i)^2 = \sum_{i=1}^N (y_i - \bar{y})^2 - b^2 \sum_{i=1}^N (x_i - \bar{x})^2. \quad (14.5)$$

Using (14.4) and (14.5) leads to

$$\frac{S^2}{\sum_{i=1}^N (y_i - \bar{y})^2} = 1 - r^2$$

or, alternatively,

$$\frac{r}{\sqrt{1 - r^2}} = \frac{b \sqrt{\sum_{i=1}^N (x_i - \bar{x})^2}}{S}. \quad (14.6)$$

Equation (14.6) provides the means to determine the distribution function of $r/\sqrt{1-r^2}$. First, notice that the variance of the linear regression parameter b is given by

$$\sigma_b^2 = \frac{N\sigma^2}{N \sum_{i=1}^N x_i^2 - \left(\sum_{i=1}^N x_i \right)^2} = \frac{\sigma^2}{\sum_{i=1}^N (x_i - \bar{x})^2},$$

as shown in Sect. 11.5, where σ^2 is the common data variance. Assuming that the true parameter for the slope of the distribution is β , then

$$\frac{b - \beta}{\sigma_b} = \frac{b - \beta}{\sqrt{\frac{N\sigma^2}{\sum_{i=1}^N (x_i - \bar{x})^2}}} \sim N(0, 1)$$

is therefore distributed like a standard Gaussian. Replacing the unknown value of σ^2 with its unbiased estimator (i.e., the model sample variance $\hat{s}^2 = S^2/(N - 2)$) leads to a statistic

$$\frac{b - \beta}{\hat{\sigma}_b} = \frac{(b - \beta)/\sigma_b}{\hat{\sigma}_b/\sigma_b}$$

where $\hat{\sigma}_b/\sigma_b = (S^2/\sigma^2)/(N - 2)$. Since $S^2/\sigma^2 \sim \chi^2(N - 2)$ under the hypothesis that the linear model applies, and enforcing the null hypothesis that the variables X and Y are uncorrelated ($\beta = 0$), one obtains a new variable that is distributed like a t distribution with $f = N - 2$ degrees of freedom, namely

$$\frac{b}{\hat{\sigma}_b} = \frac{b \sqrt{f \cdot \sum_{i=1}^N (x_i - \bar{x})^2}}{S} \sim t(N - 2).$$

Equation (14.6) can now be used to conclude that the following function of r is also t -distributed,

$$v(r) = \frac{r \sqrt{f}}{\sqrt{1 - r^2}} \sim t(N - 2). \quad (14.7)$$

The statistic $v(r)$ is a monotonic function of r , and therefore the distribution of r itself is obtained via a simple change of variables, following the method described in Sect. 4.5.1. Starting with

$$f_T(v) = \frac{1}{\sqrt{\pi f}} \frac{\Gamma((f + 1)/2)}{\Gamma(f/2)} \left(1 - \frac{v^2}{f}\right)^{-\frac{f+1}{2}}$$

and using

$$\frac{dv}{dr} = \frac{\sqrt{f}}{(1 - r^2)^{3/2}},$$

the equation of change of variables $f_r(r) = f_T(v)dv/dr$ yields (14.3) with a few steps of algebra.

The shape of the r distribution changes according to the number of degrees of freedom, as illustrated in Fig. 14.2. A dataset with just $N = 2$ datapoints has zero degrees of freedom for the calculation of the two slopes b and b' , and the correlation coefficient will always be such that $r^2 = 1$. The smallest dataset for which a meaningful r coefficient can be calculated is for $N = 3$ datapoints, or $f = 1$, leading to a probability distribution that is heavily weighted towards the $r = \pm 1$ values. The distribution becomes uniform for $f = 2$, and then gradually develops a sharper peak at $r = 0$ as f increases.

It is convenient to point out that the symmetric beta distribution is related to the (*standard*) beta distribution with parameters a and b , which has a probability distribution function

$$f_\beta(x) = \frac{1}{B(a, b)} x^{a-1} (1 - x)^{b-1} \quad (14.8)$$

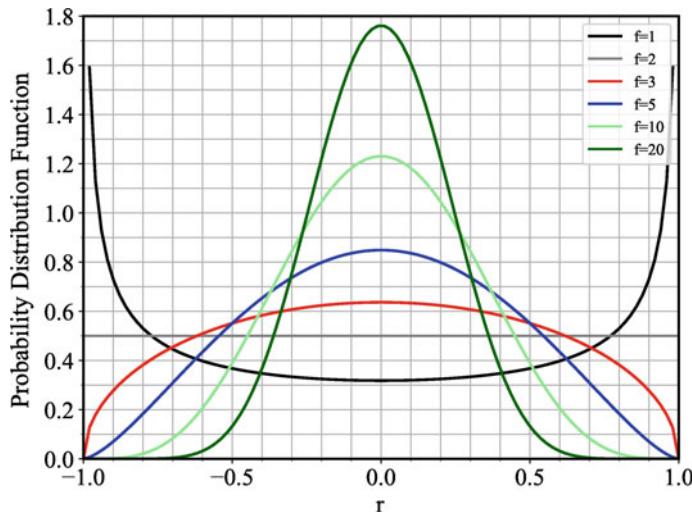


Fig. 14.2 Probability distribution functions of the linear correlation coefficient r for selected values of the number of degrees of freedom f . The probability distribution of r , under the null hypothesis of no correlation, is a symmetric beta distribution

where a and b are two parameters, and the distribution is defined for $0 \leq x \leq 1$. A variable X following the standard beta distribution can be shown to have

$$\begin{cases} E[X] = \frac{a}{a+b} \\ \text{Var}(X) = \frac{ab}{(a+b)^2(a+b+1)}. \end{cases} \quad (14.9)$$

A useful property relating the standard and symmetric beta distributions is the following: if X is a standard beta distribution with parameters $a = b = f/2$, the variable $Y = 2X - 1$ has a symmetric beta distribution of (14.3) with parameter f .¹ This property can be used to calculate the expectation and variance of the symmetric β distribution (or r -distribution) with parameter f , using the moments (14.9) of the standard beta distribution. The result is that the key moments of an r -distribution with parameter f , and therefore those of the name-sake statistic, are

$$\begin{cases} E[r] = 2E[X] - 1 = 2 \cdot \frac{f/2}{f} - 1 = 0 \\ \text{Var}(r) = 4\text{Var}(X) = \frac{1}{f+1}. \end{cases} \quad (14.10)$$

¹ Note that this property, which is proven in Appendix A.3, is used here simply as a matter of convenience, since there is no implied statistical relationship between the beta-distributed X and the r -distributed Y .

The standard and symmetric beta distributions are further described in Appendix A.3.

The distribution function for the correlation coefficient r , obtained under the null hypothesis of no correlation, can be used to test for the presence of linear relationship between the two variables. Since the null hypothesis is that there is no correlation, a two-tailed test defines the critical value r_{crit} via

$$P(|r| > r_{crit}) = 1 - \int_{-r_{crit}}^{r_{crit}} f_r(r) dr = 1 - p, \quad (14.11)$$

where p is the usual level of confidence (e.g., $p = 0.9$ or 90% confidence). Critical values of r for various probability levels are listed in Table A.24. If the measured value of r exceeds the critical value, the null hypothesis must be rejected. Given that the rejecting the null hypothesis means rejecting the absence of correlation, a value of r that exceeds the critical value can be interpreted as evidence for the presence of a linear relationship between the two quantities.

As a matter of good practice, the linear correlation coefficient test should be performed before attempting a regression between the two variables. A linear regression is meaningful only if the r coefficient is sufficiently large, for the number of degrees of freedom of the data. When the correlation coefficient is sufficiently close to zero and thus consistent with the hypothesis of no correlation, then the linear regression is not particularly meaningful, since the best-fit slope is expected to be consistent with zero.

14.4 Distribution of the Coefficient of Determination R^2 and of r^2

The coefficient of determination of a multiple linear regression between the dependent variable Y and m multiple independent variables X_k is

$$R^2 = \frac{S_e^2}{S^2} = \frac{\sum_{i=1}^N (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^N (y_i - \bar{y})^2}, \quad (14.12)$$

where S^2 is the total variance, S_e^2 the explained variance and \hat{y}_i the best-fit model with $m + 1$ adjustable parameters. It is immediately apparent that R^2 is equal to r^2 when the best-fit model is that of a simple linear regression with $m = 2$ parameters. To derive the sampling distribution of R^2 , and therefore of r^2 as a special case, it is convenient to use the statistic

$$F = \frac{S_e^2/(m)}{S_r^2/(N-m-1)}$$

that was introduced in Sect. 13.6.2, where $S_r^2 = S^2 - S_e^2$ is the residual variance. Recall that $S^2 \sim \chi^2(N - 1)$ and $S_r^2 \sim \chi^2(N - m - 1)$, since there are $m + 1$ free parameters, and their independence implies that $S_e^2 \sim \chi^2(m)$ (see Sect. 13.6.2). The F statistic is, by construction, distributed as an F variable with $m, N - m - 1$ degrees of freedom, under the null hypothesis that there is no linear correlation between Y and the independent variables. With a few steps of algebra, the R^2 statistic can be written as a function of F ,

$$R^2 = \frac{m \cdot F}{(N - m - 1) + m \cdot F}.$$

This expression makes it possible to use a known property of the F distribution with ν_1 and ν_2 degrees of freedom, which ensures that $\nu_1 F / (\nu_2 + \nu_1 F)$ is distributed like a standard beta distribution with parameters $a = \nu_1/2$ and $b = \nu_2/2$. The use of this property leads to the following distribution for R^2 :

$$f_{R^2}(R^2) = \frac{1}{B(a, b)} (R^2)^{a-1} (1 - R^2)^{b-1}, \quad (14.13)$$

which is a standard β distribution with parameters $a = m/2$ and $b = (N - m - 1)/2$. For the simple linear regression, the distribution of r^2 under the hypothesis of no correlation is given by the same formula with $m = 1$, therefore (14.13) simplifies to

$$f_{r^2}(r^2) = \frac{1}{B(1/2, f/2)} (r^2)^{-1/2} (1 - r^2)^{f/2-1}, \quad (14.14)$$

which is a standard beta distribution with parameters $a = 1/2$ and $b = f/2$, with $f = N - 2$.

Although they were derived independently, the two distributions (14.3) for r and (14.14) for r^2 must clearly be equivalent, in that testing for the absence of linear correlation using either r or r^2 must give rise to the same p -value. The equivalence can be immediately seen as follows. Since the transformation from r to r^2 is not monotonic in the $-1 \leq r \leq 1$ interval, the method of change of variables to find the distribution of r^2 from r does not apply in the entire range of the statistic. It is, however, possible to consider the distribution of r as the sum of two monotonic distributions in the $-1 \leq r \leq 0$ and in the $0 \leq r \leq 1$ ranges separately, and the method of change of variables leads to

$$f_{r^2}(r^2) = 2 \times \frac{1}{2r} f_r(r)$$

as the sum of two identical contributions from each of the two intervals. The distribution of r^2 is, therefore, related to that of r by $f_{r^2}(r^2) = f_r(r) r^{-1}$, which is in fact the relationship between the two distributions (14.3) and (14.14).

This result can alternatively be established using the methods of Sect. 4.5.2, whereby the transformation of variables $Y = R^2$ is such that

$$P(Y > y) = P(R > \sqrt{y}) + P(R < -\sqrt{y}) = 1 - F(\sqrt{y}) + F(-\sqrt{y}),$$

where F is the cumulative distribution of the r -distributed variable R . The cumulative distribution of Y is therefore

$$G(y) = P(Y \leq y) = F(\sqrt{y}) - F(-\sqrt{y}),$$

and from this the probability distribution function of $Y = R^2$ is found by derivative and simple algebraic manipulations, with

$$g(y) = \frac{dG(y)}{dy} = 2f_r(\sqrt{y}) \frac{d\sqrt{y}}{dy} = f_{r^2}(r^2).$$

The distribution of r^2 for selected values of the number of degrees of freedom is shown in Fig. 14.3. An example of the distribution of R^2 for the case of multiple linear regression with $m = 2$ is shown in Fig. 14.4. The distributions of R^2 as a function of f follow a similar pattern to those of r^2 in Fig. 14.3, with the distributions becoming more concentrated towards smaller values of R^2 as the number of data points increases. For both R^2 and r^2 , one-sided confidence intervals can be constructed in the usual way as

$$P(R^2 \geq R_{crit}^2) = 1 - \int_0^{R_{crit}^2} f_{R^2}(R^2) dR^2 = 1 - p$$

where p is the usual level of confidence, e.g., $p = 0.9$ or 90%. As the number of data points N increases, and for a fixed value of m , the critical values become smaller, as the distributions have a higher probability density towards small values of R^2 .

Example 14.3 (*Critical values of r and r^2 for Pearson's biometric data*) The linear correlation coefficient for these data is $r = 0.285$ (with $r^2 = 0.081$), as reported in Example 14.2. For $N = 1,079$ datapoints, $m = 1$ and $f = N - m - 1 = 1077$ degrees of freedom, the critical value for r at 99% confidence is 0.0784, and therefore the hypothesis of no correlation between the two quantities must be discarded at >99% confidence, since the measured value of r well exceeds the critical value. As a result, the conclusion is that the two quantities are likely to be truly correlated. Testing the hypothesis of correlation between the two variables could be equivalently carried out with a one-sided confidence interval for r^2 . Using the standard β distribution $B(1/2, 1077/2)$ according to Equation 14.13, the one-sided critical value at 99% confidence is $r_{crit}^2 = 0.00614$, whose square root is in fact identical to the two-sided critical value obtained above from the distribution of r . Hypothesis testing with r and r^2 therefore leads to the same conclusion. The p -value associated with this measurement is exceedingly small, $p < 10^{-15}$. ◇

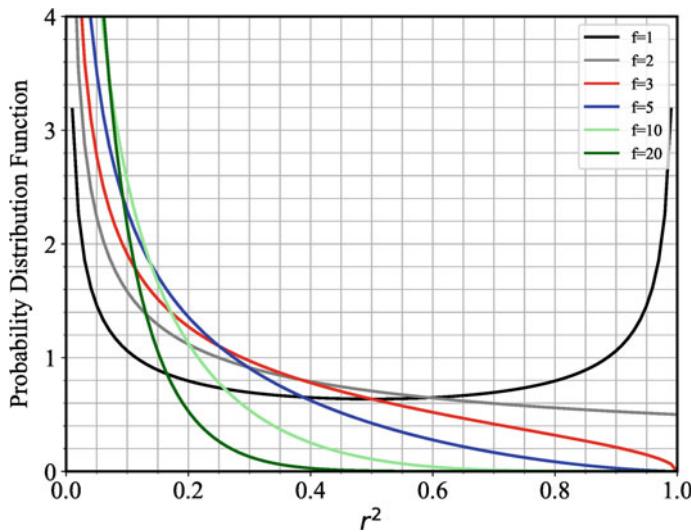


Fig. 14.3 Probability distribution functions of r^2 , for selected values of the number of degrees of freedom $f = N - m - 1$. The r^2 statistic follows a standard beta distribution

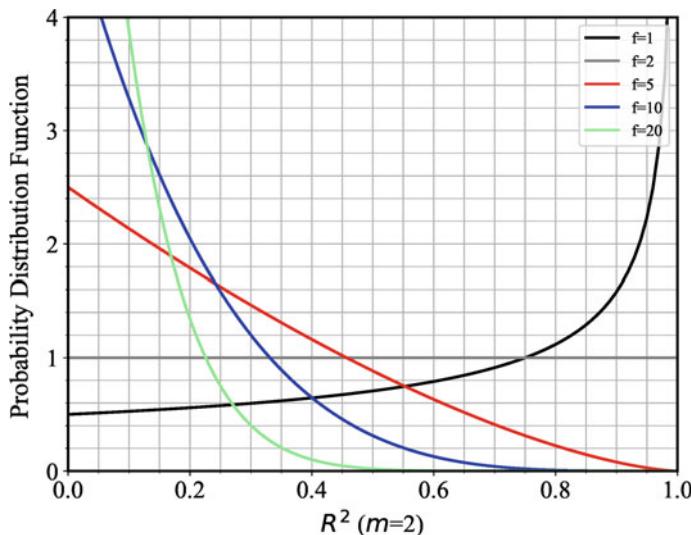


Fig. 14.4 Probability distribution functions of the coefficient of determination R^2 for the multiple linear regression, with $m = 2$. Distributions shown are for selected values of the number of degrees of freedom, $f = N - m - 1$. The R^2 statistic, just like r^2 , follows a standard beta distribution

In the case of the simple linear regression with just one independent variable, $y = a + bx$, the coefficient of determination R^2 is equal to r^2 . In this case, it is possible to test the significance of the linear model using either the correlation coefficient r or the F statistic (13.21) with $m = 1$. The two tests are equivalent.

Example 14.4 (*Linear fit to the iris setosa data using a single independent variable*) Example 13.4 showed that the coefficients of multiple regression for the variables petal length and petal width were not statistically significant, according to the t test. Excluding these two columns of data, a fit to the function $y = a + bx$, where Y is the sepal length and X the sepal width, yields best-fit value $a = 2.64$ and $b = 0.69$, a correlation coefficient of $r = 0.7425$, and a value of

$$F = \frac{S_e^2/1}{S_r^2/(N-1)} = 58.99$$

for 1 and 49 degrees of freedom. The p value associated with the r or r^2 statistics is $\sim 7 \times 10^{-10}$, and it is the same as the p value for the F statistic. These small p -values indicate that the hypothesis of no correlation between X and Y must be rejected. The value of $r^2 = 0.551$ is nearly identical to the value of $R^2 = 0.575$ obtained from the full fit using $m = 3$ independent variables. The fact that the reduction in the fraction of explained variance is minimal between the $m = 3$ and the $m = 1$ cases is an indication that the sepal length can be predicted with nearly the same precision using just the sepal width as an indicator, instead of using also the petal length and width. ◇

Summary of Key Concepts for this Chapter

Linear correlation coefficient: The linear correlation coefficient is related to the slopes of the regression of Y on X and of X on Y ,

$$r^2 = b b'.$$

Under the null hypothesis of no correlation between X and Y , the expectation of r is zero.

Distribution of r : Under the null hypothesis of no correlation between X and Y , the linear correlation coefficient r is distributed like a *symmetric β distribution* with parameter f , where $f = N - 1$ is the number of degrees of freedom. The distribution has a mean of zero and a variance that decreases as the number of degrees of freedom f increases.

Distribution of R^2 : The coefficient of multiple determination R^2 is distributed like a *standard β distribution* with parameters $a = m/2$ and $b = f/2$, where $f = N - m - 1$ and m is the number of independent variables in the regression.

Distribution of r^2 : The distribution of r^2 is obtained as a special case of the R^2 distribution, for $m = 1$. Hypothesis testing with r or r^2 is equivalent.

Problems

14.1 ■ Consider the biometric data in Pearson's experiment (Sect. 2.7). Calculate the *average* father height (X variable) for each value of the mother's height (Y variable), and the *average* mother height for each value of the father's height. Using these two averaged datasets, perform a linear regression of Y on X , where Y is the average value you have calculated, and, similarly, the linear regression of X on Y . Calculate the best-fit parameters a, b (regression of Y on X) and a', b' (regression of X on Y), assuming that each datapoint in your two sets has the same uncertainty. This problem is an alternative method to perform the linear regressions of Fig. 14.1, and it yields similar results to the case of a fit to the “raw” data, i.e., without averaging.

14.2 ■ Consider the biometric data in Pearson's experiment (Sect. 2.7).

- (a) Calculate the linear correlation coefficient between the father's height and the mother's height from the 1,079 measurements.
- (b) Determine whether there is a statistically significant linear relationship between the two quantities.

14.3 ■ Calculate the linear correlation coefficient for the data of Hubble's experiment in Sect. 11.6 (logarithm of velocity, and magnitude m). Determine whether the hypothesis of uncorrelation between the two quantities can be rejected at the 99% confidence level.

14.4 ■ Consider the fit of the *iris setosa* data using the function $y = a + bx$, where Y is the sepal length and X the sepal width. For this fit, ignore the data associated with the petal.

- (a) Determine the best-fit parameters of the linear model and their errors.
- (b) Determine whether there is a statistically significant correlation between the two quantities. Use a confidence level of 99% to draw your conclusions.

14.5 ■ Consider the *iris versicolor* data of Fig. 13.1, and the sepal length as the dependent Y variable.

- (a) Calculate the coefficient of determination when the remaining three variables (sepal width, petal length and petal width) are used as independent variables, and its p -value.
- (b) Calculate the coefficient of determination when only the petal length is used as the independent variable in the simple linear regression, and its p -value.
- (c) By comparison of the results in (a) and (b), discuss whether use of the full linear model in (a) is warranted, or whether the simpler model in (b) is sufficient to model the data variance.

14.6 A multi-variable linear regression with $m = 5$ predictor variables and $N = 20$ datapoints results in a coefficient of determination of $R^2 = 0.5$. Determine whether the null hypothesis of no correlation between the Y variable and the X_k independent variables can be rejected at the 90% confidence level.

Chapter 15

Low-Count Poisson Data and the *Cash* Statistic



Abstract Counting experiments often result in integer-valued measurements that follow a Poisson distribution. As shown in Chap. 3, a Poisson distribution is well approximated by a Gaussian distribution when the mean is sufficiently large. In the large-count limit it is therefore convenient to approximate a Poisson distribution with a normal distribution, thus enabling the use of χ^2 as a fit statistic (Chap. 11). For low-count data, however, the approximation of a Poisson distribution with a Gaussian is poor, and one should not use χ^2 to test the fit of low-count data. This chapter introduces the *Cash* or C statistic as the appropriate statistic for the likelihood of count data, both in the large-count and the low-count regimes. It is shown that the C statistic is asymptotically distributed like a χ^2 distribution in the large-count limit, and its properties in the low-count limit are also presented.

15.1 Poisson Data with Integer-Valued Variables

Experiments often result in the collection of a set of integer-valued numbers, each representing the value of a random variable Y as a function of an independent variable X . For example, one may collect the number of photons from a source as a function of time, or the number of ballots cast for a given candidate as a function of the number of eligible voters in that precinct. The independent variable X is assumed to be known exactly, while the dependent variable Y is naturally modeled with a Poisson distribution. The data are therefore of the type (x_i, y_i) , where $y_i \sim \text{Poisson}(\mu_i)$ with the mean μ_i representing the expected number of counts for that value of the independent variable. The Poisson distribution is the natural model for an integer-valued variable y_i , since it represents the probability of occurrence of a number of events, under the hypothesis that the rate of occurrence is μ_i (see Sect. 3.3.3).

The analysis of two-dimensional Poisson data starts with a parametric model $y(x)$ that describes the relationship between the two variables, the same as in the case of Gaussian data. The model is used to calculate the parent mean of each Poisson

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-981-19-0365-6_15.

distribution as $\mu_i = y(x_i)$; for example, a linear model has $\mu_i = a + bx_i$. Compared to the case of a Gaussian dataset described in Chap. 11, Poisson data are inherently simpler since the Poisson distribution is a one-parameter distribution with its variance equal to the mean. In principle, one may model an integer-valued variable with more complex distribution functions that provide flexibility in the relationship between the mean and the variance, such as the Conway–Maxwell–Poisson distribution [19, 88, 91]. The use of more complex distributions to model over-dispersed and under-dispersed data, defined respectively as data with variance smaller or larger than that of the Poisson distribution, comes however at the expense of theoretical and numerical simplicity, and it will not be treated in this book.

15.2 Likelihood of Poisson Data and the Cash Statistic

The likelihood of N independent Poisson measurements with a parent model $y(x)$ is calculated using the Poisson distribution as

$$\mathcal{L} = \prod_{i=1}^N \frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!}, \quad (15.1)$$

where $\mu_i = y(x_i)$. In place of the likelihood itself, it is convenient to work with a function of the logarithm of the likelihood, namely

$$-2 \ln \mathcal{L} = 2 \sum_{i=1}^N (\mu_i - y_i \ln \mu_i + \ln y_i!). \quad (15.2)$$

The reason for the negative sign and the factor of two will become apparent when seeking to characterize its distribution. Since the logarithm is a monotonic function of its argument, minimization of this function is equivalent to the maximization of the likelihood. Further, it is also convenient to modify this function by adding and removing certain terms that are only the function of the data. Accordingly, the C statistic is defined as

$$C \equiv -2 \ln \mathcal{L} - B = 2 \sum_{i=1}^N (\mu_i - y_i + y_i \ln(y_i/\mu_i)), \quad (15.3)$$

where

$$B = 2 \sum_{i=1}^N (y_i - y_i \ln y_i + \ln y_i!)$$

is a quantity that is independent of the model. Since the terms in B are independent of the model, minimization of the statistic C with respect to the variable model

parameters of the function $y(x)$ remains equivalent to the maximization of the original Poisson likelihood of Eq. 15.1. The statistic C defined in (15.3) is known as the *Cash* or C statistic, and it was originally proposed by W. Cash in [18] to analyze X-ray observations of astronomical sources. An equivalent statistic was also studied by S. Baker and R.D. Cousins [6].

The C statistic is approximately equal to the χ^2 statistic in the limit of large Poisson means. In fact, it is possible to show that

$$C \simeq \sum_{i=1}^N \frac{(y_i - \mu_i)^2}{\mu_i} \cdot \frac{\mu_i}{y_i}, \quad (15.4)$$

where each term in the sum differs from a $\chi^2(1)$ term by a factor μ_i/y_i that is of order unity, under the hypothesis that the data follow the model.

The asymptotic behavior of the C statistic can be obtained by ignoring terms of the third order in the Taylor series expansion of the term

$$\ln \left(\frac{\mu_i}{y_i} \right) = \ln \left(1 - \frac{d_i}{y_i} \right) \simeq -\frac{d}{y_i} - \frac{1}{2} \left(\frac{d_i}{y_i} \right)^2, \quad (15.5)$$

where $d_i = y_i - \mu_i$ is the deviation of observed counts from the parent model. This approximation is accurate only in the large-count limit, where it is expected that $d_i \ll \mu_i$, according to the Poisson distribution for y_i . From (15.5) it follows that

$$\begin{aligned} -2 \ln \mathcal{L} &= 2 \sum_{i=1}^N \left(\mu_i - y_i \ln \left(1 - \frac{d}{y_i} \right) - y_i \ln y_i + \ln y_i! \right) \\ &\simeq 2 \sum_{i=1}^N \left(\mu_i - y_i \left(-\frac{d}{y_i} - \frac{1}{2} \left(\frac{d}{y_i} \right)^2 \right) - y_i \ln y_i + \ln y_i! \right) \\ &= 2 \sum_{i=1}^N \left(\mu_i + (y_i - \mu_i) + \frac{1}{2} \frac{(y_i - \mu_i)^2}{y_i} - y_i \ln y_i + \ln y_i! \right). \end{aligned}$$

With this, the C statistic is approximated as

$$C \simeq \sum_{i=1}^N \frac{(y_i - \mu_i)^2}{y_i} = \sum_{i=1}^N \frac{(y_i - \mu_i)^2}{\mu_i} \cdot \frac{\mu_i}{y_i}$$

showing that Eq. 15.4 applies, with $\mu_i = \sigma_i^2$ the variance of the parent Poisson distribution. An equivalent approximation in the $d_i \ll \mu_i$ limit is

$$C \simeq \sum_{i=1}^N \frac{(y_i - \mu_i)^2}{\mu_i} \left(1 - \frac{d_i}{\mu_i}\right).$$

Each left-hand factor in the sum of Eq. 15.4 is distributed as a χ^2 variable with one degree of freedom,

$$\frac{(y_i - \mu_i)^2}{\mu_i} \sim \chi^2(1)$$

since $\mu_i = \sigma_i^2$ is the parent variance. Notice however that this is true only when the model $y(x_i) = \mu_i$ is fully specified, and it does not apply to the general case of models with free parameters that are estimated from the data. The additional factor in Eq. 15.4 can be used to estimate the accuracy of the C statistic with a χ^2 distribution, as a function of the mean of the parent Poisson distribution. According to the parent Poisson distribution, one can approximate the characteristic deviation from the parent means with the standard deviation of the distribution, $|d| \simeq \sqrt{y_i}$. With this approximation, each term in Eq. 15.4 differs from a $\chi^2(1)$ distribution by a factor

$$\frac{\mu_i}{y_i} = \left(1 - \frac{d_i}{y_i}\right) \simeq \left(1 \pm \frac{1}{\sqrt{y_i}}\right).$$

This term is asymptotically equal to unity in the large-count limit. The term remains significant when y_i is small, compared to the expectation of the other factor in Eq. 15.4,

$$E \left[\frac{(y_i - \mu_i)^2}{\mu_i} \right] = 1.$$

Even for $y_i = 10$, ignoring the μ_i/y_i factor leads to an error of approximately 30% for each term in the statistic. The approximation of the C statistic with a χ^2 distribution is therefore accurate only when the number of counts is significantly larger than 10. A more quantitative assessment of the approximation of the C statistic with a χ^2 distribution is provided in the following sections.

15.3 Distribution of the Cash Statistic for a Fully Specified Model

The C statistic can be summarized as the sum of N terms

$$C = \sum_{i=1}^N C_i,$$

where

$$C_i = 2(\mu_i - y_i + y_i \ln(y_i/\mu_i)). \quad (15.6)$$

When the model $y(x)$ is fully specified, the parent means μ_i are known and they do not require to be estimated from the data. In this case, each term C_i is independent from the others, and the null hypothesis that the data follow the model implies that all terms C_i are identically distributed. As a result, the probability distribution function of C can be studied as the sum of N independent and identically distributed terms C_i . A fully specified model is not a common occurrence in data analysis, since usually models have adjustable parameters that need to be estimated from the data. It is, however, necessary to begin the study of the C statistic with this simple case, similar to the case of a χ^2 distribution with fixed means (see Sect. 9.3). The N independent terms C_i contribute to the parent distribution of C in such a way that

$$\begin{cases} E[C] = \sum_{i=1}^N E[C_i] \\ \text{Var}(C) = \sum_{i=1}^N \text{Var}(C_i). \end{cases} \quad (15.7)$$

Equation 15.7 assumes the independence of the C_i , and therefore it applies to any value of N .

Unlike the case of the χ^2 distribution, an analytical form for the distribution of the C statistic is not known. The main hurdle towards an analytical evaluation of the properties of C_i is the term $y_i \ln y_i$. Since $y_i \sim \text{Poisson}(\mu_i)$ is a discrete variable, the simple method of change of variables for continuous variables described in Sect. 4.5.1 is not applicable to find the distribution of $y_i \ln y_i$. Even the mean and variance of C_i cannot be evaluated analytically. Fortunately it is possible to find simple approximations for the mean and variance as a function of the parent mean μ_i , including asymptotic values for a large value of the parent mean.

15.3.1 Asymptotic Values for the Mean and Variance

It is possible to show that the asymptotic values of the mean and variance of each term C_i are the same as those of a $\chi^2(1)$ distribution

$$\begin{cases} \lim_{\mu \rightarrow \infty} E[C_i] = 1 \\ \lim_{\mu \rightarrow \infty} \text{Var}(C_i) = 2. \end{cases} \quad (15.8)$$

This result implies that for N independent data points, the mean and variance of the C statistic are therefore

$$\begin{cases} \lim_{\mu \rightarrow \infty} E[C] = N \\ \lim_{\mu \rightarrow \infty} \text{Var}(C) = 2N \end{cases} \quad (15.9)$$

which are the same as those for a $\chi^2(N)$ distribution with N degrees of freedom. This result is expected since the C statistic is asymptotically distributed as a χ^2 distribution, as shown above.

The expectation of the random variable $y_i \ln y_i$ is calculated according to

$$\mathbb{E}[y_i \ln y_i] = \sum_{k=0}^{\infty} k \ln k e^{-\mu_i} \frac{\mu_i^k}{k!}$$

for which a simple analytical solution cannot be found. It is, however, possible to show that

$$\lim_{\mu_i \rightarrow \infty} \mathbb{E}[y_i \ln y_i] = \mu_i \ln \mu_i + \frac{1}{2}$$

From this, it is a simple exercise to obtain the asymptotic limit of $\mathbb{E}[C_i]$.

For the variance, one needs to evaluate the expectation

$$\mathbb{E}[C_i^2] = 2 \sum_{k=0}^{\infty} \left(\mu_i - k + k \ln \frac{k}{\mu_i} \right)^2 \frac{\mu_i^k e^{-\mu_i}}{k!}$$

which again does not have a simple analytical solution. One can show the asymptotic limit by performing a numerical evaluation of the series. A complete proof Eq. 15.8 is presented in [10].

15.3.2 Analytical Approximations for the Mean and Variance

The mean and variance of C_i can be evaluated numerically for any value of the parent Poisson mean. It is convenient to provide analytical approximations using the following heuristic formulas from [10]:

$$\begin{cases} \mathbb{E}[C_i] = (A + B\mu + C(\mu - D)^2)e^{-\alpha\mu} + Ee^{-\beta\mu} + F \\ \text{Var}(C_i) = (A + B\mu^2 + C(\mu - D)^2)e^{-\alpha\mu} + (E + F\mu + G(\mu - H)^2)e^{-\beta\mu} + I. \end{cases} \quad (15.10)$$

These formulas are a convenient analytical approximation of the mean and variance of C_i , and the best-fit parameters were obtained from a fit of these function to the expectation and variance evaluated numerically via the respective series (see Sect. 15.3.1). The parameters of these functions are reported in Table 15.1. Two sets of parameters are provided: one that describes the functions over the range of the mean between 0.01 and 100, as shown in Fig. 15.1; and another set that describes the functions between the narrower range between 0.1 and 10, with higher accuracy. The results show that the asymptotic limits are obtained for $\mu_i \gg 10$. For smaller values of the parent mean, the deviation from the asymptotic values are significant for both

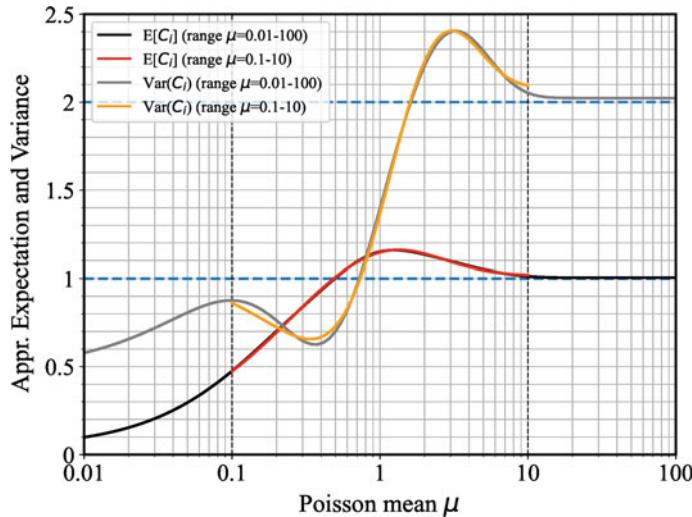


Fig. 15.1 Approximations for the mean and variance of each term C_i of the C statistic, from [10]. The approximations are obtained via a fit to empirical functions to numerical evaluations of the mean and variance

Table 15.1 Parameters for the functions of Eq. 15.10

A	B	C	D	E	F	G×10 ³	H	I	α	β
Parameters for $E[C_i]$ in range $\mu=0.01-100$										
0.065672	-6.9461	-8.0124	0.40165	0.261037	1.00512	-	-	-	5.5178	0.34817
Parameters for $E[C_i]$ in range $\mu=0.1-10$										
-0.56709	-2.7336	-2.3603	0.52816	0.33133	1.0174	-	-	-	3.9375	0.48446
Parameters for $Var(C_i)$ in range $\mu=0.01-100$										
-2.4637	1.5109	-1.5109	0.60509	1.4761	18.358	0.87316	-0.08592	2.02343	0.62652	7.8187
Parameters for $Var(C_i)$ in range $\mu=0.1-10$										
-3.1971	1.5118	-1.5118	0.79384	1.9294	6.1740	22.360	-7.2981	2.08378	0.750315	4.49654

the mean and the variance. In particular, both the mean and the variance become substantially smaller than those of the asymptotic $\chi^2(1)$ distribution, represented by dashed lines. These approximations are useful for the purpose of hypothesis testing in the low-count regime, when the C statistic is significantly different from a χ^2 distribution.

15.3.3 Other Useful Formulas for the Moments

The main hurdle towards an analytical expression for the distribution and moments of the C statistic is provided by the term $y_i \ln y_i$ in Eq. 15.3. With $y_i \sim \text{Poisson}(\mu_i)$,

even its expectation cannot be computed exactly and analytically as a function of μ_i (see Sects. 15.3.1 and 15.3.2). It is possible, however, to evaluate these moments numerically and to define two functions that describe their behavior as a function of the Poisson mean.

The expectation and variance of $y_i \ln y_i$ can be calculated according to

$$\begin{cases} E[y_i \ln y_i] = \sum_{k=0}^{\infty} k \ln k e^{-\mu_i} \frac{\mu_i^k}{k!} \\ E[(y_i \ln y_i)^2] = \sum_{k=0}^{\infty} (k \ln k)^2 e^{-\mu_i} \frac{\mu_i^k}{k!} \\ \text{Var}(y_i \ln y_i) = E[(y_i \ln y_i)^2] - E[y_i \ln y_i]^2. \end{cases}$$

These series can be evaluated numerically as a function of the Poisson mean. It is possible to show that, similar to the case of the mean, the variance of $y_i \ln y_i$ has a simple asymptotic limit

$$\lim_{\mu_i \rightarrow \infty} \text{Var}(y_i \ln y_i) = \mu_i(1 + \ln \mu_i)^2 + \frac{1}{2}.$$

It may therefore be convenient to define the functions

$$\begin{cases} e(\mu) = E[y_i \ln y_i] - \mu \ln \mu - \frac{1}{2} \\ v(\mu) = \text{Var}(y_i \ln y_i) - \mu(1 + \ln \mu)^2 - \frac{1}{2}, \end{cases} \quad (15.11)$$

where $e(\mu)$ represents the difference between the expectation of $y_i \ln y_i$ and its asymptotic value, and likewise $v(\mu)$ is the difference between the variance and its asymptotic value. These functions can be used to quantify the difference between the mean and variance of $y_i \ln y_i$ and their asymptotic values.

15.4 Hypothesis Testing with the *C* Statistic

The *C* statistic defined in Eq. 15.3 assumes that the Poisson measurements follow a fully specified $y(x)$ model that represents the null hypothesis. This section provides a method for hypothesis testing that applies to any number of counts and any number of data points, continuing with the assumption of a fully specified model. The treatment of adjustable parameters with the *C* statistic is deferred to Chap. 16. Hypothesis testing is more complex than for the case of Gaussian data since the *C* statistic does not have a simple analytical formula for its distribution function, unlike

the χ^2 distribution for Gaussian data. Since the properties of the C statistic depend on the number of data points N and on the number of counts in each measurement, it is necessary to consider separately the following regimes:

(a) *Large-count regime*, when all Poisson measurements have approximately more than 20 counts, and for any value of N . In the large-count limit, and for any number of data points N , the C statistic follows approximately the χ^2 distribution, as shown in Chap. 15. It is a subject of debate to quantify the exact number of counts that are sufficient to ensure the accuracy of this approximation. Some analysts would choose ≥ 20 counts as a good rule-of-thumb, given the good approximation of a Poisson distribution with mean $\mu = 20$ with a Gaussian of same mean and variance, as shown in Sect. 3.4. Certainly, it is not appropriate to use the Gaussian approximation when $\mu \leq 10$, given that the error in the approximation of each C statistic term is large, as shown in Sect. 15.2. In this case, the critical values of the C statistic can be approximated with the corresponding critical values of the $\chi^2(N)$ statistic.

(b) *Large- N regime*, for any number of counts. This case is partially overlapping with case (a). When the number of independent Poisson measurements N is sufficiently large, the central limit theorem ensures that the C statistic is approximately *normally distributed, regardless of the number of counts in each measurement*, as pointed out by [10, 59]. In the large- N limit, it is therefore possible to test the null hypothesis using a one-sided confidence interval of a Gaussian distribution. The critical value for the C statistic can be therefore approximated as

$$C_{crit} = E[C] + q\sqrt{\text{Var}(C)}, \quad (15.12)$$

where the parameter q takes values of, e.g., $q = 0.5, 1.3, 2.3$ for, respectively, a probability of $p = 0.68, 0.90, 0.99$ (see Table A.3). Equation 15.12 represents the one-sided critical value with probability p of a normal distribution with mean $E[C]$ and variance $\text{Var}(C)$, and it applies to any number of counts, including in the low-count limit. It is important to emphasize that this normal approximation to the C statistic applies when N is large, according to the central limit theorem, and it should not be confused with the χ^2 approximation that applies to the large-count limit, regardless of the value of N . The value of the parent Poisson mean comes into play in the calculation of $E[C_i]$ and $\text{Var}(C_i)$. For large N and in the large-count limit, Eq. 15.12 yields the same result as using a χ^2 distribution, and the tables of critical values of a χ^2 distribution (Table A.7) can be also used for the C statistic.

Example 15.1 (*Critical value of the C statistic for large N , in the large-count limit*) Consider the case of $N = 100$ Poisson measurements, all in the large-count regime with $\mu = 100$ as the parent mean. In this case, $E[C_i] = 1$ and $\text{Var}(C_i) = 2$, and the C statistic is approximated by a Gaussian distribution with mean $\mu = 100$ and variance $\sigma^2 = 200$. Equation 15.12 gives, for $p = 0.9$, a critical value of

$$C_{crit} = 118.1$$

which is in excellent agreement with the critical value of a $\chi^2(N)$ distribution which is given, according to Table A.7, by

$$\chi_{crit}^2(100) = 118.5.$$

Therefore the Gaussian approximation of the C statistic, using the asymptotic mean and variance for each term C_i , is equivalent to the use of the χ^2 distribution. ◇

When the Poisson mean is in the low-count regime, the normal approximation and Eq. 15.12 can be still applied, but with the mean and variance according to the values approximated by Eq. 15.10 and shown in Fig. 15.1. For values of the mean $\mu \ll 1$, the mean of C_i is substantially below the asymptotic value of 1, and the critical value will be correspondingly smaller than that of a χ^2 distribution with the same number of degrees of freedom. The following example shows how substantial is the difference between the accurate critical value for the C statistic in the low-count regime and the χ^2 approximation.

Example 15.2 (*Critical value of the C statistic for large N , in the low-count regime*) Consider $N = 100$ Poisson data points, all with a parent mean of $\mu = 0.1$. In this case, $E[C_i] \simeq 0.48$ and $\text{Var}(C_i) \simeq 0.86$, according to the approximations of Eq. 15.10. The critical value for $p = 0.9$ is therefore

$$C_{crit} \simeq 60$$

which is substantially smaller than in the large-count limit ($C_{crit} = 118.4$), for the same number of measurements. ◇

(c) *Low- N and low-count regime*, where neither of the approximations available for the earlier cases apply. In this case, approximations are needed for the evaluation of the critical values of the C statistic. When the number of measurements N is small, the use of the central limit theorem that leads to Eq. 15.12 becomes less accurate. In the high-count regime, this is of no consequence, since the C statistic is well approximated by χ^2 regardless of the value of N , and the critical values of χ^2 can be used with high accuracy. The case of low N and low counts therefore remains the only one where an accurate approximation for the critical value of the C statistic is not available, and it requires further discussion. Given that the model is fully specified, each term C_i in the C statistic is independent of each other, and expectations and variances of each C_i still add according to Eq. 15.7.

However, hypothesis testing in the low- N and low-count regime is complicated by the fact that Eq. 15.12 *cannot* be used because the C statistic is not well approximated by a normal distribution. In this regime, the critical values of the C statistic can be evaluated numerically as a function of the parent mean for each C_i . A simple procedure to accomplish this task starts with the simulation of several random samples from a Poisson distribution of mean μ_i . These samples are then used in Eq. 15.6 to generate samples of C_i under the null assumption that the measurements follow

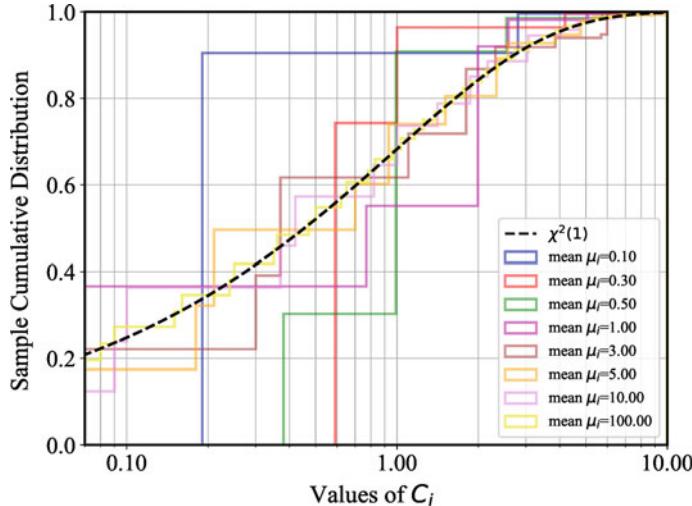


Fig. 15.2 Sample cumulative distribution of C_i for selected values of the Poisson mean. The reference statistic $\chi^2(1)$ is well approximated by C_i in the large-count limit

the parent mean, and finally calculate the quantile (i.e., the critical value) of that sample distribution. This procedure is illustrated in the following example.

Example 15.3 (*Numerical calculation of the critical value of the C statistic in the low- N and low-count regime*) The numerical evaluation of the critical value of the C statistic in the low- N regime, and for each individual C_i as a special case ($N = 1$), can be performed with standard simulation tools available in most software packages. Figure 15.2 shows the results of a numerical evaluation of the sample distributions of C_i for selected values of the parent Poisson mean, using a large number of Poisson samples ($n=100,000$ samples). The key feature of this calculation is the *discretization* of C_i , especially in the low-count regime. For example, for $\mu = 0.1$, the first four possible values of the statistic are

$$C_i = 0.2, 2.805, 8.183, 14.607.$$

These values correspond to, respectively, $n = 0, 1, 2, 3$ counts drawn from a Poisson of mean 0.1. The corresponding cumulative distribution in Fig. 15.2 has discrete jumps at those values of C_i . The critical values of C_i are reported in Table 15.2. As the Poisson mean increases, the quantiles of C_i approach those of a $\chi^2(1)$ distribution. \diamond

When $N > 1$ but still a small number, Table 15.2 is not applicable. In this case, critical values of C are obtained by drawing n sets of N Poisson number, and for each set calculate the C statistic. The resulting sample distribution of C can then be used to determine the critical values. It is to be expected that, as N increases, the discretization of the C statistic becomes less severe, and that an approximation of the

Table 15.2 Quantiles or critical values of the C_i statistic, based on a numerical calculation with $n=100,000$ Poisson samples. Exact quantiles of the reference $\chi^2(1)$ distribution are also shown

Value of parent mean	Approximate p -quantiles or critical values				
(μ_i)	$p = 0.683$	$p = 0.90$	$p = 0.95$	$p = 0.99$	$p = 0.999$
0.1	0.2	0.2	2.8	2.8	8.2
0.3	0.6	1.0	1.0	4.2	8.4
0.5	1.0	1.0	2.5	5.8	9.6
1	2.0	2.0	2.6	5.1	8.1
3	1.1	2.3	6.0	6.0	10.1
5	0.9	2.6	4.8	7.0	10.0
10	1.0	3.0	4.0	6.8	10.7
100	1.0	2.7	3.9	6.7	10.8
$\chi^2(1)$	1.0	2.7	3.8	6.6	10.8

critical values with those of a $\chi^2(N)$ becomes more accurate. In this low-count and low- N regime, tables such as Table 15.2 can be used for hypothesis testing. Critical values of the C statistic in this low- N and low-count regime are usually *smaller* than those of a χ^2 distribution with a corresponding number of degrees of freedom, since the expectation of each C_i is less than one. As a result, using the critical values of the χ^2 distribution results in a *conservative* method to reject a hypothesis. On the other hand, if the hypothesis is acceptable based on the approximate χ^2 critical values, care needs to be exercised because the conclusion might be affected by this approximation.

Summary of Key Concepts for this Chapter

C or *Cash* statistic: The test statistic for Poisson measurements obtained from $-2 \ln \mathcal{L}$ of the Poisson data with a fully specified model:

$$C = 2 \sum_{i=1}^N (\mu_i - y_i + y_i \ln(y_i/\mu_i)) = 2 \sum_{i=1}^N C_i.$$

The statistic is asymptotically distributed as $\chi^2(N)$ in the large-count limit. *Mean and variance of C_i :* The distribution of each term C_i of the C statistic cannot be evaluated analytically. The mean and variance of C_i can be evaluated numerically, also with the aid of simple analytical formulas for all values of the parent Poisson mean.

Approximation of C statistic for large N : According to the central limit theorem, the statistic is approximated as a Gaussian distribution when N is large. The mean of the distribution is the sum of the means of the N terms C_i , and the variance is the sum of the variances.

Low-count and low- N Poisson regime: In this regime, critical values of the C statistic should be calculated numerically since the χ^2 or normal approximations may not be accurate.

Problems

15.1 Show that the asymptotic limit of the expectation of each term C_i of the C statistic in (15.6) is

$$\lim_{\mu_i \rightarrow \infty} E[C_i] = 1,$$

where μ_i is the parent mean of the Poisson distribution.

15.2 Consider a dataset with N measurements drawn from a Poisson distribution with fixed mean μ .

- Use the approximate formulas (15.10) to estimate the expected value and the standard deviation of the C statistic with a parent model with constant Poisson mean $\mu = 0.2$.
- Calculate the expected value and standard deviation of a χ^2 statistic with the same number of degrees of freedom.
- Assume that the data yield a C statistic of $C = 90$ for $N = 100$. Determine whether the null hypothesis that the parent model is correct can be rejected at the 90% confidence level.

15.3 Consider a dataset with N measurements drawn from a Poisson distribution with fixed mean μ .

- Estimate the expected value and the standard deviation of the C statistic with a parent model with constant Poisson mean $\mu = 3$.
- Discuss whether for this value of the Poisson mean it is possible to approximate the C statistic with the χ^2 distribution for normally distributed measurements.

15.4 Consider a Poisson dataset composed of nine measurements, consisting of three measurements for each of the integers 4, 8, and 12. Assume a constant parent model with $\mu = 8$.

- Calculate the C statistic and its 90% confidence critical value.
- Approximate the same dataset with a Gaussian dataset of the same measurements and variance equal to the measurements, and calculate the χ^2 statistic and its 90% confidence critical value.
- Determine whether the null hypothesis that the constant $\mu = 8$ represents the parent model. By comparison of the results in (a) and (b) discuss whether the normal approximation is appropriate in this case.

Chapter 16

Maximum Likelihood Methods and Parameter Estimation with the *Cash* Statistic



Abstract Application of the maximum likelihood method to Poisson data follows the same procedure as in the case of Gaussian data. For the simple linear model, it is still possible to obtain semi-analytical formulas for the best-fit parameters, although not as simple as those for Gaussian data. Properties of the C_{\min} test statistic follow those of the χ^2_{\min} statistic in the large-count case, with differences between the two statistics in the low-count regime.

16.1 Maximum Likelihood Methods for Poisson Data

For models with adjustable parameters, the likelihood of integer-valued Poisson data (Eq. 15.1) can be maximized in order to obtain best-fit parameters, following the same procedures as for Gaussian data. Instead of maximizing the likelihood itself, it is convenient to minimize the C statistic instead,

$$C = 2 \sum_{i=1}^N (\mu_i - y_i + y_i \ln(y_i/\mu_i))$$

(see Eq. 15.3), where $y_i = y(x_i)$ are functions of the adjustable parameters that need to be minimized. The minimum value of the C statistic will be referred to as C_{\min} . Minimization of the C statistic for models of any complexity can be achieved numerically, similar to the case of χ^2_{\min} . Even for the simple two-parameter linear model, it is significantly more difficult to estimate the best-fit parameters, as will be shown in Sect. 16.2. In a few selected cases, a solution can be found analytically, as in the case of a simple constant model.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-981-19-0365-6_16.

Example 16.1 (*C* statistic fit to a constant model) In the case of a simple constant model, $f(x) = a$ and therefore $\mu_i = a$ for all data points, the maximum likelihood criterion requires

$$\frac{\partial C}{\partial a} = 2N - 2 \sum_{i=1}^N \frac{y_i}{a} = 0$$

which leads to the result that the best-fit parameter a is equal to the sample mean of the N measurements,

$$a = \frac{1}{N} \sum_{i=1}^N y_i.$$

The C statistic fit to a constant model therefore yields the same results as the χ^2 fit for Gaussian data with uniform variance. \diamond

Maximization of the C statistic leads to a C_{\min} statistic in which the N contributing terms C_i are now correlated to one another, unlike the case of a fully specified model. This correlation is clearly introduced by the estimates of the adjustable parameters, as clearly shown even in the case of the simple constant model shown in Example 16.1. A similar dependence among the N terms of the χ^2_{\min} statistic was resolved by the Cramér theorem discussed in Sect. 12.1, leading to the very convenient result that χ^2_{\min} is still distributed like a χ^2 variable, only with its number of degrees of freedom reduced to $N - m$, where m is the number of free parameters of the model. The Cramér theorem, however, only applies to the χ^2 distribution, and it is not applicable to the C statistic. This, in turn, means that the results obtained in Chap. 15 for a fully specified C statistic *do not* apply exactly to the C_{\min} for the fit to a model with free parameters. In the large-count case, the C statistic tends to the χ^2 distribution, and this makes it possible to continue using the C statistic for hypothesis testing and parameter estimation in much the same way as the χ^2 distribution. The low-count case, however, will require additional considerations, as explained in Sect. 16.4.

16.2 Linear Regression with Poisson Data

It is useful to start the treatment of maximum likelihood methods with Poisson data with the simple linear model that is used in many applications. In the case of Gaussian data, the linear regression had an analytical solution (see Sect. 11.3). The use of the C statistic for Poisson data, instead of χ^2 , leads to a more complex solution that is presented in this section.

16.2.1 The Standard Linear Model

For a linear model with the usual form $y(x) = a + bx$, the maximum likelihood method requires

$$\begin{cases} \frac{\partial}{\partial a} C = 0 \\ \frac{\partial}{\partial b} C = 0 \end{cases} \quad (16.1)$$

leading to the following two equations:

$$\begin{cases} N = \sum_{i=1}^N \frac{y_i}{a + bx_i} \\ \sum_{i=1}^N x_i = \sum_{i=1}^N \frac{x_i y_i}{a + bx_i}. \end{cases} \quad (16.2)$$

Equation 16.2 are a set of coupled non-linear equations that can in general be solved numerically, although it is not possible in general to ensure that a solution exists (see Problem 16.2). of the maximum likelihood method to the C statistic for the linear model therefore leads to a more complex solution than the analytical solution available for the χ^2 distribution. The lack of an analytical solution for the best-fit parameters also prevents a simple analytical solution for the covariance matrix. Confidence intervals on the free parameters can still be determined using critical values of the C statistic, as explained in Sect. 16.4.

16.2.2 A Factorized Linear Model with a Semi-Analytical Solution

A more convenient parameterization of the linear model is

$$f(x) = \lambda \cdot (1 + a \cdot (x - x_A)) \quad (16.3)$$

where x_A is a fixed reference value for the independent variable, and λ and a are the two adjustable parameters of the linear model. The factorization of this parameterization, proposed by J.D. Scargle [87], has mathematical advantages when used in the logarithm, as required for the C statistic. This is the parameterization that will be used to find a semi-analytical solution for the best-fit parameters of the factorized linear model, as described in [14], to which the reader is referred for more details.

Consider N integer-valued measurements y_i at a value x_i of the independent variable. Instead of considering x_i as a fixed and discrete number, it is possible to consider each measurement y_i as the sum of all counts in a range $x_i \pm \Delta x_i/2$ of the independent variable, which is referred to as a *bin* of the independent variable. The

data may therefore be thought of as being composed of N bins, ranging between values x_A and x_B of the independent variable so that $R = x_B - x_A$ is the continuous range of the independent variable. The total number of counts in the data is indicated as M . Using the definition (15.3), the C statistic for the linear model (16.3) can be shown to be

$$C = 2\lambda R \left(1 + \frac{a R}{2} \right) - 2M \ln \lambda - 2 \sum_{i=1}^N y_i \ln(1 + a(x_i - x_A)) + D \quad (16.4)$$

where

$$D = \left(2 \sum_{i=1}^N y_i \ln y_i - 2M - 2 \sum_{i=1}^N y_i \ln \Delta x_i \right)$$

is a model-independent term that does not have an effect in the minimization of the statistic. For these data, the maximum likelihood solution for the best-fit parameters λ and a are given by two uncoupled equations. The parameter a is first found via solution of the equation

$$F(a) = 1 + \frac{R}{2} \left(a - \frac{M}{g(a)} \right) = 0, \quad (16.5)$$

where $F(a)$ is a function of the unknown parameter a that makes use of the function $g(a)$ defined as

$$g(a) = \sum_{i=1}^N \frac{y_i (x_i - x_A)}{1 + a (x_i - x_A)}, \quad (16.6)$$

and M is the total number of counts in the data. Then, the parameter λ is obtained using the simple analytical formula

$$\lambda(a) = \frac{M}{R \left(1 + a \frac{R}{2} \right)}. \quad (16.7)$$

A complete proof of these equations is provided in [14]. They are obtained through simple steps of algebra using the parameterization of the linear model provided in Eq. 16.3, and the following derivatives:

$$\frac{\partial C}{\partial a} = 2R + aR^2 - 2 \frac{M}{\lambda} = 0,$$

and

$$\frac{\partial C}{\partial \lambda} = \lambda R^2 - 2 \sum_{i=1}^N \frac{y_i (x_i - x_A)}{1 + a (x_i - x_A)} = 0.$$

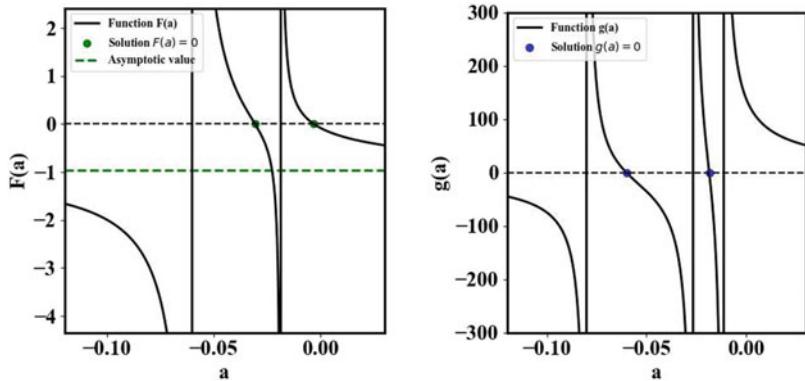


Fig. 16.1 (Left:) Function $F(a)$ for the sample dataset of Example 16.2 with $M = 3$ counts. (Right:) Function $g(a)$ for the same data set. (Reproduced from [14])

Equations 16.5 and 16.7 have the advantage of being uncoupled equations, unlike the solution of the linear regression (16.2) using the standard parameterization. Finding the zeros of the function $F(a)$, which are the possible solutions for the a parameter in the factorized linear model, does however require some care. The method to calculate the two best-fit parameters λ and a of the linear model is summarized by the following steps:

(a) Evaluate the n points of singularity for $g(a)$, where n is the number of bins with non-zero counts. According to Eq. 16.6, it is clear that there are as many points of singularity for $g(a)$ as there are non-zero y_j values. These points of singularity are simply given by

$$a_j = -\frac{1}{x_j - x_A}.$$

(b) It is possible to show that the function $g(a)$ is monotonically decreasing between its n points of singularity, and the function $F(a)$ is also monotonically decreasing between its $n - 1$ singularities. Moreover, the zeros of $g(a)$ are clearly points of singularity for $F(a)$. This leads to a characteristic pattern for the two functions $F(a)$ and $g(a)$ as shown in Fig. 16.1 for the sample data described in Example 16.2. It is therefore possible to evaluate numerically the $n - 1$ zeros of the function $g(a)$. A typical pattern for the two functions is illustrated in the example below.

Example 16.2 (*A simple dataset with $M = 3$ counts*) A sample dataset has $N = 100$ data points in the form of bins of width 1 and centered at $x_i = 0.5 + i$. Of these 100 data points, $n = 3$ bins have one count each ($y_{13} = y_{38} = y_{89} = 1$), and the other data points have zero counts. The function $F(a)$ has $n - 1 = 2$ points of discontinuity corresponding to the two zeros of $g(a)$, which were in turn found between the $n = 3$ points of singularity of $g(a)$, as shown in Fig. 16.1. The asymptotic value of $F(a)$, which is identical at both $\pm\infty$, is shown as a green dashed line. ◇

(c) It is necessary to evaluate the asymptotic value of the function $F(a)$, obtained analytically via

$$\lim_{a \rightarrow \pm\infty} F(a) = F_\infty = 1 - \frac{R}{2} \left(\frac{1}{M} \sum_{i=1}^N \frac{y_i}{x_i - x_A} \right). \quad (16.8)$$

This asymptotic limit can be either positive or negative.

(d) If the asymptotic limit is negative, then the last zero of $F(a)$ is to the right of the last point of singularity, as is the case in the example of Fig. 16.1. If the asymptotic limit is positive instead, the first zero of $F(a)$ is to the left of the first point of singularity. This zero, which can be referred to as the *external zero*, is the desired solution of the equation $F(a) = 0$. The internal zeros are those within points of singularity and they are not interesting, as shown in [14], and it is therefore not necessary to calculate them. The solution for the parameter a of the linear model is therefore the external solution of the equation $F(a) = 0$.

(e) The value of the other parameter λ is immediately calculated according to Eq. 16.7. It is necessary to point out that there are certain cases, especially with datasets with few counts, where there is no solution to Eq. 16.7. One such case is provided in the example below.

Example 16.3 (*A five-point Poisson dataset with $M = 2$ counts with no solution*) Consider a simple dataset with Poisson measurements $(0, 0, 1, 0, 1)$ taken at values $(0.5, 1.5, 2.5, 3.5, 4.5)$ of the x variable, with bin size $\Delta x = 1$. In this case, it is immediate to see that the points of singularity of $g(a)$ are at $a = -1/2.5$ and $a = -1/4.5$, with the asymptotic value of $F_\infty = 0.222$. The only zero of $F(a)$ is therefore expected to the left of the left-most point of singularity of $g(a)$, i.e., $a = -0.4$, and it is in fact found exactly at $a = -0.4$, where it causes a singularity in Eq. 16.7. This result in an infinite value of λ , which is not an acceptable solution.

16.2.3 An Extended Linear Model

The best-fit model must be positive for all values of the support of the independent variable since $\mu_i = f(x_i)$ is the parent mean of a Poisson variable and it would not be meaningful that the parent mean is negative. It is possible that, on occasion, the best-fit parameters obtained by the equations reported in this chapter lead to a non-negative model. This may be the case for certain problematic datasets with few counts and a steep distribution of counts over the support (the interested reader is referred to [14] for more details). In this case, the best-fit solution is not acceptable, and it cannot be used. Suitable alternatives are a one-parameter constant model, which is always guaranteed to provide a non-zero solution (see Example 16.1), or one-parameter linear models pivoted at each end of the support. For these problematic datasets, an *extended linear model* that is guaranteed to always have a non-negative solution is defined as follows:

(1) the standard linear model of Eq. 16.3, when such model has an acceptable non-negative solution; otherwise,

(2) the model is parameterized as one of the following three functions, namely the one that gives the smallest value of the C statistic:

(a) A one-parameter linear model *pivoted* to zero at the initial point x_A :

$$f_A(x) = \lambda_A(x - x_A), \quad (16.9)$$

for which $y_A(x_A) = 0$, and with a positive adjustable parameter $\lambda_A \geq 0$.

(b) A one-parameter linear model pivoted to zero at the final point x_B :

$$f_B(x) = \lambda_B \left(1 - \frac{x - x_A}{R} \right) \quad (16.10)$$

for which $y_B(x_B) = 0$, with an adjustable parameter $\lambda_B \geq 0$ and therefore a negative slope.

(c) A one-parameter constant model:

$$f_C(x) = \lambda_C. \quad (16.11)$$

The maximum likelihood method for the pivoted models leads to simple analytical expressions for the C statistic and the best-fit parameters.

For the linear model pivoted at x_A ,

$$C_A = \lambda_A R^2 - 2M \ln \lambda_A + D_A \quad (16.12)$$

where

$$D_A = \left(-2M + 2 \sum_{i=1}^N y_i \ln y_i - 2 \sum_{i=1}^N y_i \ln \Delta x_i - 2 \sum_{i=1}^N y_i \ln(x_i - x_A) \right) \quad (16.13)$$

and the best-fit parameter is

$$\lambda_A = \frac{2M}{R^2} > 0. \quad (16.14)$$

For the linear model pivoted at x_B ,

$$C_B = \lambda_B R - 2M \ln \lambda_B + D_B \quad (16.15)$$

with

$$D_B = \left(-2M + 2 \sum_{i=1}^N y_i \ln y_i - 2 \sum_{i=1}^N y_i \ln \Delta x_i - 2 \sum_{i=1}^N y_i \ln \left(1 - \frac{x_i - x_A}{R} \right) \right). \quad (16.16)$$

The best-fit parameter is therefore given by

$$\lambda_B = \frac{2M}{R} > 0. \quad (16.17)$$

Finally, the best-fit constant model has a C statistic of

$$C_C = 2\lambda_C R - 2M \ln \lambda_C + D_C \quad (16.18)$$

with

$$D_C = \left(-2M + 2 \sum_{i=1}^N y_i \ln y_i - 2 \sum_{i=1}^N y_i \ln \Delta x_i \right). \quad (16.19)$$

This leads to a best-fit parameter

$$\lambda_C = \frac{M}{R}, \quad (16.20)$$

which is equivalent to the sample average of the data when multiplied by a uniform bin size Δx .

Usually, the factorized linear model (16.3) provides an acceptable solution, and therefore there is no need for further analysis. Since, however, an acceptable solution cannot be ensured for all datasets, the constant and pivoted models may be convenient alternatives, although other parameterizations are also possible.

Example 16.4 (*Dataset with $M = 1$ count*) A dataset with only one count detected is a case where no solution with the factorized linear model is possible. In fact, the function $g(a)$ has a single term,

$$g(a) = \frac{x_j - x_A}{1 + a(x_j - x_A)}$$

where x_j is the location of the detected count, and accordingly

$$F(a) = 1 + \frac{R}{2} \left(a - \frac{1 + a(x_j - x_A)}{x_j - x_A} \right) = 1 - \frac{R}{2(x_j - x_A)}$$

which is no longer a function of a , and therefore cannot be used to determine the parameter. This is not surprising, since it is not meaningful to attempt to constrain a two-parameter model with a single data point. \diamond

16.2.4 Non-Uniform Bin Size and Gaps in the Data

The factorized linear model can be used with bins of any size, including non-uniform binning. In fact, in the definition of the function $g(a)$, no assumptions were made as to the size of the bins. It is also possible to use the semi-analytical solution when the data have gaps, defined as a number of intervals along the x -axis where there are no data, or the data are excluded from the fit. The non-uniform binning or the exclusion of a range of the independent variable are common situations in data analysis, and they can be accommodated with simple modifications to the equations derived earlier.

A Poisson dataset may have g non-overlapping gaps between $x_{a,j}$ and $x_{b,j}$, for $j = 1, \dots, g$, with $x_{G,j} = (x_{b,j} + x_{a,j})/2$ the mid-point of each gap and $R_{G,j} = x_{b,j} - x_{a,j}$ the length of the gap. The length of all gaps in the independent variable x is $R_G = \sum R_{G,j}$. The modifications required in this case are provided in the following. The only changes are the substitution of the range R with a modified range R_m in the definition of the function $F(a)$, and a simple modification to the function $\lambda(a)$. Similar modifications are required for the pivoted models.

When the data have gaps, the C statistic becomes

$$\begin{aligned} C = & 2\lambda R \left(1 + \frac{aR}{2} \right) - 2\lambda \sum_{j=1}^g R_{G,j} (1 + a(x_{G,j} - x_A)) \\ & - 2M \ln \lambda - 2 \sum_{i=1}^N y_i \ln(1 + a(x_i - x_A)) + D. \end{aligned} \quad (16.21)$$

Moreover, the function whose zero provides the best-fit value of a becomes

$$F(a) = 1 + a \frac{R_m}{2} - \frac{MR_m}{2g(a)} \quad (16.22)$$

where R is replaced by a modified R_m given by

$$\begin{cases} R_m = \frac{R^2 - 2S_G}{R - R_G} \\ S_G = \sum_{j=1}^g R_{G,j} (x_{G,j} - x_A), \end{cases} \quad (16.23)$$

and the best-fit solution for the parameter λ is

$$\lambda(a) = \frac{M}{R \left(1 + a \frac{R}{2} \right) - (R_G + aS_G)}. \quad (16.24)$$

These results can be proven with a few steps of algebra, and they are reported in [14]. When the data have gaps, the C statistic for the pivoted and constant models become

$$\begin{cases} C_A = \lambda_A R^2 - \lambda_A S_A^2 - 2M \log \lambda_A + D_A \\ C_B = \lambda_B R - 2\lambda_B S_B - 2M \ln \lambda_B + D_B \\ C_C = 2\lambda_C(R - R_G) - 2M \ln \lambda_C + D_C \end{cases} \quad (16.25)$$

with

$$\begin{cases} S_A^2 = \sum_{j=1}^g x_{b,j}^2 - x_{a,j}^2 \\ S_B = \sum_{j=1}^g \frac{R_{G,j}}{R} (x_B - x_{G,j}). \end{cases} \quad (16.26)$$

The best-fit model parameters then become

$$\begin{cases} \lambda_A = \frac{2M}{R^2 - 2S_G} \\ \lambda_B = \frac{2M}{R - 2S_B} \\ \lambda_C = \frac{M}{R - R_G}. \end{cases} \quad (16.27)$$

Example 16.5 (*Data with non-uniform bin sizes and a gap in the data*) The data chosen for this example span a range of the independent variable between $x_A = 0$ and $x_B = 9$, with a gap between $x_a = 3$ and $x_b = 6$. All nine measurements have a value of $y_i = 1$, with bin sizes of $\Delta x_i = 1$ for the first three data points, and $\Delta x_i = 1/2$ for the other six data points, as shown in Fig. 16.2. The data have an acceptable solution for the standard linear model (in black) with $a = 0.188$, $\lambda = 0.812$, for a best-fit statistic of $C_{cmin} = 0.078$. Given the non-uniform bin sizes, the best-fit density function $f(x)$ (black line), in units of counts-per-bin-size, has a positive slope. On the other hand, the best-fit model $y(x_i)$ (black step-wise curve) is in units of counts or counts-per-bin, and the non-uniformity of the bins leads to a non-linear step-wise bin model, although the underlying density model is linear. Given that the standard linear model has an acceptable non-negative solution, this model is expected to have a lower C statistic compared to the constant and pivoted models. The constant model has a best-fit parameter of $\lambda_C = 1.5$, according to (16.27) with $M = 9$, $R = 9$ and $R_G = 3$, for a best-fit statistic $C_C = 1.019$. The linear model pivoted at A has $\lambda_A = 0.333$ and $C_A = 2.735$, and the linear model pivoted at B has $\lambda_B = 3$, and $C_B = 14.177$.

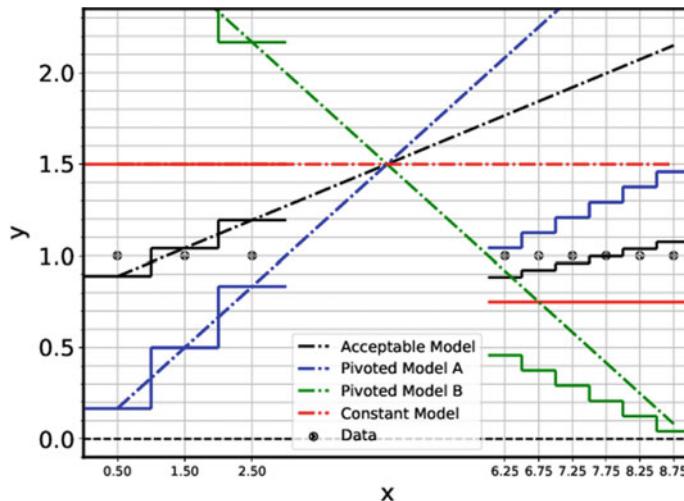


Fig. 16.2 Best-fit linear models for data with non-uniform bins and with a gap in the data. The dot-dashed curve are the density functions and the solid step-wise curves are the models $y(x_i)$ for the integer data. (Reproduced from [14])

16.3 Goodness of Fit and Hypothesis Testing with the *Cash* Statistic

Hypothesis testing with the C statistic for a fully specified model was described in Sect. 15.4. The more common situation, however, is when m adjustable parameters are estimated from the data, as in the $m = 2$ case of the linear model. For Gaussian data, the introduction of free parameters is handled with the reduction of the number of degrees of freedom of the χ^2_{\min} statistic from N to $N - m$, thanks to the Cramér theorem (see Sect. 12.1). This theorem, however, only applies to the χ^2 statistic. An extension of those considerations to the Poisson-based C statistic is provided by a key theorem by S.S. Wilks [103], which can be applied to data with virtually any distribution function, including the Poisson distribution. It is therefore necessary to study how this result applies to Poisson data and the C statistic.

16.3.1 The Wilks Theorem on the Likelihood Ratio

S.S. Wilks [103] provided an asymptotic theorem that can be used for data that follows nearly any type of distribution, including the Poisson distribution. Consider N measurements drawn from a probability distribution f , and that the data are fit to a function with m adjustable parameters. The parameters are referred to as θ_j , of which l are held fixed at their true-yet-unknown values indicated as θ_j^T , and the remaining

$m - l$ parameters are fit to the data so that the resulting likelihood is maximized. The Wilks theorem can be stated as follows, in a convenient form reported by Cash [18]:

Theorem 16.1 (The Wilks theorem) *The likelihood ratio*

$$L = \frac{\max \prod_{i=1}^N f(x_i; \theta_1^T, \dots, \theta_l^T, \theta_{l+1}, \dots, \theta_m)}{\max \prod_{i=1}^N f(x_i; \theta_1, \dots, \theta_m)}$$

is such that

$$-2 \ln L \sim \chi^2(l)$$

in the asymptotic limit of a large number of measurements N . The distribution function f can have any form, provided it meets certain conditions on the distribution of the maximum likelihood estimates of the parameters, as described in [103].

The two terms in the likelihood ratio are, respectively, the likelihood of the data with a model where only $m - l$ of the m parameters are fit to the data, and the likelihood of the data with a model where all the parameters were adjusted so as to maximize the likelihood.

In the case of a Gaussian distribution f , the logarithm of the numerator is proportional to $\chi^2(N - (m - l))$, and the logarithm of the denominator is proportional to $\chi^2(N - m)$. Since the two statistics are also independent, the Wilks theorem amounts to a simple accounting of the degrees of freedom, with $N - (m - l) - (N - m) = l$ degrees of freedom for the asymptotic χ^2 distribution. For normally distributed data, the Wilks theorem therefore provides the same result as the Cramér theorem.

The strength of the theorem, however, is that it applies not just to Gaussian likelihoods, but also to likelihoods calculated for most other distributions, including the Poisson distribution. For Poisson data, the Wilks theorem is useful in two different ways. First, when all the parameters are held fixed at their true value ($l = m$),

$$-2 \ln L = C_{\text{true}} - C_{\min} \sim \chi^2(m) \quad (16.28)$$

in the limit of a large number of measurements. The quantity C_{true} represents the C statistic evaluated for the fixed parent parameters, and therefore $C_{\text{true}} \sim \chi^2(N)$ in the large-count limit, as discussed in Chap. 15. In this limit, the previous expression therefore also ensures that, asymptotically,

$$C_{\min} \sim \chi^2(N - m), \quad (16.29)$$

which is the same result that applies to the Gaussian case. This asymptotic distribution basically extends the Cramér theorem also to the C statistic, but only in the asymptotic limits of a large number of measurements and with a large number of counts per measurement. Second, the Wilks theorem can also be used to estimate a

subset of $l < m$ interesting parameters. In this case, $-2 \ln L$ is still the difference of two C statistic as per (16.28), the first of which (C_{true}) is the C statistic evaluated by marginalizing (i.e., maximizing the likelihood) over the $m - l$ uninteresting parameters, while keeping the l interesting parameters fixed at their true values. The Wilks theorem yields

$$\Delta C = C_{\text{true}} - C_{\min} \sim \chi^2(l) \quad (16.30)$$

which requires the large- N limit, but is otherwise applicable to any number of counts, even in the low-count regime. Equation 16.30 therefore defines a ΔC *statistic* that can be used for parameter estimation in a manner similar to the $\Delta\chi^2$ statistic for normal data.

16.3.2 The Large-Count Regime

In the large-count and large- N regime, the approximation of C_{\min} with a χ^2_{\min} distribution is expected to hold according to (16.29). The large- N requirement is due to the Wilks theorem, and the large-count regime is required to ensure that $C_{\text{true}} \sim \chi^2(N)$. When these conditions are met, it is possible to minimize the parameters of the model using the C statistic, and then use critical values of the χ^2_{\min} statistic for hypothesis testing with C_{\min} . When the number of measurements N is small, the Wilks theorem becomes less accurate. Nonetheless, in the large-count regime the approximation of the C statistic with a χ^2 distribution is expected to continue to hold even when free parameters are present, according to (15.4).

The approximation of C_{\min} with a χ^2 distribution in the large-count regime is *not* equivalent to using χ^2 as a fit statistic for large-count Poisson data, although this is a common practice among data analysts. The biases introduced by this practice are discussed in Sect. 16.5. The approximation of C_{\min} with a χ^2 distribution is useful to determine the critical values of the distribution, but the analyst should use the C statistic as a fit statistic for Poisson data.

16.3.3 The Low-Count Regime

The low-count regime, regardless of the number of bins N , presents additional complications when there are free model parameters that are estimated from the data. While the Wilks theorem continues to apply in the large- N limit, thus ensuring that asymptotically

$$C_{\text{true}} - C_{\min} \sim \chi^2(m)$$

according to (16.28), in the low-count regime the C_{true} statistic is not accurately approximated by a χ^2 distribution. It is therefore not possible to use the Wilks theorem to make statements regarding the distribution of C_{\min} in the low-count regime.

Likewise, the estimation of parameters leads to a correlation among the terms C_i that was not present for a fully specified model, and it is therefore also not possible to simply use the expectations and variances for the C_i of a fully specified model (Eq. 15.10) and use them for a model with free parameters. Another consideration is that the correlation among the C_i terms prevents the application of the central limit theorem to the N terms of the C_{\min} statistic, as was done leading to Eq. 15.12 for a fully specified hypothesis. These complications preclude a simple analytical form for the distribution of C_{\min} in the low-count regime, even in the asymptotic limit of a large number of measurements. Instead, it is usually necessary to use either numerical or approximate methods to determine critical value of the statistic C_{\min} in the low-count regime.

The expectation of C_{\min} can be calculated, even in the case of a model with free parameters, as

$$E[C_{\min}] = \sum_{i=1}^N E[C_i],$$

where

$$C_i = 2(\hat{y}_i - y_i + y_i \ln(y_i/\hat{y}_i))$$

and \hat{y}_i are the values of the best-fit model, now dependent on the data. The data dependence of the estimated Poisson means \hat{y}_i makes it such that the expectations $E[C_i]$ are no longer given by (15.10), which assumed a fixed parent mean, but must be evaluated according to the best-fit model. The evaluation of $\text{Var}[C_{\min}]$ is more complex since the covariances among the C_i terms must also be accounted for. In general, critical values and other parameters of the C_{\min} statistic can be evaluated numerically, according to the model at hand. This is especially straightforward when the model $y(x)$ has simple analytical properties that make the estimates \hat{y}_i simple functions of the Poisson measurements, as in the case of the one-parameter constant model. The procedure to obtain the mean, variance and critical values of C_{\min} is illustrated in the following example.

Example 16.6 (*The distribution of C_{\min} for the one-parameter constant model*)

An example of the evaluation of the mean, variance and critical values of C_{\min} is provided by the simple one-parameter constant model $y(x) = a$. The results illustrated in this example are based on [10], which describes the use of the C statistic for the constant model. As shown in Example 16.1, the best-fit value of a is the sample mean \bar{y} of the measurements so that $\hat{y}_i = \bar{y}$. Accordingly, the best-fit statistic can be written as

$$C_{\min} = 2 \sum_{i=1}^N y_i \ln \frac{y_i}{\bar{y}} = 2 \sum_{i=1}^N y_i \ln y_i + 2M \ln N - 2M \ln M \quad (16.31)$$

where $M = \sum_{i=1}^N y_i$ is the sum of all detected counts. Equation 16.31 also shows that the N terms $y_i \ln y_i/\bar{y}$ are not independent of one another.

Moments of the C_{\min} distribution can be calculated numerically, making use of the property (2.5) for the expectation of functions of random variables. In this case, the expectation of (16.31) can be calculated by realizing that the N terms of the sum in the right-hand size are independent, and that the variable M appearing in the other terms is distributed like a Poisson with parameter $N\mu$, where μ is the parent mean of the N Poisson measurements y_i . These observations make it such the expectation of C_{\min} , under the hypothesis that the parent mean of the measurements is μ , is given by

$$\begin{aligned} E[C_{\min}] = & 2(N E[y_i \ln y_i] + N\mu \ln N - E[Ny_i \ln Ny_i]) = \\ & 2 \left(N \sum_{n=0}^{\infty} n \ln n \cdot \frac{e^{-\mu} \mu^n}{n!} + N\mu \ln N - \sum_{n=0}^{\infty} n \ln n \cdot \frac{e^{-N\mu} (N\mu)^n}{n!} \right) \end{aligned}$$

These series can be evaluated numerically as a function of the parent Poisson mean μ and of the number of measurements N . From these calculations, it can be shown that the expectation $E[C_{\min}]$ has an asymptotic limit of $N - 1$ when the parent mean is large (see Fig. 16.3 and Problem 16.3). This result is consistent with the expectation that the C_{\min} statistic converges to a $\chi^2(N - 1)$ for this one-parameter model. Moreover, the expectation of C_{\min} deviates from $N - 1$ for parent means $\mu \leq 10$ following a pattern that is qualitatively similar to that of Fig. 15.1 for the fully specified model, although with small-scale differences that are a function of the number of data points N .

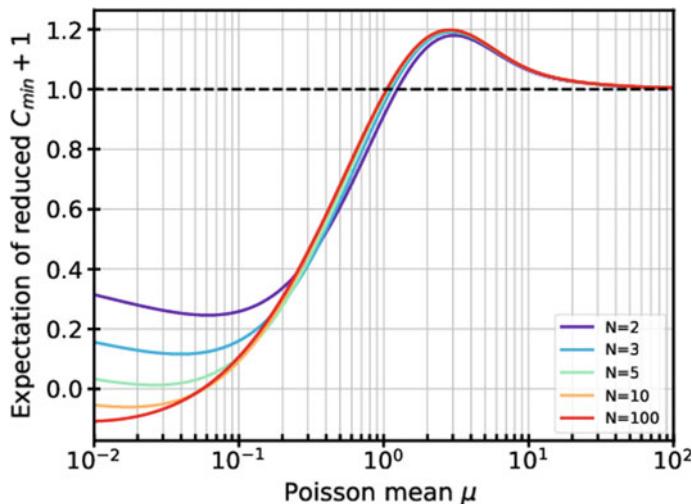


Fig. 16.3 Expectation of $C_{\min} + 1$ divided by N , for the case of the simple one-parameter constant model, as a function of the parent Poisson mean μ and the number of data points. The asymptotic expectation $N - 1$ is based on the $\chi^2(N - 1)$ distribution. For $N > 100$, the curves are virtually indistinguishable from the $N = 100$ curve (Reproduced from [10]).

The variance of C_{\min} can be evaluated numerically following similar calculations as those for the mean, using the property that $\text{Var}(X) = E[X^2] - E[X]^2$. In [10] it is shown that the asymptotic large-count limit of the variance converges to $2(N - 1)$, which is the variance of a $\chi^2(N - 1)$ distribution. As was the case for the expectation, the variance of C_{\min} differs from the asymptotic value of $2(N - 1)$ for parent means $\mu \leq 10$, following a pattern that is qualitatively similar to that of Fig. 15.1, again with small-scale differences that are a function of the number of data points N .

The overall distribution of C_{\min} and its critical values can be *simulated* by drawing random samples from N independent Poisson variables, in this case all with the same parent mean equal to \bar{y} . The N Poisson samples are then used to evaluate C_{\min} according to (16.31), and the procedure is iterated for a large number of times to obtain a sample distribution for C_{\min} . It is then straightforward to calculate critical values as the p -quantiles of the sample distribution (see Sect. 5.3). Critical values of C_{\min} obtained from the simulated sample distribution are asymptotically the same as those of a $\chi^2(N - 1)$ distribution for large μ , but they are *substantially* smaller when $\mu \ll 1$, consistent with Fig. 16.3. For parent means in the range of approximately $\mu = 1-10$, there are only small deviations from the critical values of a $\chi^2(N - 1)$ distribution.

◇

16.3.4 Approximate Methods in the Low-Count Regimes

In general, the correct procedure for hypothesis testing in the low-count regime requires the estimation of critical values of C_{\min} , following the discussion of Sect. 16.3.3 above. This procedure may require numerical calculations, given that usually a simple analytical form for the probability distribution function of C_{\min} is not available. When these calculations are unavailable, it may become necessary to use the approximate methods that are discussed in this section.

Assuming for simplicity that the C_{\min} statistic is normally distributed, one may estimate critical values according to an expression equivalent to Eq. 15.12, namely

$$C_{\min,crit} = E[C_{\min}] + q \sqrt{\text{Var}(C_{\min})} \quad (16.32)$$

where q has the usual meaning, with, e.g., $q = 1.3$ for a $p = 0.9$ critical value. This normal approximation, identical in form to that of (15.12), should become more accurate as the number of measurement increases, according to the central limit theorem and notwithstanding the correlation among terms. When estimates of $E[C_{\min}]$ and $\text{Var}(C_{\min})$ are available for the model at hand, for example by following the numerical evaluations outlined in Example 16.6, those estimates should be used. When estimates are not available, one may estimate the mean and variance according to the results for the fully specified C statistic provided in Sect. 15.3, additionally

accounting for the number of free parameters. This approximate method of hypothesis testing with the C_{\min} statistic is illustrated in the following example.

Example 16.7 (*Approximate method for hypothesis testing in the low-count regime*) Consider a sample dataset with $N = 100$ measurements, of which 50 have a measurement of 1, and the remaining 50 a measurement of 0. A fit to a one-parameter constant model yields a best-fit parameter of $a = 0.5$ for a fit statistic of $C_{\min} = 69.3$. In fact, each measurement $y_i = 0$ gives a null contribution, and each measurement $y_i = 1$ contributes $-2 \ln(0.5) = 1.386$. For a parent Poisson mean of $\mu = 0.5$, the fixed-parameter expectations are $E[C_i] \simeq 1$ and $\text{Var}(C_i) \simeq 0.7$ (see Fig. 15.1 and Eq. 15.10). The approximate critical value at 90% confidence would be estimated, according to (16.32), as

$$C_{\min,crit} = (N - 1)E[C_i] + 1.3\sqrt{(N - 1)\text{Var}(C_i)} = 110$$

The best-fit model is certainly consistent with the data at the 90% confidence level. In using the above expression for the critical value of C_{\min} the following assumptions were made: (a) the statistic is normally distributed, (b) the expectation and variance are equal to $N - m$ times the expectation and variance for the fixed-parameter C_i , where m is the number of free model parameters. As shown in Example 16.6, these approximations are reasonable for a constant model ($m = 1$), but the accuracy for more complex models is less certain. \diamond

16.4 Parameter Estimation with the C statistic

Parameter estimation with the C statistic is very similar to the case of the χ^2 statistic, thanks to the Wilks theorem. In fact, in the limit of a large number of measurements—but regardless of the number of counts in each measurement—the theorem establishes that

$$\Delta C = C_{\text{true}} - C_{\min} \sim \chi^2(l) \quad (16.33)$$

where $l \leq m$ is the number of interesting (or free) parameters in the fit, and m is the total number of available adjustable parameters. The Wilks theorem ensures that the ΔC statistic is asymptotically distributed like a χ^2 distribution, even in the low-count regime where the two constituting statistics C_{true} and C_{\min} are not χ^2 -distributed. Parameter estimation with Poisson statistic therefore conveniently follows the same rules and procedures as in the case of normally distributed data with the χ^2 statistic.

Example 16.8 (*Comparison of parameter estimation using the ΔC and $\Delta \chi^2$ statistics*) Consider an ideal dataset of N identical measurements all with values $y_i = 1$. The data are fit to a constant model $y = a$, with the goal to determine a $1-\sigma$ or $p = 0.68$ confidence interval on the free parameter. Two data models are used: one where the data follow the Poisson statistic, and another where the data are normally distributed with uniform uncertainties $\sigma_i = 1$.

In the case of Poisson data, the best-fit value of the model is clearly $a = 1$ with $C_{\min} = 0$, and it is possible to find the value δ corresponding to a change $\Delta C = 1$ according to

$$\Delta C = -2 \sum_{i=1}^N 1 \cdot \ln \frac{(1 + \delta)}{1} + 2 \sum_{i=1}^N (1 + \delta - 1) = 1$$

Using the approximation $\ln(1 + \delta) \simeq (\delta - \delta^2/2)$, this immediately leads to $\delta = \sqrt{1/N}$, and therefore the 68% confidence range can be reported as $a = 1 \pm \sqrt{1/N}$.

Using Gaussian errors instead, the best-fit model remains the same with $\chi^2_{\min} = 0$, and uncertainties in the model parameter are calculate via $\Delta\chi^2 = 1$, leading to

$$\Delta\chi^2 = \sum_{i=1}^N \left(\frac{1 - (1 + \delta)}{1} \right)^2 = 1$$

and therefore the same result as in the case of Poisson data. \diamond

A useful caveat to keep in mind, when performing parameter estimation in the low-count and/or low- N regime, is that the small number of Poisson counts detected may lead to a *discretization* of the C statistic. In this case, one may decide to perform a numerical simulation of the parent C_{\min} , and therefore estimate critical values numerically for better accuracy. An example of this discretization is provided in the following example, and additional discussion can be found in Sect. 3 of [10].

Example 16.9 (*Discretization of the C statistic in the low-count and low- N regime*) Consider $N = 10$ Poisson measurements, of which just one measurement is $y_j = 1$, and the remaining measurements have no counts. This dataset has a best-fit constant model of $a = 1/N$, and a best-fit statistic of $C_{\min} = 2 \ln N = 4.6$. According to the usual χ^2 -based criterion according to the Wilks theorem, a 68% confidence uncertainty δ_a would be estimated by requiring $\Delta C = 1$. The δ_a uncertainty is therefore estimate by solving

$$\Delta C = 2 \left(N \left(\frac{1}{N} + \delta_a \right) - 1 - \ln \left(\frac{1}{N} + \delta_a \right) - \ln N \right) = 1.$$

Since these data are in the low- N regime where the Wilks theorem may not apply, the accuracy of using $\Delta C = 1$ to obtain an estimate of the uncertainty δ could be checked by performing a numerical simulation of the distribution and therefore critical values of ΔC . This is a particularly simple task for the constant model, given the availability of an analytical best-fit model. For a parent mean $\mu = 0.1$, the Poisson probability of having measurements $y_i = 0, 1, 2, 3$ is, respectively, $P = 90.5\%$, 0.905% , 0.45% and 0.015% . Therefore, when constructing a sample distribution of ΔC via simulated samples, most datasets will have typically just a few non-zero measurements.

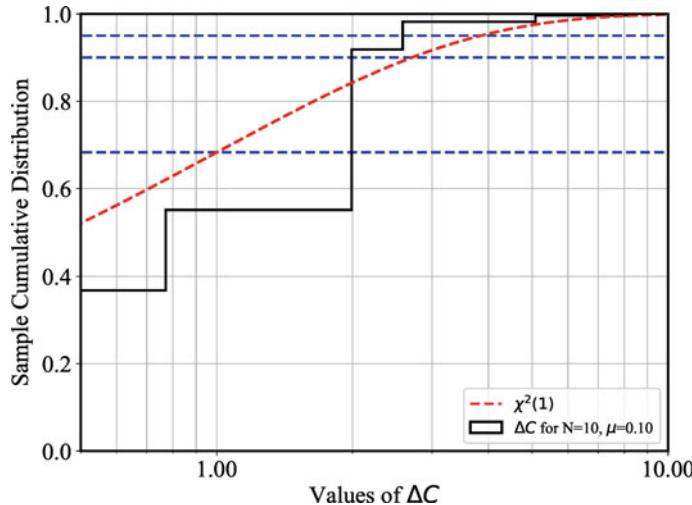


Fig. 16.4 Sample distribution of ΔC for $N = 10$ datapoints with a parent Poisson mean of $\mu = 0.1$, fit to a constant model. The sample distribution was obtained from a simulation with 10,000 Poisson samples from the parent distribution

A simple numerical simulation for the sample distribution of ΔC with $N = 10$ and $\mu = 0.1$ is reported in Fig. 16.4. The discrete nature of the distribution is not due to the sampling, but to the fact that there is only a discrete set of values for ΔC , when the product $N\mu$ is small. The most likely datasets, and the associated binomial probabilities, are reported in the table below.

Number of counts	best-fit a	C_{\min}	C_{true}	ΔC	Binomial probability
0	0	0	2	2	0.349
1	0.1	4.6	4.6	0	0.387
2	0.2	6.4	7.2	0.8	0.194
3	0.3	7.2	9.8	2.6	0.057

These values correspond to the first ‘steps’ in the sample distribution shown in Fig. 16.4, which illustrates the discrete nature of the ΔC distribution. The result of this calculation is that, for example, the 68% and 90% critical value are both estimated at $\Delta C = 2$ (instead of, respectively, 1 and 2.7). \diamond

16.5 Biases Using χ^2 for Poisson Data in the Large-Count Limit

Poisson data in the large-count limit are often fit using the χ^2 statistic. This is a common procedure that is based on the fact that the Poisson distribution is well

approximated by a normal distribution when the mean is approximately ≥ 20 counts per measurement. The maximum-likelihood method for the estimate of the best-fit parameter, however, may introduce a significant bias even in the large-count limit. The reason for this bias is the assumption that the parent variance of each Gaussian measurement is estimated by the measured variance when constructing the χ^2 statistic. This assumption was illustrated in Fig. 11.1: the parent model $y(x)$ is used to predict the parent mean of the Gaussian, $\mu_i = y(x_i)$, but the parent variance is set to the measured value (which is nonetheless indicated by the Greek letter σ_i^2). As a result, the fit statistic

$$\chi^2 = \sum_{i=1}^N \left(\frac{y_i - y(x_i)}{\sigma_i} \right)^2$$

is a weighted average of the square of the measured deviations, with weights provided by the *measured* variances. For a Poisson measurement of $y_i = n$ counts, one would use the property that the Poisson mean and variance are identical, and therefore set $\sigma_i^2 = n$. This results in low measurements having a higher weight than high measurements, as pointed out, for example, in [10, 52]. The bias resulting from approximating the Poisson variance with the measured number of count depends in general on the model being fit to the data. It is immediate to see such a bias even for a large number of counts using the simple constant model, as illustrated in the following example.

Example 16.10 (*Bias in the high-count regime using χ^2 for Poisson data*) For a simple constant model, the best-fit value is the sample mean when performing the Poisson fit using the C statistic, and it is the weighted mean when using the Gaussian approximation and χ^2 . Figure 16.5 shows the results of a numerical calculation using $N = 100$ data points drawn from a parent Poisson distribution with the mean ranging from 10 to 1,000. The C statistic fit recovers the parent mean for any value of the mean without any bias, while the χ^2 statistic is consistently biased towards *lower* values, even when $\mu \geq 20$. The bias is defined as

$$\text{bias} = \frac{\hat{a}_{\chi^2} - \mu}{\mu}$$

where \hat{a}_{χ^2} is the best-fit value from the χ^2 fit. The bias is of the order 1% even when the mean is above 100 counts.

The example above shows that the use of the χ^2_{\min} statistic to fit Poisson data results in biases that are caused by the higher weight given by the statistic to points with lower values. It is therefore good practice to use the C statistic to fit Poisson data, even in the large-count regime.

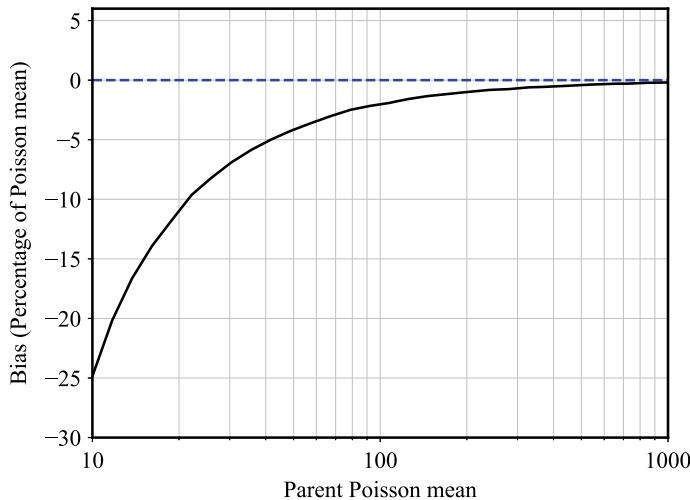


Fig. 16.5 Bias in the measurement of the best-fit value of a constant model when using χ^2 , versus the unbiased Poisson-based C statistic. Data were in the form of $N = 100$ Poisson measurements with mean in the range 10–1000. The numerical calculations used 1,000 simulated datasets for each value of the parent Poisson mean

Summary of Key Concepts for this Chapter

The Wilks theorem: The Wilks theorem states that, in the asymptotic limit of a large number of measurements, the likelihood ratio L for virtually any type of distribution is such that

$$-2 \ln L \sim \chi^2(l)$$

where l is the number of free parameters in the fit.

Goodness of fit in the large-count regime: For Poisson data in the large- N and large-count regime, the Wilks theorem extends the Cramér theorem to yield

$$C_{\min} \sim \chi^2(N - m),$$

same as for Gaussian data.

Goodness of fit in the low-count regime: For Poisson data in the low-count regime, the C_{\min} statistic cannot be accurately approximated with a χ^2 distribution. Instead, critical values can be obtained either using numerical calculations, or using approximate methods that make use of the Gaussian approximation, in the large- N limit.

Parameter estimation with the ΔC statistic: The Wilks theorem ensures that the ΔC statistic can be used in the same way as the $\Delta\chi^2$ statistic, for the purpose of parameter estimation for Poisson data. Discretization effects may lead to deviations from the χ^2 distribution in the low- N and low-count regime.

Problems

16.1 Consider the following Poisson dataset, where x is the independent variable and y the dependent Poisson variables.

x	y
0	1
1	0
2	1
3	0
4	1

- (a) Find an analytic solution for the best-fit parameters a and b of the traditional linear model, by minimizing the C statistic.
- (b) Calculate the C_{\min} statistic.

16.2 Show that minimization of the C statistic of the following Poisson data with the standard linear model leads to two equations for the best-fit parameters that do not have a solution.

x	y
0	0
1	0
2	1
3	0
4	1

16.3 Show that the expectation $E[C_{\min}]$ for the C statistic fit of a Poisson dataset of N measurements with a constant model has an asymptotic limit of $N - 1$, when the Poisson mean is sufficiently large.

16.4 Consider the following dataset where x is the independent variable and y the dependent Poisson variables:

x	y
0	1
1	1
2	1

- (a) Show that the best-fit value of the parameter a for the exponential model $y = e^{ax}$ is $\hat{a} = 0$.
- (b) Using the $\Delta C = 1$ criterion, derive a 68% confidence interval on the best-fit parameter. Discuss whether this criterion is accurate for this dataset.

16.5 Consider the same dataset as in Problem 16.4, but assume a constant fit function $y = a$. Show that the best-fit parameter according to the maximum-likelihood method

is given by $\hat{a} = 1$, and that the 68% confidence interval according to the $\Delta C = 1$ criterion is $\hat{a} \pm \sqrt{1/3}$.

16.6 Consider a five-point dataset with Poisson measurements $(0, 1, 0, 0, 1)$ taken at values $(0.5, 1.5, 2.5, 3.5, 4.5)$ of the x variable, with bin size $\Delta x = 1$. This dataset is to be fit to the factorized linear model of (16.3).

- (a) Find the points of singularity of $g(a)$.
- (b) Calculate the asymptotic value of $F(a)$ and the best-fit value of the parameter a .
- (c) Calculate the best-fit value of the parameter λ .
- (d) Calculate the value of C_{\min} .

16.7 Consider a five-point dataset with Poisson measurements $(1, 0, 1, 0, 1)$ taken at values $(0.5, 1.5, 2.5, 3.5, 4.5)$ of the x variable, with bin size $\Delta x = 1$. This dataset is to be fit to the factorized linear model of (16.3).

- (a) Find the points of singularity of $g(a)$.
- (b) Calculate the asymptotic value of $F(a)$ and the best-fit value of the parameter a .
- (c) Calculate the best-fit value of the parameter λ .
- (d) Calculate the value of C_{\min} .

16.8 Consider a Poisson dataset consisting of three measurements for each of the integers 4, 8, and 12, for a total of nine data points. Assume a constant one-parameter model $y = a$.

- (a) Determine the best-fit parameter a and its 68% confidence interval, using the $\Delta C = 1$ criterion.
- (b) Approximate the same dataset with a Gaussian dataset of same measurements and variance equal to the measurements, and determine the best-fit parameter a and its 68% confidence interval.
- (c) Discuss possible reasons for the differences between the results of (a) and (b).

Chapter 17

Systematic Errors and Intrinsic Scatter



Abstract Measurement uncertainties may arise from different sources, and certain sources of error may not be accounted for in the initial error budget. The presence of unreported sources of uncertainty may sometimes lead to a poor goodness-of-fit statistic and the rejection of the model used to fit the data. These missing sources of uncertainty may either be associated with the data themselves or with the model used to describe the data. In both cases, it is possible to account for these errors and thus ensure that the hypothesis testing is not biased by them.

17.1 What to Do When the Goodness-of-Fit Test Fails

The first step to ensure that a dataset is accurately described by a model is to test that the goodness-of-fit statistic is acceptable. For example, when the data have Gaussian errors, χ^2_{\min} can be used as the goodness-of-fit statistic. If the value of χ^2_{\min} exceeds a critical value, it is recommended that one rejects the model. At that point, the standard option is to use an alternative model and repeat the testing procedure.

There are cases when it is reasonable to try and investigate further whether the model and the dataset may still be compatible, despite the poor goodness-of-fit. The general situation when an additional effort is warranted is in the case of a model that generally follows the data without severe outliers in a clearly discernible pattern, yet the best-fit statistic (such as χ^2_{\min}) indicates that the model is not acceptable. An example of this situation is that of Fig. 12.1: the best-fit linear model follows the distribution of the data without systematic deviations, yet its high value of $\chi^2_{\min} = 60.5$ for 23 degrees of freedom cannot be formally accepted at any level of confidence.

This chapter describes two types of analysis that can be performed when the model fit is poor. The first method assumes that the model itself has a degree of uncertainty that results in an *intrinsic scatter* above and beyond the variance of the data (Sect. 17.2). The second investigates whether there are additional sources of

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-981-19-0365-6_17.

error in the data that may not have been properly accounted (Sect. 17.3). The two methods are conceptually different but result in similar modifications to the analysis.

17.2 Intrinsic Scatter and the Debiased Variance

When fitting a dataset to a model it is assumed that the data are drawn from a parent model that is described by a number of adjustable parameters. As such, it is assumed that there are true model parameters that describe the parent distribution of the data, although their value is unknown to the analyst. The available data are used to estimate the parameter values, typically through a maximum likelihood method (e.g., Chap. 11). For Gaussian data, the maximum likelihood method consists of finding the minimum of the χ^2 statistic.

It is reasonable to entertain the possibility that the model itself, although generally accurate, may have an *intrinsic scatter* or variance that needs to be accounted for in the determination of the fit statistic. In other words, the parent model may not be exact but it may feature an inherent degree of variability. The goal of this section is to provide a method to describe and measure such intrinsic scatter.

17.2.1 Direct Calculation of the Intrinsic Scatter

Each measurement in a dataset can be described as the sum of two variables,

$$y_i = \eta_i + \epsilon_i, \quad (17.1)$$

where η_i represents the unknown parent value from which the measurement y_i is drawn and ϵ_i is the variable representing the measurement error. Usually, it is assumed that $\eta_i = y(x_i)$ is a fixed number, to be estimated by the least-squares or another method. The variable ϵ_i has zero mean, and its variance is simply the measurement variance σ_i^2 so that (17.1) implies that the variance of the measurement y_i is just σ_i^2 .

This is the standard description of the data used so far.

The model η_i may however be considered as a variable with non-zero variance, as a means to describe the possibility that the model is not exact, but it has an intrinsic degree of variability as measured by its variance $\sigma_{int}^2 = \text{Var}(\eta_i)$. For simplicity, it is assumed that this model variance is constant for all points along the model. Under the assumption that the measurement error and the model are independent, variances of the variables on the right-hand side of (17.1) add, and this yields

$$\text{Var}(y_i) = \sigma_{int}^2 + \sigma_i^2. \quad (17.2)$$

In keeping with (17.1), $\text{Var}(y_i)$ is the variance of the i th variable under the hypothesis that the measurement is drawn from a parent model η_i , and it therefore represents a

residual variance from the model (see Sect. 13.6.2). Since the intrinsic variance is unknown, it is necessary to provide an estimate. For this purpose, it is meaningful to calculate the average of such variance for all the measurements y_i , assuming that each measurement is drawn from a parent mean equal to the best-fit value \hat{y}_i . As a result, (17.2) can be used to estimate the intrinsic scatter or variance of the model σ_{int}^2 as

$$\hat{\sigma}_{int}^2 = \frac{1}{N - m} \sum_{i=1}^N (y_i - \hat{y}_i)^2 - \frac{1}{N} \sum_{i=1}^N \sigma_i^2. \quad (17.3)$$

where m is the number of model parameters.

The reason for the $N - m$ factor in (17.3) is that for data with uniform variance and under the hypothesis that the model η_i is correct, the asymptotic expectation of the residual variance $S_r^2 = \sum_{i=1}^N (y_i - \hat{y}_i)^2$ is, according to Cramér's theorem,

$$E[S_r^2] = \sigma^2 (N - m),$$

where σ^2 is the common parent variance. For a uniform variance ($\sigma_i^2 = \sigma^2$) the expectation of the estimator of the intrinsic scatter is

$$E[\hat{\sigma}_{int}^2] = 0,$$

with the meaning that there is no need for an additional intrinsic scatter in the model.

The intrinsic variance (17.3) can also be referred to as the *debiased model variance*, because of the subtraction of the expected variance due to measurement errors from the residual variance. Equation (17.3) can therefore be considered an extension of the model sample variance of (12.4). It is possible that the second term in the right-hand side of (17.3) is larger than the first term, leading to a negative value for the intrinsic variance that is not meaningful. This is an indication that, within the statistical errors σ_i , there is no evidence for an intrinsic scatter of the model. Additional methods to estimate the intrinsic scatter are described in [2, 60]. It is important to remember that in calculating the intrinsic scatter the assumption was made that the model is an accurate representation of the data. This means that one can no longer test for the null hypothesis that the model represents the parent distribution—this was already assumed to be the case.

17.2.2 Alternative Method for Gaussian Data

An alternative method to measure the amount of extra variance in a fit makes use of the fact that, for a Gaussian dataset, the expected value of the reduced χ^2_{\min} is one.

A large value of the minimum χ^2 can be reduced by increasing the size of the errors until $\chi_{red}^2 \simeq 1$, or

$$\chi_{\min}^2 = \sum_{i=1}^N \frac{(y_i - \hat{y}_i)^2}{\sigma_i^2 + \hat{\sigma}_{int}^2} \simeq N - m \quad (17.4)$$

where m is the number of free model parameters, and $\hat{\sigma}_{int}$ is the estimate of the intrinsic scatter that makes the reduced χ^2 unity. In this equation, the data variance was increased by an amount equal to the estimated intrinsic variance of the model so that the effective variance becomes

$$\sigma_{eff}^2 = \sigma_i^2 + \hat{\sigma}_{int}^2. \quad (17.5)$$

This method is only approximate, in that an acceptable model need not yield exactly a value of $\chi_{red}^2 = 1$, yet it is useful as an estimate of the level of scatter present in the data. Like in the earlier method, the analyst is making the assumption that the model fits the data and that the extra variance is attributed to an intrinsic variability of the model.

Example 17.1 The example shown in Fig. 12.1 illustrates a case where the data do not show systematic deviations from a best-fit model, and yet the χ^2 test would require a rejection of the model. The quantities Energy 1 (independent variable) and Energy 2 were fit to a linear model, the best-fit linear model yielded a fit statistic of $\chi_{\min}^2 = 60.5$ for 23 degrees of freedom, and the model was therefore not acceptable. It is possible to estimate the intrinsic scatter that makes the model consistent with the data. Using (17.3), the intrinsic scatter is estimated to be $\hat{\sigma}_{int} = 2.5$. This means that the model has a typical uniform variability of 2.5 units (the units are those of the y-axis, in this case used to measure energy). Alternatively, (17.4) yields a value of $\hat{\sigma}_{int} = 1.6$ to obtain a reduced χ_{\min}^2 of unity. The two methods were not expected to provide the same answer since they are based on different assumptions. To obtain the estimate of the intrinsic scatter according to (17.4), it is necessary to step the $\hat{\sigma}_{int}$ parameter and repeat the fit for each value being tried, since the χ_{\min}^2 statistic and the best-fit model parameters change as the intrinsic scatter changes. ◇

17.3 Systematic Errors

The errors described in this book are usually referred to as *random errors* since they describe the uncertainties in the random variables of interest. An accurate value for the random error of a variable relies on the correctness of the model for the random variable itself. Consider for example a random variable S that represents the number of counts from a given source. If the random variable is measured directly, then it would seem appropriate to use a Poisson model for the variable, and use this model for making inferences from the measurements. Consider another experiment where

the same variable of interest is measured in the presence of a source of background so that $S = T - B$ is obtained from independent measurements of the total signal T and of the background B . It is clear that the distribution of S must now reflect the distributions of the two variables used for its measurement. A simple method to estimate the variance of S would be using the approximate variance formula (5.5), leading to

$$\sigma_S^2 = \sigma_T^2 + \sigma_B^2 \quad (17.6)$$

so that the variance of the random variable of interest S increases when the background is subtracted. If one assumes that there is no background, or that the background is measured without errors ($\sigma_B^2 = 0$), the random error associated with S may be erroneously underestimated. The term *statistical error* is often used as a synonym of random error. Sometimes, however, it is used to designate the leading source of random error, such as the Poisson uncertainty in a counting experiment, not including other sources of random error that are equally statistical or random in nature. Such use is not accurate, but the reader should be aware that there is no universally accepted meaning for the term “statistical error.”

The term *systematic error* is often used to designate sources of error that systematically shift the signal of interest either too high or too low. Sources of systematic errors need to be identified to correct the erroneous offset. A typical example is an instrument that is miscalibrated and systematically reports measurements that have an erroneous offset. Even after the correction for the offset, it is however quite likely that there still remains a source of error, for example associated with the fact that such correction may not be constant. If the systematic error is additive in nature, i.e., it shifts the random variable X according to $X' = X \pm S$, then the variance of the data point is to be modified according to

$$\sigma_{eff}^2 = \sigma_i^2 + \sigma_S^2. \quad (17.7)$$

The term σ_S^2 denotes the variance of the systematic error S . If the shift S is known exactly, then it would ideally have zero variance. But in all practical cases, there will be an additional source of variance from the correction of a systematic error that needs to be accounted for. The modification of the error σ_i due to the presence of a source of systematic error is therefore identical in form to the presence of intrinsic error, as can be seen by comparing (17.5) and (17.7). If the systematic error is multiplicative in nature, i.e., $X' = E \cdot X$, it may be convenient to use logarithms of the variables, $\log X' = \log X + \log E$ and then proceed as in the case of a linear offset.

Example 17.2 Continuing with Example 17.1, an additional error of magnitude $\hat{\sigma} \simeq 1.6$, added in quadrature to the reported measurement errors, yields a fit statistic of $\chi_{min,red}^2 = 1$. This means that a possible interpretation for the large initial value of χ_{min}^2 is that an additional source of error was neglected. The errors of the data in Fig. 12.1 accounted for several sources of random error, including Poisson errors in the counting of photons from these sources, the background subtraction, and for errors associated with the model used to describe the distribution of energy. The

additional error of order $\hat{\sigma}$ to be added to each data point may therefore be (a) an intrinsic error of the model, as described in Example 17.1, (b) an additional error from the correction of certain systematic errors that were performed in the process of the analysis, or (c) an additional random error that was not already included in the original error budget. The magnitude of possible errors according to scenarios (b) and (c) can be estimated based on the knowledge of the collection of the data and its analysis. If such errors cannot be as large as required to obtain an acceptable fit, the only remaining option is to attribute this error to an intrinsic variance of the model or to conclude that the model is not an accurate description of the data. ◇

17.4 Estimate of Model Parameters with Systematic Errors or Intrinsic Scatter

The method of estimation based on (17.3) assumes that it is possible to provide an estimate of the intrinsic scatter using the best-fit values \hat{y}_i obtained from the fit *without* these errors. Systematic errors or intrinsic scatter, however, do have an effect on the estimate of model parameters. The presence of systematic errors or intrinsic scatter is accounted for with the addition of another source of variance to the data according to

$$\sigma_{eff}^2 = \sigma_i^2 + \hat{\sigma}^2, \quad (17.8)$$

where $\hat{\sigma}$ is either a systematic or random error not accounted for in the initial estimate of σ_i , or the intrinsic scatter σ_{int} . Both cases lead to the same effect on the overall error budget and the χ^2 fit statistic to minimize becomes

$$\chi^2 = \sum_{i=1}^N \frac{(y_i - y(x_i))^2}{\sigma_{eff}^2}. \quad (17.9)$$

It is clear that repeating the fitting procedure with the larger effective errors instead of the original error will lead to new best-fit values and new uncertainties for the model parameters. The effect of the larger errors is to de-weight data points that have small values of σ_i and in general to provide larger confidence intervals for the model parameters. An acceptable procedure to obtain truly *best-fit* values of model parameters and their confidence intervals is to first estimate the additional source of error σ (either an intrinsic scatter or additional statistical or systematic errors) and then repeat the fit.

Example 17.3 The linear fit to the data of Table 8.1 for Energy 1 (independent variable) and Energy 2 resulted in a $\chi^2_{min} = 60.5$ for 23 degrees of freedom. The fit was not acceptable at any level of confidence. In Example 17.1 we calculated that an additional variance of $\sigma^2 = 1.6$ yields a $\chi^2_{min} = 23$. We fit the data with the addition of this error to the dependent variable and find the best-fit values of $a = -0.085 \pm 0.48$,

$b = 1.05 \pm 0.05$. For comparison, the fit obtained with the original errors returned values of $a = -0.26 \pm 0.088$, $b = 1.04 \pm 0.27$. These values could not be properly called “best-fit,” since the fit was not acceptable. Yet, comparison between these values and those for the $\chi^2_{red} = 1.0$ case shows that the best-fit parameters are affected by the additional source of error and that the confidence intervals become larger with the increased errors, as expected. \diamond

Summary of Key Concepts for this Chapter

Intrinsic scatter: An inherent uncertainty of the model that increases the measurement error. An estimate of the intrinsic scatter is provided by the *debiased model variance*

$$\sigma_{int}^2 = \frac{1}{N-m} \sum (y_i - \hat{y}_i)^2 - \frac{1}{N} \sum \sigma_i^2,$$

which is a modification of the model sample variance to remove the average variance caused by measurement errors. An alternative method to estimate the intrinsic scatter is by adding a new term to the error budget,

$$\sigma_{eff}^2 = \sigma_i^2 + \sigma_{int}^2$$

and use this effective variance for χ^2 minimization until the reduced χ^2 is approximately one.

Systematic error: A type of measurement error that systematically shifts the measurements (as opposed to the *statistical error* σ_i). Even after correcting for a systematic shift in the data, there often remains a variance associated with this shift that is typically added in quadrature to the data variance so that the effective variance becomes

$$\sigma_{eff}^2 = \sigma_i^2 + \sigma_S^2,$$

leading to a similar effect to the data as the intrinsic scatter.

Problems

17.1 ■ Consider the linear fit of the data from Table 8.1 for the radius versus ratio using (see Problem 11.6).

- (a) Calculate the intrinsic scatter using the best-fit linear model.
- (b) Evaluate a new best-fit model by adding the intrinsic scatter in quadrature with the original errors, and describe the effect of the intrinsic scatter on the best-fit model.

17.2 ■ Using the same data as in Problem 17.1 provides an alternative estimate of the intrinsic scatter using the $\chi^2_{red} = 1$ method.

17.3 Justify the $1/(N - m)$ coefficient in Eq. (17.3) for the intrinsic variance.

17.4 ■ Using the data for the Hubble measurements of Sect. 11.6, assume that each measurement of $\log v$ has an uncertainty of $\sigma = 0.01$, and that m is the independent variable. Estimate the intrinsic scatter in the linear regression of m versus $\log v$.

17.5 Using the five-point Poisson data of Problem 11.2, estimate the intrinsic scatter in the linear fit of X versus Y .

Chapter 18

Regression with Bivariate Errors



Abstract The maximum likelihood method for the fit of a two-variable dataset described in Chap. 11 assumes that one of the variables (the independent variable X) has negligible errors. There are many applications where this assumption is not accurate and uncertainties in both variables must be taken into account. This chapter expands the treatment of Chap. 11 to the fit of a two-variable dataset with errors in both variables, referred to as bivariate errors. The problem becomes treatable in the case of simple linear regression. One of the methods of regression for data with bivariate error is the BCES method developed by M. Akritas and M. Bershady. An alternative method uses a modification of the χ^2 fit statistic that accounts for errors in the independent variable.

18.1 Two-Variable Data with Bivariate Errors

The traditional methods of regression of Chaps. 11 and 12 assume a simple error model where the independent variable X is known without error, and all sources of uncertainty in the fit are due to the dependent variable Y . The two-variable dataset (X, Y) was effectively treated as a sequence of random variables of values $y_i \pm \sigma_i$ at a fixed location x_i with a parent model $y(x_i)$. There are many applications, however, in which both variables have comparable uncertainties ($\sigma_x \simeq \sigma_y$) and there is no reason to treat one variable as independent. In general, a two-variable dataset is described by a sequence of measurements (x_i, y_i) that are drawn from a joint probability distribution function (see Sect. 2.5.1). If the joint distribution function is known, it is possible to determine a likelihood of the data with the model, and then proceed with the maximization of the likelihood with respect to the adjustable model parameters. In practice, it is uncommon that the experiment yields such a detailed knowledge of the joint parent model, and it is therefore useful to investigate simpler error models for data with bivariate errors.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-981-19-0365-6_18.

A simplified data model consists of data with errors in both variables, i.e., *bivariate errors*

$$(x_i \pm \sigma_{x,i}, y_i \pm \sigma_{y,i})$$

and in the presence of a covariance $\sigma_{xy,i}^2$ between the two measurements. One example is the two measurements of energy in the data in Table 8.1, where it would be appropriate to account for errors in both measurements. There is in fact no particular reason why one measurement should be considered as the independent variable and the other the dependent variable, based on the nature of the experiment.

There are several methods to deal with two-variable datasets with bivariate errors. Given the complexity of the statistical model, there is not a uniquely accepted solution to the general problem of fitting data with bivariate errors. This chapter presents two methods for the linear fit to data with two-variable errors. The first method (Sect. 18.2) applies to a linear fit and it is an extension of the least-squares method of Sect. 11.3. The second method (Sect. 18.3) is based on an alternative definition of χ^2 and it applies to any type of fit function. Although this method does not have an analytic solution, it can be easily implemented using numerical methods such as Monte Carlo Markov chains described later in this book.

18.2 Least-Squares Linear Fit to Data with Bivariate Errors

The least-squares method described in Sect. 11.3 assumes that there are no errors in the independent variable X , and either no errors or equal errors for the dependent variable Y . Following these assumptions, the linear regression yields the following estimates for the model parameters:

$$\begin{cases} b = \frac{s_{xy}^2}{s_x^2} = \frac{\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^N (x_i - \bar{x})^2} \\ a = \bar{y} - b\bar{x} = \frac{1}{N} \sum_{i=1}^N y_i - b \frac{1}{N} \sum_{i=1}^N x_i. \end{cases} \quad (18.1)$$

A generalization of this least-squares method accounts for the presence of measurement errors in the estimate of the variances and the covariance in (18.1). The methods of analysis presented in this section were developed by M. Akritas and M. Bershady [2], who introduced the *BCES estimators* (for bivariate correlated errors and intrinsic scatter), and other authors [53, 60]. Those references can be used as a source of additional information on these methods for data with bivariate errors.

The X and Y variables can be described by

$$\begin{cases} X_i = \eta_{x,i} + \epsilon_{x,i} \\ Y_i = \eta_{y,i} + \epsilon_{y,i}, \end{cases} \quad (18.2)$$

each the sum of a *parent* quantity and a measurement error, as in (17.1). Accordingly, the intrinsic variances of the parent variables can be defined by

$$\begin{cases} \text{Var}(\eta_{x,i}) = \text{Var}(x_i) - \sigma_{x,i}^2 \\ \text{Var}(\eta_{y,i}) = \text{Var}(y_i) - \sigma_{y,i}^2 \end{cases} \quad (18.3)$$

in the same way that was used to define the intrinsic scatter. This analysis suggests that, in the presence of measurements errors, the sample variances and covariance in (18.1) should be replaced with estimates of the intrinsic variances according to (18.3), and in a similar manner for the covariance. The method of analysis that led to (18.1) assumes that the variable Y depends on X ,

$$y = a_{Y/X} + b_{Y/X} x, \quad (18.4)$$

and the need to specify a dependent and independent variable remains also within this method of analysis. Modification of (18.1) with (18.3), and an equivalent formula for the covariance, leads to the following estimator for the slope and intercept of the linear Y/X model in the presence of errors in both variables and covariance between the measurements:

$$\begin{cases} b_{Y/X} = \frac{\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y}) - \sum_{i=1}^N \sigma_{xy,i}^2}{\sum_{i=1}^N (x_i - \bar{x})^2 - \sum_{i=1}^N \sigma_{x,i}^2} \simeq \frac{s_{xy}^2 - \overline{s_{xy}^2}}{s_x^2 - \overline{s_x^2}} \\ a_{Y/X} = \bar{y} - b_{Y/X} \bar{x}. \end{cases} \quad (18.5)$$

In this equation, the sample variance and covariance of (18.1) were replaced with an expression that is asymptotically equal to the corresponding intrinsic quantities, with $\sigma_{x,i}^2$ and $\sigma_{xy,i}^2$ respectively the variances and covariances of the N measurements.

A different regression is obtained if Y is considered as the independent variable. In that case, the X -given- Y (or X/Y) model is described as

$$x = a' + b'y, \quad (18.6)$$

and (18.5) apply by exchanging the two variables X and Y :

$$\begin{cases} b' = \frac{\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y}) - \sum_{i=1}^N \sigma_{xy,i}^2}{\sum_{i=1}^N (y_i - \bar{y})^2 - \sum_{i=1}^N \sigma_{y,i}^2} \\ a' = \bar{x} - b'\bar{y}. \end{cases}$$

It is convenient to compare the results of the Y/X and X/Y fits by rewriting the latter in the usual form with x as the independent variable:

$$y = a_{X/Y} + b_{X/Y}x = -\frac{a'}{b'} + \frac{x}{b'}$$

for which we find that the slope and intercept are given by

$$\begin{cases} b_{X/Y} = \frac{1}{b'} = \frac{\sum_{i=1}^N (y_i - \bar{y})^2 - \sum_{i=1}^N \sigma_{y,i}^2}{\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y}) - \sum_{i=1}^N \sigma_{xy,i}^2} \\ a_{X/Y} = \bar{y} - b_{X/Y}\bar{x}. \end{cases} \quad (18.7)$$

In general, the two estimators Y/X and X/Y will give different results for the best-fit line. This difference highlights the importance of interpreting the data to determine which variable should be considered the independent quantity.

Uncertainties in the parameters a and b and the covariance between them have been calculated in [2]. For the Y/X estimator, they can be obtained via the following variables:

$$\begin{aligned} \xi_i &= \frac{(x_i - \bar{x})(y_i - b_{Y/X}x_i - a_{Y/X}) + b_{Y/X}\sigma_{x,i}^2 - \sigma_{xy,i}^2}{\frac{1}{N}\sum_{i=1}^N(x_i - \bar{x})^2 - \frac{1}{N}\sum_{i=1}^N\sigma_{x,i}^2} \\ \zeta_i &= y_i - b_{Y/X}x_i - \bar{x}\xi_i. \end{aligned} \quad (18.8)$$

With these, the variances of a and b and the covariance are given by

$$\begin{cases} \sigma_{b_{Y/X}}^2 = \frac{1}{N}\sum_{i=1}^N(\xi_i - \bar{\xi})^2 \\ \sigma_{a_{Y/X}}^2 = \frac{1}{N}\sum_{i=1}^N(\zeta_i - \bar{\zeta})^2 \\ \sigma_{ab}^2 = \frac{1}{N}\sum_{i=1}^N(\xi_i - \bar{\xi})(\zeta_i - \bar{\zeta}). \end{cases} \quad (18.9)$$

For the X/Y estimator there are equivalent formulas for the ξ and ζ variables that need to be used in place of (18.8):

$$\begin{aligned} \xi_i &= \frac{(y_i - \bar{y})(x_i - b_{X/Y}y_i - a_{X/Y}) + b_{X/Y}\sigma_{xy,i}^2 - \sigma_{y,i}^2}{\frac{1}{N}\sum_{i=1}^N(x_i - \bar{x})(y_i - \bar{y}) - \frac{1}{N}\sum_{i=1}^N\sigma_{xy,i}^2} \\ \zeta_i &= y_i - b_{X/Y}x_i - \bar{x}\xi_i. \end{aligned} \quad (18.10)$$

These values can be then used to calculate variances and the covariance of the parameters as in the Y/X fit.

Example 18.1 Figure 18.1 illustrates the difference in the best-fit models when X is the independent variable (18.5) or Y is the independent variable (18.7), using the

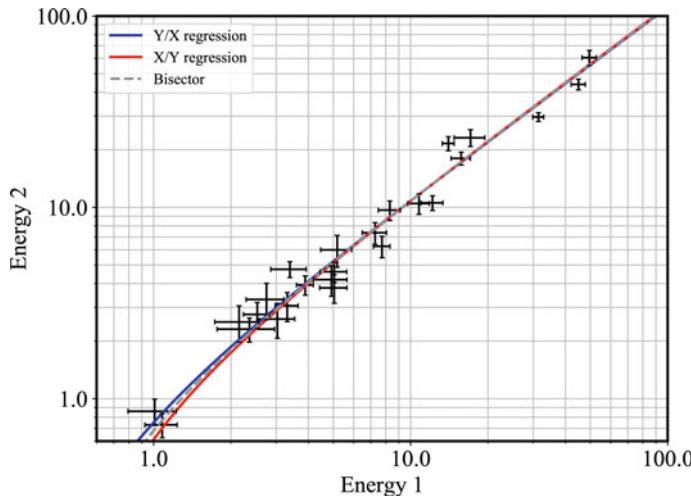


Fig. 18.1 Linear model fits to the data of Table 8.1 using the BCES method that includes errors in both variables. Note the logarithmic scale for both axes which renders the linear model as a curve

data of Table 8.1. The Y/X parameters are $a_{Y/X} = -0.367$ and $b_{Y/X} = 1.118$ and the X/Y parameters are $a_{X/Y} = -0.521$ and $b_{X/Y} = 1.132$. For these data there is no clear reason to decide which variable should be regarded as independent, and each variable could be equally treated as the independent variable and the difference between the two best-fit models is relatively small. Note that the linear model and the data were plotted in a logarithmic scale to provide a more compact figure. Also, the data of Table 8.1 do not report any covariance measurement and therefore the best-fit lines were calculated assuming independence between all measurements ($\sigma_{xy,i}^2 = 0$). Numerical implementation of the formulas for the BCES regression was provided by R.S. Nemmen [74]. \diamond

The regressions in Example 18.1 show that there is not just a single slope for the best-fit linear model, but that the results depend on which variable is assumed to be independent. In certain cases it may be appropriate to use a model that is intermediate between the two Y/X and X/Y results. This is called the *bisector* model, which consists of the linear model that bisects the two lines obtained from the Y/X and X/Y fits described above. This method is also described in [2] and by T. Isobe and colleagues [53]. The best-fit bisector line can be obtained from the following formulae:

$$\begin{cases} b_{bis} = \frac{b_{Y/X} b_{X/Y} - 1 + \sqrt{(1 + b_{Y/X}^2)(1 + b_{X/Y}^2)}}{b_{Y/X} + b_{X/Y}} \\ a_{bis} = \bar{y} - b_{bis} \bar{x}. \end{cases} \quad (18.11)$$

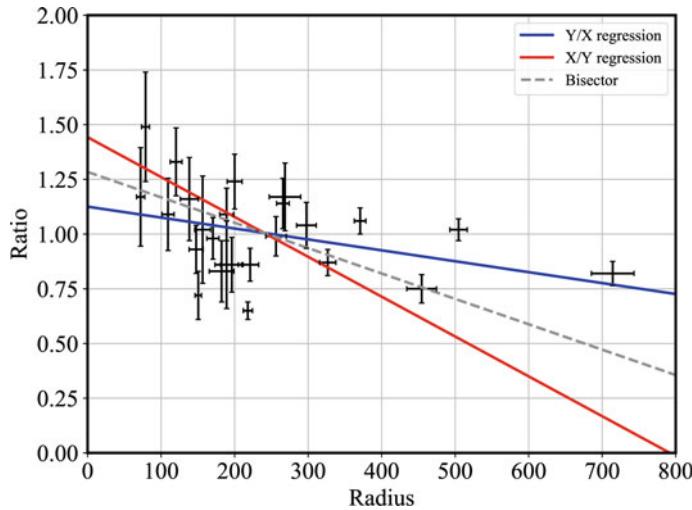


Fig. 18.2 Fit to the data of Table 8.1 using errors in both variables (see Example 18.2)

The uncertainties in the slope and intercept parameters can also be obtained using this definition for the ξ and ζ variables:

$$\begin{aligned} \xi_i &= \frac{(1 + b_{X/Y}^2) b_{bis}}{(b_{Y/X} + b_{X/Y}) \sqrt{(1 + b_{Y/X}^2)(1 + b_{X/Y}^2)}} \xi_{Y/X} + \\ &\quad \frac{(1 + b_{Y/X}^2) b_{bis}}{(b_{Y/X} + b_{X/Y}) \sqrt{(1 + b_{Y/X}^2)(1 + b_{X/Y}^2)}} \xi_{X/Y} \\ \zeta_i &= y_i - b_{bis} x_i - \bar{x} \xi_i, \end{aligned} \quad (18.12)$$

where $\xi_{Y/X}$ is the ξ variable defined in (18.8) for the Y/X fit and $\xi_{X/Y}$ is the ξ variable defined in (18.10) for the X/Y fit. The bisector method should be viewed as an entirely *ad hoc* method that aims to mitigate the choice of the independent variable, for those applications where it is unclear which variable should be designated as dependent or independent.

Example 18.2 Figure 18.2 shows the fit to the variables radius (X variable) and ratio of thermal energies (Y variable) from Table 8.1. The Y/X best-fit line has parameters $a = 1.1253$ and $b = -0.0005$, the X/Y best-fit line has $a = 1.4426$ and $b = -0.0018$, and the bisector line has $a = 1.2840$ and $b = -0.0012$. For these data the Y/X and X/Y regressions give significantly different results. This is in part due to the presence of substantial scatter in the data, which results in several data points significantly distant from the best-fit regression lines. In the other example of

regression with errors in both variables (Fig. 18.1) the Y/X and X/Y best-fit lines were in better agreement. See Problem 18.3 for other considerations regarding these data. \diamond

18.3 Linear Fit Using Bivariate Errors in the χ^2 Statistic

An alternative method to fit a dataset with errors in both variables is to re-define the χ^2 statistic to account for the presence of errors in the independent variable. In the case of a linear fit, the square of the deviation of each data point y_i from the model is given by

$$(y_i - a - bx_i)^2. \quad (18.13)$$

When there is no error in the X variable, the variance of the variable in (18.13) is simply the variance of Y , $\sigma_{y,i}^2$. In the presence of a variance term $\sigma_{x,i}^2$ for the measurements of X , the variance of the linear combination $y_i - a - bx_i$ is given by

$$\text{Var}(y_i - a - bx_i) = \sigma_{y,i}^2 + b^2\sigma_{x,i}^2,$$

where a and b are the parameters of the linear model and the variables X and Y are assumed to be independent. This suggests a new definition of the χ^2 statistic for this dataset, namely

$$\chi^2 = \sum_{i=1}^N \frac{(y_i - a - bx_i)^2}{\sigma_{y,i}^2 + b^2\sigma_{x,i}^2}. \quad (18.14)$$

Since each term at the denominator is the variance of the term at the numerator, the new *bivariate* χ^2 statistic defined in (18.14) is χ^2 -distributed with $f = N - 2$ degrees of freedom. The complication with the minimization of this function is that the unknown parameter b appears both at the numerator and the denominator of the function that needs to be minimized. As a result, an analytic solution to the maximum likelihood method cannot be given in general. Fortunately, the problem of finding the values of a and b that minimize (18.14) can be solved numerically. This method for the linear fit of two-variable data with errors in both coordinates is therefore of common use, and it is further described in the textbook by W.H. Press [80] and an application is provided in [98].

Under the assumption that the bivariate χ^2 for the linear model remains χ^2 -distributed with $f = N - 2$ degrees of freedom, parameter uncertainties can be obtained using the same considerations in Sect. 12.3, e.g., the $\Delta\chi^2 = 1$ criterion can be used for the 68% confidence interval on one interesting parameter. For models other than the linear model, (18.14) does not apply. In principle, it is possible to extend the considerations of this section to other models, starting with the model of choice and making use of the error propagation equations of Sect. 5.2 to define a χ^2 -distributed statistic that uses bivariate errors.

Summary of Key Concepts for this Chapter

Data with bivariate errors: A two-variable dataset that has errors in both variables, optionally with covariance between the measurements. For these data there is not a universally accepted method of regression, due to the complexity of the error model.

The BCES linear regressions: An extension of the traditional maximum likelihood method for the linear regression of data with bivariate error. It yields the following best-fit parameters, when x is the independent variable:

$$\begin{cases} b_{Y/X} = \frac{s_{xy}^2 - \overline{\sigma_{xy}^2}}{s_x^2 - \overline{\sigma_x^2}} \\ a_{Y/X} = \bar{y} - b_{Y/X} \bar{x}. \end{cases}$$

When it is not clear which variable is independent, the *bisector method* is an ad hoc average of the Y/X and X/Y linear regression solutions.

Modification of χ^2 for linear regression: The χ^2 statistic can also be redefined to accommodate bivariate errors according to

$$\chi^2 = \sum_{i=1}^N \frac{(y_i - a - bx_i)^2}{\sigma_{y,i}^2 + b^2 \sigma_{x,i}^2}.$$

This modification only applies to the linear regression, and it typically requires a numerical method of minimization.

Problems

18.1 ■ Use the bivariate error data of Energy 1 and Energy 2 from Table 8.1, with the errors obtained as the average of the given uncertainties.

- Calculate the best-fit parameters and errors of the linear model Y/X , where X is Energy 1 and Y is Energy 2.
- Calculate the best-fit parameters and errors of the linear model X/Y .
- Using the results of (a) and (b), calculate the bisector model.

18.2 ■ Use the bivariate error data of radius and ratio from Table 8.1, with the errors obtained as the average of the given uncertainties.

- Calculate the best-fit parameters and errors of the linear model Y/X , where X is radius and Y is ratio.
- Calculate the best-fit parameters and errors of the linear model X/Y .
- Using the results of (a) and (b), calculate the bisector model.

18.3 ■ Use the bivariate error data of radius and ratio from Table 8.1, same as in Problem 18.2.

- Given the large scatter in the data, assume that the errors are now *twice* the values used in Problem 18.2. Find the best-fit parameters and errors of the linear models Y/X and X/Y .
- The large discrepancy between the Y/X and X/Y models suggests that these data may not be well described by a linear model. Calculate the Pearson correlation coefficient between the radius and ratio (without use of the errors) and discuss whether the null hypothesis of no correlation between them can be rejected at the 99% confidence level.

18.4 ■ Use the bivariate error data of radius and ratio from Table 8.1, same as in Problem 18.3, namely with errors that are *twice* the values used in Table 8.1. Perform a numerical minimization of the bivariate χ^2 of (18.14) to find the best-fit values of the linear model and the value of χ^2_{\min} .

Chapter 19

Model and Data Comparison



Abstract The availability of multiple datasets and alternative models requires a quantitative method for comparing the goodness of fit. For Gaussian data, a lower reduced χ^2 is indicative of a better fit, but this statistic cannot be used alone to decide between competing models. For this purpose, the F -statistic can be used to compare the goodness of fit between two independent χ^2 measurements, and to determine the need for an additional “nested” model component. A related topic is the comparison of the sample distribution of a random variable with either a parent model or a second distribution from a different sample. Both tasks can be accomplished with the Kolmogorov–Smirnov statistics.

19.1 The χ^2_{\min} Statistic and the F -Test for Gaussian Data

For Gaussian data, χ^2_{\min} is the goodness of fit statistic used to determine if the fit to a given model $y(x)$ is acceptable. Although the meaning of an “acceptable” fit was discussed extensively in Sect. 9.1, it is worth reminding here that hypothesis testing can only conclusively reject a null hypothesis or model, and never conclusively prove it to be correct. Accordingly, a model is referred to as being acceptable only with the meaning that it cannot be rejected. It is possible that several different models all yield a goodness of fit that is acceptable. In this case, the data analyst is faced with the decision to determine which model best fits the experimental data. In this situation, the procedure to follow is to decide first a confidence level that is considered acceptable, for example 90 or 99%, and discard all models that do not satisfy this criterion. The remaining models are all acceptable, although a lower χ^2_{\min} certainly indicates a better fit. Further decisions regarding competing models require an analysis of statistical fluctuations of the χ^2_{\min} statistic. Although under rather restrictive

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-981-19-0365-6_19.

conditions on the models, the F statistic introduced in Sect. 9.5 may be used for this purpose. The first version of the F test applies to independent measurements of the χ^2_{\min} fit statistic, and its application is therefore limited to cases that compare different datasets, and not the same dataset to two different models. A more common application of the F test is to compare the fit of a given dataset between two models that have a *nested* component, i.e., one model is a simplified version of the other. For nested model components one can determine whether the additional component is really needed to fit the data. These two tests with the F statistic are discussed in the next sections.

19.2 F -Test for Two Independent χ^2 Measurements

Consider two χ^2_{\min} values obtained by fitting data to two different model functions, $y_1(x)$ and $y_2(x)$. If both models equally well describe the data, one would expect that the two values of χ^2 would be similar, after taking into consideration the different number of degrees of freedom, but if one is a better model then its reduced χ^2_{\min} would be significantly lower than for the other model. The statistic to use to compare the two values of χ^2_{\min} must certainly take into account the numbers of degrees of freedom, which is related to the number of model parameters used in each determination of the fit statistic. In fact, a larger number of model parameters may in fact result in a lower value of χ^2_{\min} , simply because of the larger flexibility that the model has in following the data. For example, a dataset of N points will always be fitted perfectly by a polynomial having N terms, but this does not mean that a simpler model may not be consistent with the data.

Following the theory of Sect. 9.5, the F -statistic is defined as the ratio of reduced χ^2 statistics,

$$F = \frac{\chi^2_{1,\min}/f_1}{\chi^2_{2,\min}/f_2}, \quad (19.1)$$

where f_1 and f_2 are the degrees of freedom of $\chi^2_{1,\min}$ and $\chi^2_{2,\min}$. Assuming that the two χ^2 statistics are *independent*, then F will be distributed like the F -statistic with f_1 , f_2 degrees of freedom, having a mean of approximately 1 [see (9.22) and (9.24)]. There is an ambiguity in the definition of which of the two models is labeled as 1 and which as 2, since two statistics can be constructed that are the reciprocal of each other. The usual form of the F -test uses a statistic that is greater than one so that a one-tailed test of the null hypothesis with significance p is given by

$$P(F \geq F_{crit}) = \int_{F_{crit}}^{\infty} f_F(x) dx = 1 - p. \quad (19.2)$$

The null hypothesis is that the two values of χ^2_{\min} are distributed following a χ^2 distribution, meaning that both fit functions used in the χ^2_{\min} are acceptable.

An F -statistic that exceeds its critical value means that the ratio of the two measurements is not consistent with typical ratios from two χ^2 distributions, with the implication that one measurement (the one at the numerator) is significantly larger than the other (the one at the denominator), although both are consistent with the respective χ^2 distribution. The F -test therefore aims to detect differences between the two statistics that the χ^2 distribution alone could not. When the F -test fails, it is necessary to investigate the reason for the large value of the statistic. Example 19.1 discusses how to interpret the F -test in practice.

A crucial caveat that significantly limits the use of the F -test according to (19.1) is the hypothesis of independence between the two χ^2 statistics. Consider for example fitting a given dataset to two different models, say a linear model and an exponential model. The two fits will lead to two χ_{\min}^2 statistics that are clearly dependent because of the use of the same data in the regression. It is therefore *not* possible to apply the F -test to determine whether the linear or the exponential models are to be preferred. If either model resulted in a χ_{\min}^2 statistic that is not acceptable, then that model would be discarded purely based on the properties of the χ^2 distribution. Fortunately, there are a class of models with a nested component to which a modified version of the F -test is applicable for fits to the same data. They are discussed in Sect. 19.3. In the present form, the F -test can only be applied to independent measurements of χ_{\min}^2 .

Example 19.1 Consider the radius versus ratio data of Table 8.1. The linear fit to the entire dataset is not acceptable, and therefore a linear model for all measurements must be discarded. Consider the following subsets of the data, measurements 1–5, 6–10, and 21–25. A linear fit to these subsets results in the values of best-fit parameters and χ^2 shown in the table, along with the probability to exceed the value of the fit statistic.

Measurements	a	b	χ_{\min}^2	p-value
1–5	0.969 ± 0.086	-0.00029 ± 0.000101	5.31	0.150
6–10	1.269 ± 0.214	-0.00072 ± 0.00105	6.68	0.083
21–25	0.975 ± 0.111	0.00009 ± 0.00027	0.15	0.985

All three samples provide acceptable fits to a linear model, with the last fit having an exceptionally low χ_{\min}^2 statistic. The F -statistic to compare the first and second measurements is

$$F = \frac{\chi_{\min}^2(6 - 10)}{\chi_{\min}^2(1 - 5)} = 1.26.$$

Both χ_{\min}^2 statistics have the same number of degrees of freedom (3), and the p-value of the F -statistic is 0.428, indicating that the ratio is well within reasonable values. This was expected since both χ_{\min}^2 statistics have reasonable values compared to the parent $\chi^2(3)$ distribution. In this case, the F -test does not add any additional information to the individual χ_{\min}^2 tests.

Consider now the F -statistic between the second and third datasets,

$$F = \frac{\chi_{\min}^2(1 - 5)}{\chi_{\min}^2(21 - 25)} = 35.32,$$

for a p -value of 0.0077. The small p -value indicates that the ratio of χ^2 measurements is unlikely to happen as a result of random fluctuations. In this case, it is clear that the reason for the unusually large F value is the small denominator, rather than the numerator. It would therefore be incorrect to discard the hypothesis that the first range of data is well fit by a linear model, which it is. Rather, it would be reasonable to either question the unusually small value of χ^2_{\min} for the last set of data, perhaps investigating whether the errors were calculated correctly, or accept the small p -value as a result of the exceptionally good fit of the 21–25 data points. \diamond

19.3 F-Test for an Additional Model Component

A class of models of interest in data analysis is *nested models*, whereby a simplified model can be obtained by fixing certain parameters of the more general model. An example of nested models are polynomial functions. For example, the general model can be taken as a polynomial of second order,

$$y(x) = a + bx + cx^2$$

and the simpler model as a linear model,

$$\bar{y}(x) = a + bx.$$

The simpler model is obtained from the general model with $c = 0$ and has one fewer free parameters than the general model. An analyst often wishes to find evidence that the simpler model is not sufficient to explain the data, while the more general model provides an acceptable fit. In general, the full model $y(x)$ has m adjustable parameters, and the simpler model $\bar{y}(x)$ is obtained by fixing p of the m parameters to a reference (fixed) value. The task is to determine whether the additional p parameters, associated with the additional *nested component* of the general model, are required to fit the data.

The Cramér Theorem 12.1 can be used to find the asymptotic distribution of the fit statistics for the two models:

$$\begin{cases} \chi^2_{\min} \sim \chi^2(N - m) & \text{(full model)} \\ \bar{\chi}^2_{\min} \sim \chi^2(N - m + p) & \text{(simplified model).} \end{cases} \quad (19.3)$$

It is clearly true that $\chi^2_{\min} < \bar{\chi}^2_{\min}$ because of the additional free parameters used in the determination of χ^2_{\min} . It is therefore possible to define the statistic

$$\Delta\chi^2 = \bar{\chi}^2_{\min} - \chi^2_{\min} \sim \chi^2(p), \quad (19.4)$$

in the same way in which $\Delta\chi^2 = \chi_{\text{true}}^2 - \chi_{\min}^2 \sim \chi^2(m)$ was defined in Sect. 12.3, under the null hypothesis that both the full and nested models are accurate descriptions of the data. Properties (19.3) and (19.4) are a consequence of the nested nature of the two models. In fact, the χ_{true}^2 statistic assumes that any adjustable parameters of the model have been fixed at the true-yet-unknown values, in the same way as the additional parameters of the nested component for $\bar{\chi}_{\min}^2$. In the example of the second-order polynomial component, fixing the parameter $c = 0$ for the statistic $\bar{\chi}_{\min}^2$ assumes that $c = 0$ is the parent value. Following the same arguments as in Sect. 12.3, it is also true that $\Delta\chi^2$ and χ_{\min}^2 continue to be independent also for the present case of a nested model component.

It is therefore possible to define the *F*-statistic

$$F = \frac{\Delta\chi^2/p}{\chi_{\min}^2/(N-m)} \quad (19.5)$$

to test for the necessity of the additional nested component with p additional free parameters. In this case, the null hypothesis is that the full model $y(x)$ with m parameter and the simplified model $\bar{y}(x)$ are equivalent, i.e., adding the p parameters does *not* constitute a significant change or improvement to the model. The probability distribution of F under the null hypothesis is an *F* distribution with $f_1 = p$, $f_2 = N - m$ degrees of freedom. A rejection of the hypothesis indicates that the full and simplified models are not equivalent, therefore providing positive evidence that the additional model parameters in the nested component *are* needed to model the data. A common situation is when there is a single additional model parameter, $p = 1$, and the corresponding critical values of F are reported in Table A.8.

The use of the *F*-test according to (19.5) is also subject to additional constraints that stem from certain hypotheses required for the applicability of Cramér's theorem, as stated in Sects. 30.3 and 33.2 of Cramér's book [21]. It is useful to investigate these conditions in more detail, starting with Cramér's definition of the χ^2 statistic of his Sect. 30.3. The asymptotic distribution of χ_{\min}^2 is obtained by minimizing the statistic

$$\chi^2 = \sum_{i=1}^N \frac{(\nu_i - n p_i(\alpha_1, \dots, \alpha_m))^2}{n p_i(\alpha_1, \dots, \alpha_m)}$$

where ν_i is the measured number of occurrence of events in the i th group or measurement, n is the total number of tries of the experiment, and p_i is the probability of occurrence of an event in the i th group, which is function of the usual m parameters α_j . In the notation used in this book, $y(x_i) = n p_i$ represents the expectation of model function and $y_i = \nu_i$ is the measurement of the random variable. In the process of minimizing χ^2 , Cramér uses a so-called *modified χ^2 minimum method* which results in estimating the best-fit model parameters via the simplified equation

$$-\frac{1}{2} \frac{\partial \chi^2}{\partial \alpha_j} \simeq \sum_{i=1}^N \frac{\nu_i - n p_i}{p_i} \cdot \frac{\partial p_i}{\partial \alpha_j} = 0.$$

Using the fact that the sum of the probabilities p_i is one, Cramér points out that this condition is equivalent to

$$\sum_{i=1}^N \frac{\nu_i}{p_i} \frac{\partial p_i}{\partial \alpha_j} = \frac{1}{\mathcal{L}} \frac{\partial \mathcal{L}}{\partial \alpha_j} = 0$$

where $\mathcal{L} = p_1^{\nu_1} \cdot \dots \cdot p_N^{\nu_N}$ is the likelihood, showing that estimating parameters via minimization of χ^2 is part of a broader class of maximum likelihood estimates, as also shown in Chap. 11.

When estimating best-fit parameters using the maximum likelihood method, the probability distribution of the measured variables must obey a number of regularity conditions. Cramér enumerates these conditions in Sect. 33.2 of his book, and a similar treatment of the topic is also provided in the textbook by R.J. Serfling (see Sect. 4.2 [89]). One of the key conditions is that the distribution function has continuous first and second derivatives with respect to the parameters, in an *open interval* of the fit parameters. As pointed out by R. Prottassov [81], this condition requires that, in order to use Cramér's theorem and therefore the F -test of (19.5),

the null values of the additional parameters may not be on the boundary of the set of possible parameter values.

A practical case of *improper* use of the F -test to detect an additional nested component is that of an emission or absorption line model. In fact, such model would have a parameter that is constrained to be in an interval of the type $\alpha \geq 0$ (say, if the parameter represents the intensity of an emission line), with $\alpha = 0$ the null value for the simpler model. Since the null parameter value falls on the boundary of the parameter range, the distribution function does not have continuous derivatives at that point, therefore invalidating one of the hypotheses of Cramér's theorem.

Example 19.2 The data of Example 12.2 and Fig. 12.2 are well fit by a two-parameter linear model. Using a one-parameter constant model yields a fit statistic of $\chi^2_{\min} = 39.77$, and therefore the constant model is not a good fit to all measurements. Using only the middle three measurements, it is interesting to compare the goodness of fit to a linear model, and that to a constant model, and determine whether the addition of the b parameter provides a significant improvement to the fit. The best-fit linear model has a $\chi^2_{\min} = 5.42$ for 1 degree of freedom, for a p -value of 0.20, and the constant model $\bar{\chi}^2_{\min} = 7.96$ for 2 degrees of freedom and a p -value of 0.019. The fact that the two χ^2_{\min} statistics are virtually equivalent in terms of their p -values is already an indication that the constant model is sufficient to explain the middle three data points. The F -statistic to test for the additional slope term is

$$F = \frac{\Delta\chi^2/1}{\chi_{\min}^2/1} = 0.47,$$

with $\Delta\chi^2 = \bar{\chi}_{\min}^2 - \chi_{\min}^2 = 2.53$. The F -statistic has a p -value of 0.62, which shows that there is clearly no evidence for the need of the additional linear component. \diamond

The F test for an additional component illustrates the principle of simplicity or parsimony in the analysis of data. When choosing between two models, both with an acceptable fit statistic at a similar confidence level, one should prefer the simpler model with fewer parameters. This general guiding principle of parsimony is often referred to as *Occam's razor*, after the Middle Ages philosopher and Franciscan friar William of Occam, who is credited with establishing this logical principle. In his *Summa Logicae* [75], Occam describes this concept as *Frustra fit per plura quod potest fieri per pauciora*, which can be translated as “It is futile to do with more things than which can be done with fewer.” Similar principles had been stated by a multitude of earlier thinkers, including Aristotle’s *ceteris paribus* principle (“all other things being equal”) [4], and Occam’s fellow scholastic philosopher John Duns Scotus, who stated that *Pluralitas non est ponenda sine necessitate*, which can be translated as “plurality should not be posited without necessity” [24]. The F -test for additional nested components described in this section embodies this principle, and frames it within a quantitative hypothesis testing method.

19.4 Kolmogorov–Smirnov Tests

A. Kolmogorov and N. Smirnov devised two statistics that are useful to compare the sample distribution of a variable to a model, or for the comparison of two sample distributions to one another [23, 62, 94]. The tests based on these statistics make use of the cumulative distribution function of a sample, and are applicable to measurements of a single variable, for example to determine if the variable is distributed like a Gaussian or any other distribution. The greatest advantage of the Kolmogorov–Smirnov tests is that they do not require the data to be binned, and, for the case of the comparison between two datasets, it does not require any parameterization of the data. These advantages come at the expense of a more complicated mathematical treatment to find the null distribution function of the test statistic. Fortunately, numerical tables and analytical approximations make these tests manageable and very useful.

19.4.1 Comparison of Data to a Model

Consider a random variable X with cumulative distribution function $F(x)$. The data consist of N measurements that can be arranged in increasing order, $x_1 \leq x_2 \leq$

$\dots \leq x_N$. This condition can be achieved by re-labeling the measurements, while preserving the statistical properties of the data. The goal is to construct a statistic that describes the difference between the sample distribution of the data and a specified distribution, to test whether the data are compatible with this distribution. Recall that the cumulative sample distribution is defined as

$$F_N(x) = \frac{(\text{Number of measurements} \leq x)}{N}, \quad (19.6)$$

and by definition it is $0 \leq F_N(x) \leq 1$. The test statistic proposed by Kolmogorov is defined as

$$D_N = \max_x |F_N(x) - F(x)|, \quad (19.7)$$

where $F(x)$ is the parent cumulative distribution, and the maximum value of the difference between the parent distribution and the sample distribution is calculated for all values in the support of the random variable X . One of the remarkable properties of the statistic D_N is that it has the same distribution for any underlying distribution of X , provided X is a continuous variable. The proof that D_N has the same distribution regardless of the distribution of X illustrates the properties of the cumulative distribution and of the quantile function presented in Sect. 5.3.

To show that the statistic D_N is independent of the distribution $F(x)$, assume that $F(x)$ is continuous and strictly increasing. This is certainly the case for a Gaussian distribution, or any other distribution that does not have intervals where the distribution function is $f(x) = 0$. It is convenient to make a change of variables $y = F(x)$, so that the measurement x_k corresponds to $y_k = F(x_k)$. This change of variables is such that

$$F_N(x) = \frac{(\text{Number of } x_i < x)}{N} = \frac{(\text{Number of } y_k < y)}{N} = U_N(y)$$

where $U_N(y)$ is the sample cumulative distribution of Y and $0 \leq y \leq 1$. The cumulative distribution of Y is

$$U(y) = P(Y < y) = P(X < x) = F(x) = y,$$

showing that Y is a uniform distribution between 0 and 1. As a result, the statistic D_N is equivalent to

$$D_N = \max_{0 \leq y \leq 1} |U_N(y) - U(y)|$$

where Y is a uniform distribution. Since this is true no matter the original distribution X , D_N has the same distribution for any X . Note that this derivation relies on the continuity of X , and this assumption must be verified to apply the resulting Kolmogorov–Smirnov test.

Table 19.1 Critical values of the Kolmogorov statistic D_N using the approximation (19.8) for large values of N , and using the accurate numerical solution for a reference value of $N = 100$

Confidence level p	$\sqrt{N} \times D_{N,crit}$	$D_{N,crit}$ (for $N = 100$)
0.50	0.828	0.081
0.60	0.895	0.088
0.70	0.973	0.096
0.80	1.073	0.106
0.90	1.224	0.121
0.95	1.358	0.134
0.99	1.628	0.161

The distribution of the statistic D_N was determined by Kolmogorov in 1933 [62], and although it is not easy to evaluate analytically, it can be computed numerically (see, e.g., [68, 93]). The asymptotic distribution of the D_N statistic was provided by W. Feller [27]. In the limit of large N , the cumulative distribution of D_N is given by

$$\lim_{N \rightarrow \infty} P(D_N < z/\sqrt{N}) = \sum_{r=-\infty}^{\infty} (-1)^r e^{-2r^2 z^2} = \Phi(z). \quad (19.8)$$

The function $\Phi(z)$ can also be used to approximate the probability distribution of D_N for small values of N , using

$$P(D_N < z/(\sqrt{N} + 0.12 + 0.11/\sqrt{N})) \simeq \Phi(z). \quad (19.9)$$

This approximation was suggested by M.A. Stephens [96]. The probability distribution of D_N can be used to test whether a sample distribution is consistent with the model distribution. Given that the Kolmogorov statistic is non-negative, it is meaningful to use one-sided critical values with probability p according to

$$P(D_N \geq D_{N,crit}) = 1 - p \quad (19.10)$$

with the usual meaning that values of the statistic D_N larger than the critical values are expected to occur with a small probability $1 - p$, under the null hypothesis that the sample distribution follows the model. Critical values of the Kolmogorov statistic are shown in Table 19.1 in the limit of large N . The critical values depend on N , and critical values of the D_N statistic as a function of N are provided in Table A.25. If the measured value for D_N is greater than the critical value, then the null hypothesis must be rejected, and the data are not consistent with the model. The test allows no free parameters, i.e., the distribution that represents the null hypothesis must be fully specified.

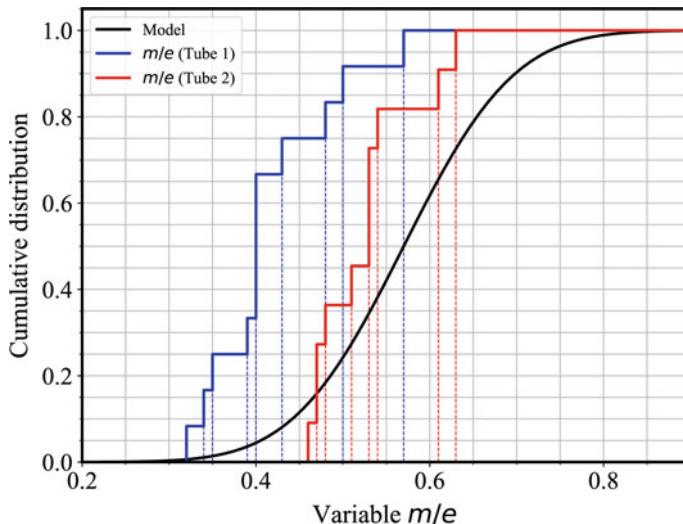


Fig. 19.1 Kolmogorov–Smirnov test applied to the measurements of the ratio m/e from Thomson’s experiments described on Sect. 2.4. The blue curve is the cumulative sample distribution of m/e for Tube 1, and the red curve for Tube 2 (measurements have been multiplied by 10^7). The black curve is the cumulative distribution of a Gaussian with $\mu = 5.7$ and $\sigma^2 = 1$

Example 19.3 Consider the data from Thomson’s experiment to measure the ratio m/e of an electron (Sect. 2.4). The Kolmogorov statistic D_N can be used to test whether either of the two sets of measurement of the variable m/e is consistent with a given hypothesis. It is necessary to realize that the Kolmogorov–Smirnov test applies to a fully specified hypothesis, i.e., the parent distribution $F(x)$ cannot have free parameter that is to be determined by a fit to the data. It is assumed that the hypothesis is that the ratio is described by a Gaussian distribution of $\mu = 5.7$ (the true value in units of 10^7 g C^{-1} , though the units are unnecessary for this test), and a fiducial variance of $\sigma^2 = 1$. Figure 19.1 illustrates the method of analysis for the Kolmogorov test for the comparison of data with a model. The first step is the evaluation of the cumulative sample distribution, in this case two separate distributions for the measurements from Tube 1 and Tube 2 separately. For example, the 12 individual measurements from Tube 1 are sorted from smaller to larger and indicated as vertical dashed lines. Some measurements are identical and correspond to larger jumps in the sample distribution. The statistic is $D_N = 0.675$ for the measurements of Tube 1, which corresponds to a p -value of less than 10^{-5} for $N = 12$. The measurements of Tube 1 are therefore not consistent with the hypothesis. The measurements from Tube 2 show a better agreement, with a value of $D_N = 0.436$ for $N = 11$ that corresponds to a p -value of 0.02. See Problem 19.1 for a quantitative analysis of the results. ◇

19.4.2 Two-Sample Kolmogorov–Smirnov Test

N. Smirnov modified the Kolmogorov statistic for the purpose of comparing two samples of data to each other. Consider two independent sets of measurements of a given variable. The two-sample Kolmogorov–Smirnov statistic is

$$D_{NM} = \max_x |F_N(x) - G_M(x)| \quad (19.11)$$

where $F_N(x)$ is the cumulative sample distribution of a set of N observations, and $G_M(x)$ is the cumulative sample distribution of another independent set of M observations. In this case, there is no parent model used in the testing, and the test is completely non-parametric. The statistic D_{NM} measures the maximum deviation between the two cumulative distributions, and by nature it is a discrete distribution. It is possible to show that the distribution of the statistic is the same as in (19.9), provided that the size N of the Kolmogorov D_N statistic is replaced with $MN/(M + N)$, which can be considered as the effective number of data points of the two distributions. As N and M become large, the null distribution of the D_{NM} statistic approaches the following function:

$$\lim_{N,M \rightarrow \infty} P\left(D_{NM} < z / \sqrt{\frac{MN}{M+N}}\right) = \Phi(z). \quad (19.12)$$

where $\Phi(z)$ is the same function defined in (19.8).

It was already shown that for a sample distribution with N measurements,

$$F_N(x) - F(x) = U_N(y) - U(y),$$

where U is a uniform distribution in $(0,1)$. Since

$$F_N(x) - G_M(x) = F_N(x) - F - (G_M(x) - G),$$

where $F = G$ is the parent distribution, it follows that

$$F_N(x) - G_M(x) = U_N - V_M,$$

where U_N and V_M are the sample distributions of two uniform variables. Therefore the statistic

$$D_{NM} = \max_x |F_N(x) - G_M(x)|$$

is independent of the parent distribution, same as for the statistic D_N .

Next it is useful to show how the term $M N / (M + N)$ originates. It is clear that the expectation of $F_N(x) - G_M(x)$ is zero, at least in the limit of large N and M ; the second moment can be calculated as

$$\mathrm{E}[(F_N - G_M)^2] = \mathrm{E}[(F_N - F)^2] + \mathrm{E}[(G_M - G)^2].$$

In fact, since $F_N(x) - F(x)$ is independent of $G_M(x) - G(x)$, their covariance is zero. The two expectations can be evaluated as

$$\begin{aligned} \mathrm{E}[(F_N(x) - F(x))^2] &= \mathrm{E}\left[\frac{1}{N}(\{\text{Number of } x_i \text{'s} < x\} - N F(x))^2\right] = \\ &\quad \frac{1}{N^2} \mathrm{E}[(\{\text{Number of } x_i \text{'s} < x\} - \mathrm{E}[\{\text{Number of } x_i \text{'s} < x\}])^2]. \end{aligned}$$

For a fixed value of x , the variable $\{\text{Number of } x_i \text{'s} < x\}$ is a binomial distribution in which “success” is represented by one measurement being $< x$, and the probability of success is $p = F(x)$. The expectation in the equation above is therefore equivalent to the variance of a binomial distribution with N tries, for which $\sigma^2 = N p(1 - p)$, leading to

$$\mathrm{E}[(F_N(x) - F(x))^2] = \frac{1}{N} F(x)(1 - F(x)).$$

It follows that

$$\mathrm{E}[(F_N(x) - G_M(x))^2] = \left(\frac{1}{N} + \frac{1}{M}\right) F(x)(1 - F(x))$$

A simple way to make the mean square of $F_N(x) - G_M(x)$ independent of N and M is to divide it by $\sqrt{1/N + 1/M}$. This factor therefore arises as a necessary condition for the variable $\sqrt{NM/(N+M)} D_{NM}$ to be independent of N and M .

Finally, it can be shown that $\sqrt{NM/(N+M)} D_{NM}$ is distributed in the same way as $\sqrt{N} D_N$, at least in the asymptotic limit of large N and M . Using the results from the D_N distribution derived in the previous section,

$$\max_x \left| \sqrt{\frac{MN}{M+N}} (F_N(x) - G_M(x)) \right| = \max_{0 \leq y \leq 1} \left| \sqrt{\frac{MN}{M+N}} (U_N - V_M) \right|,$$

where the variable in the right-hand side term can be rewritten as the sum of the two variables

$$\sqrt{\frac{M}{M+N}} (\sqrt{N}(U_N - U)) + \sqrt{\frac{N}{M+N}} (\sqrt{M}(V_M - V)).$$

Using the central limit theorem, it can be shown that the two variables $\alpha = \sqrt{N} (U_N - U)$ and $\beta = \sqrt{M} (V_M - V)$ have the same distribution, which tends to a Gaussian in the limit of large number of measurements. It follows that the statistic of interest can be written as

$$\sqrt{\frac{MN}{M+N}}(F_N(x) - G_M(x)) = \sqrt{\frac{M}{M+N}}\alpha + \sqrt{\frac{N}{M+N}}\beta$$

For two independent and identically distributed Gaussian variables α and β , the variable $a \cdot \alpha + b \cdot \beta$ is distributed like α , provided that $a^2 + b^2 = 1$. It is therefore possible to conclude that, in the asymptotic limit of a large number of measurements,

$$\sqrt{\frac{MN}{M+N}} \cdot \max_x |(F_N(x) - G_M(x))| \sim \sqrt{N} \cdot \max_x |(U_N - U)|.$$

For the two-sample Kolmogorov–Smirnov test it is therefore possible to use the same critical values as in the Kolmogorov–Smirnov one-sample test, provided N in (19.8) is substituted with $M N / (M + N)$, and that both numbers are sufficiently large.

Example 19.4 The two-sample Kolmogorov–Smirnov statistic D_{NM} can be used to compare the data from Tube 1 and Tube 2 of Thomson’s experiment to measure the ratio m/e of an electron. The statistic is $D_{NM} = 0.75$, indicating that the two sets of measurements are not in agreement with one another. See Problem 19.2 for a quantitative analysis of this test. ◇

Summary of Key Concepts for this Chapter

F-Test: The F -statistic

$$F = \frac{\chi_1^2/f_1}{\chi_2^2/f_2}$$

is used to compare two *independent* χ^2 statistics with f_1 and f_2 degrees of freedom.

F-Test for an additional component: The significance of an additional model component with p parameters can be tested using the statistic

$$F = \frac{\Delta\chi^2/p}{\chi_{\min}^2/(N-m)}$$

when the additional component is *nested* within the general model with m parameters, i.e., the simplified model can be obtained by fixing one or more parameters of the more general model.

Kolmogorov–Smirnov test to compare data to a model: A test statistic to compare the sample distribution of a one-variable dataset to a fixed model,

$$D_N = \max_x |F_N(x) - F(x)|.$$

Two-sample Kolmogorov–Smirnov test: A non-parametric test statistic to compare two sample distributions,

$$D_{NM} = \max_x |F_M(x) - G_N(x)|.$$

Problems

19.1 ■ Using the data from Thomson's experiment of Sect. 2.4, determine the values of the Kolmogorov–Smirnov statistics D_N for the measurements of Tube 1 and for the measurements of Tube 2, when compared with a Gaussian model for the measurement with $\mu = 5.7$ and $\sigma^2 = 1$. Determine at what confidence level the hypothesis that the two measurements are consistent with the model can be rejected.

19.2 ■ Using the data from Thomson's experiment of Sect. 2.4, determine the values of the two-sample Kolmogorov–Smirnov statistic D_{NM} for comparison between the two measurements. Determine at what confidence level the hypothesis that the two measurements are consistent with one another can be rejected.

19.3 ■ Use the five-point data of Example 12.2, and a linear model for the last three data points. Use the F -test for an additional model component to test whether the simpler constant model for the last three data points can be rejected at the 99% confidence level. In this problem the first two data points are ignored.

19.4 A dataset with $N = 5$ points is fit to a linear model with a best-fit statistic of $\bar{\chi}_{\min}^2$. When adding one nested parameter to the model ($p = 1$), the fit statistic is χ_{\min}^2 . Determine the minimum value of the ratio $\bar{\chi}_{\min}^2/\chi_{\min}^2$ that is required for the additional parameter to be significant at the 90% confidence level.

19.5 A Gaussian dataset is fit to a parametric model and it results in a minimum χ^2 fit statistic of $\chi_1^2 = 10$ for 5 degrees of freedom. The same dataset is also fit to a different model, with $\chi_2^2 = 5$ for 4 degrees of freedom. Determine which model is acceptable at the 90% confidence, and whether the F -test can be used to choose which of the two models is a better fit.

19.6 A Gaussian dataset with N measurements is fit to a parametric model with $m - 1$ free parameters, and it yields a fit statistic $\bar{\chi}_{\min}^2$ that is acceptable at the 90% confidence level. The model is then augmented with a one-parameter nested component for a total of m parameters, resulting in a fit statistic χ_{\min}^2 . Determine a sufficient condition for the value of the minimum $\Delta\chi^2 = \bar{\chi}_{\min}^2 - \chi_{\min}^2$ that, in the limit of a large number of degrees of freedom, provides a 90% degree of confidence that the additional parameter is significant.

Part III

Monte Carlo Methods

Chapter 20

Monte Carlo and Re-sampling Methods



Abstract The term *Monte Carlo* refers to the use of random variables to evaluate quantities such as integrals or parameters of fit functions that are typically too complex to evaluate via other analytic methods. This chapter presents elementary Monte Carlo methods that are of common use in data analysis and statistics, in particular the bootstrap and jackknife methods to estimate parameters of fit functions.

20.1 What is a Monte Carlo Analysis?

The term *Monte Carlo* derives its name from a locality in the Principality of Monaco, known for its resorts and casinos. In statistics and data analysis, Monte Carlo is used as an umbrella term to indicate the use of computer-assisted numerical methods, typically with the aid of random numbers. Although electronic computers have become available only since the latter part of the twentieth century, the idea of repeating an experiment to determine a sample distribution is at the core of the theory of probability and statistic. A famous experiment that can be used to illustrate the use of Monte Carlo methods is the *Buffon coin drop*, named after the eighteenth-century scientist G.L. Le Clerc *comte de Buffon*. Buffon proposed a simple experiment where a coin is dropped on a floor covered with square tiles, with the goal to calculate the probability that the coin crosses the gaps between the tiles [65]. Once the sizes of the coin and of the tiles are specified, the probability can be calculated analytically using simple geometric arguments, and confirmed experimentally by repeating the experiment many times. Alternatively, one may perform a *numerical simulation* of the experiment by randomizing the landing coordinates of the coin, and at each step in the simulation evaluate analytically whether the coin crosses the gaps. These

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-981-19-0365-6_20.

simulated repetitions of the experiment are then used to determine the probability of interest, same as if the experiment had been performed in reality.

The origin of the expression “Monte Carlo” to signify such computer-assisted numerical calculations is usually attributed to the *Los Alamos Laboratories* scientists N. Metropolis and S. Ulam who developed this technique after World War II [72], who also attributed this term to Ulam’s uncle who would borrow money from relatives because he “just had to go to Monte Carlo” [70]. Traditional Monte Carlo methods include numerical integration of functions that can be graphed but that don’t have a simple analytic solution. Another problem that benefits by the use of random numbers is the estimation of uncertainties in the best-fit parameters of analytical models used to fit data, in cases when an analytical solution for the error in the parameters is not available. In many of those cases, the bootstrap or the jackknife methods can be used to obtain reliable estimates for those uncertainties. Among many other applications, the Markov chain Monte Carlo method stand out as a class of Monte Carlo methods that is now commonplace across many fields of research. The theory of Markov chains (Chap. 21) dates to the early twentieth century, yet only recently it has found widespread use as Monte Carlo Markov chains (Chap. 22) because of the computational power necessary to implement the method.

20.2 Traditional Monte Carlo Integration

A common numerical task is the evaluation of the integral of a function $g(x)$ for which an analytic solution is either unavailable or too complicated to calculate exactly,

$$I = \int_A g(x)dx. \quad (20.1)$$

The goal is to derive a method to approximate this integral by randomly drawing N samples from the support A . For simplicity it is assumed that the support of the function is the interval between a and b . The method starts with the drawing of N samples from a uniform distribution between these two values, therefore assuming that the variable X has a probability distribution function $f(x)$ given by

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{if } a \leq x \leq b \\ 0 & \text{otherwise.} \end{cases} \quad (20.2)$$

Recall that for a random variable X with continuous distribution $f(x)$, the expectation is defined according to (2.7) as

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f(x)dx$$

and it can be estimated via its sample mean

$$\mathbb{E}[X] \simeq \frac{1}{N} \sum_{i=1}^N x_i$$

using N independent samples of the variable. For the purpose of evaluating the integral (20.1), the expectation of a function $g(x)$ of the random variable X is

$$\mathbb{E}[g(x)] = \int_{-\infty}^{\infty} g(x) f(x) dx,$$

which can be estimated using the law of large numbers (Sect. 2.3.2) as

$$\mathbb{E}[g(x)] \simeq \frac{1}{N} \sum_{i=1}^N g(x_i). \quad (20.3)$$

In the limit of a large number of measurements, these equations can be thus used to approximate the integral of interest as a simple sum:

$$I = (b - a) \times \mathbb{E}[g(x)] \simeq \frac{b - a}{N} \times \sum_{i=1}^N g(x_i). \quad (20.4)$$

Equation 20.4 can be implemented by drawing N random uniform samples x_i from the support, then calculating the samples $g(x_i)$ and evaluating the sum. This is the basic *Monte Carlo integration* method, and it can be easily implemented by using a random number generator for the uniform distribution. The method can be generalized to more than one dimension. If the support A has an n -dimensional volume V , then the integration of an n -dimensional function $g(x)$ is given by the following sum:

$$I \simeq \frac{V}{N} \times \sum_{i=1}^N g(x_i). \quad (20.5)$$

It is clear that the precision in the evaluation of the integral depends on the number of samples drawn. The error made by this method of integration can be estimated using the following interpretation of (20.4) or (20.5). Since I is estimated via the sample mean of the function $g(x)$, the variance of I is proportional to the variance of the sample mean of $g(x)$, which can be estimated via

$$s_{\bar{g}}^2 = \frac{s_g^2}{N} = \frac{1}{N} \left(\frac{1}{N-1} \sum_{i=1}^N (g(x_i) - \bar{g})^2 \right).$$

The term in parenthesis is the sample variance of the function $g(x)$, which is an unbiased estimate of the parent variance and immediately calculated from the N samples. The additional factor of $1/N$ accounts for the fact that the variable of interest is the sample mean of N measurements. According to (20.5), the sample variance of I is therefore estimated as

$$s_I^2 = V \times s_g^2, \quad (20.6)$$

with the estimated standard error s_I therefore decreasing with sample size like the square root of N .

Example 20.1 (*Estimate of π using Monte Carlo integration*) The number π is usually defined as the ratio of the circumference to the diameter of a circle, or as the ratio of the area to the square of its radius. For example, a circle of unit radius is represented by $x^2 + y^2 = 1$ and, given the symmetry of a circle, it is possible to set

$$\frac{\pi}{4} = \int_0^1 \sqrt{1 - x^2} dx \quad (20.7)$$

where the integral represents the area of one-fourth of a circle of unit radius. This integral can be evaluated with a simple Monte Carlo method by drawing N samples x_i from a uniform distribution between 0 and 1 (the support of the integral), resulting in an estimate of the number as

$$\pi \simeq 4 \cdot \frac{1}{N} \sum_{i=1}^N g(x_i),$$

where $g(x_i) = \sqrt{1 - x_i^2}$ is a random sample for the integrand function. The variance of the estimate is approximately

$$s_\pi^2 = \frac{4}{N} \cdot \frac{\sum_{i=1}^N (g(x_i) - \bar{g})^2}{N - 1}$$

For example, a sample of $N = 100$ random numbers yields an estimate of 3.263 ± 0.080 , a sample of $N=1,000$ yields 3.192 ± 0.028 , or a sample of $N=10,000$ yields 3.1397 ± 0.0090 , to be compared with the known value of approximately 3.1416. Of course, different samples will yield different estimates, the numbers reported above refer to two specific samples drawn with the aid of a computer code that implements the equations shown in this example.

Note that the integral in (20.7) can be evaluated analytically with the substitution $x = \cos \theta$, and realizing that the integrals of $\sin^2 \theta$ and $\cos^2 \theta$ between 0 and $\pi/2$ are identical, due to the symmetry of the trigonometric functions. However, such analytic solution does not resolve the problem of finding the numerical value of π , which can be accomplished with the simple Monte Carlo integration presented in this example.



20.3 Hit-or-Miss Monte Carlo Methods

Another method to integrate a function, or to perform related mathematical operations, relies on the geometry of the problem and can be shown by way of an example. Consider the measurement of the area of a circle of radius R . One can draw a random sample of N values from two independent uniform distributions between $-R$ and R , as shown in Fig. 20.1, and count all the points that fall within the circle (n) according to the condition $x_i^2 + y_i^2 \leq R^2$. Accordingly, the area of the circle is estimated as

$$A = \frac{n}{N} \times V \quad (20.8)$$

in which $V = R^2$ is the area or volume sampled by the two random variables. In the case of a circle of radius $R = 1$, the area in parameter space is $V = 4$. Since the area is defined as $A = \pi R^2$, this method provides an approximation to the number π . This method is usually referred to as a *hit-or-miss Monte Carlo*, or *dart* Monte Carlo.

An estimate of the uncertainty associated with this Monte Carlo method is provided by the fact that n can be thought of as the number of successes in N tries of a binary experiment. The parent probability of success p is related to the quantity to estimate (in this case, the area A), and of course its exact value is, in general, unknown. A simple estimate is provided by the estimated rate of success $\hat{p} \simeq n/N$, leading to an estimate for the variance of n as

$$\hat{\sigma}_n^2 = N \cdot \frac{n}{N} \cdot \frac{N-n}{N},$$

according to the usual formula for the variance of a binomial distribution. This estimated variance can be used as a measure of the uncertainty in the Monte Carlo estimate of A ,

$$\hat{\sigma}_A^2 = \frac{V^2}{N^2} \times \hat{\sigma}_n^2 = \frac{V^2}{N^2} \cdot \frac{n(N-n)}{N}, \quad (20.9)$$

with the variance of A decreasing with sample size approximately like $1/N$, similar to the variance (20.6) for the case of the traditional Monte Carlo integration.

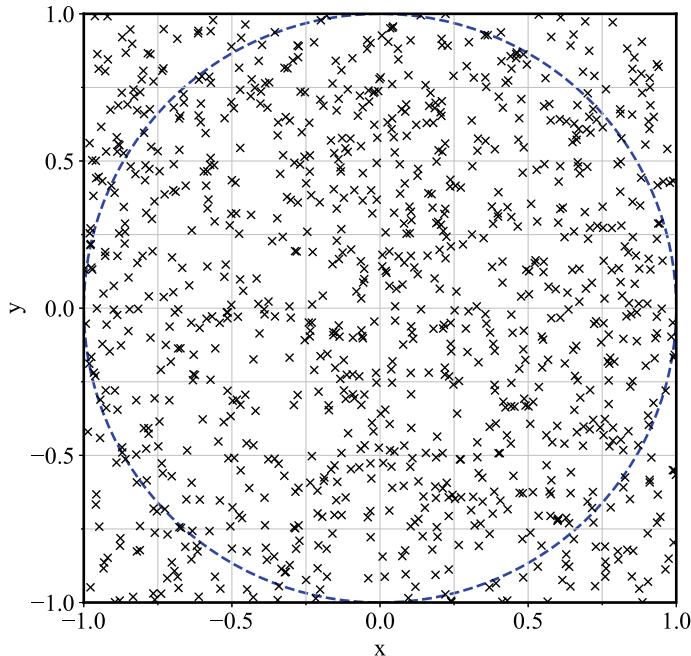


Fig. 20.1 Hit-or-miss Monte Carlo simulation to calculate the area of a circle (also a simulation of the number π), with $N=1,000$ simulated points. In this realization, 806 of the 1,000 points fell within the boundaries of the circle

Example 20.2 (*Estimate of π using hit-or-miss Monte Carlo*) Figure 20.1 shows a Monte Carlo simulation of the area of a circle of unit radius, using 1,000 random numbers drawn in a box of linear size 2. The realization shown in the figure has $n = 806$ points within the unit circle, resulting in an estimate of the area of the circle of $A \simeq 3.224 \pm 0.050$. This Monte Carlo simulation is also an estimate of the number π . \diamond

It is instructive to compare the relative error in the estimates of π using the traditional Monte Carlo integration provided in Example 20.1 and the estimate using the hit-or-miss method in Example 20.2. For the same number of samples drawn, namely $N=1,000$, the variance of the traditional method (20.6) is significantly smaller than that of the hit-or-miss method (20.9), with a ratio of variances of approximately $(0.028/0.50)^2 = 0.32$. This comparison shows that the hit-or-miss method Monte Carlo method is generally less efficient than the traditional method. Additional discussion on the comparison of Monte Carlo methods and their efficiency can be found, for example, in the textbook by J.M. Hammersley and D.C. Handscomb [47].

20.4 Simulation of Random Variables

A method for the simulation of a random variable was discussed in Sect. 5.3. Since the generation of random samples from a uniform random variable was involved, this method also falls under the category of Monte Carlo simulations. The method is based on the relationship (5.24),

$$X = F^{-1}(U),$$

where F^{-1} represents the inverse of the cumulative distribution of the target variable X , and U represents a uniform random variable between 0 and 1. The practical use of this equation is straightforward: N independent samples u_i are drawn from a standard uniform variable, and each is transformed to x_i according to (5.24). The transformed set of x_i values is a random sample from the variable of interest. This method relies on the availability of a cumulative distribution F and its inverse. Sect. 5.3 provided an example on how to use (5.24) to simulate an exponential distribution, which has a simple analytic function for its cumulative distribution. A Monte Carlo simulation of the exponential distribution based on those equations is illustrated in the following example.

Example 20.3 (*Simulation of an exponential distribution*) The cumulative distribution of an exponential distribution with parameter λ is

$$F(x) = 1 - e^{-\lambda x}$$

with $x \geq 0$ representing possible values of the exponential variable. Its inverse, also known as the quantile function, is

$$x = -\frac{\ln(1-u)}{\lambda}$$

where $0 \leq u \leq 1$ represents possible values of the standard uniform variable. As an example, using 1,000 random values u_i drawn from a uniform variable yields the sample distribution function for the exponential variable X shown in Fig. 20.2, in agreement with the theoretical distribution function. ◇

The Gaussian distribution is perhaps the most common variable in many statistical applications, and its generation cannot be accomplished by (5.24) since the cumulative distribution is a special function and $F(x)$ does not have a close form. A method to overcome this limitation was discussed in Sect. 5.3.2, and it consists of using two uniform random variables U and V to simulate two standard Gaussians X and Y of zero mean and unit variance via (5.27),

$$\begin{cases} X = \sqrt{-2 \ln(1-U)} \cdot \cos(2\pi V) \\ Y = \sqrt{-2 \ln(1-U)} \cdot \sin(2\pi V). \end{cases} \quad (20.10)$$

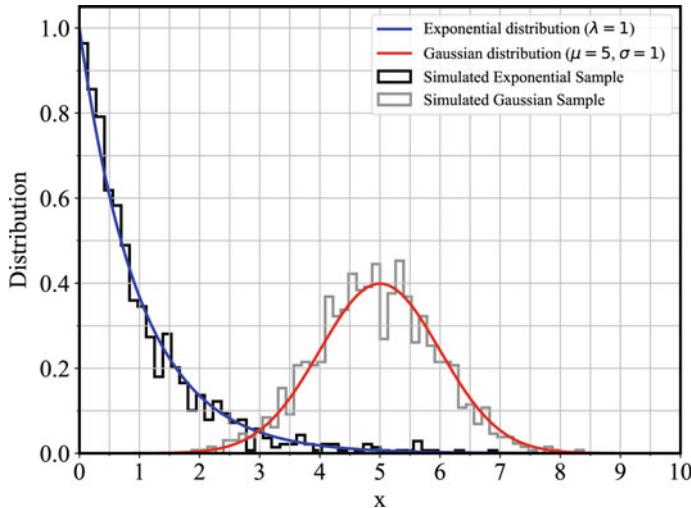


Fig. 20.2 Monte Carlo simulation of the probability distribution function of an exponential distribution and of a Gaussian distribution, using 1000 samples for each

A Gaussian X' of mean μ and variance σ^2 is related to the standard Gaussian X by the transformation

$$X = \frac{X' - \mu}{\sigma},$$

and therefore it can be simulated via

$$X' = \left(\sqrt{-2 \ln(1 - U)} \cdot \cos(2\pi V) \right) \sigma + \mu. \quad (20.11)$$

Figure 20.2 shows a simulation of a Gaussian distribution function using (20.11).

20.5 Re-sampling Methods

The use of a sample of N measurements to estimate parameters of a distribution, or parameters of a parent model, is a common occurrence in statistics. For example, the measurements can be used to estimate the mean of a parent random variable using the method of maximum likelihood (e.g., Sect. 6.2), or to estimate parameters of a fit function with a regression (Chap. 11). A known issue with certain methods of estimation is that the statistic used for the estimate, or estimator, is *biased*, in that the expectation of the estimator is not equal to the parameter being estimated. An example of this problem is the maximum likelihood estimator of the variance of a Gaussian variable (see Sect. 6.2.1), which requires the replacement of the factor of

N with $N - 1$ to avoid a bias in the estimate. In other situations, removing the bias is a more complex task.

An idea to remove the bias in an estimator, while at the same time providing uncertainties on the estimate, is that of *re-sampling* the original measurements, for example by using a subset of the sample or using randomly drawn measurements from the original sample. This idea was pioneered by H.M. Quenouille [82, 83] and by J.W. Tukey [99] in what became the *jackknife method* (for a review, see [73]). The two methods presented in this section, the jackknife and a more general *bootstrap* method, are among the most common re-sampling techniques. Among their numerous applications, these methods are especially useful to estimate best-fit parameters and their uncertainties in the fit to two-variable datasets. It was shown in Chap. 11 that the best-fit parameters and their uncertainties can be estimated analytically, for example, in the case of a linear regression with known errors in the dependent variable. In those cases, the exact analytical solution is typically the most straightforward to implement. These re-sampling methods can be useful to obtain uncertainties in the best-fit parameters for two-dimensional data with no known variance, a situation where the maximum likelihood method offered no direct guidance (see Sect. 12.2).

Re-sampling methods such as the bootstrap and the jackknife can be applied in many ways, and readers interested in this topic can refer to the vast literature on the subject. It may be useful to provide a historical perspective on this subject with a quote from Quenouille's *Notes on Bias in Estimation* [83] that aptly summarizes the typical estimation framework:

1. One of the commonest problems in statistics is, given a series of observations x_1, x_2, \dots, x_n , to find a function of these, $t_n(x_1, x_2, \dots, x_n)$, which should provide an estimate of an unknown parameter θ .

The desirable properties of estimation procedures have been discussed fully elsewhere. They are:

- (a) That the estimator should be efficient according to some definition of efficiency previously arranged. Most commonly, the reciprocal of the variance of the estimates is taken as a measure of its efficiency, as this is most useful where central limit theory may be relevant.
- (b) That the estimator should utilize all the information contained in the observations, x_1, x_2, \dots, x_n concerning the parameter θ . This is not always possible, but, if such an estimator exists, it is called sufficient.
- (c) That the estimator should be consistent, i.e., t_n converges in some probabilistic sense to 0, usually $\lim_{n \rightarrow \infty} t_n \rightarrow \theta$.
- (d) That the estimator should be unbiased, i.e., $E(t_n) = \theta$.

The method of maximum likelihood is popular in that it satisfies properties (a) to (c), whence, by evaluating $E(t_n)$, an unbiased statistic may be derived. That such evaluation is necessary is obvious when it is remembered that $\psi(t_n)$ is, by the same theory, the estimator of $\psi(\theta)$, and, since in general $E[\psi(t_n)] \neq \psi[E(t_n)]$, it will be the exception rather than the rule for a maximum likelihood estimator to be unbiased. Provided the exceptions may be simply evaluated no real difficulty arises. However, often the complexity of the evaluation presents a major drawback and some simple approach is then desirable.

20.6 The Jackknife Method

The Quenouille–Tukey jackknife method is one of the earliest resampling methods, and it makes use of just N re-sampled datasets to estimate parameters of interest from the given set of N measurements of one or more random variables. A case of particular interest for the application of the jackknife method is that of a two-dimensional dataset used for regression analysis, such as the linear regression, where the parameters of interest are the adjustable model parameters. The relative simplicity of the jackknife method makes it a very efficient method to estimate regression parameters and their uncertainties. The jackknife can be used for any type of estimation, not just the regression, but its use is illustrated for the case of regression with two-dimensional data because of the relevance of this application.

The jackknife method consists of the following steps, where Z refers to the original sample of N measurements:

1. Generate a re-sampled dataset Z_j , obtained by deleting the j th element from the dataset. Each of the N re-sampled dataset has therefore dimension $N - 1$.
2. Each dataset Z_j is used to estimate the parameters of interest in the same way as from the original data Z . For example, if the method is used to estimate the two parameters of the linear regression, the re-sampled dataset Z_j yields the best-fit values of the linear model, a_j and b_j . The parameters estimated from Z_j will be cumulatively referred to as $\hat{\theta}_j$.
3. The parameters of interest are also calculated from the full-dimensional dataset Z . The best-fit parameters from Z are referred to as $\hat{\theta}$.
4. For each dataset Z_j , the *pseudo-values* θ_j^* are defined as

$$\theta_j^* = N\hat{\theta} - (N - 1)\hat{\theta}_j \quad (20.12)$$

5. The jackknife estimator of the parameters of interest and their uncertainty is given by the following equations:

$$\begin{cases} \theta^* = \frac{1}{N} \sum_{j=1}^N \theta_j^* \\ \sigma_{\theta^*}^2 = \frac{1}{N(N-1)} \sum_{j=1}^N (\theta_j^* - \theta^*)^2. \end{cases} \quad (20.13)$$

Steps 1–5 constitute the jackknife methods to estimate the unknown parameters of the model. The simplicity of this method therefore makes it a popular choice for a number of applications.

In addition to its simplicity, the jackknife estimator is designed to remove a type of bias proportional to $1/N$ that might be present in the original estimator $\hat{\theta}$,

$$E[\hat{\theta}] = \theta + \frac{a}{N} + \mathcal{O}\left(\frac{1}{N^2}\right), \quad (20.14)$$

where a is a constant. The removal of the a/N bias term can be immediately seen by evaluating the expectation of θ^* ,

$$E[\theta^*] = N E[\hat{\theta}] - \frac{N-1}{N} \sum_{j=1}^N E[\hat{\theta}_j].$$

According to (20.14), the statistic $\hat{\theta}_j$, which is calculated from $N-1$ of the N independent measurements, has an expectation

$$E[\hat{\theta}_j] = \theta + \frac{a}{N-1} + \mathcal{O}\left(\frac{1}{N^2}\right),$$

and this leads to

$$E[\theta^*] = \theta + \mathcal{O}\left(\frac{1}{N^2}\right)$$

to show that the jackknife estimator is unbiased within terms of order $1/N^2$.

It is possible to show that the jackknife is usually an unbiased estimator for the linear regression, in the asymptotic limit of a large number of measurements, although the variance of the estimator may be larger than that of a least-square estimate [105]. Bias may be introduced when the X measurements are *unbalanced*, in the sense of an uneven coverage of the variable range. In general, it is therefore possible that the jackknife estimator may be biased, and as such it should be used with caution. There are modifications to the simple jackknife method of (20.13) that attempt to further reduce the bias, such as the *second-order jackknife* (for example, see the review by R. Miller [73]).

Example 20.4 (Jackknife on the Hubble data) The Hubble dataset Z consists of $N = 10$ measurements of the magnitude m and the logarithm of the velocity $\log v$, as shown in Fig. 11.2. Given that error bars on the dependent variable $\log v$ were not given, it is assumed that the uncertainties have a common value for all measurement (and therefore the value of the error is irrelevant for the determination of the best-fit parameters). The jackknife method can be used to estimate the best-fit parameters of the fit to a linear model of m versus $\log v$, using N re-sampled datasets Z_j with size $N-1$. For each re-sampled dataset, the linear regression yields the best-fit values of the parameters a_j and b_j . According to (20.13), the estimates of the model parameters are $a^* = 0.52 \pm 0.13$ and $b^* = 0.199 \pm 0.008$, where the uncertainties are the standard deviations of the estimated parameters. These estimates are in very

good agreement with the direct fit to the original dataset (with parameters $a = 0.548$ and $b = 0.197$), which, however, did not provide uncertainties in the fit parameters.



20.7 The Bootstrap Method

The bootstrap is a re-sampling method that, like the jackknife, can be used to estimate parameters of interest that depend on the data points. The data under consideration are the usual N measurements of one or more random variables. The bootstrap method was introduced by B. Efron [25], and in its primitive form it consists in generating a *bootstrap sample* of N measurements, sampled *with replacement* from the original data. This implies that a number of the original measurements are likely to be repeated, and some omitted, in the re-sampled data. The bootstrap sample is then used to estimate the sampling distribution of the parameter of interest via the bootstrap distribution, i.e., the distribution of the parameter of interest evaluated for the bootstrap sample. In principle, the distribution can be evaluated by theoretical means, but more often it is evaluated by means of a Monte Carlo approximation that consists of using a large number of bootstrap samples. The repetition of many bootstrap samples by numerical means is the most common way to implement a bootstrap analysis.

One of the applications of interest to the data analyst is the fit of two-dimensional data, where the parameters to estimate are the adjustable parameters of the model. A review by C.F.J. Wu [105] provides an introduction to the use of re-sampling methods for regression analysis, including the bootstrap. A typical bootstrap regression analysis consists of the following steps:

1. Draw at random N data points from the original set Z , with replacement, to form a synthetic bootstrap dataset Z_i . The new dataset, therefore, has the same dimension as the original set, but a few of the original points may be repeated, and a few are missing.
2. From the bootstrap dataset Z_i , calculate the parameter(s) of interest θ_i . For example, the two parameters of a linear regression $\theta_i = (a_i, b_i)$ can be calculated using the maximum likelihood method.
3. Steps (1) and (2) are repeated a large number of times, say $i = 1, \dots, N_{boot}$.
4. At the end of the process, the parameter values θ_i are used to approximate the sampling distribution of the parameters, and therefore obtain an estimate of the best-fit values and confidence intervals.

One advantage of the bootstrap method for regression analysis, also shared with the jackknife, is that it can be used even when the errors on the data points are not available, which is a very common occurrence. In this situation, the direct maximum likelihood method applied to the original set Z alone would not provide uncertainties in the best-fit parameters, as explained in Chap. 11. Since at each Monte Carlo iteration the best-fit parameters alone must be evaluated, a dataset without errors in

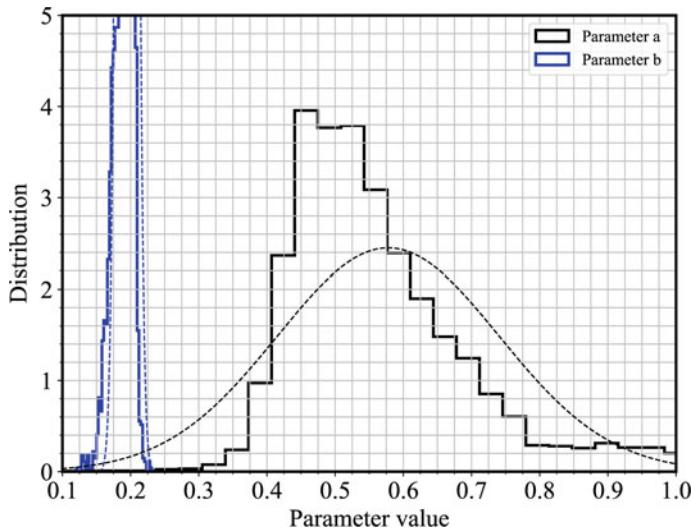


Fig. 20.3 Results of the Monte Carlo bootstrap method applied to the data from Hubble’s experiment. The dashed curves are Gaussian distributions evaluated for the sample mean and variance of the distributions

the dependent variable can still be fit to find the best-fit parameters, and the bootstrap method will provide an estimate of the uncertainties.

Example 20.5 (*Bootstrap Analysis of Hubble’s Data*) To perform a Monte Carlo bootstrap of the Hubble data, $N_{boot}=10,000$ synthetic datasets were randomly drawn from the original dataset with $N = 10$ measurements, with replacement. This means that each synthetic Monte Carlo dataset Z_i typically has a few of the original data points repeated, and a few are missing. Given the small number of data points in the data, there is only a relatively small number of possible synthetic datasets that can be drawn with replacement, and therefore many of the synthetic datasets are identical. For each dataset Z_i , a linear regression is used to estimate the best-fit values of the parameters a_i and b_i . The resulting sample distributions of the parameters are shown in Fig. 20.3, along with their normal approximations (with sample means and standard deviations $a = 0.578 \pm 0.163$ and $b = 0.195 \pm 0.010$). The sample distribution of parameter a has a median of $a = 0.538$ and a 68% central range of 0.450–0.696, and the sample distribution of b has a median of $b = 0.197$ and a central range of 0.188–0.202. The figure shows that the sample distributions are not well approximated by a normal distribution, implying that the 68% confidence intervals that are inferred by the normal approximation are biased. It is therefore appropriate, in this case, to use the median of the distribution as the “best-fit” value for the parameter, and the corresponding 68% central confidence interval. The asymmetry in the sample distributions does not improve with a larger number of bootstrap samples, since there is only a finite number of synthetic datasets that can be generated at random, with replacement, from the original dataset (see Problem 9.1). ◇

For the linear regression with equal variances, the bootstrap is an unbiased method to estimate the parameters and their variance. However, similar to the jackknife, bootstrap estimates for more complex models may become biased [25, 105]. The possibility of bias and uncertainties in the error estimates should therefore always be considered when interpreting results from a bootstrap analysis.

It is useful to illustrate the bootstrap of a one-dimensional dataset to estimate the mean of a random variable X from N independent and identically distributed measurements. Usually the mean would be directly estimated via a direct calculation of the sample mean from the given sample, but this illustration serves the purpose to describe the bootstrap method in more detail. The sample mean for a given bootstrap dataset Z_i can be written as

$$\bar{x}_i = \frac{1}{N} \sum_{j=1}^N x_j n_{ij} \quad (20.15)$$

where n_{ij} is the number of occurrences of data point x_j in the synthetic set Z_i . A value of $n_{ij} = 0$ means that x_j was not selected for the set, $n_{ij} = 1$ it means that there is just one occurrence of x_j (as in the original set), and so on. The number n_{ij} is a random variable that is distributed like a binomial with $p = 1/N$, since the drawing is done at random and with replacement. Its expectation and variance are therefore

$$\begin{cases} E[n_{ij}] = Np = 1 \\ \text{Var}(n_{ij}) = Np(1 - p) = \frac{N-1}{N}. \end{cases} \quad (20.16)$$

It follows that \bar{x}_i is an unbiased estimator of the sample mean,

$$E[\bar{x}_i] = \frac{1}{N} \sum_{j=1}^N x_j E[n_{ij}] = \bar{x}. \quad (20.17)$$

The expectation operator used in the equation above relates to the way in which a specific synthetic dataset can be drawn, i.e., indicates an “average” over a specific dataset with respect to the binomial distribution for n_{ij} . The operation of expectation should also be repeated to average over all possible datasets Z consisting of N measurements of the random variable X , i.e., the expectation is relative to the independent and identically distributed x_j measurements. This expectation results in

$$E[\bar{x}] = \mu, \quad (20.18)$$

proving that the bootstrap estimator of the mean is unbiased. Although the same symbol is used for the expectation of (20.17) and (20.18), the two operations are therefore different in nature. Calculation of the variance of the sample mean of dataset Z_i is complicated by the fact that the random variables n_{ij} are not independent. In fact, they are related by

$$\sum_{j=1}^N n_{ij} = N, \quad (20.19)$$

and this enforces a negative correlation between the variables that vanishes only in the limit of very large N . It can be shown that the covariance between n_{ij} and n_{ik} , where $j \neq k$ and i labels the dataset, is given by

$$\sigma_{jk}^2 = -\frac{1}{N}. \quad (20.20)$$

The proof of (20.20) is left as an exercise, and it is based on the use of (20.19) and (4.3) (see Problem 20.2). The variance of \bar{x}_i can be calculated using (4.3), since \bar{x}_i is a linear combination of N random variables n_{ij} :

$$\text{Var}(\bar{x}_i) = \frac{1}{N^2} \left(\frac{N-1}{N} \sum_{j=1}^N x_j^2 - \frac{2}{N} \sum_{j=1}^N \sum_{k=j+1}^N x_j x_k \right)$$

where (20.16) and (20.20) were used. Finally, it is necessary to calculate the expectation of this variance, in the sense of varying the dataset Z itself:

$$\mathbb{E}[\text{Var}(\bar{x}_i)] = \frac{N-1}{N^3} \mathbb{E} \left[\sum_{j=1}^N x_j^2 \right] - \frac{2}{N^3} \left(\frac{1}{2} \sum_{j \neq k} \mathbb{E}[x_j x_k] \right) \quad (20.21)$$

The last sum in the equation above is over all pairs (j, k) ; the factor $1/2$ takes into account the double-counting of terms, and the sum contains a total of $N(N-1)$ identical terms. Since the measurements x_j, x_k are independent and identically distributed, $\mathbb{E}[x_j x_k] = \mathbb{E}[x_j]^2$, and it follows that

$$\mathbb{E}[\text{Var}(\bar{x}_i)] = \frac{N-1}{N^2} (\mathbb{E}[x_j^2] - \mathbb{E}[x_j]^2) = \frac{N-1}{N^2} \sigma^2 = \frac{N-1}{N} \sigma_\mu^2 \quad (20.22)$$

where σ^2 is the variance of the random variable X , and $\sigma_\mu^2 = \sigma^2/N$ is the variance of the sample mean of N measurements. The bootstrap estimator of the mean \bar{x}_i therefore yields an unbiased estimator of the parent mean, with an expected variance that is nearly identical to that of the sample mean.

Summary of Key Concepts for this Chapter

Monte Carlo: Any numerical method that makes use of random variables to perform calculations that are too complex to be performed analytically, such as Monte Carlo integration and hit-or-miss methods.

Jackknife method: A re-sampling method that makes use of N re-sampled datasets Z_j consisting of $N - 1$ measurements with the j th measurement deleted from the dataset Z .

Bootstrap method: A re-sampling method that consists of generating a synthetic dataset Z_i where N measurements are drawn at random, and with replacement, from the N original measurements of the dataset Z . The bootstrap method is commonly used as a Monte Carlo method by generating many datasets Z_i to estimate parameters of interest.

Problems

20.1 Calculate how many unique synthetic bootstrap datasets can be generated from a dataset Z with N data points. Notice that the order in which the data points appear in the dataset is irrelevant.

20.2 For a bootstrap dataset Z_i constructed from a set Z of N independent measurements of a variable X , show that the covariance between the number of occurrence n_{ij} and n_{ik} is given by (20.20),

$$\sigma_{jk}^2 = -\frac{1}{N}.$$

20.3 Perform a numerical simulation of the number π using the traditional Monte Carlo method of integration, and determine how many samples are sufficient to achieve an average precision of 0.1%. The first six significant digits of the number are $\pi = 3.14159$.

20.4 Perform a numerical simulation of the number π using the hit-or-miss Monte Carlo method. Determine how many samples are sufficient to achieve an average precision of 0.1%.

20.5 ■ Perform a bootstrap simulation of the linear fit of the Hubble data of Table 11.1, and find the 68% central confidence intervals on the parameters a and b .

20.6 ■ Consider the five-point dataset of data of Example 12.2.

- (a) Run a bootstrap simulation with $N=1,000$ iterations for the fit to a linear model, and plot the sample probability distribution function of the parameters a and b . Notice that re-sampled datasets with five identical points cannot be used for the linear fit and should be excluded from the simulation.
- (b) Calculate the mean and standard deviation from the a and b samples.
- (c) Calculate the median and 68% confidence intervals on the fit parameters, and describe why they differ from the mean and standard deviations calculated in (b).

20.7 ■ Consider the five-point dataset of data of Example 12.2, assuming that the errors in the dependent variable y are unknown. Estimate the values of a and b to the fit to a linear model using a jackknife method.

20.8 Given two uniform random variables U_1 and U_2 between $-R$ and $+R$, provide an analytic expression to simulate a Gaussian variable of mean μ and variance σ^2 .

20.9 The jackknife method is used to estimate the sample mean of N measurements. Show that the jackknife estimator is the same as the sample mean itself, and that therefore the method is unbiased for this application.

Chapter 21

Introduction to Markov Chains



Abstract The theory of Markov chains is rooted in the work of Russian mathematician Andrey Markov and has an extensive body of literature to establish its mathematical foundations. The availability of computing resources has recently made it possible to use Markov chains to analyze a variety of scientific data, making Markov chain Monte Carlo one of the most popular methods of data analysis. This chapter presents the key mathematical properties of Markov chains, necessary to understand its implementation as Markov chain Monte Carlo.

21.1 Stochastic Processes and Markov Chains

This section presents key mathematical properties of Markov chains. The treatment is somewhat theoretical but necessary to ensure that the applications to the analysis of data are consistent with the mathematics of Markov chains, which can be very complex. The goal is therefore that of defining and understanding a basic set of definitions and properties necessary to use Markov chains for the analysis of data, especially via the Monte Carlo simulations. There is an extensive literature on the subject of stochastic processes and Markov chain, and the interested reader is referred to textbooks on stochastic processes such as the ones by S. Ross [85] or D.R. Cox and H.D. Miller [20].

Markov chains are a specific type of stochastic processes, or sequence of random variables. A typical example of Markov chains is the *random walk*, where at each time step a person randomly takes a step in one of two possible directions, for example forward or backward. As time progresses, the location of the person is the random variable of interest, and the collection of such random variables forms a Markov chain. The ultimate goal of a Markov chain is to determine the *stationary*

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-981-19-0365-6_21.

distribution of the random variable. For the random walk, the goal is to determine the probability that, at a given time, the person is located n steps away from the starting point.

An application of interest to the data analyst is the regression or fit of a dataset to a parametric model with adjustable parameters. The goal is to create a Markov chain for each parameter of the model, in such a way that the stationary distribution for each parameter is the distribution function of the parameter. The chain will therefore result in the knowledge of the best-fit value of the parameter, and of confidence intervals, making use of the information provided by the dataset.

21.2 Mathematical Properties of Markov Chains

A stochastic process is defined as a sequence of variables X_t ,

$$\{X_t, \text{ for } t \in T\} \quad (21.1)$$

where t labels the sequence. The domain for the index t is indicated as T , to signify “time.” The domain is usually a subset of the real numbers ($T \subset \mathbb{R}$) or of the natural numbers ($T \subset \mathbb{N}$). As time progresses, the random variables X_t change value, and the stochastic process describes this evolution. A Markov chain is a particular stochastic process that satisfies the following properties:

1. The time domain is the natural numbers ($T \subset \mathbb{N}$), and each random variable X_t can have values in a countable set, e.g., the natural numbers or even a p -dimensional space (\mathbb{N}^p), but not real numbers (\mathbb{R}^p). A typical example of a Markov chain is one in which $X_n = \varepsilon_i$, where both n (the time index) and i (the index of the random variable) are natural numbers. Therefore a Markov chain takes the form of a sequence of random variables, identified by an integer:

$$X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow \dots \rightarrow X_n \rightarrow \dots$$

The random variable X_n describes the state of the system at time $t = n$. The fact that Markov chains must be defined by way of countable sets may appear an insurmountable restriction for many applications, since it would appear that the domain for a p -dimensional parameter space is \mathbb{R}^p , since parameters usually span a range of real numbers. While a formal extension of Markov chains to \mathbb{R}^p is also possible, this is not a complication for any practical application, since any parameter space can be somehow “binned” into a finite number of states. For example, the position of the person in a random walk was “binned” into a number of finite (or infinite but countable) positions, and a similar process can be applied to virtually any parameter of interest for a given model. This means that the variable under consideration can occupy one of a countable multitude

of states $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_i, \dots$, and the random variable X_n identifies the state of the system at time step n , for example $X_n = \varepsilon_i$.

2. A far more important property that makes a stochastic process a Markov chain is the fact that subsequent steps in the chain are only dependent on the current state of the chain, and not on any of its previous history. This “short memory” property is known as the *Markovian property*, and it is the key into the construction of Markov chains for the purpose of data analysis. In mathematical terms, given the present time $t = n$, the future state of the chain X_{n+1} at a time $t = n + 1$ depends only on the state at the present time (X_n), but not on past history. Much of the efforts in the construction of a Markov chain lie in the identification of a *transition probability* from state ε_i to state ε_j between consecutive time steps,

$$p_{ij} = P(X_{n+1} = \varepsilon_j / X_n = \varepsilon_i). \quad (21.2)$$

A Markov chain requires that this probability be time-independent, and therefore a Markov chain has the property of time homogeneity. Chapter 22 will show how the transition probability takes into account the likelihood of the data Z with the model.

The two properties described above result in the fact that Markov chain is a sequence of states determined by transition probabilities p_{ij} (also referred to as the *transition kernel*) that are fixed in time. The ultimate goal is to determine the probability to find the system in each of the allowed states. With an eye towards future applications for the analysis of data, each state may represent values of one or many parameters, and therefore a Markov chain makes it possible to reconstruct the probability distribution of the parameters.

Example 21.1 (*A binary Markov chain*) A binary Markov chain has only two possible states, ε_1 and ε_2 . The transition probabilities can be described by a transition matrix

$$P = \begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix} = \begin{bmatrix} 1 - \alpha & \alpha \\ \beta & 1 - \beta \end{bmatrix},$$

where α represents the probability of going from state 1 to state 2, and β the probability of going from state 2 to 1, between two successive states. ◇

Example 21.2 (*The Random Walk*) The *random walk* is a Markov chain that represents the location of a person who randomly takes a step of unit length forward with probability p , or a step backward with probability $q = 1 - p$ (typically $p = q = 1/2$). The state of the system is defined by the location i where the person is located at time $t = n$,

$$X_n = \{\text{Location } i \text{ along the } \mathbb{N}^+ \text{ axis}\},$$

where \mathbb{N}^+ indicates all positive and negative integers, including zero. For this chain, the time domain is the set of positive numbers ($T = \mathbb{N}$), and the position can be an integer of either sign (\mathbb{N}^+). The transition probability describes the fact that the person can only take either a step forward or backward:

$$p_{ij} = \begin{cases} p & \text{if } j = i + 1, \text{ or move forward} \\ q & \text{if } j = i - 1, \text{ or move backward} \\ 0 & \text{otherwise.} \end{cases} \quad (21.3)$$

The chain satisfies the Markovian property, since the transition probability depends only on its present position, and not on previous history. \diamond

Example 21.3 (The Ehrenfest Chain) Another case of a Markov chain is a simple model of diffusion, known as the *Ehrenfest chain*. Consider two boxes with a total of m balls. At each time step, one selects a ball at random from either box and replaces it in the other box. The state of the system can be defined via the random variable

$$X_n = \{\text{Number of balls in the first box}\}.$$

The random variable can have only a finite number of values $(0, 1, \dots, m)$. At each time step, the transition probability is

$$p_{ij} = \begin{cases} \frac{m-i}{m} & \text{if } j = i + 1 \text{ (box had } i \text{ balls, now has } i+1) \\ \frac{i}{m} & \text{if } j = i - 1 \text{ (box had } i \text{ balls, now has } i-1). \end{cases} \quad (21.4)$$

For example, in the first case one of the $m - i$ balls was chosen from the second box and placed in the first box. The transition probabilities depend only on the number of balls in the first box at any given time and are completely independent of how the box came to have that many balls. This chain therefore satisfies the Markovian property. \diamond

21.3 Recurrent and Transient States

It is of interest to know how often a state is visited by the chain and, in particular, whether a given state can be visited infinitely often. The frequency of a visit to a specific state is in fact an indicator of the probability of occurrence of that state. Assume that the system is initially in state ε_i and define two probabilities: the probability u_k that the system returns to the initial state in *exactly* k time steps, and the probability v_n that the system returns to the initial state at time n , with the possibility that it may have returned there other times prior to n . Clearly, it is true that $v_n \geq u_n$.

To determine whether a state is recurrent or transient, it is useful to define

$$u = \sum_{n=1}^{\infty} u_n \quad (21.5)$$

as the probability of the system returning the initial state ε_i for the first time at some time n . The state can be classified as *recurrent* or *transient* according to the probability of returning to that state:

$$\begin{cases} u = 1 & \text{state is recurrent;} \\ u < 1 & \text{state is transient.} \end{cases} \quad (21.6)$$

Therefore a recurrent state is one that will certainly be visited again by the chain at some time in the future. Notice that no indication is given as to the time when the system will return to the initial state.

We also state a few theorems that are relevant to the understanding of recurrent states. Proofs of these theorems can be found, for example, in the textbook by Ross [85] or other textbooks on stochastic processes, and are not reported here.

Theorem 21.1 *With v_n the probability that the system returns to a state ε_i at time n , state ε_i is recurrent if and only if*

$$\sum_{n=1}^{\infty} v_n = \infty.$$

This theorem states that if the system does return to a given state, then it will do so infinitely often. Also, since this is a necessary and sufficient condition, any transient state will not be visited by the chain an infinite number of times. This means that transient states will not be visited anymore after a given time, i.e., they are only visited during an initial period. The fact that recurrent states are visited infinitely often means that it is possible to construct a sample distribution function for recurrent states with a precision that is function of the length of the chain. No information is provided on the timing of the return to a recurrent state.

It is also necessary to introduce the definition of *accessible* states: a state ε_j is said to be accessible from state ε_i if $p_{ij}(m) > 0$ for some natural number m , meaning that there is a non-zero probability of reaching this state from another state in m time steps. The following theorems establish properties of accessible states, and how the property of accessibility relates to that of recurrence.

Theorem 21.2 *If a state ε_j is accessible from a recurrent state ε_i , then ε_j is also a recurrent state, and ε_i is accessible from ε_j .*

This theorem states that once the system reaches a recurrent state, the states visited previously by the chain must also be recurrent, and therefore will be visited again infinitely often. This means that recurrent states form a network, or class, of states that share the property of recurrence, and these are the states that the chain will sample over and over again as function of time.

Theorem 21.3 *If a Markov chain has a finite number of states, then each state is accessible from any other state, and all states are recurrent.*

This theorem ensures that all states in a finite chain will be visited infinitely often, and therefore the chain will sample all states as function of time. This property is of special relevance for Markov chain Monte Carlo methods where the states of the chain are possible values of the parameters. As the chain progresses, all values of the parameters are accessible and will be visited in proportion of the posterior distribution of the parameters.

Example 21.4 (*Recurrence of States of the Random Walk*) Consider the random walk with transition probabilities given by (21.3). It is of interest to determine whether the initial state of the chain is a recurrent or a transient state for the chain. The probability of returning to the initial state in k steps is clearly given by the binomial distribution,

$$p_{ii}(k) = \begin{cases} 0 & \text{if } k \text{ is odd} \\ \binom{k}{n} p^n q^n & \text{if } k = 2n \text{ is even.} \end{cases} \quad (21.7)$$

Using Stirling's first-order approximation for the factorial function in the binomial coefficient,

$$n! \simeq \sqrt{2\pi n} n^n e^{-n},$$

the probability to return at time $k = 2n$ to the initial state becomes

$$v_k = p_{ii}(k) = \binom{2n}{n} p^n q^n = \frac{(2n)!}{(n!)^2} p^n q^n \simeq \frac{\sqrt{4\pi n}}{2\pi n} \frac{(2n)^{2n} e^{-2n}}{n^{2n} e^{-2n}} p^n q^n = \frac{(4pq)^n}{\sqrt{\pi n}}$$

which holds only for k even. This equation can be used in conjunction with Theorem 21.1 to see if the initial state is transient or recurrent. Consider the series

$$\sum_{n=1}^{\infty} v_n = \sum_{n=1}^{\infty} \frac{1}{\sqrt{\pi n}} (4pq)^n.$$

According to Theorem 21.1, the divergence of this series is a necessary and sufficient condition to prove that the initial state is recurrent.

(a) If $p \neq q$, it follows that $x = 4pq < 1$ and

$$\sum_{n=1}^{\infty} \frac{1}{\sqrt{\pi n}} (4pq)^n < \sum_{n=1}^{\infty} x^n = \frac{x}{1-x};$$

since $x < 1$, the series converges and therefore the state is transient. This means that the system may return to the initial state, but only for a finite number of times, even after an infinite time. Notice that as time progresses the state of the system will drift in the direction that has a probability $> 1/2$.

(b) If $p = q = 1/2$, then $4pq = 1$. The series becomes

$$\frac{1}{\sqrt{\pi}} \sum_{n=1}^{\infty} \frac{1}{n^{1/2}} \quad (21.8)$$

which is a divergent series (see Problem 21.1). Therefore a random walk with the same probability of taking a step to the left or to the right will return to the origin infinitely often. \diamond

21.4 Limiting Probabilities and Stationary Distribution

The ultimate goal of a Markov chain is to calculate the probability that a system occupies a given state ε_i after a large number n of steps. This probability is called the *limiting probability*. According to the frequentist approach defined in (1.3), the limiting probability is given by

$$p_j^* = \lim_{n \rightarrow \infty} p_j(n), \quad (21.9)$$

where $p_j(n)$ is the probability of the system to be found in state ε_j at time $t = n$. With the aid of the total probability theorem, the probability of the system to be in state ε_j at time $t = n$ is

$$p_j(n) = \sum_k p_k(n-1) \cdot p_{kj}, \quad (21.10)$$

where k labels all possible states of the system. This formula can be used to calculate recursively the probability $p_j(n)$ using the probability at the previous step and the transition probabilities p_{kj} , which do not vary with time. Equation (21.10) can be written in a different form if the system is known to be in state ε_i at an initial time $t = 0$:

$$p_{ij}(n) = P(X_n = \varepsilon_j) = \sum_k p_{ik}(n-1) \cdot p_{kj}, \quad (21.11)$$

where $p_{ij}(n)$ is the probability of the system going from state ε_i to ε_j in n time steps. The probabilities $p_j(n)$ and $p_{ij}(n)$ change as the chain progresses. The limiting probabilities p_j^* , on the other hand, are independent of time, and they form the *stationary distribution* π_j of the chain. General properties for the stationary distribution can be given for Markov chains that have certain specific properties.

Additional definitions that are useful to characterize Markov chains and to determine the stationary distribution of the chain are now introduced. A number of states that are accessible from each other, meaning there is a non-zero probability to reach one state from the other ($p_{ij} > 0$), are said to *communicate*, and all states that communicate are part of the same *class*. The property of communication (\leftrightarrow) is an equivalence relation, meaning that it obeys the following three properties:

- (a) The reflexive property: $i \leftrightarrow i$;
- (b) The symmetric property: if $i \leftrightarrow j$, then $j \leftrightarrow i$; and
- (c) The transitive property: if $i \leftrightarrow j$ and $j \leftrightarrow k$, then $i \leftrightarrow k$. Therefore, each class is separate from any other class of the same chain. A chain is said to be *irreducible* if it has only one class, and thus all states communicate with each other.

Another property of Markov chains is periodicity. A state is said to be *periodic* with period T if $p_{ii}(n) = 0$ when n is not divisible by T , and T is the largest such integer with this property. This means that the return to a given state must occur in multiples of T time steps. A chain is said to be *aperiodic* if $T = 1$, and return to a given state can occur at any time. It can be shown that all states in a class share the same period.

The uniqueness of the stationary distribution, and an equation that can be used to determine it, are established by the following theorems.

Theorem 21.4 *An irreducible aperiodic Markov chain belongs to either of the following two classes:*

1. *All states are positive recurrent. In this case, $\pi_i = p_i^*$ is the stationary distribution, and this distribution is unique. A Markov chain with these properties is said to be ergodic.*
2. *All states are transient or null recurrent, and there is no stationary distribution.*

This theorem establishes that a “well-behaved” *ergodic* Markov chain, i.e., one with positive recurrent states, does have a stationary distribution, and that this distribution is unique. This is the type of Markov chains that is useful in practical applications. Positive recurrent states, defined in Sect. 21.3, are those with a finite expected time to return to the same state, while the time to return to a transient or null recurrent state is infinite. This theorem also ensures that, regardless of the starting point of the chain, the same stationary distribution will eventually be reached.

The results presented in the foregoing lead to a key theorem that enables the evaluation of the limiting distribution of probabilities of a Markov chain:

Theorem 21.5 *The limiting probabilities p_i^* of a Markov chain with transition probabilities p_{ij} are the solution of the system of linear equations*

$$p_j^* = \sum_i p_i^* \cdot p_{ij}, \quad (21.12)$$

where i labels the states of the system.

In fact, according to the recursion formula (21.10),

$$p_j(n) = \sum_i p_i(n-1) \cdot p_{ij}, \quad (21.13)$$

and therefore a proof of the theorem follows by taking the limit for n going to infinity. For a chain with a probability distribution that, at a time t_0 , satisfies

$$p_j(t_0) = \sum_i p_i(t_0) \cdot p_{ij}, \quad (21.14)$$

the probability distribution of the states satisfies (21.12), and therefore the chain has reached the stationary distribution $\pi_j = p_j^*$. Theorem 21.5 guarantees that, from that point on, the chain will maintain its stationary distribution. The importance of a stationary distribution is that, as time elapses, the chain samples this distribution. The sample distribution of the chain, e.g., a histogram plot of the occurrence of each state, can therefore be used as an approximation of the posterior distribution.

Example 21.5 (*Stationary distribution of the binary chain*) The stationary distribution of the binary chain with transition probabilities α and β is

$$\pi = \frac{1}{\alpha + \beta} (\beta, \alpha).$$

In fact, it can be immediately proven that this distribution satisfies (21.12) using the transition probabilities in Example 21.1. For example,

$$\pi_1 = p_{11} \cdot \pi_1 + p_{21} \cdot \pi_2 = (1 - \alpha) \cdot \frac{\beta}{\alpha + \beta} + \beta \cdot \frac{\alpha}{\alpha + \beta} = \frac{\beta}{\alpha + \beta} = \pi_1.$$

It is also useful to find the n -step transition matrix, or the probabilities of transition between states in n steps. For this purpose, the transition matrix of the binary chain can be also written as

$$\mathbf{P} = \begin{bmatrix} \pi_1 & \pi_2 \\ \pi_1 & \pi_2 \end{bmatrix} + \frac{1 - \alpha - \beta}{\alpha + \beta} \begin{bmatrix} \alpha & -\alpha \\ -\beta & \beta \end{bmatrix}.$$

By induction, it is possible to show that the n -step transition matrix is given by

$$\mathbf{P}(n) = \begin{bmatrix} \pi_1 & \pi_2 \\ \pi_1 & \pi_2 \end{bmatrix} + \frac{(1 - \alpha - \beta)^n}{\alpha + \beta} \begin{bmatrix} \alpha & -\alpha \\ -\beta & \beta \end{bmatrix}. \quad (21.15)$$

This transition matrix may be useful when evaluating the asymptotic properties of the chain. \diamond

Example 21.6 (*Stationary Distribution of the Ehrenfest Chain*) The goal is to find a distribution function $\pi_j = p_j^*$ that is the stationary distribution of the Ehrenfest chain.

This case is of interest because the finite number of states makes the calculation of the stationary distribution easier to achieve analytically. The condition for a stationary distribution is

$$p_j^* = \sum_{i=1}^N p_i^* \cdot p_{ij},$$

where N is the number of states of the chain. The condition can also be written in matrix notation. Recall that the transition probabilities for the Ehrenfest chain are

$$p_{ij} = \begin{cases} \frac{m-i}{m} & \text{if } j = i+1 \\ \frac{i}{m} & \text{if } j = i-1, \end{cases}$$

and they can be written as a transition matrix \mathbf{P}

$$\mathbf{P} = [p_{ij}] = \begin{bmatrix} 0 & 1 & 0 & 0 & \dots & 0 & 0 \\ \frac{1}{m} & 0 & \frac{m-1}{m} & 0 & \dots & 0 & 0 \\ 0 & \frac{2}{m} & 0 & \frac{m-2}{m} & \dots & 0 & 0 \\ 0 & 0 & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & \dots & 1 & 0 \end{bmatrix}. \quad (21.16)$$

Notice that the sum of each line is one, since

$$\sum_{j=0}^m p_{ij} = 1$$

is the probability of going from state ε_i to *any* state ε_j . In (21.16) it is convenient to regard the vertical index to be $i = 0, \dots, m$, and the horizontal index $j = 0, \dots, m$.

The way in which (21.12) is used is to verify whether a distribution is the stationary distribution of the chain. In the case of the Ehrenfest chain, it is reasonable to expect that the binomial distribution is the stationary distribution,

$$\pi_i = \binom{m}{i} p^i q^{m-i} \quad i = 0, \dots, m,$$

where p and q represent the probability of finding a ball in either box. At equilibrium one expects $p = q = 1/2$, since even an initially uneven distribution of balls between the two boxes should result in an even distribution at later times. To prove this hypothesis, consider $\pi = [\pi_0, \pi_1, \dots, \pi_m]$ as a row vector of dimension $m + 1$, and verify the equation

$$\pi = \pi \mathbf{P}, \quad (21.17)$$

which is the matrix notation for the condition of a stationary distribution. For the Ehrenfest chain, this condition is

$$(\pi_0, \pi_1, \dots, \pi_m) = (\pi_0, \pi_1, \dots, \pi_m) \begin{bmatrix} 0 & 1 & 0 & 0 & \dots & 0 & 0 \\ \frac{1}{m} & 0 & \frac{m-1}{m} & 0 & \dots & 0 & 0 \\ \frac{m}{m} & \frac{m}{m} & \frac{m}{m} & \frac{m-2}{m} & \dots & 0 & 0 \\ 0 & \frac{2}{m} & 0 & \frac{m-2}{m} & \dots & 0 & 0 \\ 0 & \frac{m}{m} & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & \dots & 1 & 0 \end{bmatrix}.$$

For a given state i , only two terms (at most) contribute to the sum,

$$\pi_i = \pi_{i-1} \cdot p_{i-1,i} + \pi_{i+1} \cdot p_{i+1,i}. \quad (21.18)$$

From this, it is easy to prove that the $p = q = 1/2$ binomial is the stationary distribution of the Ehrenfest chain (see Problem 10.6). \diamond

21.5 Ergodic Averages and Variance Estimates

An ergodic Markov chain has the desirable property of positive recurrence of all the states, and it features the stationary distribution that is necessary to make inferences on the variables that describe the state of the system (i.e., the unknown model parameters in a regression). After an ergodic Markov chain reaches its stationary distribution, samples from the chain can be used to estimate the quantities of interest, here referred to as the variable X . For example, the sample mean of n links of the chain,

$$\overline{X_n} = \frac{1}{n} \sum_{i=1}^n X_i, \quad (21.19)$$

is referred to as an *ergodic average*, and it can be shown to be an asymptotically unbiased estimator of the parent mean of X . This is an identical result to the law of large numbers (2.10), and it applies also to any function of the random variable X ,

$$\overline{f_n(X)} = \frac{1}{n} \sum_{i=1}^n f(X_i). \quad (21.20)$$

The applicability of the law of large numbers to stationary stochastic processes should not be surprising, since the expectation of the sample mean of N measurements is equal to the expectation of the variable itself, regardless of the correlation among measurements. These ergodic averages satisfy the following fundamental property that goes under the name of the *ergodic theorem*.

Theorem 21.6 (The Ergodic theorem) *For an ergodic Markov chain X with stationary distribution π , the ergodic average (21.20), in the asymptotic limit of a large number n of links, converges to the expectation of the function f under the stationary distribution,*

$$\lim_{n \rightarrow \infty} \overline{f_n(X)} = E[f(X)]. \quad (21.21)$$

This theorem is discussed, for example, by G. Roberts in [43], and by D. Gamerman in [37], and it is equivalent to the law of large numbers (see Sect. 2.3.2). The ergodic theorem constitutes a key tool for the analysis of Markov chains, since it establishes the unbiasedness of the ergodic average of a function of the Markov chain state X .

The ergodic theorem does not provide information, however, on the actual distribution of the ergodic average. Correlation among the n links of a Markov chain comes into play when estimating the variance and, in general, the probability distribution of an ergodic average. For the sum of independent random variables, the central limit theorem established the normality of the distribution and the linear addition of variances (see Sect. 4.4). But links of a Markov chain are dependent on one another, and therefore the assumption of independence is not satisfied. Fortunately, it is possible to provide a generalization of the central limit theorem to stationary stochastic processes such as ergodic Markov chains.

Theorem 21.7 (Central limit theorem for stationary stochastic processes) *In the limit of a large number n of identically distributed random variables X_i that are part of a stationary stochastic process, such as an ergodic Markov chain, the distribution of the ergodic average \overline{X}_n converges to*

$$\lim_{n \rightarrow \infty} \frac{\overline{X}_n - E[X]}{1/\sqrt{n}} \sim N(0, \sigma_{as}^2),$$

where σ_{as}^2 is the asymptotic variance given by

$$\sigma_{as}^2 = \text{Var}(X_i) + 2 \sum_{k=1}^{\infty} \text{Cov}(X_i, X_{i+k}). \quad (21.22)$$

The mathematical conditions for this theorem can be found, for example, in the article by G. Jones [57], which provides a review of the literature on the subject, and in [37]. In the absence of correlation among the variables, the variance of the ergodic mean is $\text{Var}(X_i)/n$, which leads to the usual central limit theorem. The series of covariances provides the correction that applies to ergodic means of correlated stochastic processes.

Example 21.7 (Asymptotic variance of a stationary binary process) It is possible to evaluate the asymptotic variance (21.22) for the binary Markov chain of Example 21.1. If the two states of the chain are assumed to be $\epsilon_1 = 0$ and $\epsilon_2 = 1$, the expectation of the random variable is

$$\mathbb{E}[X_i] = \frac{\alpha}{\alpha + \beta},$$

and the variance is

$$\text{Var}(X_i) = \frac{\alpha\beta}{(\alpha + \beta)^2},$$

since X_i is a Bernoulli variable (i.e., a binomial for $n = 1$ trial). To evaluate the covariance between neighboring variables X_i and X_{i+1} , it is immediate to see that

$$\mathbb{E}[X_i X_{i+1}] = \frac{\alpha}{\alpha + \beta} \times (1 - \beta)$$

since the system is in the state $\epsilon_2 = 1$ with probability $\alpha/(\alpha + \beta)$ at a given time i , and $(1 - \beta)$ is the probability of remaining in that state at the next time step. As a result, the covariance between consecutive variables is

$$\text{Cov}(X_i, X_{i+1}) = \mathbb{E}[X_i X_{i+1}] - \mathbb{E}[X_i]^2 = \frac{\alpha\beta(1 - \alpha - \beta)}{(\alpha + \beta)^2} = \text{Var}(X_i)(1 - \alpha - \beta).$$

The covariance between two variables separated by a time $k \geq 1$ can be likewise evaluated as

$$\text{Cov}(X_i, X_{i+k}) = \text{Var}(X_i)(1 - \alpha - \beta)^k,$$

and therefore (21.22) contains a simple geometric series with ratio $(1 - \alpha - \beta) < 1$ which converges to

$$\sum_{k=1}^{\infty} (1 - \alpha - \beta)^k = \left(\frac{1}{\alpha + \beta} - 1 \right) = \frac{1 - \alpha - \beta}{\alpha + \beta}.$$

Finally, the asymptotic variance of the binary Markov chain with two states $\epsilon_1 = 0$ and $\epsilon_2 = 1$ is evaluated as

$$\sigma_{as}^2 = \frac{\alpha\beta(2 - \alpha - \beta)}{(\alpha + \beta)^3}. \quad (21.23)$$

It is worth noting that the asymptotic variance differs from the Bernoulli variance by a factor of $(2 - \alpha - \beta)/(\alpha + \beta)$, which is attributable to the correlation among variables in the Markov chain. The asymptotic variance of the ergodic mean \bar{X}_n is equal to σ_{as}^2/n . Equation 21.23 is also derived in the textbook by D.R. Cox and H.D. Miller [20] without making explicit use of the central limit theorem for Markov chains. ◇

The ergodic average for a parameter or, more generally, for a function of the parameters as defined in (21.20) should be viewed as an *estimator* for the parameter or function of interest using the (correlated) Markov chain links. The ergodic theorem determines that the estimator is unbiased, at least in the asymptotic limit of a

large number of links. Moreover, the central limit theorem for stationary stochastic processes provides the asymptotic distribution of the estimator, which is needed for hypothesis testing and confidence intervals. The complication with this sampling distribution is that its variance, indicated by σ_{as}^2 , is generally unknown. In the case of Example 21.7, it was possible to evaluate it based on a knowledge of the variance and covariances of the variables X_i , but for most applications the variance and covariances are unknown. To make use of the central limit theorem for the ergodic averages, it is therefore necessary to estimate this variance σ_{as}^2 , which represents the sampling variance of the estimator.

Correlation among the samples generally prevents the use of the “direct” sample variance of the Markov chain links. Overcoming this correlation for the purpose of estimating the variance of the ergodic average is a critical issue for the analysis of Markov chains (and, of course, also for the simulated Markov chain Monte Carlo discussed in Chap. 22). There are a number of methods available for this purpose that include *spectral estimation* of the variance or the use of *batch means*. These methods are described in detail, for example, in [35, 44, 58], and in the textbook by D. Gamerman [37].

The method of batch means is especially easy to implement, and in its simplest form it consists of dividing a chain of length N into a non-overlapping batches of length b so that $N = a \times b$. The basic idea is that if the batches are separated by a sufficient number of iterations, they become approximately independent of each other. For a quantity of interest X , which can be a parameter of the chain or a function of the parameters, first it is necessary to define the mean in each batch of length b ,

$$\bar{X}_k = \frac{1}{b} \cdot \sum_{i=(k-1)b+1}^{kb} X_i$$

where $k = 1, \dots, a$ labels the batch number, and then estimate the variance as

$$\hat{\sigma}_X^2 = \frac{b}{a-1} \sum_{k=1}^a (\bar{X}_k - \bar{X})^2, \quad (21.24)$$

where \bar{X} is the sample mean over the entire chain. The estimator is therefore equal to the sample variance of the a batch means, multiplied by the length of each batch. Accordingly, the estimator of the variance of the sample mean of X is

$$\hat{\sigma}_{\bar{X}}^2 = \frac{\hat{\sigma}_X^2}{N} = \frac{1}{a(a-1)} \sum_{k=1}^a (\bar{X}_k - \bar{X})^2.$$

Equation 21.24 is a rather crude way to estimate the variance, but its overall simplicity makes it a popular estimator. This estimate can be used as the asymptotic variance (21.23) for the central limit theorem for Markov chains.

An alternative method to ameliorate the presence of correlation is to *thin* the chain by using only every n th link (e.g., every 10th or 100th iteration). In this case, instead of averaging over a batch, the analyst simply uses a subset of the chain in an effort to sample the posterior distribution with a reduced amount of correlation. If the chain is thinned sufficiently so that Markov chain links are approximately independent, then the variance of the ergodic mean can be approximated by means of the usual sample variance.

Summary of Key Concepts for this Chapter

Markov chain: A stochastic process or sequence of random variables that vary as a function of an integer time variable.

Markovian property: It is the key property of Markov chains, stating that the state of the system at a given time depends only on the state at the previous time step, but not on previous history.

Recurrent and transient state: A recurrent state occurs infinitely often while a transient state only occurs a finite number of times in the Markov chain.

Stationary distribution: It is the asymptotic distribution of each variable of interest, obtained after a large number of time steps of the Markov chain.

Ergodic averages: They are sample averages of a function of n links of a Markov chain, and they are unbiased estimators.

Problems

21.1 Consider the Ehrenfest chain described in Example 21.6. Show that the stationary distribution is the binomial distribution with $p = q = 1/2$.

21.2 Show that the random walk with $p = q = 1/2$ returns to the origin infinitely often, and therefore the origin is a recurrent state of the chain.

21.3 For the random walk with $p \neq q$, show that the origin is a transient state of the Markov chain.

21.4 The diffusion model of Example 21.3 is modified in such a way that, at each time step, one has the option to choose one box at random from which to replace a ball in the other box, regardless of the number of balls in either box.

- Determine the transition probabilities p_{ij} for this process.
- Determine whether this process is a Markov chain.

21.5 Using the model of diffusion of Problem 21.4, determine whether the binomial distribution with $p = q = 1/2$ is the stationary distribution.

Chapter 22

Markov Chain Monte Carlo



Abstract Markov chain Monte Carlo methods have become popular with the availability of modern-day computing resources. The basic idea behind Markov chain Monte Carlo is to estimate quantities of interest, such as model parameters, by repeatedly querying the data in order to generate a Markov chain that can then be analyzed to estimate parameters and their confidence intervals. The data analyst will find them an essential tool that permits tasks that are simply not possible with other methods, such as the simultaneous estimate of parameters for multi-parametric models of virtually any level of complexity.

22.1 Introduction to Markov Chain Monte Carlo Methods

A common data analysis problem is the fit of data to a model with several adjustable parameters. Chapter 11 presented the maximum-likelihood method to determine the best-fit values and confidence intervals for the model parameters. The simple linear regression of Sect. 11.3 or the multiple linear regressions (Chap. 13) have an analytic solution for the best-fit parameters and its uncertainties, but the regression to non-linear functions do not have analytic solutions at all. When an analytic solution is not available, the χ^2_{\min} method to search for best-fit parameters and their confidence intervals is still applicable, as described in Sect. 12.3. The main complication is the computational cost of sampling the parameter space in search of χ^2_{\min} and surfaces of constant $\Delta\chi^2$, especially when there is a large number of adjustable parameters. Consider, for example, a model with 10 free parameters: even a very coarse sampling of 10 values for each parameter will result in 10^{10} evaluations of the likelihood to cover the entire parameter space, at a significant computational cost. It is not always possible to overcome this limitation by searching for just a few interesting parameters

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-981-19-0365-6_22.

at a time, e.g., fixing the value of the background while searching for the flux of the source. In fact, there may be a correlation among parameters and this requires that the parameters be estimated simultaneously.

The *Markov chain Monte Carlo* (MCMC) methods presented in this chapter provide a way to bypass altogether the need for a uniform sampling of parameter space. This is achieved by constructing a Markov chain that only samples the interesting region of parameters space, i.e., the region near the maximum of the likelihood. The method is so versatile and computationally efficient that MCMC techniques have become one of the leading methods of data analysis. Although MCMC applications can be applied to a variety of data analysis tasks (see, for example, the book by Gilks, Richardson and Spiegelhalter [43]), the main application described in this textbook is to data regression.

22.2 Markov Chain Monte Carlo for Regression Analysis

The typical Markov chain Monte Carlo application to data regression makes use of a dataset Z and a model with m adjustable parameters, referred to as $\theta = (\theta_1, \dots, \theta_m)$, for which it must be possible to calculate the likelihood

$$\mathcal{L} = P(Z|\theta) \quad (22.1)$$

for all possible parameter values.¹ The calculation of the likelihood is usually the most computationally intensive task of a MCMC, and its complexity depends on the type of data and model at hand. According to Bayesian statistics, there is a *prior* knowledge on the parameters (see Sect. 1.6) that may come from experiments that were conducted beforehand, or from any other type a priori belief on the parameters. The prior probability distribution will be referred to as $p(\theta)$, and in its simplest form it consists of a uniform distribution that takes the form of hard limits on possible values of the parameter.

The information sought through the MCMC is the probability distribution of the model parameters *after* the measurements are made, i.e., the posterior distribution $P(\theta|Z)$. According to Bayes' theorem (see Sect. 1.6), the posterior distribution is given by

$$P(\theta|Z) = \frac{p(\theta) \cdot P(Z|\theta)}{P(Z)} = \frac{p(\theta) \cdot \mathcal{L}}{P(Z)}, \quad (22.2)$$

where the quantity $P(Z) = \int p(\theta) \cdot \mathcal{L} d\theta$ is a normalization constant. In general (22.2) can be very complicated, as it requires a multi-dimensional integration of the term $P(Z)$. The alternative provided by the Markov chain Monte Carlo method is the construction of a sequence of *dependent* samples for the parameters θ in the form

¹ For the linear regression of Chap. 11, the adjustable parameters were indicated with the usual Latin letters a and b , or a_k , and they are equivalent to the Greek letter parameters used in this chapter.

of a Markov chain where each parameter value appears in the chain in proportion to this posterior distribution. After the chain is run for a large number of iterations, the posterior distribution is obtained via the sample distribution of the parameters in the chain.

At the heart of a MCMC is a sampling method that creates a Markov chain with the desired stationary distribution. There are several algorithms to sample the parameter space that satisfy the requirement of having the posterior distribution of the parameters $P(\theta/Z)$ as the stationary distribution of the chain. A very common algorithm that can be used in most applications is the one developed by N. Metropolis and W. K. Hastings [48, 71]. The Metropolis–Hastings algorithm is surprisingly easy to implement and constitutes a reference for any MCMC implementation. Another algorithm is the Gibbs sampler [40], but its use is limited by certain specific requirements on the distribution function of the parameters. Both algorithms, presented in this chapter, provide a way to sample values of the parameters and describe a way to accept them into the Markov chain.

22.3 The Metropolis–Hastings MCMC

The Metropolis–Hastings algorithm [48, 71] was devised in 1953 by N. Metropolis as a “general method, suitable for fast electronic computing machines, of calculating the properties of any substance which may be considered as composed of interacting individual molecules,” and it was generalized in 1970 by W. K. Hastings [48]. The basic idea is to draw random choices for parameter values and determine whether they are accepted into the Markov chain, or not, based on the likelihood of the data with the selected parameters. The method consists of the following steps:

1. The Markov chain starts with an arbitrary choice of the initial values of the model parameters, $\theta_0 = (\theta_1^0, \dots, \theta_m^0)$. This initial set of parameters is automatically accepted into the chain. As will be explained later, some of the initial links in the MCMC will later be discarded to offset the arbitrary choice of the starting point. The index $n = 0$ refers to the initial timestep in the Markov chain.
2. A *candidate* θ' for the next link of the chain is drawn from a *proposal* or *auxiliary distribution* $q(\theta'/\theta_n)$, where θ_n is the current link in the chain. This distribution is the probability of drawing a given candidate θ' , given that the chain is in state θ_n . There is a large amount of freedom in the choice of the auxiliary distribution, which can depend on the current state of the chain θ_n , according to the Markovian property, but not on its prior history. One of the simplest choices for a proposal distribution is an m -dimensional uniform distribution of fixed width in the neighborhood of the current parameter. A uniform prior is very simple to implement, and it is the default choice in many applications.
3. A *prior distribution* $p(\theta)$ is required before a decision can be made whether the candidate is accepted into the chain or rejected. The Metropolis–Hastings algorithm gives freedom on the choice of the prior distribution as well. A typical

choice of prior is another uniform distribution between two hard limits, enforcing a prior knowledge that a given parameter may not exceed certain boundaries. Sometimes the boundaries are set by nature of the parameter itself, e.g., certain parameter may only be positive numbers, or in a fixed interval range. Other priors may be more restrictive. Consider as an example the measurement of the slope of the curve in the Hubble data presented in Sect. 11.6. It is clear that, after a preliminary examination of the data, the slope parameter b will not be a negative number, and will not be larger than, say, $b = 2$ (in the given units of measure). Therefore one can assume a prior on this parameter equal to $p(b) = 1/2$, for $0 \leq b \leq 2$. Much work on priors has been done by H. Jeffreys [56], in search of mathematical functions that express the lack of prior knowledge, known as *Jeffreys priors*. For many applications, though, simple uniform prior distributions are often sufficient.

4. After drawing a random candidate θ' , a decision must be made whether to accept it into the chain, or reject it. This choice is made according to the following *acceptance probability*, which is the heart of the Metropolis–Hastings algorithm:

$$\alpha(\theta'/\theta_n) = \min \left\{ \frac{\pi(\theta')q(\theta_n/\theta')}{\pi(\theta_n)q(\theta'/\theta_n)}, 1 \right\}. \quad (22.3)$$

The acceptance probability $\alpha(\theta'/\theta_n)$ is a number between 0 and 1 that determines the probability of accepting θ' as the new link, where $q(\theta'/\theta_n)$ is the proposal distribution, and $\pi(\theta') = P(\theta'/Z)$ is the intended stationary distribution of the chain. Equation (22.3) means that the probability of going to a new value in the chain, and therefore having $\theta_{n+1} = \theta'$, is proportional to the ratio of the posterior distribution of the candidate to that of the previous link. The acceptance probability can also be re-written by making use of Bayes' theorem (22.2) as

$$\alpha(\theta'/\theta_n) = \min \left\{ \frac{p(\theta')P(Z/\theta')q(\theta_n/\theta')}{p(\theta_n)P(Z/\theta_n)q(\theta'/\theta_n)}, 1 \right\} \quad (22.4)$$

In this form, the acceptance probability can be calculated based on known quantities. The product $p(\theta_n)q(\theta'/\theta_n)$ at the denominator represents the probability of occurrence of a given candidate θ' : in fact, the first term is the prior probability of the n -th link in the chain, and the second term is the probability of generating the candidate, once the chain is in that state. The other term is the likelihood $\mathcal{L} = P(Z/\theta_n)$ of the current link in the chain. At the numerator, all terms have reverse order of conditioning between the current link and the candidate. Therefore, thanks to Bayes' theorem, all quantities in (22.4) are now known, since $p(\theta_n)$ and $q(\theta'/\theta_n)$ (and their conjugates) are chosen by the analyst, and the likelihood can be calculated for all model parameters.

The acceptance probability means that the candidate is accepted in the chain in proportion to the value of $\alpha(\theta'/\theta_n)$. Two cases are possible:

- (a) $\alpha = 1$: This means that the candidate will always be accepted in the chain, since the probability of acceptance is 100%. The candidate becomes the next link in the chain, $\theta_{n+1} = \theta'$. The min operator guarantees that the probability is never greater than 1, which would not be meaningful.
- (b) $\alpha < 1$: This means that the candidate can only be accepted in the chain with a probability α . To enforce this probability of acceptance, it is sufficient to draw a random number $0 \leq u \leq 1$ and then accept or reject the candidate according to the following criterion:

$$\begin{cases} \text{if } u \leq \alpha \Rightarrow \text{candidate is accepted, and } \theta_{n+1} = \theta' \\ \text{if } u > \alpha \Rightarrow \text{candidate is rejected, and } \theta_{n+1} = \theta_n. \end{cases} \quad (22.5)$$

In fact, since u is a uniformly distributed random number between 0 and 1, $u \leq \alpha$ will occur with probability α . It is important to notice that if the candidate is rejected, then the chain doesn't move from its current location and a new link equal to the previous one is added to the chain. This means that at each time step in the chain a new link is added, either by repeating the last link (if the candidate is rejected) or by adding a different link (if the candidate is accepted).

The logic of the Metropolis–Hastings algorithm can be easily understood in the case of uniform prior and auxiliary distributions. In that case, the candidate is accepted in proportion to just the ratio of the likelihoods, since all other terms in (22.3) cancel out:

$$\alpha(\theta'/\theta_n) = \min \left\{ \frac{\mathcal{L}(\theta')}{\mathcal{L}(\theta_n)}, 1 \right\}. \quad (22.6)$$

If the candidate has a higher likelihood than the current link, it is automatically accepted. If the likelihood of the candidate is lower than the likelihood of the current link, then it is accepted in proportion to the ratio of the likelihoods of the candidate and of the current link. The possibility of accepting a parameter of *lower* likelihood permits a sampling of the parameter space, instead of a simple search for the point of maximum likelihood that would only result in a point estimate. The acceptance probability also means that a Metropolis–Hastings MCMC features an *acceptance rate* that is defined as the ratio of candidates accepted over the total number of links in the chain. A lower acceptance rate indicates that more candidates were discarded and that therefore there is a larger number of repeated link values in the chain.

It is necessary to show that use of the Metropolis–Hastings algorithm creates a Markov chain that has $\pi(\theta_n) = P(\theta_n/Z)$ as its stationary distribution. For this purpose it is sufficient to show that the posterior distribution of the parameters satisfies the relationship

$$\pi(\theta_n) = \sum_j \pi(\theta_j) \cdot p_{jn} \quad (22.7)$$

according to Theorem 21.5, where p_{jn} are the transition probabilities of the Markov chain, and the index j runs over all possible states.

To prove that the Metropolis–Hastings algorithm leads to a Markov chain with the desired stationary distribution, consider the original chain and the time-reversed chain:

$$\begin{array}{ll} \text{original chain:} & X_0 \rightarrow X_1 \rightarrow X_2 \rightarrow \dots \rightarrow X_n \rightarrow \rightarrow X_{n+1} \dots \\ \text{time-reversed chain:} & X_0 \leftarrow X_1 \leftarrow \dots \leftarrow X_n \leftarrow X_{n+1} \leftarrow \dots \end{array}$$

The time-reversed chain is defined by the transition probability p_{ij}^* :

$$\begin{aligned} p_{ij}^* &= P(X_n = \varepsilon_j / X_{n+1} = \varepsilon_i) = \frac{P(X_n = \varepsilon_j, X_{n+1} = \varepsilon_i)}{P(X_{n+1} = \varepsilon_i)} \\ &= \frac{P(X_{n+1} = \varepsilon_i / X_n = \varepsilon_j)P(X_n = \varepsilon_j)}{P(X_{n+1} = \varepsilon_i)}, \end{aligned}$$

leading to the following relationship with the transition probability p_{ij} of the original chain:

$$p_{ij}^* = p_{ji} \cdot \frac{\pi(\theta_j)}{\pi(\theta_i)}. \quad (22.8)$$

If the original chain is *time-reversible*, then $p_{ij}^* = p_{ij}$, and the time-reversed process is also a Markov chain. In this case, the stationary distribution will follow the relationship

$$\pi(\theta_i) \cdot p_{ij} = p_{ji} \cdot \pi(\theta_j) \quad (22.9)$$

known as the equation of *detailed balance*. The detailed balance is the hallmark of a time-reversible Markov chain, stating that the probability to move forward and backwards is the same, once the stationary distribution is reached. Therefore, if the transition probability of the Metropolis–Hastings algorithm satisfies this equation, with $\pi(\theta) = P(\theta/Z)$, then the chain is time reversible, and with the desired stationary distribution. In fact, (22.7) follows from the detailed balance equation (22.9) by summing over j , with $i = n$. Moreover, Theorem 21.4 ensures that this distribution is unique.

The Metropolis–Hastings algorithm enforces a specific transition probability between states θ_i and θ_j ,

$$p_{ij} = q(\theta_j/\theta_i)\alpha(\theta_j/\theta_i) \quad \text{if } \theta_i \neq \theta_j, \quad (22.10)$$

where q is the probability of generating the candidate (or proposal distribution), and α the probability of accepting it. The probability of remaining at the same state θ_i is therefore

$$p_{ii} = 1 - \sum_{j \neq i} p_{ij},$$

where the sum is over all possible states. The transition probability (22.3) is

$$\alpha(\theta_j/\theta_i) = \min \left\{ \frac{p(\theta_j)P(Z/\theta_j)q(\theta_i/\theta_j)}{p(\theta_i)P(Z/\theta_i)q(\theta_j/\theta_i)}, 1 \right\} = \min \left\{ \frac{\pi(\theta_j)q(\theta_i/\theta_j)}{\pi(\theta_i)q(\theta_j/\theta_i)}, 1 \right\}$$

where $\pi(\theta_i) = p(\theta_i/Z) = P(Z/\theta_i)p(\theta_i)/P(Z)$ is the posterior distribution. Notice that the probability $P(Z)$ cancels out, therefore its value does not play a role in the construction of the chain and it needs not be evaluated. It is clear that, if $\alpha(\theta_j/\theta_i) < 1$, then $\alpha(\theta_i/\theta_j) = 1$, thanks to the min operation. Assuming, without loss of generality, that $\alpha(\theta_i, \theta_j) < 1$, it follows that

$$\alpha(\theta_j/\theta_i) = \frac{\pi(\theta_j)q(\theta_i/\theta_j)}{\pi(\theta_i)q(\theta_j/\theta_i)}$$

leading to

$$\alpha(\theta_j/\theta_i) \cdot \pi(\theta_i)q(\theta_j/\theta_i) = \pi(\theta_j)q(\theta_i/\theta_j) \cdot \alpha(\theta_i/\theta_j).$$

Now, since it was assumed that $\alpha(\theta_j/\theta_i) < 1$, the operation of min becomes redundant. Using (22.10) the previous equation simplifies to

$$p_{ij} \cdot \pi(\theta_i) = p_{ji} \cdot \pi(\theta_j),$$

which shows that the Metropolis–Hastings algorithm satisfies the detailed balance equation; it thus generates a time-reversible Markov chain, with stationary distribution equal to the posterior distribution.

Example 22.1 (*Metropolis–Hastings MCMC on Hubble’s data*) The data from Hubble’s experiment (Sect. 11.6) can be used to run a Monte Carlo Markov chain to obtain the posterior distribution of the parameters a and b . A linear regression to these data was also presented in Sect. 11.6. The MCMC is constructed using uniform priors on the two fit parameters a and b :

$$\begin{cases} p(a) = \frac{10}{7} & \text{for } 0.2 \leq b \leq 0.9 \\ p(b) = 10 & \text{for } 0.15 \leq a \leq 0.25. \end{cases}$$

The uniform prior simply enforces that the candidates cannot exceed these hard bounds. The proposal distributions are also uniform distributions, respectively, of fixed width 0.2 and 0.04 for a and b , and centered at the current value of the parameters:

$$\begin{cases} p(a'/a_n) = 5 & \text{for } a_n - 0.1 \leq a' \leq a_n + 0.1 \\ p(b'/b_n) = 25 & \text{for } b_n - 0.02 \leq b' \leq b_n + 0.02, \end{cases}$$

where a_n and b_n are, respectively, the n -th links of the chain, and a' and b' are the two candidates for the $(n+1)$ -th link of the chain. The choice of these ranges for the candidates is somewhat arbitrary and it reflects the Bayesian nature of the MCMC.

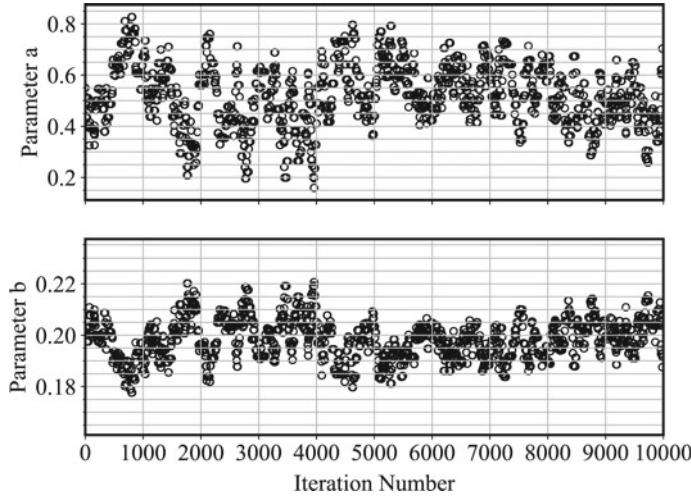


Fig. 22.1 MCMC for parameters a, b of linear model fit to the Hubble data in Table 11.1. The chain was run for 10,000 iterations

In practice, once the choice of a uniform distribution with fixed width is made, the actual value of the prior and proposals distributions are not used explicitly. In fact, the acceptance probability becomes simply a function of the ratio of the likelihoods:

$$\alpha(\theta'/\theta_n) = \min \left\{ \frac{\mathcal{L}(\theta')}{\mathcal{L}(\theta_n)}, 1 \right\} = \min \left\{ e^{\frac{\chi^2(\theta_n) - \chi^2(\theta')}{2}}, 1 \right\},$$

where $\chi^2(\theta_n)$ and $\chi^2(\theta')$ are the statistics calculated, respectively, using the n -th link of the chain and the candidate parameters. Use of the χ^2 statistics implies that the Hubble data are normally distributed, according to (6.1).

Figure 22.1 shows that there are several occasions where two or more consecutive links in the chain are identical. This is an indication that the candidate parameter drawn at that iteration was rejected, and the previous link was therefore repeated. In fact, the overall acceptance rate of candidates was less than 10%, meaning that on average a new candidate was rejected 10 times before a new value is accepted in the chain. Figure 22.2 shows the sample distributions of the two fit parameters. These distributions can be used to provide estimates of the parameters and their confidence intervals.

◇

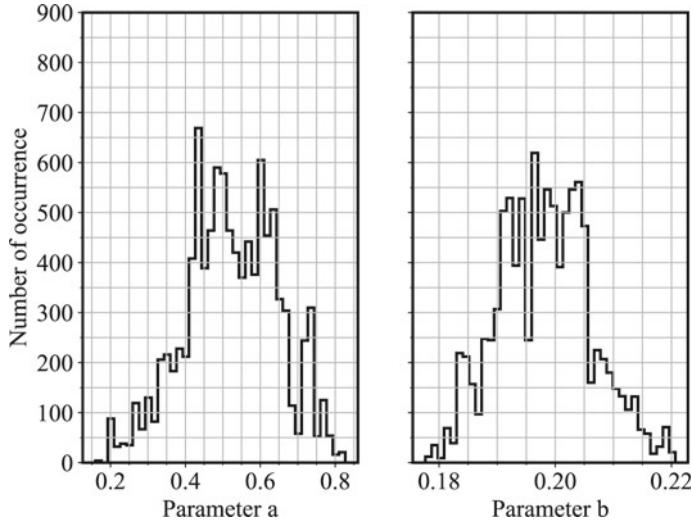


Fig. 22.2 Sample distribution function for parameters a and b , constructed using the same MCMC on the Hubble data as Fig. 22.1

22.4 The Gibbs Sampler

The Gibbs sampler is another method to create a Markov chain with the posterior distribution of the parameters as its stationary distribution. The method was proposed by S. Geman and D. Geman [40] as an application to image analysis that makes use of the *Gibbs distribution* of common use in thermodynamics. Its algorithm is based on the availability of the *full conditional distribution*,

$$\pi_i(\theta_i) = \pi(\theta_i | \theta_{j \neq i}), \quad (22.11)$$

which is the (posterior) distribution of a given parameter, given that the values of all other parameters are known. If the full conditional distributions are known and can be sampled from, then a simple algorithm can be implemented:

1. Start the chain at a given value of the parameters, $\theta_0 = (\theta_0^1, \dots, \theta_0^m)$, where the superscripts now indicate the parameter, and the subscripts the time.
2. Obtain a new value in the chain through successive generations:

$$\begin{aligned} \theta_1^1 &\text{ drawn from } \pi(\theta_1 | \theta_0^2, \theta_0^3, \dots) \\ \theta_1^2 &\text{ drawn from } \pi(\theta_2 | \theta_1^1, \theta_0^3, \dots) \\ &\dots \\ \theta_1^m &\text{ drawn from } \pi(\theta_m | \theta_1^1, \theta_1^2, \dots, \theta_1^{m-1}). \end{aligned}$$

For example, θ_1^2 is the second (of m) parameters at iteration 1, which is the one after the initial set of parameters.

3. Iterate until convergence to stationary distribution is reached.

A justification of this method can be found in the textbook by D. Gamerman [37]. In the case of data fitting with a dataset Z and a model with m adjustable parameters, it is not always possible to know the full conditional distributions, thus this method is not as common as the Metropolis–Hastings algorithm. The great advantage of the Gibbs sampler, however, is that the acceptance rate is 100%, since there is no rejection of candidates for the Markov chain, unlike in the case of the Metropolis–Hastings algorithm.

Example 22.2 (*A Markov chain with the Gibbs sampler*) This example reproduces an application presented by B. Carlin and colleagues [17], and illustrates a possible application where the knowledge of the full conditional distribution results in the possibility of implementing a Gibbs sampler. Consider a Poisson dataset of n numbers y_i , with $i = 1, \dots, n$, fit to a step-function model:

$$y = \begin{cases} \lambda & \text{if } i \leq m \\ \mu & \text{if } i > m. \end{cases} \quad (22.12)$$

The model therefore has three parameters, the values λ, μ of the function, and the point of discontinuity, m . This situation could be representative of a quantity that suddenly changes its value at an unknown time. Assume that the priors on the parameters are, respectively, a gamma distributions for λ and μ , respectively, $p(\lambda) = f_\gamma(\lambda; \beta, \alpha)$ and $p(\mu) = f_\gamma(\mu; \delta, \gamma)$, where the Greek letters α through δ are the parameters of the gamma distribution defined in (9.6). The parameter m , on the other hand, is assumed to follow a uniform distribution, $p(m) = 1/n$. According to Bayes' theorem, the posterior distribution is proportional to the product of the likelihood and the priors:

$$\pi(\lambda, \mu, m) \propto P(y_1, \dots, y_n | \lambda, \mu, m) \cdot p(\lambda)p(\mu)p(m). \quad (22.13)$$

The posterior is therefore given by

$$\pi(\lambda, \mu, m) \propto \prod_{i=1}^m e^{-\lambda} \lambda^{y_i} \prod_{i=m+1}^n e^{-\mu} \mu^{y_i} \cdot (\lambda^{\alpha-1} e^{-\beta\lambda}) \cdot (\mu^{\gamma-1} e^{-\delta\mu}) \cdot \frac{1}{n}$$

leading to

$$\pi(\lambda, \mu, m) \propto \lambda^{\left(\alpha-1 + \sum_{i=1}^m y_i\right)} e^{-\lambda(\beta+m)} \cdot \mu^{\left(\gamma-1 + \sum_{i=m+1}^n y_i\right)} e^{-\mu(\delta+n-m)},$$

where terms that are independent of the parameters λ, μ and m are ignored. The equation above indicates that the conditional posteriors, obtained by fixing all parameters except one, are given by

$$\begin{cases} \pi_\lambda(\lambda) = f_\gamma\left(\lambda; \alpha + \sum_{i=1}^m y_i, \beta + m\right) \\ \pi_\mu(\mu) = f_\gamma\left(\mu; \gamma + \sum_{i=m+1}^n y_i, \delta + n - m\right) \\ \pi_m(m) = \frac{\pi(\lambda, \mu, m)}{\sum_{l=1}^n \pi(\lambda, \mu, l)}. \end{cases} \quad (22.14)$$

This is therefore a case where the conditional posterior distributions are known, and therefore the Gibbs algorithm is applicable. All three conditional distributions can be sampled using the methods described in Sect. 5.3. \diamond

22.5 Convergence of Markov Chain Monte Carlo

The ease of implementation of a MCMC, for example, via the Metropolis–Hastings algorithm, comes with the fundamental caveat that the properties of the chain depend on the arbitrary choices made in its setup, primarily the choice of initial values and the parameters of the prior and proposal distributions. It is therefore necessary to ensure that the MCMC has reached *convergence* to the stationary distribution and that the chain has been run for a sufficiently large number of iteration, before inferences on the posterior distribution can be made. The primary issues of concern when analyzing a MCMC, and the tools needed to address them are discussed in the following.

(a) *Time to reach convergence*: Convergence indicates that the chain has started to sample the posterior distribution, so that the MCMC samples are representative of the distribution of interest. The period of time required for the chain to reach convergence goes under the name of *burn-in* period and varies from chain to chain according to a variety of factors, such as the choice of prior and proposal distributions. It is therefore necessary to identify and remove such initial period from the chain prior to further analysis. The *Geweke z-score test* (Sect. 22.6) and the *Gelman–Rubin test* (Sect. 22.7) are two of the most common tests used to identify the burn-in period.

(b) *Length of the chain*: Another important consideration is that the chain must be run for a sufficient number of iterations, so that the sample distribution becomes a good approximation of the true posterior distribution. It is clear that the larger the number of iterations after the burn-in period, the more accurate will be the estimates of the parameters of the posterior distribution. In practice it would be convenient to know the minimum *stopping time* that enables to estimate the posterior distribution with the required precision. This is usually difficult to estimate with precision, but the *Raftery–Lewis diagnostic* (Sect. 22.8) is designed to give an approximate estimate of both the burn-in time and the minimum required stopping time.

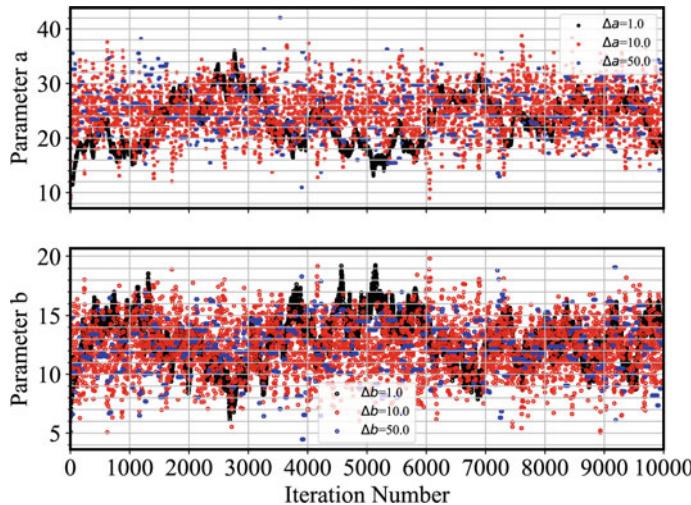


Fig. 22.3 Time evolution of the MCMC for parameters a, b of the linear model fit to the data in Example 12.2, using uniform proposal distributions. The chain started at $a = 12$ and $b = 6$. The acceptance rates for the three chains were, respectively, 91.8%, 38.6%, and just 3.5% for the chain with the largest proposal distributions

Example 22.3 (*MCMC on the 5-point data of Example 12.2*) Typical considerations concerning the burn-in period and the stopping time of a chain can be illustrated with three chains based on the data from Example 12.2. The chains were run, respectively, with a uniform proposal distribution of widths 1, 10, and 50 for both parameters of the linear model, and starting at the same point (Figs. 22.4 and 22.3). The simple linear regression on those data yielded best-fit parameter values of $a = 25.44 \pm 4.26$ and $b = 12.06 \pm 2.11$, and the initial points chosen for the chains are both lower than the best-fit values. The chain with narrower proposal distributions ($\Delta a = \Delta b = 1$) requires a longer time to reach the stationary value of the parameters, as can be seen in Fig. 22.3 with an interval of approximately 1,000 iterations where the a chain samples only lower values of the distribution. This is in part because, at each time interval, the candidate can be chosen in just a limited neighborhood of the previous link, thus limiting the progress of the chain toward the maximum of the likelihood. Moreover, the sampling of parameter space remains less uniform for the duration of the chain, because the chain requires longer time to span the entire parameter range. The intermediate values for the proposal distribution ($\Delta a = \Delta b = 10$) result in an almost immediate convergence, and the sampling of parameter space is clearly more uniform. An increase in the size of the proposal distribution, as in the $\Delta a = \Delta b = 50$ chains, may eventually lead to slow convergence and poor sampling. In this case, candidates are drawn from regions of parameter space that have very low likelihood, or large χ^2 , and therefore the chain has a tendency to remain at the same location for extended periods of time, with low acceptance rate. The result is a chain with poor coverage of parameter space and poorly determined sample distribution for their

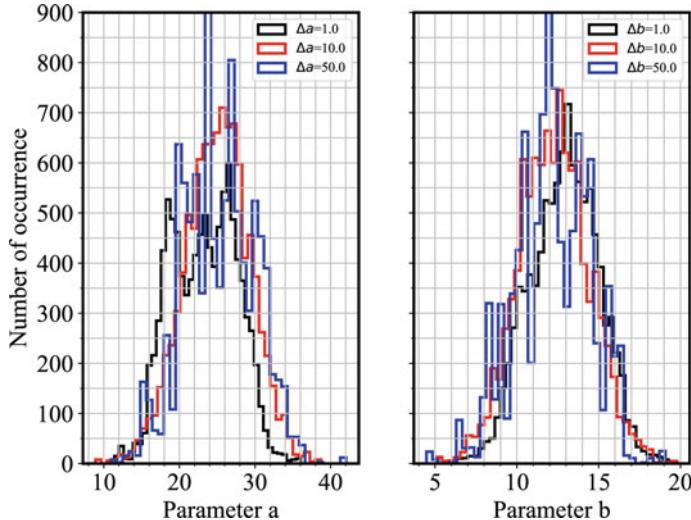


Fig. 22.4 Sample distributions of the parameters a and b , for the MCMC of the linear model fit to the data in Example 12.2, using uniform proposal distributions as in Fig. 22.4

parameters, as shown in Fig. 22.4. A smoother sample distribution is preferable, because it leads to a more accurate determination of the median, and of confidence ranges on the parameters. There is no requirement that the proposal distributions of the two parameters are identical, as was the choice for this application. Usually different parameters will have different proposal distributions, and a good amount of fine-tuning is always required to balance the requirements of convergence and sampling of parameter space. \diamond

(c) *Correlation of MCMC samples:* Another consideration is that elements in the chain are more or less correlated to one another, according to the choice of the proposal distribution, and other choices in the setup of the chain. Links in the chains are correlated by construction, since the next link in the chain typically depends on the current state of the chain. This is illustrated, for example, by repeated links in a Metropolis–Hastings MCMC with low acceptance rate. In principle a Markov chain can be constructed that does not depend on the current state of the chain, but in most cases it is convenient to make full use of the Markovian property that allows to make use of the current state of the chain. The chains in Figs. 22.3 and 22.4 illustrate how the degree of correlation varies with the proposal distribution choice. For example, the chain with the narrowest proposal distribution widths appears more correlated than the one with the intermediate choice for the width, and the chain with the largest widths has periods with the highest degree of correlation, namely, when the chain does not move for tens or hundreds of iterations. This shows that the degree of correlation is a non-linear function of the proposal distribution width, and that fine-tuning is always required to obtain a chain with good *mixing* properties.

22.6 The Geweke z -Score Convergence Test

A simple test of convergence is provided by the standardized difference of the mean of two segments of the chain, as proposed by J. Geweke [41]. Under the null hypothesis that the chain is sampling the same distribution during both segments, the sample means are expected to be drawn from the same distribution. Consider a segment A at the beginning of the chain, and a segment B at the end of the chain. If the chain is of length N , the prescription is to use an initial segment of $N_A = 0.1 N$ links and a final segment with $N_B = 0.5 N$ links, although those choices are somewhat arbitrary, and segments of different length can also be used. The mean of each parameter in the two segments A and B is calculated as usual as

$$\begin{cases} \bar{\theta}_A = \frac{1}{N_A} \sum_{j=1}^{N_A} \theta_j \\ \bar{\theta}_B = \frac{1}{N_B} \sum_{j=N-N_B+1}^N \theta_j. \end{cases} \quad (22.15)$$

and the test statistic is the *z -score* of the difference between the means of the two segments:

$$Z_G = \frac{\bar{\theta}_A - \bar{\theta}_B}{\sqrt{\sigma_{\bar{\theta}_A}^2 + \sigma_{\bar{\theta}_B}^2}}. \quad (22.16)$$

To compare the two-sample means, it is necessary to estimate their sample variances. This task is complicated by the fact that one cannot in general just use the sample variance (2.14) in the two segments, because of the correlation between links of the chain. Instead, alternative methods such as that of the batch means in (21.24), or thinning of the chain, should be used to provide a more accurate estimate of the variance of the MCMC (see also Sect. 21.5). A typical application of the Geweke z -score test of convergence is to step the start of segment A forward in time, until the Z_G scores don't exceed approximately ± 3 , which correspond to a $\pm 3\sigma$ deviation in the means of the two segments. The burn-in period that needs to be excised can be identified by inspecting the evolution of the Z_G scores.

Example 22.4 (Z_G score using the chain from Example 12.2) Figure 22.5 shows the Z_G statistic obtained by moving forward the initial 10% segment of the chain of Fig. 22.3, and comparing it with the fixed final 50% of the chain. The variance of the mean in the two segments is calculated using three methods: the sample variance of the data, the sample variance of the thinned chain using every 10-th iteration, and the batch means estimator of the variance according to (21.24). By using all links in the chain, the variance is underestimated because of the correlation among links, leading to erroneously large values of Z_G illustrated as the solid curves in Fig. 22.5. If the chain is thinned by a factor of 10, then the estimates of the variance become larger, and the resulting z -scores show that the chains converge nearly immediately, as is also clear by a visual inspection of the time-evolution of the chain. For the method

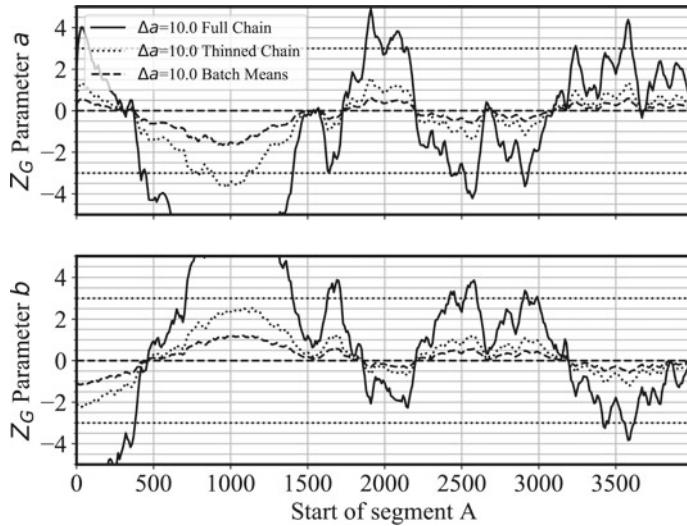


Fig. 22.5 Geweke Z scores using segments A and B of the chain in Fig. 22.3 with, respectively, 10% and 50% of the total chain length. Segment A is stepped forward from the beginning of the chain until iteration 4,000. The results correspond to the chain run with a proposal distribution width of $\Delta a = \Delta b = 10$

of batch means, segment A was divided into $a = 31$ segments of length $b = 32$, and segment B into $a = 70$ segments of length $b = 71$, as an approximation of the initial 10% and final 50% of the chain. This method returns smaller values of the Z_G scores, indicating that the chain reaches convergence almost immediately. Given the uncertainties in the estimate of the variance, it would be prudent to excise the initial 1,500 iterations or so, which are also the ones most affected by the arbitrary choice of the starting point for the chain. \diamond

22.7 The Gelman–Rubin Convergence Test

The Gelman–Rubin test investigates the effect of initial conditions on the convergence properties of the MCMC and makes use of m parallel chains starting from different initial points. Initially, the m chain will be far apart because of the different starting points. As the chains start sampling the stationary distribution, they will have the same statistical properties. The test is based on two estimates of the variance, or variability, of the chains: the *within-chain* variance for each of the m chains W , and the *between-chain* variance B . At the beginning of the chain, W will underestimate the true variance of the model parameters, because the chains have not had time to sample all possible values. On the other hand, B will initially overestimate the variance, because of the different starting points. The test, proposed by A. Gelman

and D. Rubin, [39] defines the ratio of the within-to-between variance as a test to measure convergence of the chains, to identify an initial burn-in period that should be removed because of the lingering effect of initial conditions.

For each parameter, consider m chains of N iterations each, where $\bar{\theta}_i$ is the mean of each chain $i = 1, \dots, m$ and $\bar{\theta}$ the mean of the means:

$$\begin{cases} \bar{\theta}_i = \frac{1}{N} \sum_{j=1}^N \theta_{ij} \\ \bar{\theta} = \frac{1}{m} \sum_{i=1}^m \bar{\theta}_i, \end{cases} \quad (22.17)$$

where θ_{ij} indicates the j -th link of the i -th chain. The between-chain variance B is defined as

$$B = \frac{N}{m-1} \sum_{i=1}^m (\bar{\theta}_i - \bar{\theta})^2, \quad (22.18)$$

where (22.18), B/N is the sample variance of the means $\bar{\theta}_i$. The within-chain variance W is defined as the mean of the sample variance within each chain,

$$W = \frac{1}{m(N-1)} \sum_{i=1}^m \sum_{j=1}^N (\theta_{ij} - \bar{\theta}_i)^2. \quad (22.19)$$

Finally, an overall unbiased estimator for the variance of parameter θ , under the hypothesis that the stationary distribution is being sampled, is a weighted mean of the between-chain and within-chain variances, is

$$\hat{\sigma}_\theta^2 = \left(\frac{N-1}{N} \right) W + \frac{B}{N}. \quad (22.20)$$

It was suggested by S. Brooks and A. Gelman [15] to add an additional term to this estimate of the variance, to account for the variability in the estimate of the means, so that the estimate of the MCMC variance becomes

$$\hat{V} = \hat{\sigma}_\theta^2 + \frac{B}{mN}. \quad (22.21)$$

Convergence of the MCMC can be monitored by use of the following statistic:

$$\hat{R} = \frac{\hat{V}}{W}, \quad (22.22)$$

which is expected to converge to 1 when the stationary distribution in all chains has been reached. Usually the square root of \hat{R} is used in place of the statistic itself, as a means to use a statistic that has the same units of measure as the parameter. The

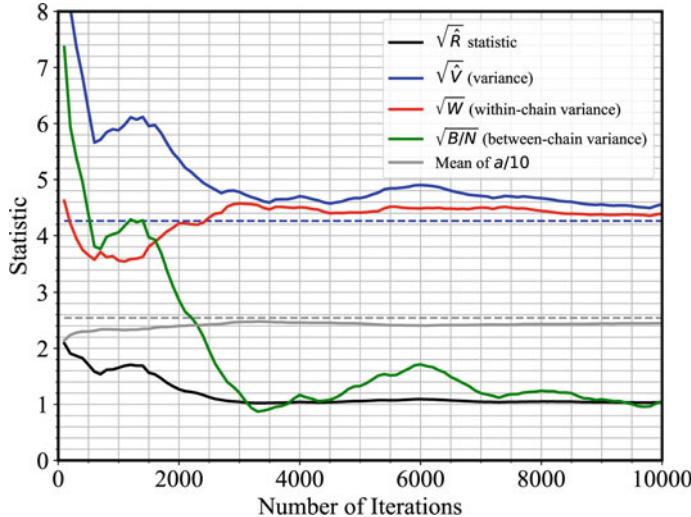


Fig. 22.6 Gelman–Rubin statistic \hat{R} and associated statistics calculated from the $m = 3$ chains with $N = 10,000$ iteration from the same MCMC as in Fig. 22.3. The statistics were calculated in batches of length $b = 100$, with the first batch including the first 100 iterations, and the last including the entire chain

\hat{R} statistic has the following meaning: if its value is significantly larger than one, either the within-chain variance W is too small, typically because the chain is only sampling a portion of the posterior distribution, or \hat{V} is too large, often because the between-chain variance B/N is affected by different initial values of the m chains.

A procedure to test for convergence of the chain using the Gelman–Rubin statistic is to divide the chain into segments of length b , such that the N iterations are divided into N/b batches. The statistic \hat{R} is evaluated for a progressively larger number of batches, so that the last iteration includes the entire chain. Optionally, an initial portion of the chain can be omitted altogether, in order to exclude an initial burn-in period.

Example 22.5 (*Gelman–Rubin statistics on the chain from Example 22.3*) Figure 22.6 shows the Gelman–Rubin statistics associated with the three chains presented in Example 22.3. In this example the chains for parameter a are analyzed, but an identical procedure can be applied to the other parameter b . The three chains have different proposal distributions that result in a different sampling of parameter space, moreover starting from initial values that are significantly offset from the expected best-fit values. The \hat{R} statistic stabilizes to a value close to unity after about 3,000 iterations. In the initial portion of the chain, the between-chain variance B/N is larger than the within-chain variance W , leading to a large value of \hat{R} . The low value for W derives from the fact that the chains are not yet sampling the entirety of parameter space, indicating that convergence of the chain and therefore its sampling has not occurred yet. After that initial portion, all variances and the

Gelman–Rubin statistic stabilize, indicating that from that point on the chains are sampling the posterior distribution.

As part of convergence testing, it is also useful to monitor the value of the parameters, as illustrated also in Fig. 22.6, where the parameter a was divided by 10 for convenience. The evolution of the time average of a further confirms that the sampling from the posterior distribution stabilizes after an initial period of variability of approximately 3,000 iterations. For comparison, the linear regression yields a best-fit value of $a = 25.44 \pm 4.26$, with the corresponding horizontal lines also reported in the figure. Usually one does not have this information available when running an MCMC, but in this case these values serve the purpose to illustrate that the chain has not yet reached the known values. This is likely due to two facts: (a) the effect of the initial values, indicating that an initial burn-in period should be excised from the chain; and (b) that the chain has not been run long enough. It is therefore suggested that an initial portion of the chain, possibly 3,000 iterations or so, be removed, and that the chain is run for a longer number of iterations, before using the chain to infer parameter values. \diamond

22.8 The Raftery–Lewis Diagnostic

An ideal test for the convergence of the MCMC is one that determines the length of the burn-in period and the subsequent length of the chain that are required to achieve a given precision in the estimate of the MCMC parameters. The method proposed by A. Raftery and S. Lewis [84] provides estimates of both quantities, based on just a short test run of the chain. The ultimate goal is to provide an estimate of the q -quantile u of a parameter θ of the chain, or a function of the parameter,

$$P(\theta \leq u) = q.$$

Such quantiles of the posterior distribution of θ can then be used to estimate confidence intervals at a given level of probability, which is usually the goal of running a MCMC. The Raftery–Lewis diagnostic is based on the use of a binary Markov chain associated with the given MCMC, which features convenient properties to estimate the quantiles of interest. The diagnostic returns the number of burn-in iterations m to be excised, the subsequent number of iterations n that are required to achieve a given precision in the estimate of the quantile, and the thinning factor k required for the estimate.

Given a Markov chain X_i for the parameter or function of interest, the first step is to construct an associated stochastic process defined by

$$Z_i = \delta(X_i \leq u) = \begin{cases} 1 & \text{if } X_i \leq u \\ 0 & \text{if } X_i > u. \end{cases}$$

The Z_i process is therefore composed of binary variables that indicate whether, at each step of the Markov chain, the value is greater than the q -quantile u . Of course, the true value of u is unknown, but it can be estimated from the test run of the chain. This binary process is not guaranteed to be a Markov chain, in the sense that it may not obey the Markovian property that the state Z_i is only, at most, dependent on the previous state of the chain Z_{i-1} . Raftery and Lewis argue that it is possible to make Z_i an approximate Markov chain by thinning the process to every k -th iteration when k is sufficiently large, and they also provide a prescription to estimate this thinning factor. Such thinned binary chain may be referred to as $Z_i^{(k)}$, but for the sake of simplicity the earlier notation of Z_i is retained to indicate the binary process that has been sufficiently thinned to approximate it to a binary Markov chain.

Properties of the Z_i process associated with the original chain can now be studied using standard methods of a Markov chain that has a binary Bernoulli distribution as its stationary distribution, as shown in Examples 21.1 and 21.5, with the parameters α and β of the binary Markov chain also estimated from the test run. First, to ensure that the chain has started sampling the posterior stationary distribution π after m steps, the following conditions should be satisfied:

$$\begin{cases} |P(Z_{m+1} = 0/Z_1 = j) - \pi_1| \leq \epsilon \\ |P(Z_{m+1} = 1/Z_1 = j) - \pi_2| \leq \epsilon, \end{cases}$$

where $j = 1, 2$ are the two possible states. The two m -step probabilities represent, respectively, the first and second row of the m -step transition probability matrix (21.15), and the conditions therefore simplify to

$$\begin{cases} \left| \frac{(1 - \alpha - \beta)^m}{\alpha + \beta} \begin{bmatrix} \alpha \\ -\alpha \end{bmatrix} \right| \leq \begin{bmatrix} \epsilon \\ \epsilon \end{bmatrix} \\ \left| \frac{(1 - \alpha - \beta)^m}{\alpha + \beta} \begin{bmatrix} -\beta \\ \beta \end{bmatrix} \right| \leq \begin{bmatrix} \epsilon \\ \epsilon \end{bmatrix}. \end{cases}$$

From this, it is immediate to see that the number m of iterations required to reach convergence is given by

$$|(1 - \alpha - \beta)^m| \leq \frac{\epsilon(\alpha + \beta)}{\max(\alpha, \beta)}. \quad (22.23)$$

A solution of (22.23) therefore yields the first part of the Raftery–Lewis diagnostic, namely, the number of initial iterations m that must be excised from the chain.

Next, the number of iterations required to achieve a given precision in the estimate of the quantile u can be obtained from the asymptotic distribution of the ergodic mean (see Example 21.7) of the associated binary chain Z_n , since

$$\bar{Z}_n = \frac{1}{n} \sum_{k=m+1}^{m+n} Z_n$$

represents the average number of times the chain X had values *lower* than the estimated q -quantile u . The precision in the estimate is specified by requiring that

$$P(q - r \leq \bar{Z}_n \leq q + r) = s,$$

meaning that there is a probability s (e.g., $s = 0.95$) that the estimate \bar{Z}_k of the q -quantile is within $\pm r$ (e.g., $r = 0.01$) of the true value u . Since the distribution of $(\bar{Z}_n - q)$ is a Gaussian of zero mean and variance σ_{as}^2/n , the requirement is equivalent to

$$\int_{-r\sqrt{n}/\sigma_{as}}^{+r\sqrt{n}/\sigma_{as}} f(x)dx = 2 \left(B\left(\frac{r\sqrt{n}}{\sigma_{as}}\right) - \frac{1}{2} \right) = s,$$

where $f(x)$ is the probability distribution of the standard Gaussian and $B(x)$ the associated cumulative distribution (see Eq. (3.13)), using the symmetry of $B(x)$ around $x = 0$. Upon inversion of the cumulative distribution function, the number of iterations n can therefore be estimated as

$$n = \frac{\alpha\beta(2 - \alpha - \beta)}{(\alpha + \beta)^3} \times \frac{\left(B^{-1}\left(\frac{s+1}{2}\right)\right)^2}{r^2}, \quad (22.24)$$

where B^{-1} is the inverse of the cumulative distribution function of the standard Gaussian, and therefore $B^{-1}((s+1)/2)$ is the $(s+1)/2$ -quantile of the standard Gaussian. Equation (22.24) provides the second part of the Raftery–Lewis diagnostic. For the method of estimation of the thinning factor k , the reader is referred to [84].

As an example of a typical application of the Raftery–Lewis diagnostic, consider the estimate of the 90% confidence interval of a parameter θ . In this case, the two quantiles to estimate from the MCMC are the $q = 0.05$ quantile θ_1 and the $q = 0.95$ quantile θ_2 , so that the interval (θ_1, θ_2) contains approximately 90% of the posterior probability for that parameter. First, the trial run of the MCMC can be used to excise a number m of initial iterations according to (22.23) and to determine a thinning factor k for the chain. The burn-in period that needs to be excised from the chain is therefore $m \times k$ links. Then, (22.24) provides an estimate of the number n of thinned links required to estimate each quantile with the specified precision r . The un-thinned chain of length $n \times k$ can be used for the estimate of the quantiles instead, and in that case the solution provided is conservative. A typical choice is $r = 0.01$, meaning that the quantiles are, respectively, 0.05 ± 0.01 and 0.95 ± 0.01 , with a probability

of, say, $s = 0.95$ or 95%. According to (22.24), the number of iterations is strongly dependent on r ($n \propto r^{-2}$), and only mildly on s . The dependence of n and m on the quantile of choice is embedded in the transition probabilities α and β that are estimated based on the choice of the p -quantile u .

Example 22.6 (*Raftery–Lewis diagnostic on the MCMC of Example 22.3*) The Raftery–Lewis diagnostic is run on the chains from Example 22.3, which were used to determine the parameters a and b of the linear regression to the 5-point data from Example 12.2. The three chains start at the same location, $a = 12$ and $b = 6$, with increasingly larger proposal distributions ($\Delta a = \Delta b = 1, 10$, and 50). The chains result in acceptance probabilities of 91.8, 38.6, and 3.5 %. The Raftery–Lewis diagnostic can be implemented using the `raftery.diag` function in the `coda R` package. This was adapted from the original software to implement it, written by A. Raftery and S. Lewis as the Fortran code `gibbsit`. A run of the code on the three chains requires the specification of three tolerances which were chosen at the defaults proposed by the authors ($r = 0.0125$, $s = 0.95$ and $\epsilon = 0.001$), and the desired quantiles, in this case chosen as $q = 0.05$ and 0.95 , so that the chain may be used to estimate a central 90% confidence interval. The test provides the following results, where k is the suggested thinning factor, m is the number of iterations to remove as part of the burn-in, and n the number of iterations required to achieve the specified precision in the estimate of the quantiles. The method also provides estimates for the thinning k_{ind} required to achieve an *independence chain* (e.g., a Markov chain that does not depend on the current state, as opposed to a conventional first-order Markov chain that follows the usual Markovian property) and the number of iterations n_{ind} required for such chain.

Quantile	k	m	n	k_{ind}	n_{ind}
Chain a , proposals $\Delta a = \Delta b = 1$					
$q = 0.05$	9	450	131742	114	1168
$q = 0.95$	9	216	77679	67	1168
Chain a , proposals $\Delta a = \Delta b = 10$					
$q = 0.05$	4	44	14828	18	1168
$q = 0.95$	1	30	10025	16	1168
Chain a , proposals $\Delta a = \Delta b = 50$					
$q = 0.05$	1	136	45487	40	1168
$q = 0.95$	1	127	42882	43	1168
Chain b , proposals $\Delta a = \Delta b = 1$					
$q = 0.05$	6	126	40386	46	1168
$q = 0.95$	12	120	38364	39	1168
Chain b , proposals $\Delta a = \Delta b = 10$					
$q = 0.05$	1	18	6098	8	1168
$q = 0.95$	1	20	6564	12	1168
Chain b , proposals $\Delta a = \Delta b = 50$					
$q = 0.05$	1	157	57480	50	1168
$q = 0.95$	1	149	49315	43	1168

The Raftery–Lewis test provides the following indications:

- (a) The chains with the smallest proposal distribution need to be thinned substantially, and run for substantially longer time, in order to estimate the quantiles with the required precision. For example, using the $q = 0.05$ quantile of the first chain, $m \times k \simeq 4,000$ initial iterations need to be excised, and the chain needs to be run for a subsequent $n \times k \simeq 1,000,000$ iterations.
- (b) The chains with the intermediate and larger proposal distributions have substantially smaller values of $m \times k$ and $n \times k$, indicating that a larger proposal distribution is a better means to estimate the parameters with fewer iterations.
- (c) Different quantiles require chains of different lengths. The reason for this difference is that for extreme quantiles (e.g., to estimate a 99% or a 99.9% confidence interval), the transition probabilities α and β become closer to 0 and 1, while for quantiles closer to the median they are closer to $1/2$, with the corresponding effect of making n in (22.23) smaller or larger, respectively. An equivalent way to look at this result is that the variance of a Bernoulli distribution is with the probability of success α is $\alpha(1 - \alpha)$, which is maximum when $\alpha = 1/2$. \diamond

22.9 Inference with MCMC

The main purpose of a MCMC for regression analysis is to make inferences on the parameters $\theta_1, \dots, \theta_m$ that describe the state of the chain, and to constrain associated functions of the parameters. The key feature of the chain is that, after an initial burn-in period that needs to be excised from further analysis, the chain samples the posterior distribution of the parameters. The shape of this distribution is usually unknown, and therefore the analyst is typically left with the task of estimating the expectation and quantiles of the distribution. For this purpose, the ergodic theorem (see Sect. 21.5) guarantees that the expectation of a parameter θ can be estimated via the respective ergodic mean,² e.g.,

$$\text{E}[\theta] \simeq \frac{1}{n} \sum_{i=1}^n \theta_i, \quad (22.25)$$

where the sum extends over n links of the chain *after* the initial burn-in. Similar expectations can be evaluated for functions of the parameters, $f(\theta)$, when such functions are of interest. This is a practical situation of common occurrence in data analysis, and the ergodic theorem makes it such the expectation of such secondary variables can be obtained immediately as

² In this section, the symbol θ is used to represent one of the m parameters, or, in the case of the function $f(\theta)$, one or several of them. This choice is made to keep the notation as simple as possible.

$$\mathbb{E}[f(\theta)] \simeq \frac{1}{n} \sum_{i=1}^n f(\theta_i), \quad (22.26)$$

which is equivalent to (21.20). Consider, as an example, a MCMC whose state is described by a few parameters that model the shape and the thermodynamic properties of an astronomical source. From these parameters it is also possible to infer secondary properties, such as the distance or the mass of the same source, via an analytic function $f(\theta)$ that does not introduce additional sources of uncertainty. The MCMC can therefore be simply *augmented* by adding, link-by-link, the value of the secondary parameter $f(\theta)$, and then using the ergodic averages to estimate its expectation. An application of this type of data analysis was presented by the author in [12]. This ability to augment a chain to obtain ergodic averages of secondary quantities is a very useful feature of Markov chain Monte Carlo methods.

In addition to the point estimate provided by the ergodic averages (22.25) or (22.26), it is also necessary to provide confidence intervals for the parameters. The sampling distribution of the ergodic averages, discussed in Sect. 21.5, makes it possible to use asymptotic normality of the ergodic average, together with an estimated variance (e.g., by thinning the chain or by batch means), to provide confidence intervals. Alternatively, confidence intervals can be estimated using the property that the MCMC samples the posterior distribution of the parameters, after the initial burn-in period is excised, and therefore the sample distribution of the parameters approximates the true posterior distribution. Confidence intervals can accordingly be determined via the appropriate quantiles of the sampling distribution of the estimators, say calculating the 0.05 and 0.95 quantiles for the purpose of estimating a 90% confidence interval. The Raftery–Lewis diagnostic provides a useful prescription for the estimate of quantiles with a specified precision, especially in terms of the length of the chain required to achieve a given accuracy in the estimate of the quantiles (see Sect. 22.8). The availability of the sample distribution of a parameter makes it such it is also possible to estimate the median, which can be regarded as an alternative estimate of the best-fit value of that parameter.

Example 22.7 (*Confidence intervals on MCMC chains from Example 22.3*) The Raftery–Lewis diagnostic for the MCMC of the 5-point data indicated (see Example 22.6) that the chains should be run for longer than the 10,000 iterations used in Example 22.3, in order to estimate the quantiles with the required precision. The same three chains were therefore re-run for a length of 100,000 iterations, of which the first 10,000, where somewhat conservatively, excised as part of the burn-in period. The sample distributions of the parameters from these chains are shown in Fig. 22.7.

The remaining 90,000 iterations were used to estimate the sample mean of all the links, which is the ergodic average of the parameters, the standard deviation of all the links and for the chains thinned to every 10-th iteration, and the $p = 0.16$ and $p = 0.84$ quantiles to approximate the 68% central confidence interval:

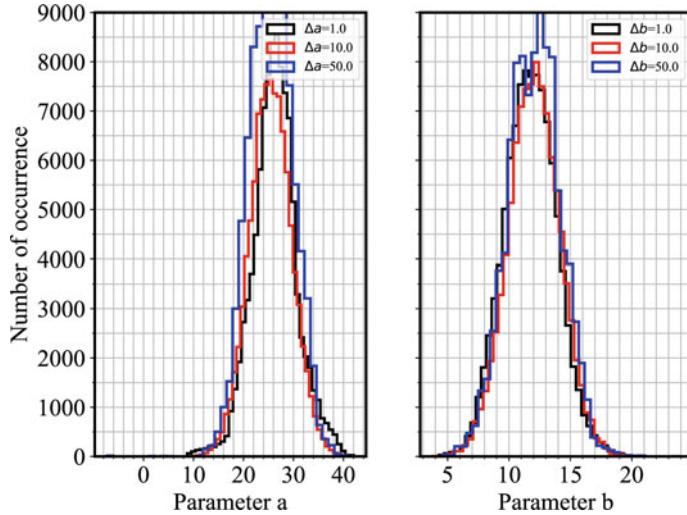


Fig. 22.7 Sample distributions of the parameters a and b as in Fig. 22.4, but for chains with 100,000 iterations

Chain	Mean	St. Dev.		Median	68% C. I.
		All	Thinned		
$a (\Delta a = \Delta b = 1)$	26.2	4.38	4.38	26.3	22.1–30.0 (26.1 ± 3.9)
$a (\Delta a = \Delta b = 10)$	25.4	4.18	4.22	25.4	21.3–29.6 (25.4 ± 4.2)
$a (\Delta a = \Delta b = 50)$	25.5	4.25	4.24	25.6	21.2–29.7 (25.4 ± 4.3)
$b (\Delta a = \Delta b = 1)$	11.8	2.12	2.12	11.7	9.7–13.8 (11.8 ± 2.1)
$b (\Delta a = \Delta b = 10)$	12.1	2.10	2.11	12.1	10.0–14.1 (12.1 ± 2.1)
$b (\Delta a = \Delta b = 50)$	12.1	2.10	2.10	12.2	10.0–14.1 (12.1 ± 2.1)

The best-fit parameters from the traditional linear regression were obtained as $a = 25.44 \pm 4.26$ and $b = 12.06 \pm 2.11$ (see Example 12.2). The chain with the narrowest proposal distributions provides estimates that are somewhat offset from the other two, consistently with the Raftery–Lewis diagnostic that in fact suggested a longer run of that chain. The coincidence between the sample mean and median of the distributions, and of the 68% confidence intervals calculated from the $p = 0.16$ and 0.84 quantiles and from the mean and standard deviation, indicates that both parameters have an approximately normal distribution, and that either can be used as the point estimate.

It is also useful to point out that the variance estimated from all the links, and the one estimated with the thinned chain, are virtually identical for this MCMC. This agreement can be informally explained with the large number of MCMC links used

for the sample variance. Although for a shorter segment of the chain—as was the case for Example 22.4—the variance would not be accurately estimated using all the links, this problem is ameliorated by the length of this chain. Samples do remain correlated to one another, as illustrated in Fig. 22.3, but the averaging of the squares of the deviations over a large number of links results in a more accurate estimate of the variance. ◇

Summary of Key Concepts for This Chapter

Markov chain Monte Carlo (MCMC): A numerical method to implement a Markov chain, with the goal of estimating the posterior distribution of model parameters via

$$P(\theta/Z) \propto p(\theta)\mathcal{L}.$$

Metropolis–Hastings algorithm: A commonly used sampling method to draw and accept or reject candidates for the MCMC. It is based on an acceptance probability that simplifies to a ratio of likelihoods,

$$\alpha(\theta' / \theta_n) = \min \left\{ \frac{\mathcal{L}(\theta')}{\mathcal{L}(\theta_n)}, 1 \right\}$$

when the priors and proposal distributions are uniform.

The Gibbs Sampler: An alternative algorithm to create a MCMC that makes use of the full conditional distribution of each parameter.

Convergence tests: Tests to ensure that the MCMC is sampling the intended posterior distribution. They typically require to excise a *burn-in* time when the MCMC has not yet reached the stationary distribution.

Geweke z-score test: A simple test of convergence that makes use of *z*-scores of two segments of the chain.

Gelman–Rubin test: A convergence test that requires multiple parallel chains and makes use of *between-chain* and *within-chain* variances.

Raftery–Lewis diagnostic: A convergence test that compares a sample of the MCMC to an uncorrelated chain to determine thinning, burn-in time, and required length of the chain to specify quantiles of the distribution with a given precision.

Problems

22.1 ■ Consider the Hubble data of $\log m$ and velocity from Table 11.1.

- (a) Construct a Monte Carlo Markov chain for the fit to a linear model with 10,000 iterations. Use uniform distributions for the prior and proposal distributions of the two model parameters a and b , the latter with widths of 0.2 and 0.04, respectively, for a and b in the neighborhood of the current value. Start the chain at values of $a = 0.2$ and $b = 0.9$.
- (b) After completion of the chain, plot the sample distribution of the two model parameters.

22.2 A one-parameter chain is constructed such that in two intervals A and B the following values are accepted into the chain:

$$\begin{aligned}A &: (10, 11, 13, 11, 10) \\B &: (7, 8, 1, 11, 10, 8),\end{aligned}$$

where A is an initial interval, and B an interval at the end of the chain. Not knowing how the chain was constructed, use the Geweke z score to test whether the chain *might* have converged.

22.3 ■ Consider the 5-point data of Example 12.2.

- (a) Construct a Monte Carlo Markov chain for the parameters of the linear model, with 10,000 iterations. Use uniform distributions for the prior and proposal distributions, the latter with a width of 10 for both parameters. Start the chain at $a = 12$ and $b = 6$.
- (b) After completion of the chain, plot the sample distribution of the two model parameters.
- (c) Estimate the best-fit values of the two parameters and their $1-\sigma$ uncertainties.

22.4 Consider the following portions of two one-parameter chains, run in parallel and starting from different initial positions:

$$\begin{aligned}\theta_1 &= (7, 8, 10, 11, 10, 8) \\ \theta_2 &= (11, 11, 8, 10, 9, 12).\end{aligned}$$

Using two segments of length $b = 3$, calculate the Gelman–Rubin statistic $\sqrt{\hat{R}}$ for the first segment and then for the entire chain, under the hypothesis of uncorrelated samples.

22.5 Consider a Poisson dataset with n measurements, and the 3-parameter step-function model described in Example 22.2. Assuming that the priors on the three parameters λ , μ and m are uniform, show that the full conditional distributions are given by

$$\begin{cases} \pi_\lambda(\lambda) = f_\gamma\left(\lambda; m, \sum_{i=1}^m y_i + 1\right) \\ \pi_\mu(\mu) = f_\gamma\left(\mu; n-m, \sum_{i=m+1}^n y_i + 1\right) \\ \pi_m(m) = \frac{e^{-m\lambda} \lambda^{\sum_{i=1}^m y_i} \cdot e^{-(n-m)\mu} \mu^{\sum_{i=m+1}^n y_i}}{\sum_{l=1}^n \left(e^{-l\lambda} \lambda^{\sum_{i=1}^l y_i} \cdot e^{-(n-l)\mu} \mu^{\sum_{i=l+1}^n y_i}\right)}, \end{cases} \quad (22.27)$$

where $f_\gamma(x; \alpha, r)$ represents the gamma distribution defined in (9.6), with x the independent variable of choice, α the rate parameter and r the shape parameter of the distribution.

22.6 Consider the step-function model described in Example 22.2, and a dataset consisting of the following five measurements:

$$0, 1, 3, 4, 2.$$

Start a Metropolis–Hastings MCMC at $\lambda = 0$, $\mu = 2$ and $m = 1$, and use uniform priors on all three parameters. Assume for simplicity that all parameters can only be an integer, and use uniform proposal distributions that span the ranges $\Delta\lambda = \pm 2$, $\Delta\mu = \pm 2$ and $\Delta m = \pm 2$. The following numbers are drawn in the first three iterations, with u the random number needed to the acceptance of candidates:

Iteration	$\Delta\lambda$	$\Delta\mu$	Δm	u
1	+1	-1	+1	0.5
2	+1	+2	+1	0.7
3	-1	-2	+1	0.1

With this information, calculate the first four links of the Metropolis–Hastings MCMC.

22.7 Consider a Monte Carlo Markov chain constructed with a Metropolis–Hastings algorithm, using uniform prior and proposal distributions. At a given iteration, the chain is at the point of maximum likelihood or, equivalently, minimum χ^2 . Calculate the probability of acceptance of a candidate that corresponds to a likelihood of, respectively, $\Delta\chi^2 = 1, 2$, and 10 .

Chapter 23

Numerical Methods and `python` Codes



Abstract This chapter describes the `python` codes that are provided with this edition of the textbook. The codes are used for the solution of all problems that require numerical methods, and to reproduce numerical tables, examples, and other results presented in the book.

23.1 Analytical and Numerical Methods

Mathematics, statistics, and other sciences generally aim to provide simple analytical solutions to problems. An example of what is meant by the phrase *analytical* is the derivation of the distribution of the χ^2 statistic, defined by (9.5) as the sum of the square of N standardized deviations. Using theoretical arguments, it is possible to prove that its probability distribution function follows (9.7),

$$f_{\chi^2}(x) = \left(\frac{1}{2}\right)^{N/2} \frac{1}{\Gamma(N/2)} e^{-x/2} x^{N/2-1}.$$

Such equation has the convenient properties of being compact and therefore easy to use for applications. This is the type of expression that is generally regarded as analytical.

Upon closer inspection, though, it becomes apparent that it is difficult to quantify exactly the meaning of the word “analytical.” For example, the Gamma function, used in the distribution of the χ^2 statistic, is defined by

$$\Gamma(r) = \int_0^\infty x^{r-1} e^{-x} dx$$

and it has the property that $\Gamma(1/2) = \sqrt{\pi}$. The Gamma function is therefore simply a convenient symbol for an integral that cannot be in general reduced to a simpler form. Another example is the integral defined in (A.2),

$$\operatorname{erf}(z) = \frac{1}{\sqrt{\pi}} \int_{-z}^z e^{-t^2} dt,$$

which is called the *error function* and it is represented by a compact symbol. Yet, the integral itself cannot be immediately simplified to, say, an algebraic form that is simpler to evaluate for a given value of the variable z . The error function is also related to the *incomplete Gamma function*, defined in (7.19) as

$$\gamma(z, M) = \int_0^M x^{z-1} e^{-x} dx$$

via the equation

$$\operatorname{erf}(x) = \gamma(1/2, x^2)/\Gamma(1/2).$$

The integral in the incomplete Gamma function, and therefore the error function, cannot be further simplified, but it *can* be solved with the aid of a series expansion, leading to a relatively simple algorithm for its numerical evaluation (e.g., see the textbook *Numerical Recipes* by W.H. Press [80]). These examples should serve as an illustration that there is no clear line of demarcation between *analytical* and *numerical* solutions. A numerical method can be broadly defined as a series of steps, or an algorithm, that yields the desired solution, as in the case of a series expansion for the incomplete Gamma function. Indeed, the algorithm that provides a numerical solution must itself be analytical, in that it must be coded exactly step-by-step to obtain the result.

Over the years, numerical solutions have become virtually synonymous with computer-aided calculations, given the emphasis on performing repetitive and time-consuming mathematical operations with the aid of computers. That, in turn, requires a programming language that can efficiently implement numerical methods. The data analyst should certainly strive to provide a simple analytical solution when possible, but also be ready to use numerical methods when needed. Analytical and numerical methods are therefore to be treated with the same importance—both are needed to solve problems.

23.2 Introduction to python

The python language was invented by G. van Rossum in 1989,¹ and it has become one of the most popular programming languages. It has a number of convenient features that makes it an excellent language for statistics and many other applications.

First, the language is easy to learn and to use, in that there is a very limited amount of overhead in the creation of basic structures such as multi-dimensional arrays, or

¹ See van Rossum's Foreword in M. Lutz's first edition of *Programming Python* [66].

to perform mathematical operations. For example, in one of the scripts provided (`5points.py`), the following lines

```
dataTypes=[('number','I'),('x','f'),('y','f')]
data=np.genfromtxt('5points.dat',skip_header=1,dtype=dataTypes)
x=data['x']
y=data['y']
yerr=y**0.5
```

are used to read the file `5points.dat`, which contains the following:

n	X	Y
1	0.0	25
2	1.0	36
3	2.0	64
4	3.0	49
5	4.0	81

The code skips the first line of input (`skip_header=1`), and lets the user access its columns of data by given names (e.g., `x=data['x']`) that were specified via the `dataTypes` list. Moreover, the last column performs the square root on all terms of the list (`yerr=y**0.5`), without the need of a `for` loop.

Second, the language has a vast support in the form of packages and libraries that provide plotting functions, built-in mathematical and statistical functions, and a variety of functions that implement numerical methods. For example, the following lines from the same script,

```
popt,pcov=curve_fit(linear,x,y,sigma=yerr,absolute_sigma=True)
ymodel=linear(x,popt[0],popt[1])
```

perform the linear regression according to (11.5) using the function `curve_fit`, returning the best-fit values (`popt`) and the covariance matrix (`pcov`) in one simple step. The same function can also be used for non-linear regression (see, e.g., Sect. 11.7), when the `linear` function is replaced by another user-defined function. In that case, `curve_fit` provides a numerical solution to the maximum-likelihood method. In particular, the `numpy` and `scipy` libraries contain an extensive collection of functions, including virtually all distribution functions and their related quantities (e.g., quantiles, moments, etc.). In addition, the `matplotlib` library provides excellent support for plots and graphs—all figures for the book were made using functions from this package.

Finally, the `python` language is free and easy to install for any operating system, and it has an extensive network of volunteer developers and debuggers that ensures a great user support system. These qualities are the reasons why this language was chosen for performing numerical methods associated with this textbook. Moreover, readers who prefer other languages may still find the codes a useful reference for the development of their own codes, thanks to the overall ease with which `python` scripts can be parsed even by a non-expert user.

It is not the purpose of this textbook to provide a complete introduction to the python language or to numerical methods, especially since there are many tutorials available both in print (e.g., [66] for python and [80] for numerical methods) and online. Rather, this chapter can be used to learn about the general features of the codes (Sect. 23.3) and the specific codes provided to solve problems of this textbook (Sect. 23.4), so that the user can immediately proceed with their use. The user is also encouraged to parse and edit the codes as they wish.

23.3 General Features of **python** Codes for this Textbook

First of all it is necessary to provide a note on the development of the python language. At the time of writing the current python version was 3.8, and that is the version that was used to develop the codes provided with the textbook. In most Unix/Linux/macOS systems, the software is conveniently accessed from the command line of a shell via

```
python3 [python file]
```

which executes the python commands contained in the file. For example,

```
python3 5points.py
```

executes the commands contained in the `5points.py` code. In non-unix systems, python can sometimes be accessed through additional software, and the codes provided will be run accordingly. Note that the “`python3`” command is a reminder that, at some point in the past, python evolved into two parallel lines, of which version 2 (accessible as “plain” python) is considered a legacy version, and `python3` is regarded as the version that is being more actively developed.

23.3.1 Structure of the Codes

Most of the codes described below in Sect. 23.4 provide solutions to multiple problems, according to a common topic or data. For example, `thomson.py` is used to solve several problems that are presented through the book over multiple chapters, all having in common the same two Thomson datasets described in Sect. 2.4. When executing a code, e.g.,

```
python3 thomson.py
```

the output will generally be of this type:

```
For problems 2.3 and 5.2
=====
Table 1: 12 measurements; table 2: 11 measurements
Table1: w/q = 13.275 +- 8.456, I = 312.917 +- 93.358
(mean and std. dev.)
...
For problem5.2
Table 1 m/e [0.57 0.34 0.43 0.32 0.48 0.4 0.4 0.35 0.5
0.4 0.4 0.39]
Table1: m/e = 0.415 +- 0.071 (+- 0.021 for mean), 90 C.I.:
0.297 - 0.533
```

The user should be able to immediately identify the solution of the specific problem at hand, and parse the code using the problem identifiers (e.g., `problem5.2`) provided in output.

All the data presented in the book are also provided in electronic format. For example, the two files `thomson1.dat` and `thomson2.dat` contain the Thomson data in the form

Gas	w/q	I	m/e	m/e*	v	v*
air	4.6	230	0.57	0.58	4	4.000
air	18	350	0.34	0.34	10	10.286
air	6.1	230	0.43	0.43	5.4	5.304
air	25	400	0.32	0.32	12	12.500

...

These data are automatically read into the relevant code by appropriate functions, so that the user does not have to specify them directly. All the user has to do is to ensure that the data files needed for a given code (as described in Sect. 23.4) are in the path of the `python` script.

In a few cases, such as the *iris* data introduced in Sect. 13.2, the code (`iris.py`) require a minimal level of user editing before execution. For example, Problems 14.4 and 14.5 are similar in that they both require a linear regression of the same dependent variable (the sepal length) using a different independent variable, also for two different data (*iris setosa* for Problem 14.4 and *iris versicolor* for Problem 14.5). Instead of adding an extra layer of user input to the code, the code was kept “bare” and non-interactive, and the user is required instead to identify the lines that need to be commented/uncommented for a specific problem. The user should find the task simple; for example, in `iris.py`, the initial setting is for the *iris versicolor*:

```
# =====
# ID: (select one that applies)
#     1 - Iris Setosa
#     2 - Iris versicolor
#     3 - Iris Virginica
IDLabel=['Iris Setosa','Iris versicolor','Iris Virginica']
ID=2
```

To solve problems related to *iris setosa* instead, the only change required will be to modify the corresponding line to

```
...
ID=1
```

In so doing, the user is encouraged to understand the methods being used in the codes, instead of using the codes as a “black box,” and to edit the code for any additional specific needs.

23.3.2 Functions, Library Import and Settings

In order to keep the codes as simple as possible, a number of common python commands have been relegated to an auxiliary file named `imports.py`. As a result, scripts begin with a line that executes its commands,

```
exec(open('imports.py').read())
```

This file contains three types of commands, illustrated in the following.

- (a) The import of functions from libraries, such as

```
import numpy as np
from matplotlib import pyplot as plt
...
```

which makes functions such as `np.genfromtxt` available in the script. The `numpy` and `matplotlib` libraries may come with the local `python` installation, or may have to be installed by the user. Installation instructions will vary by operating system but are generally easy and free of charge.

(b) The definition of new functions. Although the vast majority of functions are from existing libraries, there are a few functions that were developed by the author specifically for this book. An example is the function that determines the intrinsic scatter according to (17.3),

```
def intrinsicScatter(y, ymodel, yerr, dof):
    result=0
    N=len(y)
    for i in range(N):
        result+=(y[i]-ymodel[i])**2/(N-dof)-yerr[i]**2/N
    return result**0.5
```

or the C statistic defined in (15.3), which was coded as

```
def CContribTwo(Ni, mu):
    if (Ni==0):
        return 2*mu
```

```
if (Ni>0):
    return 2*(mu - Ni +Ni*np.log(Ni/mu)).
```

In a few cases, functions were defined within the specific code that uses them. This was the case, for example, of the code `fatherMother.py`, where a few specific functions were needed to deal with the binned nature of the Pearson data.

(c) The definition of certain variables or settings of common use, such as

```
plt.rcParams["font.family"] = "Times New Roman"
plt.rcParams.update({'font.size': 18})
plt.rcParams['axes.linewidth'] = 2.0
```

This also illustrates how python provides the means to set a variety of variables that apply to all subsequent codes.

23.3.3 Data Associated with the Codes

As already noted in Sect. 23.3.1, all data used in this book and in the codes are provided to the user. A typical example of a code that makes use of data is `hubble.py`, which is used to solve problems related to the Classic Experiment in Sect. 11.6. For this experiment, the data are in `hubbleHumasonData.dat`:

Nebula	v	N	m
Virgo	890	7	12.5
Pegasus	3810	5	15.5
Pisces	4630	4	15.4
Cancer	4820	2	16.0
Perseus	5230	4	16.4
Coma	7500	3	17.0
UMajor	11800	1	18.0
Leo	19600	1	19.0
unnamed	2350	16	13.8
unnamed	630	21	11.6

These data are read within the `hubble.py` code with

```
dataTypes=[('name','S8'),('v','f'),('N','I'),('m','f')]
data=np.genfromtxt('hubbleHumasonData.dat',skip_header=1,
    dtype=dataTypes)
v=data['v']
m=data['m']
```

The user therefore can simply proceed to the use of the code, for example by issuing the command

```
python3 hubble.py
```

In some cases, the data are in the form of just a few data points, and therefore certain codes (e.g., `10points.py` for Problem 9.5) contain the data themselves, without the need to import them from an external file.

23.4 Description of Codes

This section provides a short description of all the codes provided in this textbook. For each code, there is a description of what data files are required, and what problems are solved. Most codes can be executed directly without any user input. Special instructions for certain codes are also noted.

File name: `thomson.py`

Uses data files: `thomson1.dat`, `thomson2.dat`

Problems: 2.3, 2.4, 5.1, 5.2, 7.1, 7.5, 9.4, 19.1, 19.2

File name: `fatherMother.py`

Uses data files: `fatherMotherStature.dat`

Problems: 2.7, 14.1, 14.2

File name: `mendel.py`

Uses data files: `mendel.dat`

Problems: 3.8, 3.9, 6.2, 9.7, 9.8

File name: `gehrels.py`

Uses data files: (None)

Problems: 7.2

File name: `pressureRatio.py`

Uses data files: `pressureRatio.dat`

Problems: 8.1, 8.2, 8.4, 8.5, 11.6, 12.6, 17.1, 17.2, 18.1, 18.2, 18.3, 18.4

Note: For Problems 18.2–18.3, the user is required to edit the script to determine which data to use. The selection is done in the following lines:

```
print('BCES fits, Radius vs. Ratio')
print('Problems 18.1, 18.2 and 18.3')
corrFactor=2.0 # for problem 18.3
#corrFactor=1.0 # for problem 18.2
```

where the user needs to comment/uncomment the relevant line (the # symbol marks a line that has been commented). These problems also require installation of the `bces.bces` package, which is freely available.

File name: `10points.py`

Uses data files: (None)

Problems: [9.5](#)

File name: `prob9.6.py`

Uses data files: (None)

Problems: [9.6](#)

File name: `ContingencyTables.py`

Uses data files: (None)

Problems: [10.1, 10.3, 10.5, 10.6](#)

Note: This code provides a variety of tests (in particular the χ^2 test with and without the Yule correction, and the Fisher exact test) and it can be used for any contingency table. A few tables, including those used in the problems for Chap. 10, have been defined at the beginning of the code, and the user must select the table of choice prior to running the code. The selection is done by uncommenting the table of choice, e.g.,

```
# Begin choice of contingency table =====
#problem 10.1 (Simple N=10 table, to illustrate method)
print('Problem 10.1')
a=2
b=3
c=1
d=4
tabLabel='Problem 10.1'
# =====
'''
# problem 10.3 and Example 10.4 of textbook
print('Problem 10.3')
a=1
b=4
c=5
d=0
tabLabel='Problem 10.3 and Example 10.4'
# Invert columns to satisfy m1<m2, equivalent table
a=0
b=5
c=4
d=1
'''
```

Notice how the triple quotation symbol marks the beginning of a multiple-line comment. The default is the contingency table for Problem 10.1, which is uncommented.

File name: clinicalTestingContingency.py
Uses data files: (None)
Problems: 10.8, 10.9, 10.10, 10.11

File name: ch3Distributions.py
Uses data files: (None)
Problems: 10.12

Note: This code is also used to reproduce many of the results in Chap. 3. The user may find this code convenient to learn how `scipy` handles distribution functions.

File name: hubble.py
Uses data files: hubbleHumasonData.dat
Problems: 11.1, 12.1, 14.3, 17.4, 20.5

File name: 5points.py
Uses data files: 5points.dat
Problems: 11.2, 12.3, 17.5, 19.3, 20.6, 20.7

File name: 6points.py
Uses data files: 6points.dat
Problems: 11.3

File name: iris.py
Uses data files: irisData.dat
Problems: 13.1, 13.2, 13.6, 13.7, 13.8, 14.4, 14.5
Note: This code requires the user to select the type of *iris* of choice, by editing the script. In addition to the selection of the dataset of choice, which was illustrated in Sect. 23.3.1, the user is required to select the independent variable for the regression:

```
print('Problem 14.4 and 14.5')
# x1: sepal width
# x2: petal length
#xReg=x1 #using sepal width as predictor variable, for problem
#       14.4
xReg=x2 # using petal length as predictor variable, for problem
#       14.5
results=stats.linregress(xReg,y)
```

The default is set to the solution of Problem 14.4.

File name: `rDist.py`

Uses data files: (None)

Problems: 14.6

File name: `CiAsympt.py`

Uses data files: (None)

Problems: 15.1

File name: `CiApprox.py`

Uses data files: (None)

Problems: 15.2, 15.3

File name: `cstatChi.py`

Uses data files: (None)

Problems: 15.4, 16.8

File name: `linearCash.py`

Uses data files: (None)

Problems: 16.1

File name: `cstatLinear.py`

Uses data files: `codesBonamenteSpenceBook.tar`

Problems: 16.6, 16.7

Note: Problems 16.6 and 16.7 are based on the paper *A semi-analytical solution to the maximum-likelihood fit of Poisson data to a linear model using the Cash statistic*, by M. Bonamente and D. Spence [14]. For that paper, the authors developed python software that is also used for these problems, and for other material in Chap. 16 of the textbook. The code, ancillary data, and instructions for their use are provided in `codesBonamenteSpenceBook.tar`.

File name: `monteCarlo.py`

Uses data files: (None)

Problems: 20.3, 20.4

File name: `mcmc.py`

Uses data files: `hubbleHumasonData.dat`

Problems: [22.1](#)

File name: `prob22.2.py`

Uses data files: (None)

Problems: [22.2](#)

File name: `mcmc5points.py`

Uses data files: `5points.dat`

Problems: [22.3](#)

Note: This code can be used as a template for running MCMC on any data. The output from this code is a MCMC that can be analyzed with the script `mcmcInference.py` below.

File name: `mcmcInference.py`

Uses data files: (output from `mcmc5points.py`)

Problems: [22.3](#)

Note: This script assumes that a MCMC was generated first with the code `mcmc5points.py`.

File name: `prob22.4.py`

Uses data files: (None)

Problems: [22.4](#)

23.5 Numerical Methods for Tables in Appendix

All numbers provided in Tables A.1 through A.25 can be reproduced with the help of appropriate python routines. This section illustrates how the numbers can be obtained. For each section of the appendix, a short python code is provided with an example of how the relevant functions are used (e.g., `chi2` for the χ^2 distribution) and a numerical example that reproduces numbers from the tables. The user can follow these examples for their specific purposes.

Appendix: [A.1](#)

Topic: *Gaussian distribution and error function*

Functions used: `scipy.stats.norm`

Codes provided: `appendixA1.py`

The `norm` distribution, like other continuous or discrete distributions, has a number

of so-called *methods* associated with it in python. One of these is the probability distribution function, accessible as `norm.pdf`, and the cumulative distribution function, `norm.cdf`. These functions are used to reproduce the numbers in the tables for Appendix A.1. For example,

```
z=1.0
B=norm.cdf(z,loc,scale)
A=2*B-1
print('Area within +-%2.1f: %3.4f'%(z,A))
```

illustrates how the integrals $A(z)$ and $B(z)$ are evaluated numerically.

Appendix: A.2

Topic: *Poisson distribution and Gehrels approximation*

Functions used: `scipy.stats.poisson`, `muupGehrels`, `muloGehrels`

Codes provided: `appendixA2.py`

The `poisson` distribution can be used in a similar way to `norm`. The Gehrels approximations for upper and lower limits are implemented by two functions (`muupGehrels` and `muloGehrels`) that are defined in `imports.py`.

Appendix: A.3

Topic: *The gamma, beta and r distributions*

Functions used: `scipy.stats.gamma`, `beta`, `rdist`

Codes provided: `rDist.py`

The gamma distribution is implemented as `gamma`, the beta distribution as `beta`, and the symmetric beta distribution, or *r*-distribution, as `rdist`. The `rDist.py` code, also provided for Problem 14.6, illustrates the use of `beta` and `rdist`.

Appendix: A.4

Topic: *The χ^2 distribution*

Functions used: `scipy.stats.chi2`

Codes provided: `appendixA4.py`

The χ^2 distribution is available as `chi2`. The code illustrates its use to reproduce the critical values of the distribution, as in Table A.7. The one-sided critical value is also the quantile or percentile, and it is the inverse of the cumulative distribution. The quantile is accessed directly via the `chi2.ppf` (percent point function) method,

```
dof=20
p=0.90
chiCrit=chi2.ppf(p,dof)
```

Appendix: A.5

Topic: *The F distribution*

Functions used: `scipy.stats.f`

Codes provided: `appendixA5.py`

The *F* distribution is available as `f`, and the one-sided critical values can be reproduced with the help of the code provided. This code can be used to obtain additional values that were not reported in the tables.

Appendix: A.6

Topic: *The Student t distribution*

Functions used: `scipy.stats.t`

Codes provided: `appendixA6.py`

The code can be used to calculate integrals of the *t* distribution (`t`) reported in the appendix, and also the critical values reported in Table 9.1.

Appendix: A.7

Topic: *The r-distribution and the linear correlation coefficient*

Functions used: `scipy.stats.rdist`

Codes provided: `appendixA7.py`

The linear correlation coefficient r is distributed as an *r*-distribution (`rdist`), also known as the symmetric beta distribution. The square of the linear correlation coefficient, r^2 , is distributed like a standard beta distribution. The code provided illustrates critical values for the distributions of r and r^2 , and the equivalence between the two methods of hypothesis testing.

Appendix: A.8

Topic: *Kolmogorov–Smirnov statistics*

Functions used: `scipy.stats.kstwo, kstwobign`

Codes provided: `appendixA8.py`

The distributions of the D_N statistic defined in (19.7) and of the D_{NM} statistic defined in (19.11) can be evaluated with the `kstwo` function. The `kstwobign` can be used to evaluate the asymptotic distribution of $\sqrt{N}D_N$ for large N . (Note that the `ksone` implements a different Kolmogorov–Smirnov test not discussed in this textbook.)

Appendix: Numerical Tables

A.1 The Gaussian Distribution and the Error Function

The Gaussian or normal distribution was defined in (3.8) as

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}},$$

where μ is the mean and σ^2 the variance of the distribution. The standard Gaussian is a Gaussian distribution with zero mean and unit variance, and it is obtained from (3.8) with a change of variable

$$z = \frac{x - \mu}{\sigma},$$

and it has a probability distribution function

$$g(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}.$$

The maximum value of a Gaussian is obtained for $x = \mu$, and the value of x , where the Gaussian is $a \leq 1$ times the peak value is given by

$$z = \frac{x - \mu}{\sigma} = \sqrt{-2 \ln a}. \quad (\text{A.1})$$

Figure A.1 shows a standard Gaussian normalized to its peak value, and values of a times the peak value are tabulated in Table A.1. The Half Width at Half Maximum (HWHM) has a value of approximately 1.18σ , as can also be seen from the intersection of the normalized distribution function with the horizontal line at a value of 0.5 in Fig. A.1.

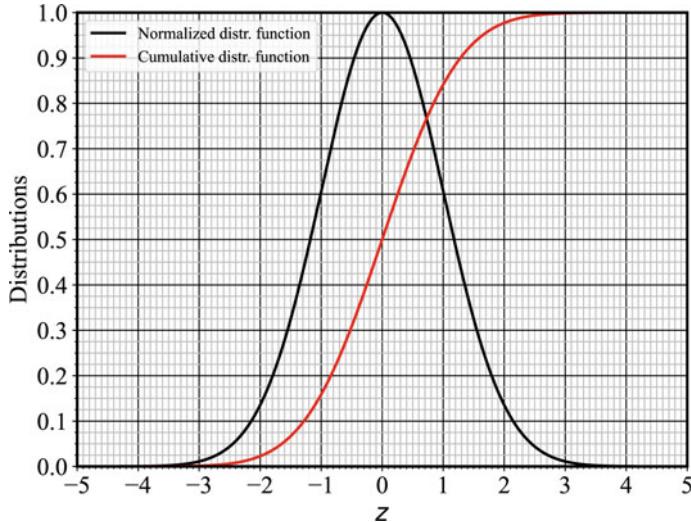


Fig. A.1 Normalized probability distribution function of a standard Gaussian ($\mu = 0$ and $\sigma = 1$), and the associated cumulative distribution function

Table A.1 Values of a times the peak value for a Gaussian distribution

a	z								
0.980	0.201	0.960	0.286	0.940	0.352	0.920	0.408	0.900	0.459
0.880	0.506	0.860	0.549	0.840	0.591	0.820	0.630	0.800	0.668
0.780	0.705	0.760	0.741	0.740	0.776	0.720	0.811	0.700	0.845
0.680	0.878	0.660	0.912	0.640	0.945	0.620	0.978	0.600	1.011
0.580	1.044	0.560	1.077	0.540	1.110	0.520	1.144	0.500	1.177
0.480	1.212	0.460	1.246	0.440	1.281	0.420	1.317	0.400	1.354
0.380	1.391	0.360	1.429	0.340	1.469	0.320	1.510	0.300	1.552
0.280	1.596	0.260	1.641	0.240	1.689	0.220	1.740	0.200	1.794
0.180	1.852	0.160	1.914	0.140	1.983	0.120	2.059	0.100	2.146
0.080	2.248	0.060	2.372	0.040	2.537	0.020	2.797	0.010	3.035

The *error function* is usually defined as

$$\text{erf}(z) = \frac{1}{\sqrt{\pi}} \int_{-z}^z e^{-t^2} dt \quad (\text{A.2})$$

and it is related to the integral $A(z)$ of the Gaussian distribution defined in (3.11),

$$A(z) = \int_{\mu-z\sigma}^{\mu+z\sigma} f(x)dx = \frac{1}{\sqrt{2\pi}} \int_{-z}^z e^{-\frac{t^2}{2}} dt.$$

The relationship between the two integrals is given by

$$\operatorname{erf}\left(z/\sqrt{2}\right) = A(z). \quad (\text{A.3})$$

The function $A(z)$ is the integrated probability of a Gaussian distribution between $\mu - z\sigma$ and $\mu + z\sigma$. The number z therefore represents the number of standard deviations by which the interval extends in each direction. The function $A(z)$ is tabulated in Table A.2, where each number in the table corresponds to a number z given by the number in the left column (e.g., 0.0, 0.1, etc.), and with the second decimal digit given by the number in the top column (e.g., the number 0.007979 in the table corresponds to $z = 0.01$).

The cumulative distribution of a standard Gaussian function was defined in (3.13) as

$$B(z) = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt,$$

and it is therefore related to the integral $A(z)$ by

$$B(z) = \frac{1}{2} + \frac{A(z)}{2}.$$

The values of $B(z)$ are tabulated in Table A.3. Each number in the table corresponds to a number z given by the number in the left column (e.g., 0.0, 0.1, etc.), and with the second decimal digit is given by the number in the top column (e.g., the value of 0.503990 corresponds to $z = 0.01$).

Critical values of the standard Gaussian distribution functions corresponding to selected values of the integrals $A(z)$ and $B(z)$ are shown in Table A.4. They indicate the value of the variable z required to include a given probability, and are useful for either two-sided or one-sided confidence regions in the hypothesis testing.

Table A.2 Values of the integral $A(z)$ as a function of z , the number of standard errors σ

	0	1	2	3	4	5	6	7	8	9
0.0	0.000000	0.007979	0.015957	0.023933	0.031907	0.039878	0.047844	0.055806	0.063763	0.071713
0.1	0.079656	0.087591	0.095517	0.103434	0.111340	0.119235	0.127119	0.134990	0.142847	0.150691
0.2	0.158519	0.166332	0.174129	0.181908	0.189670	0.197413	0.205136	0.212840	0.220523	0.228184
0.3	0.235823	0.243439	0.251032	0.258600	0.266144	0.273661	0.281153	0.288618	0.296055	0.303464
0.4	0.310844	0.318194	0.325515	0.332804	0.340063	0.347290	0.354484	0.361645	0.368773	0.375866
0.5	0.382925	0.389949	0.396937	0.403888	0.410803	0.417681	0.424521	0.431322	0.438086	0.444810
0.6	0.451494	0.458138	0.464742	0.471306	0.477828	0.484308	0.490746	0.497142	0.503496	0.509806
0.7	0.516073	0.522296	0.528475	0.534610	0.540700	0.546746	0.552746	0.558700	0.564609	0.570472
0.8	0.576289	0.582060	0.587784	0.593461	0.599092	0.604675	0.610211	0.615700	0.621141	0.625534
0.9	0.631880	0.637178	0.642428	0.647629	0.652783	0.657888	0.662945	0.667954	0.672914	0.677826
1.0	0.682690	0.687505	0.692272	0.696990	0.701660	0.706282	0.710856	0.715381	0.719858	0.724287
1.1	0.728668	0.733001	0.737287	0.741524	0.745714	0.749856	0.753952	0.757999	0.762000	0.766954
1.2	0.769861	0.773721	0.777555	0.781303	0.785025	0.788701	0.792331	0.795916	0.799455	0.802950
1.3	0.806399	0.809805	0.813165	0.816482	0.819755	0.822984	0.826170	0.829313	0.832414	0.835471
1.4	0.838487	0.841461	0.84493	0.847283	0.850133	0.852942	0.855710	0.858439	0.861127	0.863776
1.5	0.866386	0.868957	0.871489	0.873984	0.876440	0.878859	0.881240	0.883585	0.885894	0.888166
1.6	0.890402	0.892603	0.894768	0.896899	0.898995	0.901057	0.903086	0.905081	0.907043	0.908972
1.7	0.910869	0.912735	0.914568	0.916370	0.918141	0.919882	0.921593	0.923273	0.924924	0.925546
1.8	0.928140	0.929705	0.931241	0.932750	0.934232	0.935687	0.937115	0.938517	0.939892	0.941242
1.9	0.942567	0.943867	0.945142	0.946394	0.947621	0.948824	0.950005	0.951162	0.952297	0.955409
2.0	0.954500	0.955569	0.956617	0.957644	0.958650	0.959636	0.960602	0.961548	0.962475	0.963383
2.1	0.964272	0.965142	0.965994	0.966829	0.967646	0.968445	0.969228	0.969994	0.970743	0.971476
2.2	0.972194	0.972895	0.973582	0.974253	0.974909	0.975551	0.976179	0.976793	0.977393	0.977979
2.3	0.978552	0.979112	0.979660	0.980194	0.980717	0.981227	0.981725	0.982212	0.982688	0.983152
2.4	0.983605	0.984048	0.984480	0.984902	0.985313	0.985715	0.986107	0.986489	0.986862	0.987226

(continued)

Table A.2 (continued)

Table A.3 Values of the integral $B(z)$ as a function of z

	0	1	2	3	4	5	6	7	8	9
0.0	0.500000	0.503990	0.507979	0.511967	0.515954	0.519939	0.523922	0.527903	0.531882	0.535857
0.1	0.533828	0.543796	0.547759	0.551717	0.555670	0.559618	0.563560	0.567495	0.571424	0.575346
0.2	0.579260	0.583166	0.587065	0.590954	0.594835	0.598707	0.602568	0.606420	0.610262	0.614092
0.3	0.617912	0.621720	0.625516	0.629300	0.633072	0.636831	0.640577	0.644309	0.648028	0.651732
0.4	0.655422	0.659097	0.662758	0.666402	0.670032	0.673645	0.677242	0.680823	0.684387	0.687933
0.5	0.691463	0.694975	0.698469	0.701944	0.705402	0.708841	0.712261	0.715661	0.719043	0.722405
0.6	0.725747	0.729069	0.732371	0.735653	0.738914	0.742154	0.745373	0.748571	0.751748	0.754903
0.7	0.758037	0.761148	0.764238	0.767305	0.770350	0.773373	0.776373	0.779350	0.782305	0.785236
0.8	0.788145	0.791030	0.793892	0.796731	0.799546	0.802338	0.805106	0.807850	0.810571	0.813267
0.9	0.815940	0.818589	0.821214	0.823815	0.826392	0.828944	0.831473	0.833977	0.836457	0.838913
1.0	0.841345	0.843753	0.846136	0.848495	0.850830	0.853141	0.855428	0.857691	0.859929	0.862144
1.1	0.864334	0.866501	0.868643	0.870762	0.872857	0.874928	0.876976	0.879000	0.881000	0.882977
1.2	0.884931	0.886661	0.888768	0.890652	0.892513	0.894351	0.896166	0.897958	0.899728	0.901475
1.3	0.903200	0.904902	0.906583	0.908241	0.9099878	0.911492	0.913085	0.914657	0.916207	0.917736
1.4	0.919244	0.920731	0.922197	0.923642	0.925067	0.926471	0.927855	0.929220	0.930564	0.931888
1.5	0.933193	0.934479	0.935745	0.936992	0.938220	0.939430	0.940620	0.941793	0.942947	0.944083
1.6	0.945201	0.946301	0.947384	0.948450	0.949498	0.950529	0.951543	0.952541	0.953522	0.954486
1.7	0.955435	0.956367	0.957284	0.958185	0.959071	0.959941	0.960797	0.961637	0.962462	0.963273
1.8	0.964070	0.964853	0.965621	0.966375	0.967116	0.967844	0.968558	0.969259	0.969946	0.970621
1.9	0.971284	0.971934	0.972571	0.973197	0.973811	0.974412	0.975003	0.975581	0.976149	0.976705
2.0	0.977250	0.977785	0.978309	0.978822	0.979325	0.979818	0.980301	0.980774	0.981238	0.981692
2.1	0.982136	0.982571	0.982997	0.983415	0.983823	0.984223	0.984614	0.984997	0.985372	0.985738
2.2	0.986097	0.986448	0.986791	0.987127	0.987455	0.987776	0.988090	0.988397	0.988697	0.988990
2.3	0.989276	0.989556	0.989830	0.990097	0.990359	0.990614	0.990863	0.991106	0.991344	0.991576
2.4	0.991803	0.992024	0.992240	0.992451	0.992657	0.992858	0.993054	0.993341	0.993613	

(continued)

Table A.3 (continued)

Table A.4 Table of critical values of the standard Gaussian distribution that include a given probability, for two-sided confidence intervals $(-z, z)$ of the integral $A(z)$, and for one-sided intervals $(-\infty, z)$ of the integral $B(z)$

Probability	Two-sided z	One-sided z
0.01	0.013	-2.326
0.05	0.063	-1.645
0.10	0.126	-1.282
0.20	0.253	-0.842
0.30	0.385	-0.524
0.40	0.524	-0.253
0.50	0.674	-0.000
0.60	0.842	0.253
0.70	1.036	0.524
0.80	1.282	0.842
0.90	1.645	1.282
0.95	1.960	1.645
0.99	2.576	2.326
0.999	3.290	3.090
0.9999	3.890	3.718

A.2 Upper and Lower Limits for a Poisson Distribution

Estimates of the parent mean of a Poisson distribution from the single measurement of n_{obs} counts was discussed in Sect. 7.5. The Gehrels approximation [38] can be used to calculate upper and lower limits for the Poisson mean according to (7.8),

$$\begin{cases} \mu_{up} = n_{obs} + \frac{S^2 + 3}{4} + S \sqrt{n_{obs} + \frac{3}{4}} \\ \mu_{lo} = n_{obs} \left(1 - \frac{1}{9n_{obs}} - \frac{S}{3\sqrt{n_{obs}}} \right)^3, \end{cases}$$

where n_{obs} is the number of counts. The confidence level is described by the Poisson parameter S , corresponding to the number of standard deviations σ for a Gaussian distribution. For example, $S = 1$ corresponds to an 84.1% confidence level, $S = 2$ to a 97.7%, and $S = 3$ corresponds to 99.9%; see Table 7.2 for correspondence between values of S and probability. Tables A.5 and A.6 report the upper and lower limits to the Poisson mean as a function of the number of counts, and for selected values of the parameter S . More accurate approximations are described in [38].

Table A.5 Selected upper limits for the Poisson mean using the Gehrels approximation

n_{obs}	Gehrels Upper Limits		
	Poisson parameter S or confidence level		
	$S = 1$ ($1-\sigma$, or 84.1%)	$S = 2$ ($2-\sigma$, or 97.7%)	$S = 3$ ($3-\sigma$, or 99.9%)
0	1.87	3.48	5.60
1	3.32	5.40	7.97
2	4.66	7.07	9.97
3	5.94	8.62	11.81
4	7.18	10.11	13.54
5	8.40	11.55	15.19
6	9.60	12.95	16.79
7	10.78	14.32	18.35
8	11.96	15.67	19.87
9	13.12	16.99	21.37
10	14.28	18.31	22.84
20	25.56	30.86	36.67
30	36.55	42.84	49.64
40	47.38	54.52	62.15
50	58.12	66.00	74.37
60	68.79	77.34	86.38
70	79.41	88.57	98.23
80	89.99	99.72	109.96
90	100.53	110.80	121.58
100	111.04	121.82	133.11

Table A.6 Selected lower limits for the Poisson mean using the Gehrels approximation

n_{obs}	Lower limits		
	Poisson parameter S or confidence level		
	$S = 1$ (1- σ , or 84.1%)	$S = 2$ (2- σ , or 97.7%)	$S = 3$ (3- σ , or 99.9%)
1	0.17	0.01	0.00
2	0.71	0.21	0.03
3	1.37	0.58	0.17
4	2.09	1.04	0.42
5	2.85	1.57	0.75
6	3.63	2.14	1.13
7	4.42	2.75	1.56
8	5.24	3.38	2.02
9	6.06	4.04	2.52
10	6.90	4.71	3.04
20	15.57	12.08	9.16
30	24.56	20.07	16.16
40	33.70	28.37	23.63
50	42.96	36.88	31.40
60	52.28	45.53	39.38
70	61.66	54.28	47.52
80	71.08	63.13	55.79
90	80.53	72.04	64.17
100	90.02	81.01	72.63

A.3 The Gamma and Beta Distributions and Functions

The *gamma distribution* is defined as

$$f_\gamma(x) = \frac{\alpha(\alpha x)^{r-1} e^{-\alpha x}}{\Gamma(r)}$$

where α, r are positive numbers, and $x \geq 0$, see (9.6). The parameter r is usually referred to as the shape parameter, and α as the rate parameter. The Gamma function, defined in (7.17), serves as a normalization constant for the distribution,

$$\Gamma(r) = \int_0^\infty x^{r-1} e^{-x} dx,$$

and it has the property that for integer values of its argument it is $\Gamma(n) = (n - 1)!$, or $\Gamma(n + 1) = n\Gamma(n)$. The mean and the variance of a random variable X with a gamma distribution with parameters r and α are

$$\begin{cases} \text{E}[X] = \frac{r}{\alpha} \\ \text{Var}(X) = \frac{r}{\alpha^2}. \end{cases}$$

The *beta distribution* is defined as

$$f_\beta(x) = \frac{1}{B(a, b)} x^{a-1} (1-x)^{b-1} \quad (\text{A.4})$$

and it is defined in the interval $0 \leq x \leq 1$, with a and b , respectively, referred to as the left and right parameter of the distribution. The distribution is also referred to as the *standard beta distribution*, and it was introduced in (14.13) as the distribution function of r^2 . The Beta function is defined as

$$B(a, b) = \int_0^1 u^{a-1} (1-u)^{b-1} du \quad (\text{A.5})$$

and it serves as the normalization for the beta distribution, see (9.23). It is possible to prove that the Beta function is related to the Gamma function via

$$B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}.$$

The mean and the variance of a random variable X with a beta distribution of parameters a and b are

$$\begin{cases} \text{E}[X] = \frac{a}{a+b} \\ \text{Var}(X) = \frac{ab}{(a+b)^2(a+b+1)}. \end{cases}$$

There are several useful relationships between the beta and gamma distributions, and the χ^2 and F statistics:

- (a) A χ^2 statistic with f degrees of freedom has a gamma distribution with parameters $\alpha = 1/2$ and $b = f/2$. This result was used in Sect. 9.3.1.
- (b) If a random variable X has a gamma distribution with shape parameter a and rate parameter r , and a random variable Y has a gamma distribution with the same rate parameter r and shape parameter b , with X – and Y -independent of one another, then

$$V = \frac{X}{X+Y}$$

has a beta distribution with parameters a and b .

(c) If X has an F distribution with n degrees of freedom at the numerator and m degrees of freedom at the denominator, than

$$Y = \frac{(n/m)X}{1 + (n/m)Y}$$

follows a beta distribution with left parameter $n/2$ and right parameter $m/2$.

This property can be easily proven starting with the fact that Y can be written as the ratio of two independent χ^2 -distributed variables $U \sim \chi^2(n)$ and $V \sim \chi^2(m)$,

$$X = \frac{U/n}{V/m},$$

and from this it follows that

$$Y = \frac{U/V}{1 + U/V} = \frac{U}{U + V}.$$

Since the χ^2 distribution is a special case of gamma distribution, property (b) applies to the Y variable, leading to the result that Y follows a beta distribution with the given left and right parameters.

The appearance of the Beta function at the denominator of both the beta and gamma distributions suggests the following alternative way to define the Beta function:

$$B(a, b) = \int_0^\infty \frac{t^{a-1}}{(1+t)^{a+b}} dt. \quad (\text{A.6})$$

The equivalence of this integral with (9.23) can be proven with the substitution $u = t/(1+t)$.

A related distribution is the *symmetric beta distribution*, or r -distribution, defined in the interval $-1 \leq y \leq 1$ as

$$f_r(y) = \frac{1}{B(1/2, f/2)} (1 - y^2)^{f/2-1}. \quad (\text{A.7})$$

The mean and variance of a variable R that follows the symmetric beta or r distribution are

$$\begin{cases} E[R] = 0 \\ \text{Var}(R) = \frac{1}{f+1}. \end{cases} \quad (\text{A.8})$$

A relationship between the standard and symmetric beta distributions is the following: if X is a standard beta distribution with parameters $a = b = f/2$, the variable $Y = 2X - 1$ has a symmetric beta distribution of (14.3) with parameter f (see Sect. 14.3).

This property can be proven with the method of change of variables, since the transformation is monotonic between $0 \leq x \leq 1$ and $-1 \leq y \leq 1$, with

$$g(y) = f_\beta(x) \frac{dx}{dy} = \frac{1}{B(f/2, f/2)} \left(\frac{y+1}{2} \right)^{f/2-1} \left(1 - \frac{y+1}{2} \right)^{f/2-1} \frac{1}{2}.$$

The *Legendre duplication formula* for the Gamma function

$$\Gamma(z)\Gamma(z + 1/2) = 2^{1-2z}\Gamma(1/2)\Gamma(2z)$$

for $z = f/2$ can now be used to show that the distribution of Y ,

$$g(y) = \frac{2^{f-1}}{B(1/2, f/2)} \frac{(1-y^2)^{f/2-1}}{2 \cdot 4^{f/2-1}} = f_r(y)$$

is indeed the r or symmetric beta distribution.

This r distribution applies to the linear correlation coefficient statistic r , as shown in Sect. 14.3, whereby the statistic evaluated for N independent datapoints follows a symmetric beta distribution with parameter $f = (N-2)/2$, where $N-2$ is the number of degrees of freedom, under the null hypothesis of no correlation between the two variables used for the statistic. In Sect. 14.4 it was also found that the distribution of r^2 under the same null hypothesis is a standard beta distribution with parameters $a = 1/2$ and $b = f/2$,

A.4 The χ^2 Distribution

The probability distribution function for a χ^2 variable is defined as

$$f_{\chi^2}(x) = \left(\frac{1}{2} \right)^{f/2} \frac{1}{\Gamma(f/2)} e^{-x/2} x^{f/2-1}, \quad (\text{A.9})$$

where f is the number of degrees of freedom. A χ^2 distribution with f degrees of freedom is equivalent to a gamma distribution with shape parameter $r = f/2$ and rate parameter $\alpha = 1/2$. When a χ^2 statistic is obtained as the sum of the square of N z -scores from a fully specified hypothesis (see Sect. 9.3), then the number of degrees of freedom is equal to the number of measurements, $f = N$, as in (9.7). More generally, when a χ^2_{\min} statistic is obtained from a model with m free parameters, the Cramér theorem (Sect. 12.1) usually implies that the asymptotic distribution of χ^2_{\min} is that of a χ^2 distribution with $f = N - m$ degrees of freedom. Therefore (A.9) can be used for a fully specified hypothesis and for a model with adjustable parameters, provided the number of degrees of freedom f is chosen accordingly.

The critical value or p -quantile of the distribution is given by (9.15),

$$P(\chi^2 \geq \chi_{crit}^2) = \int_{\chi_{crit}^2}^{\infty} f_{\chi^2}(x) dx = 1 - p.$$

The critical value is a function of the number of degrees of freedom f and the level of probability p . The probability p is intended as a large number, such as 0.68, 0.90, or 0.99, with the meaning that there is a small probability that the statistic has values higher than the critical value χ_{crit}^2 .

A random variable X that follows a χ^2 distribution with f degrees of freedom has the following mean and variance:

$$\begin{cases} E[X] = f \\ \text{Var}(X) = 2f \end{cases}$$

(see Sect. 9.3). It is convenient to tabulate the critical values of the reduced χ^2 ,

$$\chi_{red}^2 = \frac{\chi_{crit}^2}{f},$$

that corresponds to a given probability level, as function of the number of degrees of freedom. Selected critical values of the χ^2 distribution are reported in Table A.7. When using this table, it is necessary to multiply the tabulated reduced χ^2 by the number of degrees of freedom f to obtain the critical value of χ^2 .

If X is a χ^2 -distributed variable with f degrees of freedoms, the asymptotic distribution tends to a normal distribution according to

$$\lim_{f \rightarrow \infty} \frac{X - f}{\sqrt{2f}} = N(0, 1). \quad (\text{A.10})$$

In fact, a χ^2 variable is obtained as the sum of independent distributions (Sect. 9.3), to which the central theorem limit applies (Sect. 4.4). Figure A.2 illustrates the difference between the χ^2 distribution and the normal distribution, where the χ^2 distributions were rescaled according to (A.10). For example, the $x = 0$ value of a $\chi^2(5)$ distribution corresponds to a standardized value of -1.58 , which is the lowest value in the figure. For a large number of degrees of freedom, the standard Gaussian distribution can be used to supplement Table A.7 according to (A.10). For example, for $p = 0.99$, the one-sided critical value of the standard Gaussian is approximately 2.326, according to Table A.4. Using this value into (A.10) for $f = 200$ would give a critical value for the χ^2 distribution of 1.2326 (compare to 1.247 from Table A.7). The values of $f = \infty$ in Table A.7 are obtained using the Gaussian approximation, according to (A.10).

Table A.7 Critical values of the χ^2 distribution with f degrees of freedom

f	Probability p to have a value of reduced χ^2 below the critical value											
	0.01	0.05	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	0.95
1	0.00016	0.00390	0.01580	0.0642	0.1485	0.2750	0.4549	0.7083	1.0742	1.6424	2.7055	3.8415
2	0.0101	0.0513	0.1054	0.2231	0.3567	0.5108	0.6931	0.9163	1.2040	1.6094	2.3026	2.9957
3	0.0383	0.11173	0.1948	0.3351	0.4746	0.6231	0.7887	0.9821	1.2216	1.5472	2.0838	2.6049
4	0.0743	0.1777	0.2659	0.4122	0.5487	0.6882	0.8392	1.0112	1.2196	1.4972	1.9449	2.3719
5	0.1109	0.2291	0.3221	0.4685	0.6000	0.7311	0.8703	1.0264	1.2129	1.4578	1.8473	2.2141
6	0.1454	0.2726	0.3674	0.5117	0.6379	0.7617	0.8914	1.0351	1.2052	1.4263	1.7741	2.0986
7	0.1770	0.3096	0.4047	0.5460	0.6673	0.7847	0.9065	1.0405	1.1976	1.4005	1.7167	2.0096
8	0.2058	0.3416	0.4362	0.5742	0.6909	0.8028	0.9180	1.0438	1.1906	1.3788	1.6702	1.9384
9	0.2320	0.3695	0.4631	0.5978	0.7104	0.8174	0.9270	1.0460	1.1841	1.3602	1.6315	1.8799
10	0.256	0.394	0.487	0.618	0.727	0.830	0.934	1.047	1.178	1.344	1.599	1.831
11	0.278	0.416	0.507	0.635	0.741	0.840	0.940	1.048	1.173	1.330	1.570	1.789
12	0.298	0.436	0.525	0.651	0.753	0.848	0.945	1.049	1.168	1.318	1.546	1.752
13	0.316	0.453	0.542	0.664	0.764	0.856	0.949	1.049	1.163	1.307	1.524	1.720
14	0.333	0.469	0.556	0.676	0.773	0.863	0.953	1.049	1.159	1.296	1.505	1.692
15	0.349	0.484	0.570	0.687	0.781	0.869	0.956	1.049	1.155	1.287	1.487	1.666
16	0.363	0.498	0.582	0.697	0.789	0.874	0.959	1.049	1.151	1.279	1.471	1.643
17	0.377	0.510	0.593	0.706	0.796	0.879	0.961	1.048	1.148	1.271	1.457	1.623
18	0.390	0.522	0.604	0.714	0.802	0.883	0.963	1.048	1.145	1.264	1.444	1.604
19	0.402	0.532	0.613	0.722	0.808	0.887	0.965	1.048	1.142	1.258	1.432	1.587
20	0.413	0.543	0.622	0.729	0.813	0.890	0.967	1.048	1.139	1.252	1.421	1.571
30	0.498	0.616	0.687	0.779	0.850	0.915	0.978	1.044	1.118	1.208	1.342	1.459

(continued)

Table A.7 (continued)

f	Probability p to have a value of reduced χ^2 below the critical value							
	0.01	0.05	0.10	0.20	0.30	0.40	0.50	0.60
40	0.554	0.663	0.726	0.809	0.872	0.928	0.983	1.041
50	0.594	0.695	0.754	0.829	0.886	0.937	0.987	1.038
60	0.625	0.720	0.774	0.844	0.897	0.944	0.989	1.036
70	0.649	0.739	0.790	0.856	0.905	0.949	0.990	1.034
80	0.669	0.755	0.803	0.865	0.911	0.952	0.992	1.032
90	0.686	0.768	0.814	0.873	0.917	0.955	0.993	1.031
100	0.700	0.780	0.823	0.880	0.921	0.958	0.993	1.030
200	0.782	0.841	0.874	0.915	0.945	0.971	0.997	1.022
300	0.820	0.870	0.897	0.931	0.956	0.977	0.998	1.019
400	0.843	0.887	0.910	0.940	0.962	0.980	0.998	1.016
500	0.86	0.90	0.92	0.95	0.97	0.98	1.00	1.01
1000	0.90	0.93	0.94	0.96	0.98	0.99	1.00	1.01
∞	0.90	0.93	0.94	0.96	0.98	0.99	1.00	1.01

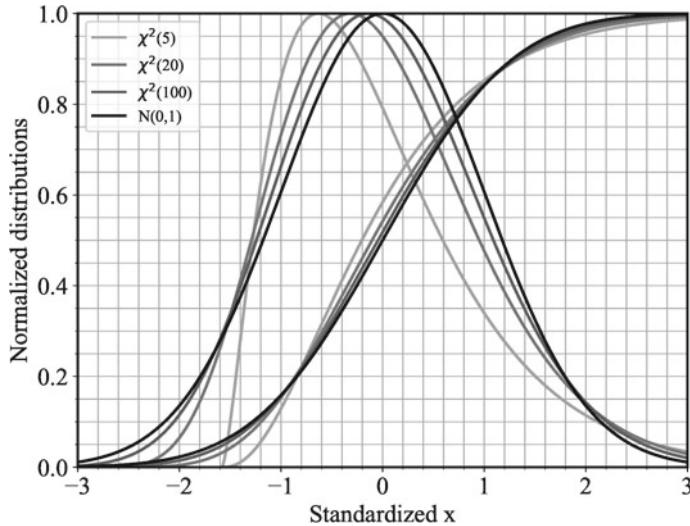


Fig. A.2 Normalized probability distribution functions of χ^2 distributions, standardized according to (A.10), and the associated cumulative distribution functions. Although the χ^2 distribution is defined for positive values only, the x axis extends to negative values according to (A.10)

A.5 The F Distribution

The F distribution with f_1, f_2 degrees of freedom is defined in (9.22) as

$$f_F(x) = \frac{f_1/f_2}{B(f_1/2, f_2/2)} \frac{\left(x \frac{f_1}{f_2}\right)^{f_1/2-1}}{\left(1 + x \frac{f_1}{f_2}\right)^{f_1/2+f_2/2}},$$

where $B(f_1/2, f_2/2)$ is the Beta function. The shape of the F distribution varies significantly as a function of the two parameters f_1 and f_2 . The mean and the variance of the F statistic are given by (9.24), which is repeated here for convenience:

$$\begin{cases} E[F] = \frac{f_2}{f_2 - 2} & \text{(for } f_2 > 2\text{)} \\ \text{Var}(F) = \frac{2 f_2^2 (f_1 + f_2 - 2)}{f_1 (f_2 - 2)^2 (f_2 - 4)} & \text{(for } f_2 > 4\text{)} \end{cases}$$

with a mean that approaches unity as f_2 increases. Figure A.3 illustrates the change in shape of the distribution when $f_1 = f_2$ increases.

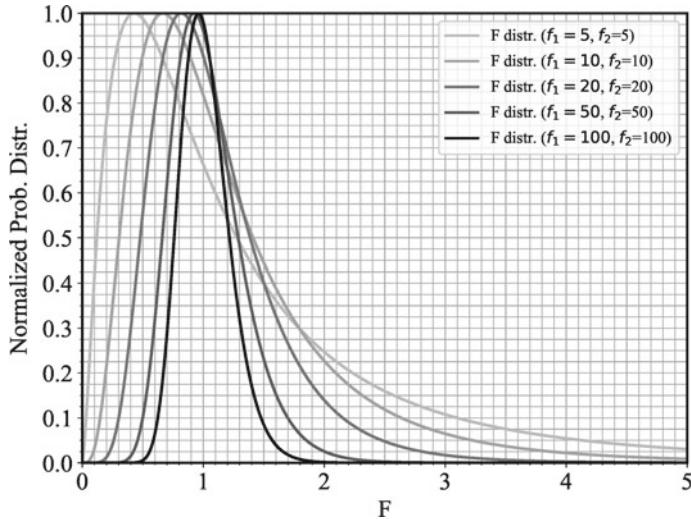


Fig. A.3 Normalized probability distribution functions of the F distribution for selected values of the number $f_1 = f_2$ of degrees of freedom

The critical value F_{crit} that includes a probability p is given by (9.26),

$$P(F > F_{crit}) = \int_{F_{crit}}^{\infty} f_F(x) dx = 1 - p.$$

and it is a function of the number of degrees of freedom f_1 and f_2 . Table A.8 reports the critical values for various probability levels p , for a fixed value $f_1 = 1$, and as a function of f_2 . Tables A.9, A.10, A.11, A.12, A.13, A.14, and A.15 report the critical values of the F statistic as a function of both f_1 and f_2 .

Asymptotic values when f_1 and f_2 approach infinity can be found using the limits (9.25),

$$\begin{cases} \lim_{f_2 \rightarrow \infty} f_F(z, f_1, f_2) = f_{\chi^2}(x, f_1) \text{ where } x = f_1 z \\ \lim_{f_1 \rightarrow \infty} f_F(z, f_1, f_2) = f_{\chi^2}(x, f_2) \text{ where } x = f_2/z. \end{cases}$$

For example, the critical values of the F distribution for $f_1 = 1$ and in the limit of large f_2 are obtained from the first row of Table A.7.

Table A.8 Critical values of F statistics for $f_1 = 1$ degrees of freedom

f_2	Probability p to have a value of F below the critical value						
	0.50	0.60	0.70	0.80	0.90	0.95	0.99
1	1.000	1.894	3.852	9.472	39.863	161.448	4052.182
2	0.667	1.125	1.922	3.556	8.526	18.513	98.503
3	0.585	0.957	1.562	2.682	5.538	10.128	34.116
4	0.549	0.885	1.415	2.351	4.545	7.709	21.198
5	0.528	0.846	1.336	2.178	4.060	6.608	16.258
6	0.515	0.820	1.286	2.073	3.776	5.987	13.745
7	0.506	0.803	1.253	2.002	3.589	5.591	12.246
8	0.499	0.790	1.228	1.951	3.458	5.318	11.259
9	0.494	0.780	1.209	1.913	3.360	5.117	10.561
10	0.490	0.773	1.195	1.883	3.285	4.965	10.044
20	0.472	0.740	1.132	1.757	2.975	4.351	8.096
30	0.466	0.729	1.112	1.717	2.881	4.171	7.562
40	0.463	0.724	1.103	1.698	2.835	4.085	7.314
50	0.462	0.721	1.097	1.687	2.809	4.034	7.171
60	0.461	0.719	1.093	1.679	2.791	4.001	7.077
70	0.460	0.717	1.090	1.674	2.779	3.978	7.011
80	0.459	0.716	1.088	1.670	2.769	3.960	6.963
90	0.459	0.715	1.087	1.667	2.762	3.947	6.925
100	0.458	0.714	1.085	1.664	2.756	3.936	6.895
200	0.457	0.711	1.080	1.653	2.731	3.888	6.763
∞	0.455	0.708	1.074	1.642	2.706	3.842	6.635

Table A.9 Critical values of F statistic that include $p = 0.50$ probability

f_2	f_1									
	2	4	6	8	10	20	40	60	80	100
1	1.500	1.823	1.942	2.004	2.042	2.119	2.158	2.172	2.178	2.182
2	1.000	1.207	1.282	1.321	1.345	1.393	1.418	1.426	1.430	1.433
3	0.881	1.063	1.129	1.163	1.183	1.225	1.246	1.254	1.257	1.259
4	0.828	1.000	1.062	1.093	1.113	1.152	1.172	1.178	1.182	1.184
5	0.799	0.965	1.024	1.055	1.073	1.111	1.130	1.136	1.139	1.141
6	0.780	0.942	1.000	1.030	1.048	1.084	1.103	1.109	1.113	1.114
7	0.767	0.926	0.983	1.013	1.030	1.066	1.085	1.091	1.094	1.096
8	0.757	0.915	0.971	1.000	1.017	1.053	1.071	1.077	1.080	1.082
9	0.749	0.906	0.962	0.990	1.008	1.043	1.061	1.067	1.070	1.072
10	0.743	0.899	0.954	0.983	1.000	1.035	1.053	1.059	1.062	1.063
20	0.718	0.868	0.922	0.950	0.966	1.000	1.017	1.023	1.026	1.027
30	0.709	0.858	0.912	0.939	0.955	0.989	1.006	1.011	1.014	1.016
40	0.705	0.854	0.907	0.934	0.950	0.983	1.000	1.006	1.008	1.010
50	0.703	0.851	0.903	0.930	0.947	0.980	0.997	1.002	1.005	1.007
60	0.701	0.849	0.901	0.928	0.945	0.978	0.994	1.000	1.003	1.004
70	0.700	0.847	0.900	0.927	0.943	0.976	0.993	0.998	1.001	1.003
80	0.699	0.846	0.899	0.926	0.942	0.975	0.992	0.997	1.000	1.002
90	0.699	0.845	0.898	0.925	0.941	0.974	0.991	0.996	0.999	1.001
100	0.698	0.845	0.897	0.924	0.940	0.973	0.990	0.996	0.998	1.000
200	0.696	0.842	0.894	0.921	0.937	0.970	0.987	0.992	0.995	0.997
∞	0.693	0.839	0.891	0.918	0.934	0.967	0.983	0.989	0.992	0.993

Table A.10 Critical values of F statistic that include $p = 0.60$ probability

f_2	f_1									
	2	4	6	8	10	20	40	60	80	100
1	2.625	3.093	3.266	3.355	3.410	3.522	3.579	3.598	3.608	3.613
2	1.500	1.718	1.796	1.835	1.859	1.908	1.933	1.941	1.945	1.948
3	1.263	1.432	1.489	1.518	1.535	1.570	1.588	1.593	1.596	1.598
4	1.162	1.310	1.359	1.383	1.397	1.425	1.439	1.444	1.446	1.448
5	1.107	1.243	1.287	1.308	1.320	1.345	1.356	1.360	1.362	1.363
6	1.072	1.200	1.241	1.260	1.272	1.293	1.303	1.307	1.308	1.309
7	1.047	1.171	1.209	1.227	1.238	1.257	1.266	1.269	1.270	1.271
8	1.030	1.150	1.186	1.203	1.213	1.231	1.239	1.241	1.242	1.243
9	1.016	1.133	1.168	1.185	1.194	1.210	1.217	1.219	1.220	1.221
10	1.006	1.120	1.154	1.170	1.179	1.194	1.200	1.202	1.203	1.204
20	0.960	1.064	1.093	1.106	1.112	1.122	1.124	1.124	1.124	1.124
30	0.945	1.046	1.074	1.085	1.090	1.097	1.097	1.097	1.096	1.096
40	0.938	1.037	1.064	1.075	1.080	1.085	1.084	1.083	1.082	1.081
50	0.933	1.032	1.058	1.068	1.073	1.078	1.076	1.074	1.073	1.072
60	0.930	1.029	1.054	1.064	1.069	1.073	1.070	1.068	1.066	1.065
70	0.928	1.026	1.052	1.061	1.066	1.069	1.066	1.064	1.062	1.061
80	0.927	1.024	1.049	1.059	1.064	1.067	1.063	1.060	1.059	1.057
90	0.926	1.023	1.048	1.057	1.062	1.065	1.061	1.058	1.056	1.054
100	0.925	1.021	1.047	1.056	1.060	1.063	1.059	1.056	1.054	1.052
200	0.921	1.016	1.041	1.050	1.054	1.055	1.050	1.046	1.043	1.041
∞	0.916	1.011	1.035	1.044	1.047	1.048	1.041	1.036	1.032	1.029

Table A.11 Critical values of F statistic that include $p = 0.70$ probability

f_2	f_1									
	2	4	6	8	10	20	40	60	80	100
1	5.056	5.830	6.117	6.267	6.358	6.544	6.639	6.671	6.687	6.697
2	2.333	2.561	2.640	2.681	2.705	2.754	2.779	2.787	2.791	2.794
3	1.847	1.985	2.028	2.048	2.061	2.084	2.096	2.100	2.102	2.103
4	1.651	1.753	1.781	1.793	1.800	1.812	1.818	1.819	1.820	1.821
5	1.547	1.629	1.648	1.656	1.659	1.665	1.666	1.666	1.667	1.667
6	1.481	1.551	1.565	1.570	1.571	1.572	1.570	1.570	1.569	1.569
7	1.437	1.499	1.509	1.511	1.511	1.507	1.504	1.502	1.501	1.501
8	1.405	1.460	1.467	1.468	1.466	1.460	1.455	1.452	1.451	1.450
9	1.380	1.431	1.436	1.435	1.433	1.424	1.417	1.414	1.413	1.412
10	1.361	1.408	1.412	1.409	1.406	1.395	1.387	1.384	1.382	1.381
20	1.279	1.311	1.305	1.297	1.290	1.268	1.252	1.245	1.242	1.240
30	1.254	1.280	1.271	1.261	1.253	1.226	1.206	1.197	1.192	1.189
40	1.241	1.264	1.255	1.243	1.234	1.205	1.182	1.172	1.167	1.163
50	1.233	1.255	1.245	1.233	1.223	1.192	1.167	1.156	1.150	1.146
60	1.228	1.249	1.238	1.226	1.215	1.183	1.157	1.146	1.139	1.135
70	1.225	1.245	1.233	1.221	1.210	1.177	1.150	1.138	1.131	1.127
80	1.222	1.242	1.230	1.217	1.206	1.172	1.144	1.132	1.125	1.120
90	1.220	1.239	1.227	1.214	1.203	1.168	1.140	1.127	1.120	1.115
100	1.219	1.237	1.225	1.212	1.200	1.165	1.137	1.123	1.116	1.111
200	1.211	1.228	1.215	1.201	1.189	1.152	1.121	1.106	1.097	1.091
∞	1.204	1.220	1.205	1.191	1.178	1.139	1.104	1.087	1.076	1.069

Table A.12 Critical values of F statistic that include $p = 0.80$ probability

f_2	f_1									
	2	4	6	8	10	20	40	60	80	100
1	12.000	13.644	14.258	14.577	14.772	15.171	15.374	15.442	15.477	15.497
2	4.000	4.236	4.317	4.358	4.382	4.432	4.456	4.465	4.469	4.471
3	2.886	2.956	2.971	2.976	2.979	2.983	2.984	2.984	2.984	2.984
4	2.472	2.483	2.473	2.465	2.460	2.445	2.436	2.433	2.431	2.430
5	2.259	2.240	2.217	2.202	2.191	2.166	2.151	2.146	2.143	2.141
6	2.130	2.092	2.062	2.042	2.028	1.995	1.976	1.969	1.965	1.963
7	2.043	1.994	1.957	1.934	1.918	1.879	1.857	1.849	1.844	1.842
8	1.981	1.923	1.883	1.856	1.838	1.796	1.770	1.761	1.756	1.753
9	1.935	1.870	1.826	1.798	1.778	1.732	1.704	1.694	1.689	1.686
10	1.899	1.829	1.782	1.752	1.732	1.682	1.653	1.642	1.636	1.633
20	1.746	1.654	1.596	1.558	1.531	1.466	1.424	1.408	1.399	1.394
30	1.699	1.600	1.538	1.497	1.468	1.395	1.347	1.328	1.318	1.312
40	1.676	1.574	1.509	1.467	1.437	1.360	1.308	1.287	1.276	1.269
50	1.662	1.558	1.492	1.449	1.418	1.338	1.284	1.262	1.249	1.241
60	1.653	1.548	1.481	1.437	1.406	1.324	1.268	1.244	1.231	1.223
70	1.647	1.540	1.473	1.429	1.397	1.314	1.256	1.231	1.218	1.209
80	1.642	1.535	1.467	1.422	1.390	1.306	1.247	1.222	1.208	1.199
90	1.639	1.531	1.463	1.418	1.385	1.300	1.240	1.214	1.200	1.191
100	1.636	1.527	1.459	1.414	1.381	1.296	1.234	1.208	1.193	1.184
200	1.622	1.512	1.443	1.396	1.363	1.274	1.209	1.180	1.163	1.152
∞	1.609	1.497	1.426	1.379	1.344	1.252	1.182	1.150	1.130	1.117

Table A.13 Critical values of F statistic that include $p = 0.90$ probability

f_2	f_1									
	2	4	6	8	10	20	40	60	80	100
1	49.500	55.833	58.204	59.439	60.195	61.740	62.529	62.794	62.927	63.007
2	9.000	9.243	9.326	9.367	9.392	9.441	9.466	9.475	9.479	9.481
3	5.462	5.343	5.285	5.252	5.230	5.184	5.160	5.151	5.147	5.144
4	4.325	4.107	4.010	3.955	3.920	3.844	3.804	3.790	3.782	3.778
5	3.780	3.520	3.404	3.339	3.297	3.207	3.157	3.140	3.132	3.126
6	3.463	3.181	3.055	2.983	2.937	2.836	2.781	2.762	2.752	2.746
7	3.257	2.960	2.827	2.752	2.703	2.595	2.535	2.514	2.504	2.497
8	3.113	2.806	2.668	2.589	2.538	2.425	2.361	2.339	2.328	2.321
9	3.006	2.693	2.551	2.469	2.416	2.298	2.232	2.208	2.196	2.189
10	2.924	2.605	2.461	2.377	2.323	2.201	2.132	2.107	2.095	2.087
20	2.589	2.249	2.091	1.999	1.937	1.794	1.708	1.677	1.660	1.650
30	2.489	2.142	1.980	1.884	1.820	1.667	1.573	1.538	1.519	1.507
40	2.440	2.091	1.927	1.829	1.763	1.605	1.506	1.467	1.447	1.434
50	2.412	2.061	1.895	1.796	1.729	1.568	1.465	1.424	1.402	1.389
60	2.393	2.041	1.875	1.775	1.707	1.544	1.437	1.395	1.372	1.358
70	2.380	2.027	1.860	1.760	1.691	1.526	1.418	1.374	1.350	1.335
80	2.370	2.016	1.849	1.748	1.680	1.513	1.403	1.358	1.334	1.318
90	2.362	2.008	1.841	1.739	1.671	1.503	1.391	1.346	1.320	1.304
100	2.356	2.002	1.834	1.732	1.663	1.494	1.382	1.336	1.310	1.293
200	2.329	1.973	1.804	1.701	1.631	1.458	1.339	1.289	1.261	1.242
∞	2.303	1.945	1.774	1.670	1.599	1.421	1.295	1.240	1.207	1.185

Table A.14 Critical values of F statistic that include $p = 0.95$ probability

f_2	f_1	2	4	6	8	10	20	40	60	80	100
1	199.500	224.583	233.986	238.883	241.882	248.013	251.143	252.196	252.724	253.041	
2	19.000	19.247	19.330	19.371	19.396	19.446	19.471	19.479	19.483	19.486	
3	9.552	9.117	8.941	8.845	8.786	8.660	8.594	8.572	8.561	8.554	
4	6.944	6.388	6.163	6.041	5.964	5.803	5.717	5.688	5.673	5.664	
5	5.786	5.192	4.950	4.818	4.735	4.558	4.464	4.431	4.415	4.405	
6	5.143	4.534	4.284	4.147	4.060	3.874	3.774	3.740	3.722	3.712	
7	4.737	4.120	3.866	3.726	3.636	3.444	3.340	3.304	3.286	3.275	
8	4.459	3.838	3.581	3.438	3.347	3.150	3.043	3.005	2.986	2.975	
9	4.256	3.633	3.374	3.230	3.137	2.936	2.826	2.787	2.768	2.756	
10	4.103	3.478	3.217	3.072	2.978	2.774	2.661	2.621	2.601	2.588	
20	3.493	2.866	2.599	2.447	2.348	2.124	1.994	1.946	1.922	1.907	
30	3.316	2.690	2.421	2.266	2.165	1.932	1.792	1.740	1.712	1.695	
40	3.232	2.606	2.336	2.180	2.077	1.839	1.693	1.637	1.608	1.589	
50	3.183	2.557	2.286	2.130	2.026	1.784	1.634	1.576	1.544	1.525	
60	3.150	2.525	2.254	2.097	1.992	1.748	1.594	1.534	1.502	1.481	
70	3.128	2.503	2.231	2.074	1.969	1.722	1.566	1.504	1.471	1.450	
80	3.111	2.486	2.214	2.056	1.951	1.703	1.545	1.482	1.448	1.426	
90	3.098	2.473	2.201	2.043	1.938	1.688	1.528	1.464	1.429	1.407	
100	3.087	2.463	2.191	2.032	1.927	1.677	1.515	1.450	1.414	1.392	
200	3.041	2.417	2.144	1.985	1.878	1.623	1.455	1.385	1.346	1.321	
∞	2.996	2.372	2.099	1.938	1.831	1.571	1.394	1.318	1.273	1.243	

Table A.15 Critical values of F statistic that include $p = 0.99$ probability

f_2	f_1	2	4	6	8	10	20	40	60	80	100
1	4999.500	5624.583	5858.986	5981.070	6055.847	6208.730	6286.782	6313.030	6326.197	6334.110	
2	99.000	99.249	99.333	99.374	99.399	99.449	99.474	99.482	99.487	99.489	
3	30.817	28.710	27.911	27.489	27.229	26.690	26.411	26.316	26.269	26.240	
4	18.000	15.977	15.207	14.799	14.546	14.020	13.745	13.652	13.605	13.577	
5	13.274	11.392	10.672	10.289	10.051	9.553	9.291	9.202	9.157	9.130	
6	10.925	9.148	8.466	8.102	7.874	7.396	7.143	7.057	7.013	6.987	
7	9.547	7.847	7.191	6.840	6.620	6.155	5.908	5.824	5.781	5.755	
8	8.649	7.006	6.371	6.029	5.814	5.359	5.116	5.032	4.989	4.963	
9	8.022	6.422	5.802	5.467	5.257	4.808	4.567	4.483	4.441	4.415	
10	7.559	5.994	5.386	5.057	4.849	4.405	4.165	4.082	4.039	4.014	
20	5.849	4.431	3.871	3.564	3.368	2.938	2.695	2.608	2.563	2.535	
30	5.390	4.018	3.473	3.173	2.979	2.549	2.299	2.208	2.160	2.131	
40	5.178	3.828	3.291	2.993	2.801	2.369	2.114	2.019	1.969	1.938	
50	5.057	3.720	3.186	2.890	2.698	2.265	2.006	1.909	1.857	1.825	
60	4.977	3.649	3.119	2.823	2.632	2.198	1.936	1.836	1.783	1.749	
70	4.922	3.600	3.071	2.777	2.585	2.150	1.886	1.784	1.730	1.696	
80	4.881	3.563	3.036	2.742	2.551	2.115	1.849	1.746	1.690	1.655	
90	4.849	3.535	3.009	2.715	2.524	2.088	1.820	1.716	1.659	1.623	
100	4.824	3.513	2.988	2.694	2.503	2.067	1.797	1.692	1.634	1.598	
200	4.713	3.414	2.893	2.601	2.411	1.971	1.695	1.584	1.521	1.481	
∞	4.605	3.319	2.802	2.511	2.321	1.878	1.592	1.473	1.404	1.358	

A.6 The Student t Distribution

The Student t distribution is given by (9.33),

$$f_T(t) = \frac{1}{\sqrt{f\pi}} \frac{\Gamma(f/2 + 1/2)}{\Gamma(f/2)} \left(1 + \frac{t^2}{f}\right)^{-\frac{f+1}{2}},$$

where f is the number of degrees of freedom. The mean and variance of the distribution are given by (9.36), which is repeated here for convenience:

$$\begin{cases} E[T] = 0 & (\text{for } f > 1) \\ \text{Var}(T) = \frac{f}{f-2} & (\text{for } f > 2), \end{cases}$$

where the case of $f = 1$ corresponds to the Cauchy distribution (4.13), for which the expectation is undefined. As the number of degrees of freedom increases, the distribution tends to a standard normal. An illustration of the change in the shape of the t distribution as a function of the number of degrees of freedom is provided in Fig. A.4.

The probability p that the absolute value of a t -distributed variable T exceeds a critical value T_{crit} is given by (9.37),

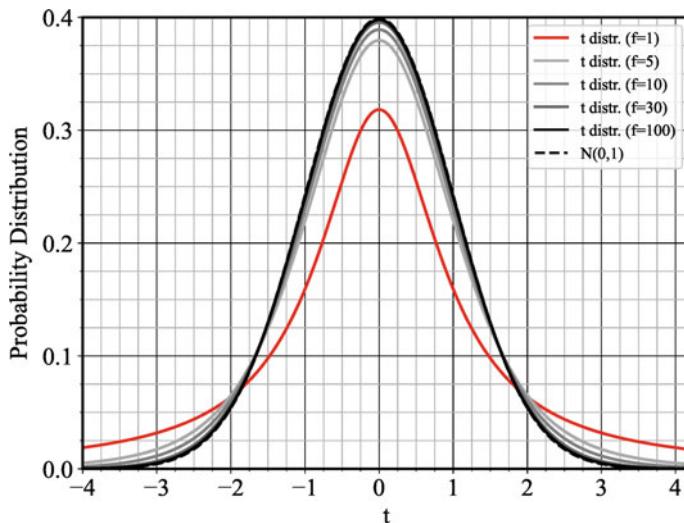


Fig. A.4 Probability distribution functions of the t distribution for selected values of the number of degrees of freedom

$$P(|T| \geq T_{crit}) = 1 - \int_{-T_{crit}}^{T_{crit}} f_T(t) dt = 1 - p.$$

These two-sided critical values are tabulated in Tables A.16, A.17, A.18, A.19, A.20, A.21 and A.22 for selected values of f , as a function of the critical value T_{crit} . In these tables, the left column indicates the value of T_{crit} to the first decimal digit, and the values on the top column are the second decimal digit. Table A.23 provides a comparison of the probability p for five critical values, $T_{crit} = 1$ through 5, as a function of the number of degrees of freedom. The case of $f = \infty$ corresponds to a standard Gaussian.

Table A.16 Integral of Student's function ($f = 1$), or probability p , as function of critical value T_{crit}

	0	1	2	3	4	5	6	7	8	9
0.0	0.000000	0.006366	0.012731	0.019093	0.025451	0.031805	0.038151	0.044491	0.050821	0.057142
0.1	0.063451	0.069748	0.076031	0.082299	0.088551	0.094786	0.101003	0.107201	0.113378	0.119533
0.2	0.125666	0.131775	0.137860	0.143920	0.149953	0.155958	0.161936	0.167884	0.173803	0.179691
0.3	0.185547	0.191372	0.197163	0.202921	0.208645	0.214334	0.219988	0.225605	0.231187	0.236731
0.4	0.242238	0.247707	0.253138	0.258530	0.263883	0.269197	0.274472	0.279706	0.284900	0.290054
0.5	0.295167	0.300240	0.305272	0.310262	0.315212	0.320120	0.324987	0.329813	0.334597	0.339340
0.6	0.344042	0.348702	0.353321	0.357899	0.362436	0.366932	0.371387	0.375801	0.380175	0.384508
0.7	0.388800	0.395053	0.397266	0.401438	0.405572	0.409666	0.413721	0.417737	0.421714	0.425653
0.8	0.429554	0.433417	0.437242	0.441030	0.444781	0.448495	0.452173	0.455814	0.459420	0.462990
0.9	0.466525	0.470025	0.473490	0.476920	0.480317	0.483680	0.487010	0.490306	0.493570	0.496801
1.0	0.500000	0.503167	0.506303	0.509408	0.512481	0.515524	0.518537	0.521520	0.524474	0.527398
1.1	0.530293	0.533159	0.535997	0.538807	0.541589	0.544344	0.547071	0.549772	0.552446	0.555094
1.2	0.557716	0.560312	0.562883	0.565429	0.567950	0.570447	0.572919	0.575367	0.577792	0.580193
1.3	0.582571	0.584927	0.587259	0.589570	0.591858	0.594124	0.596369	0.598592	0.600795	0.602976
1.4	0.605137	0.607278	0.609398	0.611499	0.613580	0.615641	0.617684	0.619707	0.621712	0.623698
1.5	0.625666	0.627616	0.629548	0.631462	0.633359	0.635239	0.637101	0.638947	0.640776	0.642589
1.6	0.644385	0.646165	0.647930	0.649678	0.651411	0.653129	0.654832	0.656519	0.658192	0.659850
1.7	0.661494	0.663124	0.664739	0.666340	0.667928	0.669502	0.671062	0.672669	0.674143	0.675663
1.8	0.677171	0.678666	0.680149	0.681619	0.683077	0.684522	0.685956	0.687377	0.688787	0.690185
1.9	0.691572	0.692947	0.694311	0.695664	0.697006	0.698337	0.699657	0.700967	0.702266	0.703555
2.0	0.704833	0.706101	0.707359	0.708607	0.709846	0.711074	0.712293	0.713502	0.714702	0.715893
2.1	0.717074	0.718246	0.719410	0.720564	0.721709	0.722846	0.723974	0.725093	0.726204	0.727307
2.2	0.728401	0.729487	0.730565	0.731635	0.732696	0.733750	0.734797	0.735835	0.736866	0.737889
2.3	0.738905	0.739914	0.740915	0.741909	0.742895	0.743875	0.744847	0.745813	0.746772	0.747723
2.4	0.748668	0.749607	0.750539	0.751464	0.752383	0.753295	0.755101	0.755994	0.756881	

(continued)

Table A.16 (continued)

	0	1	2	3	4	5	6	7	8	9
2.5	0.757762	0.758638	0.759507	0.760370	0.761227	0.762078	0.762924	0.763764	0.764598	0.765427
2.6	0.766250	0.767068	0.767880	0.768687	0.769488	0.770284	0.771075	0.771861	0.772642	0.773417
2.7	0.774188	0.774953	0.775714	0.776469	0.777220	0.777966	0.778707	0.779443	0.780175	0.780902
2.8	0.781625	0.782342	0.783056	0.783765	0.784469	0.785169	0.785865	0.786556	0.787243	0.787926
2.9	0.788605	0.789279	0.789949	0.790616	0.791278	0.791936	0.792590	0.793240	0.793887	0.794529
3.0	0.795168	0.795802	0.796433	0.797060	0.797684	0.798304	0.798920	0.799532	0.800141	0.800746
3.1	0.801348	0.801946	0.802541	0.803133	0.803720	0.804305	0.804886	0.805464	0.806039	0.806610
3.2	0.807178	0.807743	0.808304	0.808863	0.809418	0.809970	0.810519	0.811065	0.811608	0.812148
3.3	0.812685	0.813219	0.813750	0.814278	0.814803	0.815325	0.815845	0.816361	0.816875	0.817386
3.4	0.817894	0.818400	0.818903	0.819403	0.819900	0.820395	0.820887	0.821376	0.821863	0.822348
3.5	0.822829	0.823308	0.823785	0.824259	0.824731	0.825200	0.825667	0.826131	0.826593	0.827053
3.6	0.827510	0.827965	0.828418	0.828868	0.829316	0.829761	0.830205	0.830646	0.831085	0.831521
3.7	0.831956	0.832388	0.832818	0.833246	0.833672	0.834096	0.834517	0.834937	0.835354	0.835770
3.8	0.836183	0.836594	0.837004	0.837411	0.837816	0.838220	0.838621	0.839020	0.839418	0.839813
3.9	0.840207	0.840599	0.840989	0.841377	0.841763	0.842147	0.842530	0.842911	0.843290	0.843667
4.0	0.844042	0.844416	0.844788	0.845158	0.845526	0.845893	0.846258	0.846621	0.846983	0.847343
4.1	0.847701	0.848057	0.848412	0.848766	0.849118	0.849468	0.849816	0.850163	0.850509	0.850853
4.2	0.851195	0.851536	0.851875	0.852213	0.852549	0.852883	0.853217	0.853548	0.853879	0.854208
4.3	0.854535	0.854861	0.855185	0.855508	0.855830	0.856150	0.856469	0.856787	0.857103	0.857417
4.4	0.857731	0.858043	0.858353	0.858663	0.858971	0.859277	0.859583	0.859887	0.860190	0.860491
4.5	0.860791	0.861090	0.861388	0.861684	0.861980	0.862274	0.862566	0.862858	0.863148	0.863437
4.6	0.863725	0.864012	0.864297	0.864582	0.864865	0.865147	0.865428	0.865707	0.865986	0.866263
4.7	0.866539	0.866815	0.867089	0.867362	0.867633	0.867904	0.868174	0.868442	0.868710	0.868976
4.8	0.869242	0.869506	0.869769	0.870031	0.870292	0.870553	0.870812	0.871070	0.871327	0.871583
4.9	0.871838	0.872092	0.872345	0.872597	0.872848	0.873098	0.873347	0.873596	0.873843	0.874089

Table A.17 Integral of Student's function ($f = 2$), or probability p , as function of critical value T_{crit}

	0	1	2	3	4	5	6	7	8	9
0.0	0.000000	0.007071	0.014141	0.021208	0.028273	0.035333	0.042388	0.049437	0.056478	0.063511
0.1	0.070535	0.077548	0.084549	0.091538	0.098513	0.105474	0.112420	0.119349	0.126261	0.133154
0.2	0.140028	0.146882	0.153715	0.160526	0.167313	0.174078	0.180817	0.187532	0.194220	0.200881
0.3	0.207514	0.214119	0.220695	0.227241	0.233756	0.240239	0.246691	0.253110	0.259496	0.265848
0.4	0.272166	0.278448	0.284695	0.290906	0.297080	0.303218	0.309318	0.315380	0.321403	0.327388
0.5	0.333333	0.339240	0.345106	0.350932	0.356718	0.362462	0.368166	0.373829	0.379450	0.385029
0.6	0.390567	0.396062	0.401516	0.406926	0.412295	0.417620	0.422903	0.428144	0.433341	0.438496
0.7	0.443607	0.448676	0.453702	0.458684	0.463624	0.468521	0.473376	0.478187	0.482956	0.487682
0.8	0.492366	0.497008	0.501607	0.506164	0.510679	0.515152	0.519583	0.523973	0.528322	0.532629
0.9	0.536895	0.541121	0.545306	0.549450	0.553555	0.557619	0.561644	0.565629	0.569575	0.573482
1.0	0.577351	0.581180	0.584972	0.588725	0.592441	0.596120	0.599761	0.603366	0.606933	0.610465
1.1	0.613960	0.617420	0.620844	0.624233	0.627588	0.630907	0.634193	0.637444	0.640662	0.643846
1.2	0.646997	0.650115	0.653201	0.656255	0.659276	0.662266	0.665225	0.668153	0.671050	0.673917
1.3	0.676753	0.679560	0.682337	0.685085	0.6887804	0.690494	0.693156	0.695790	0.698397	0.700975
1.4	0.703527	0.706051	0.708549	0.711021	0.713466	0.715886	0.718280	0.720649	0.722993	0.725312
1.5	0.727607	0.729878	0.732125	0.734348	0.736547	0.738724	0.740878	0.743009	0.745118	0.747204
1.6	0.749269	0.751312	0.753334	0.755335	0.757314	0.759273	0.761212	0.763130	0.765029	0.766908
1.7	0.768767	0.770607	0.772428	0.774230	0.776013	0.777778	0.779525	0.781254	0.782965	0.784658
1.8	0.786334	0.787993	0.789635	0.791260	0.792868	0.794460	0.796036	0.797596	0.799140	0.800668
1.9	0.802181	0.803679	0.805161	0.806628	0.808081	0.809519	0.810943	0.812352	0.813748	0.815129
2.0	0.816497	0.817851	0.819192	0.820519	0.821833	0.823134	0.824423	0.825698	0.826961	0.828212
2.1	0.829450	0.830677	0.831891	0.833094	0.834285	0.835464	0.836632	0.837788	0.838934	0.840068
2.2	0.841191	0.842304	0.844497	0.845578	0.846649	0.847710	0.848760	0.849801	0.850831	
2.3	0.851852	0.852864	0.853865	0.854858	0.855841	0.856815	0.857780	0.858735	0.859682	0.860621
2.4	0.861550	0.862471	0.863384	0.864288	0.865183	0.866071	0.866950	0.867822	0.868685	0.869541

(continued)

Table A.17 (continued)

	0	1	2	3	4	5	6	7	8	9
2.5	0.870389	0.871229	0.872061	0.872887	0.873704	0.874515	0.875318	0.876114	0.876902	0.877684
2.6	0.878459	0.879227	0.879988	0.880743	0.881490	0.882232	0.882966	0.883695	0.884417	0.885132
2.7	0.885842	0.886545	0.887242	0.887933	0.888618	0.889298	0.889971	0.890639	0.891301	0.891957
2.8	0.892608	0.893253	0.893893	0.894527	0.895156	0.895780	0.896398	0.897011	0.897619	0.898222
2.9	0.898820	0.899413	0.900001	0.900584	0.901163	0.901736	0.902305	0.902869	0.903429	0.903984
3.0	0.904534	0.905080	0.905622	0.906159	0.906692	0.907220	0.907745	0.908265	0.908780	0.909292
3.1	0.909800	0.910303	0.910803	0.911298	0.911790	0.912278	0.912762	0.913242	0.913718	0.914191
3.2	0.914660	0.915125	0.915586	0.916044	0.916499	0.916950	0.917397	0.917841	0.918282	0.918719
3.3	0.919153	0.919583	0.920010	0.920434	0.920855	0.921273	0.921687	0.922098	0.922507	0.922912
3.4	0.923314	0.923713	0.924109	0.924502	0.924892	0.925279	0.925664	0.926045	0.926424	0.926800
3.5	0.927173	0.927543	0.927911	0.928276	0.928639	0.928998	0.929355	0.929710	0.930062	0.930411
3.6	0.930758	0.931103	0.931445	0.931784	0.932121	0.932456	0.932788	0.933118	0.933445	0.933771
3.7	0.934094	0.934414	0.934733	0.935049	0.935363	0.935675	0.935984	0.936292	0.936597	0.936900
3.8	0.937201	0.937500	0.937797	0.938092	0.938385	0.938676	0.938965	0.939252	0.939537	0.939820
3.9	0.940101	0.940380	0.940657	0.940933	0.941206	0.941478	0.941748	0.942016	0.942282	0.942547
4.0	0.942809	0.943070	0.943330	0.943587	0.943843	0.944097	0.944350	0.944601	0.944850	0.945098
4.1	0.945343	0.945588	0.945831	0.946072	0.946311	0.946550	0.946786	0.947021	0.947255	0.947487
4.2	0.947717	0.947946	0.948174	0.948400	0.948625	0.948848	0.949070	0.949290	0.949509	0.949727
4.3	0.949943	0.950158	0.950372	0.950584	0.950795	0.951005	0.951213	0.951420	0.951626	0.951830
4.4	0.952034	0.952235	0.952436	0.952636	0.952834	0.953031	0.953227	0.953422	0.953615	0.953807
4.5	0.953998	0.954188	0.954377	0.954565	0.954752	0.954937	0.955121	0.955305	0.955487	0.955668
4.6	0.955848	0.956027	0.956205	0.956381	0.956557	0.956732	0.956905	0.957078	0.957250	0.957420
4.7	0.957590	0.957759	0.957926	0.958093	0.958259	0.958424	0.958587	0.958750	0.958912	0.959073
4.8	0.959233	0.959392	0.959551	0.959708	0.959865	0.960020	0.960175	0.960329	0.960481	0.960634
4.9	0.960785	0.960935	0.961085	0.961233	0.961381	0.961528	0.961674	0.961820	0.961964	0.962108

Table A.18 Integral of Student's function ($f = 3$), or probability p , as function of critical value T_{crit}

	0	1	2	3	4	5	6	7	8	9
0.0	0.000000	0.007351	0.014701	0.02049	0.029394	0.036735	0.044071	0.051401	0.058725	0.066041
0.1	0.073348	0.080645	0.087932	0.095207	0.102469	0.109718	0.116953	0.124172	0.131375	0.138562
0.2	0.145730	0.152879	0.160009	0.167118	0.174205	0.181271	0.188313	0.195332	0.202326	0.209294
0.3	0.216237	0.223152	0.230040	0.236900	0.243730	0.250531	0.257301	0.264040	0.270748	0.277423
0.4	0.284065	0.290674	0.297248	0.303788	0.310293	0.316761	0.323194	0.329590	0.335948	0.342269
0.5	0.348552	0.354796	0.361002	0.367168	0.373294	0.379380	0.385426	0.391431	0.397395	0.403318
0.6	0.409199	0.415038	0.420835	0.426590	0.432302	0.437972	0.443599	0.449182	0.454723	0.460220
0.7	0.465673	0.471083	0.476449	0.481772	0.487051	0.492286	0.497477	0.502624	0.507727	0.512786
0.8	0.517801	0.522773	0.527700	0.532584	0.537424	0.542220	0.546973	0.551682	0.556348	0.560970
0.9	0.565549	0.570085	0.574579	0.579029	0.583437	0.587802	0.592125	0.596406	0.600645	0.604842
1.0	0.608998	0.613112	0.617186	0.621218	0.625209	0.629160	0.633071	0.636942	0.640773	0.644565
1.1	0.648317	0.652030	0.655705	0.659341	0.662939	0.666499	0.670021	0.673506	0.676953	0.680364
1.2	0.683738	0.687076	0.690378	0.693644	0.696875	0.700071	0.703232	0.706358	0.709450	0.712508
1.3	0.715533	0.718524	0.721482	0.724407	0.727300	0.730161	0.732990	0.735787	0.738553	0.741289
1.4	0.743993	0.746667	0.749311	0.751925	0.754510	0.757066	0.759593	0.762091	0.764561	0.767002
1.5	0.769416	0.771803	0.774163	0.776495	0.778801	0.781081	0.783335	0.785563	0.787766	0.789943
1.6	0.792096	0.794223	0.796327	0.798406	0.800462	0.802494	0.804503	0.806488	0.808451	0.810392
1.7	0.812310	0.814206	0.816080	0.817933	0.819764	0.821575	0.823365	0.825134	0.826883	0.828611
1.8	0.830320	0.832010	0.833680	0.835331	0.836962	0.838576	0.840170	0.841747	0.843305	0.844846
1.9	0.846369	0.847874	0.849363	0.850834	0.852289	0.853727	0.855148	0.856554	0.857943	0.859316
2.0	0.860674	0.862017	0.863344	0.864656	0.865954	0.867236	0.868504	0.869758	0.870998	0.872223
2.1	0.873435	0.874633	0.875818	0.876989	0.878147	0.879292	0.880425	0.881544	0.882651	0.883746
2.2	0.884828	0.885899	0.886957	0.888004	0.889039	0.890062	0.891074	0.892075	0.893065	0.894044
2.3	0.895012	0.895969	0.896916	0.897853	0.898779	0.899695	0.900600	0.901496	0.902383	0.903259
2.4	0.904126	0.904983	0.905831	0.906670	0.907500	0.908321	0.909133	0.909936	0.910730	0.911516

(continued)

Table A.18 (continued)

	0	1	2	3	4	5	6	7	8	9
2.5	0.912294	0.913063	0.913824	0.914576	0.915321	0.916057	0.916786	0.917507	0.918221	0.918926
2.6	0.919625	0.920315	0.920999	0.921675	0.922344	0.923007	0.923662	0.924310	0.924951	0.925586
2.7	0.926214	0.926836	0.927451	0.928060	0.928662	0.929258	0.929848	0.930432	0.931010	0.931582
2.8	0.932147	0.932708	0.933262	0.933811	0.934354	0.934891	0.935423	0.935950	0.936471	0.936987
2.9	0.937498	0.938004	0.938504	0.939000	0.939490	0.939976	0.940456	0.940932	0.941403	0.941870
3.0	0.942332	0.942789	0.943241	0.943689	0.944133	0.944572	0.945007	0.945438	0.945864	0.946287
3.1	0.946705	0.947119	0.947529	0.947935	0.948337	0.948735	0.949129	0.949520	0.949906	0.950289
3.2	0.950669	0.951044	0.951416	0.951785	0.952149	0.952511	0.952869	0.953223	0.953575	0.953922
3.3	0.954267	0.954608	0.954946	0.955281	0.955613	0.955942	0.956267	0.956590	0.956909	0.957226
3.4	0.957539	0.957850	0.958157	0.958462	0.958764	0.959064	0.959360	0.959654	0.959945	0.960234
3.5	0.960519	0.960803	0.961083	0.961361	0.961637	0.961910	0.962180	0.962448	0.962714	0.962977
3.6	0.963238	0.963497	0.963753	0.964007	0.964259	0.964508	0.964755	0.965000	0.965243	0.965484
3.7	0.965722	0.965959	0.966193	0.966426	0.966656	0.966884	0.967110	0.967335	0.967557	0.967777
3.8	0.967996	0.968212	0.968427	0.968640	0.968851	0.969060	0.969268	0.969473	0.969677	0.969879
3.9	0.970079	0.970278	0.970475	0.970670	0.970864	0.971056	0.971246	0.971435	0.971622	0.971808
4.0	0.971992	0.972174	0.972355	0.972535	0.972713	0.972889	0.973064	0.973238	0.973410	0.973581
4.1	0.973750	0.973918	0.974084	0.974250	0.974413	0.974576	0.974737	0.974897	0.975055	0.975212
4.2	0.975368	0.975523	0.975676	0.975829	0.975980	0.976129	0.976278	0.976425	0.976571	0.976716
4.3	0.976860	0.977003	0.977144	0.977285	0.977424	0.977562	0.977699	0.977835	0.977970	0.978104
4.4	0.978237	0.978369	0.978500	0.978630	0.978758	0.978886	0.979013	0.979139	0.979263	0.979387
4.5	0.979510	0.979632	0.979753	0.979873	0.979992	0.980110	0.980228	0.980344	0.980460	0.980574
4.6	0.980688	0.980801	0.980913	0.981024	0.981135	0.981244	0.981353	0.981461	0.981568	0.981674
4.7	0.981780	0.981884	0.981988	0.982091	0.982194	0.982295	0.982396	0.982496	0.982596	0.982694
4.8	0.982792	0.982890	0.982986	0.983082	0.983177	0.983271	0.983365	0.983458	0.983550	0.983642
4.9	0.983733	0.983823	0.983913	0.984002	0.984091	0.984178	0.984266	0.984352	0.984438	0.984523

Table A.19 Integral of Student's function ($f = 5$), or probability p , as function of critical value T_{crit}

	0	1	2	3	4	5	6	7	8	9
0.0	0.000000	0.007592	0.015183	0.022772	0.030359	0.037942	0.045520	0.053093	0.060659	0.068219
0.1	0.075770	0.083312	0.090844	0.098366	0.105875	0.113372	0.120856	0.128325	0.135780	0.143218
0.2	0.150640	0.158043	0.165429	0.172795	0.180141	0.187466	0.194769	0.202050	0.209308	0.216542
0.3	0.223751	0.230935	0.238092	0.245223	0.252326	0.259401	0.266446	0.273463	0.280448	0.287403
0.4	0.294327	0.301218	0.308077	0.314902	0.321694	0.328451	0.335173	0.341860	0.348510	0.355124
0.5	0.361701	0.368241	0.374742	0.381206	0.387630	0.394015	0.400361	0.406667	0.412932	0.419156
0.6	0.425340	0.431482	0.437582	0.443641	0.449657	0.455630	0.461561	0.467449	0.473293	0.479094
0.7	0.484851	0.490565	0.496234	0.501859	0.507440	0.512976	0.518468	0.523915	0.529317	0.534674
0.8	0.559986	0.545253	0.550476	0.555653	0.560785	0.565871	0.570913	0.575910	0.580861	0.585768
0.9	0.590629	0.595445	0.600217	0.604944	0.609626	0.614263	0.618835	0.623404	0.627908	0.632367
1.0	0.636783	0.641154	0.645482	0.649767	0.654007	0.658205	0.662359	0.666471	0.670539	0.674566
1.1	0.678549	0.682491	0.686391	0.690249	0.694066	0.697841	0.701576	0.705270	0.708923	0.712536
1.2	0.716109	0.719643	0.723136	0.726591	0.730007	0.733384	0.736723	0.740023	0.743286	0.746512
1.3	0.749700	0.752851	0.755965	0.759043	0.762085	0.765092	0.768062	0.770998	0.773899	0.777675
1.4	0.779596	0.782394	0.785158	0.787889	0.790587	0.793252	0.795885	0.798485	0.801054	0.803591
1.5	0.806097	0.808572	0.811016	0.813430	0.815814	0.818168	0.820493	0.822789	0.825056	0.827295
1.6	0.829505	0.831688	0.833843	0.835970	0.838071	0.840145	0.842192	0.844213	0.846209	0.848179
1.7	0.850124	0.852043	0.853938	0.855809	0.857655	0.859478	0.861277	0.863053	0.864805	0.866535
1.8	0.868243	0.869928	0.871591	0.873233	0.874853	0.876452	0.878030	0.879587	0.881124	0.883640
1.9	0.884137	0.885614	0.887072	0.888510	0.889930	0.891330	0.892712	0.894076	0.895422	0.896750
2.0	0.898061	0.899354	0.900630	0.901889	0.903132	0.904358	0.905567	0.906761	0.907938	0.909101
2.1	0.910247	0.911378	0.912495	0.913596	0.914683	0.915755	0.916813	0.917857	0.918887	0.919904
2.2	0.920906	0.921896	0.922872	0.923835	0.924786	0.925723	0.926649	0.927562	0.928462	0.929351
2.3	0.930228	0.931093	0.931947	0.932789	0.933620	0.934440	0.935249	0.936048	0.936835	0.937613
2.4	0.938380	0.939136	0.939883	0.940620	0.941347	0.942064	0.942772	0.943470	0.944159	0.944839

(continued)

Table A.19 (continued)

	0	1	2	3	4	5	6	7	8	9
2.5	0.945510	0.946172	0.946826	0.947470	0.948107	0.948734	0.949354	0.949955	0.950568	0.951164
2.6	0.951751	0.952331	0.952903	0.953467	0.954024	0.954574	0.955116	0.955652	0.956180	0.956702
2.7	0.957216	0.957724	0.958225	0.958720	0.959208	0.959690	0.960166	0.960635	0.961098	0.961556
2.8	0.962007	0.962452	0.962892	0.963326	0.963754	0.964177	0.964594	0.965006	0.965412	0.965814
2.9	0.966210	0.966601	0.966987	0.967368	0.967744	0.968115	0.968482	0.968843	0.969200	0.969553
3.0	0.969901	0.970245	0.970584	0.970919	0.971250	0.971576	0.971898	0.972217	0.972531	0.972841
3.1	0.973147	0.973450	0.973748	0.974043	0.974334	0.974621	0.974905	0.975185	0.975462	0.975735
3.2	0.976005	0.976272	0.976535	0.976795	0.977051	0.977305	0.977555	0.977802	0.978046	0.978287
3.3	0.978525	0.978760	0.978992	0.979221	0.979448	0.979672	0.979893	0.980111	0.980326	0.980539
3.4	0.980749	0.980957	0.981162	0.981365	0.981565	0.981763	0.981958	0.982151	0.982342	0.982530
3.5	0.982716	0.982900	0.983081	0.983261	0.983438	0.983613	0.983786	0.983957	0.984126	0.984292
3.6	0.984457	0.984620	0.984781	0.984940	0.985097	0.985252	0.985406	0.985557	0.985707	0.985855
3.7	0.986001	0.986146	0.986288	0.986429	0.986569	0.986707	0.986843	0.986977	0.987111	0.987242
3.8	0.987372	0.987500	0.987627	0.987753	0.987877	0.987999	0.988120	0.988240	0.988359	0.988475
3.9	0.988591	0.988705	0.988818	0.988930	0.989041	0.989150	0.989258	0.989364	0.989470	0.989574
4.0	0.989677	0.989779	0.989880	0.989979	0.990078	0.990175	0.990271	0.990366	0.990461	0.990554
4.1	0.990646	0.990737	0.990826	0.990915	0.991003	0.991090	0.991176	0.991261	0.991345	0.991429
4.2	0.991511	0.991592	0.991673	0.991752	0.991831	0.991909	0.991986	0.992062	0.992137	0.992211
4.3	0.992285	0.992358	0.992430	0.992501	0.992572	0.992641	0.992710	0.992778	0.992846	0.992913
4.4	0.992979	0.993044	0.993108	0.993172	0.993236	0.993298	0.993360	0.993421	0.993482	0.993542
4.5	0.993601	0.993660	0.993718	0.993775	0.993832	0.993888	0.993944	0.993999	0.994053	0.994107
4.6	0.994160	0.994213	0.994265	0.994317	0.994368	0.994418	0.994468	0.994518	0.994567	0.994615
4.7	0.994663	0.994711	0.994758	0.994804	0.994850	0.994896	0.994941	0.994986	0.995030	0.995073
4.8	0.995117	0.995160	0.995202	0.995244	0.995285	0.995327	0.995367	0.995408	0.995447	0.995487
4.9	0.995526	0.995565	0.995603	0.995641	0.995678	0.995715	0.995752	0.995789	0.995825	0.995860

Table A.20 Integral of Student's function ($f = 10$), or probability p , as function of critical value T_{crit}

	0	1	2	3	4	5	6	7	8	9
0.0	0.000000	0.007782	0.015563	0.023343	0.031120	0.038893	0.046662	0.054426	0.062184	0.069936
0.1	0.077679	0.085414	0.093140	0.100856	0.108560	0.116253	0.123933	0.131600	0.139252	0.146889
0.2	0.154511	0.162116	0.169703	0.177272	0.184822	0.192352	0.199861	0.207350	0.214816	0.222260
0.3	0.229679	0.237075	0.244446	0.251791	0.259110	0.266402	0.273666	0.280902	0.288109	0.295286
0.4	0.302433	0.309549	0.316633	0.323686	0.330706	0.337692	0.344645	0.351563	0.358447	0.365295
0.5	0.372107	0.378882	0.385621	0.392322	0.398985	0.405610	0.412197	0.418744	0.425251	0.431718
0.6	0.438145	0.444531	0.450876	0.457179	0.463441	0.469660	0.475837	0.481971	0.488061	0.494109
0.7	0.500113	0.506072	0.511988	0.517860	0.523686	0.529468	0.535205	0.540897	0.546543	0.552144
0.8	0.557700	0.563209	0.568673	0.574091	0.579463	0.584788	0.590067	0.595300	0.600487	0.605627
0.9	0.610721	0.615768	0.620769	0.625724	0.630632	0.635494	0.640309	0.645078	0.649800	0.654477
1.0	0.659107	0.663691	0.668230	0.672722	0.677169	0.681569	0.685925	0.690235	0.694499	0.698719
1.1	0.702893	0.707023	0.711108	0.715149	0.719145	0.723097	0.727006	0.730870	0.734691	0.738469
1.2	0.742204	0.745896	0.749545	0.753152	0.756717	0.760240	0.763721	0.767161	0.770559	0.773917
1.3	0.777235	0.780511	0.783748	0.786945	0.790103	0.793221	0.796301	0.799341	0.802344	0.805308
1.4	0.808235	0.811124	0.813976	0.816792	0.819570	0.822313	0.825019	0.827690	0.830326	0.833927
1.5	0.835493	0.838025	0.840523	0.842987	0.845417	0.847815	0.850180	0.852512	0.854813	0.857081
1.6	0.859319	0.861525	0.863700	0.865845	0.867959	0.870044	0.872099	0.874125	0.876122	0.878091
1.7	0.880031	0.881943	0.883828	0.885685	0.887516	0.889319	0.891097	0.892848	0.894573	0.896273
1.8	0.897948	0.899958	0.901223	0.902825	0.904402	0.905955	0.907485	0.908992	0.910477	0.911938
1.9	0.9133378	0.914796	0.916191	0.917566	0.918919	0.920252	0.921564	0.922856	0.924128	0.925380
2.0	0.926612	0.927826	0.929020	0.930196	0.931353	0.932492	0.933613	0.934717	0.935803	0.936871
2.1	0.937923	0.938958	0.939977	0.940979	0.941965	0.942936	0.943891	0.944830	0.945755	0.946664
2.2	0.947559	0.948440	0.949306	0.950158	0.950996	0.951821	0.952632	0.953430	0.954215	0.954987
2.3	0.955746	0.956493	0.957228	0.957950	0.958661	0.959360	0.960047	0.960723	0.961388	0.962042
2.4	0.962685	0.963317	0.963939	0.964550	0.965151	0.965743	0.966324	0.966896	0.967458	0.968010

(continued)

Table A.20 (continued)

	0	1	2	3	4	5	6	7	8	9
2.5	0.968554	0.969088	0.969613	0.970130	0.970637	0.971136	0.971627	0.972110	0.972584	0.973050
2.6	0.973509	0.973960	0.974403	0.974838	0.975266	0.975687	0.976101	0.976508	0.976908	0.977301
2.7	0.977687	0.978067	0.978440	0.978807	0.979168	0.979522	0.979871	0.980213	0.980550	0.980881
2.8	0.981206	0.981526	0.981840	0.982149	0.982452	0.982750	0.983044	0.983352	0.983615	0.983893
2.9	0.984167	0.984436	0.984700	0.984960	0.985215	0.985466	0.985712	0.985955	0.986193	0.986427
3.0	0.986657	0.986883	0.987105	0.987323	0.987538	0.987749	0.987956	0.988160	0.988360	0.988556
3.1	0.988750	0.988940	0.989126	0.989310	0.989490	0.989667	0.989842	0.990013	0.990181	0.990346
3.2	0.990509	0.990668	0.990825	0.990979	0.991131	0.991280	0.991426	0.991570	0.991711	0.991850
3.3	0.991987	0.992121	0.992253	0.992383	0.992510	0.992635	0.992758	0.992879	0.992998	0.993115
3.4	0.993229	0.993342	0.993453	0.993562	0.993669	0.993774	0.993878	0.993979	0.994079	0.994177
3.5	0.994274	0.994369	0.994462	0.994554	0.994644	0.994732	0.994819	0.994905	0.994989	0.995071
3.6	0.995153	0.995232	0.995311	0.995388	0.995464	0.995538	0.995611	0.995683	0.995754	0.995824
3.7	0.995892	0.995959	0.996025	0.996090	0.996154	0.996217	0.996278	0.996339	0.996398	0.996457
3.8	0.996515	0.996571	0.996627	0.996682	0.996735	0.996788	0.996840	0.996891	0.996941	0.996991
3.9	0.997039	0.997087	0.997134	0.997180	0.997226	0.997270	0.997314	0.997357	0.997399	0.997441
4.0	0.997482	0.997522	0.997562	0.997601	0.997639	0.997677	0.997714	0.997750	0.997786	0.997821
4.1	0.997856	0.997890	0.997923	0.997956	0.997989	0.998020	0.998052	0.998082	0.998113	0.998142
4.2	0.998172	0.998201	0.998229	0.998257	0.998284	0.998311	0.998337	0.998363	0.998389	0.998414
4.3	0.998439	0.998463	0.998487	0.998511	0.998534	0.998557	0.998579	0.998601	0.998623	0.998644
4.4	0.998665	0.998886	0.998706	0.998726	0.998746	0.998765	0.998784	0.998803	0.998821	0.998840
4.5	0.998857	0.998875	0.998892	0.998909	0.998926	0.998942	0.998958	0.998974	0.998990	0.999005
4.6	0.999020	0.999035	0.999050	0.999064	0.999078	0.999092	0.999106	0.999120	0.999133	0.999146
4.7	0.999159	0.999172	0.999184	0.999196	0.999208	0.999220	0.999232	0.999243	0.999255	0.999266
4.8	0.999277	0.999288	0.999298	0.999309	0.999319	0.999329	0.999349	0.999358	0.999368	0.999386
4.9	0.999377	0.999387	0.999396	0.999404	0.999413	0.999422	0.999430	0.999439	0.999447	0.999455

Table A.21 Integral of Student's function ($f = 20$), or probability p , as function of critical value T_{crit}

	0	1	2	3	4	5	6	7	8	9
0.0	0.000000	0.007880	0.015758	0.023636	0.031510	0.039382	0.047249	0.055111	0.062968	0.070818
0.1	0.078660	0.086494	0.094320	0.102135	0.109940	0.117733	0.125514	0.133282	0.141036	0.148776
0.2	0.156500	0.164208	0.171899	0.179572	0.187227	0.194863	0.202478	0.210073	0.217647	0.225199
0.3	0.232727	0.240232	0.247713	0.255169	0.262599	0.270003	0.277379	0.284728	0.292049	0.299341
0.4	0.306604	0.313836	0.321037	0.328207	0.335345	0.342450	0.349522	0.356561	0.363565	0.370534
0.5	0.377468	0.384366	0.391228	0.398053	0.404841	0.411591	0.418302	0.424975	0.431608	0.438202
0.6	0.444756	0.451269	0.457742	0.464174	0.470563	0.476911	0.483217	0.489480	0.495700	0.501877
0.7	0.508010	0.514099	0.52045	0.526146	0.532102	0.538013	0.543879	0.549700	0.555476	0.561206
0.8	0.566889	0.572527	0.578119	0.583664	0.589162	0.594614	0.600019	0.605378	0.610689	0.615953
0.9	0.621171	0.626341	0.631463	0.636539	0.641567	0.646548	0.651482	0.656368	0.661207	0.665999
1.0	0.670744	0.675441	0.680091	0.684694	0.689250	0.693759	0.698222	0.702637	0.707006	0.711328
1.1	0.715604	0.719833	0.724016	0.728154	0.732245	0.736291	0.740291	0.744245	0.748155	0.752019
1.2	0.755839	0.759614	0.763344	0.767031	0.770673	0.774271	0.777826	0.781338	0.784807	0.788233
1.3	0.791616	0.794957	0.798256	0.801513	0.804728	0.807903	0.811036	0.814129	0.817181	0.820193
1.4	0.823165	0.826698	0.828992	0.831846	0.834662	0.837440	0.840180	0.842882	0.845546	0.848174
1.5	0.850765	0.853319	0.855837	0.858320	0.860767	0.863179	0.865556	0.867899	0.870207	0.872482
1.6	0.874723	0.876932	0.879107	0.881250	0.883361	0.885440	0.887487	0.889503	0.891489	0.893444
1.7	0.895369	0.897264	0.899130	0.900967	0.902775	0.904554	0.906305	0.908029	0.909725	0.911394
1.8	0.913036	0.914651	0.916241	0.917804	0.919342	0.920855	0.922343	0.923806	0.925245	0.926660
1.9	0.928052	0.929420	0.930765	0.932088	0.933388	0.934666	0.935922	0.937157	0.938370	0.939563
2.0	0.940735	0.941886	0.943018	0.944130	0.945222	0.946295	0.947350	0.948385	0.949402	0.950402
2.1	0.951383	0.952347	0.953293	0.954222	0.955135	0.956031	0.956911	0.957774	0.958622	0.959455
2.2	0.960272	0.961074	0.961861	0.962634	0.963392	0.964136	0.964866	0.965583	0.966286	0.966976
2.3	0.967653	0.968317	0.968969	0.969608	0.970235	0.970850	0.971453	0.972044	0.972624	0.973194
2.4	0.973752	0.974299	0.974835	0.975361	0.975877	0.976383	0.976879	0.977365	0.977842	0.978309

(continued)

Table A.21 (continued)

	0	1	2	3	4	5	6	7	8	9
2.5	0.978767	0.979216	0.979556	0.980087	0.980510	0.980924	0.981330	0.981727	0.982117	0.982499
2.6	0.982873	0.983240	0.983599	0.983951	0.984296	0.984634	0.984965	0.985289	0.985607	0.985918
2.7	0.986222	0.986521	0.986813	0.987099	0.987380	0.987654	0.987923	0.988186	0.988444	0.988696
2.8	0.988943	0.989185	0.989422	0.989654	0.989881	0.990103	0.990321	0.990534	0.990743	0.990947
2.9	0.991146	0.991342	0.991533	0.991721	0.991904	0.992083	0.992259	0.992431	0.992599	0.992764
3.0	0.992925	0.993082	0.993236	0.993387	0.993535	0.993679	0.993820	0.993959	0.994094	0.994226
3.1	0.994356	0.994482	0.994606	0.994727	0.994846	0.994962	0.995075	0.995186	0.995294	0.995400
3.2	0.995504	0.995606	0.995705	0.995802	0.995897	0.995990	0.996081	0.996169	0.996256	0.996341
3.3	0.996424	0.996505	0.996585	0.996662	0.996738	0.996812	0.996885	0.996956	0.997025	0.997093
3.4	0.997159	0.997224	0.997287	0.997349	0.997410	0.997469	0.997527	0.997583	0.997638	0.997692
3.5	0.997745	0.997797	0.997847	0.997897	0.997945	0.997992	0.998038	0.998083	0.998127	0.998170
3.6	0.998212	0.998253	0.998293	0.998333	0.998371	0.998408	0.998445	0.998481	0.998516	0.998550
3.7	0.998583	0.998616	0.998648	0.998679	0.998709	0.998739	0.998768	0.998797	0.998824	0.998851
3.8	0.998878	0.998904	0.998929	0.998954	0.998978	0.999002	0.999025	0.999047	0.999069	0.999091
3.9	0.999112	0.999132	0.999152	0.999172	0.999191	0.999210	0.999228	0.999246	0.999263	0.999280
4.0	0.999297	0.999313	0.999329	0.999345	0.999360	0.999375	0.999389	0.999403	0.999417	0.999430
4.1	0.999444	0.999456	0.999469	0.999481	0.999493	0.999505	0.999516	0.999528	0.999539	0.999549
4.2	0.999560	0.999570	0.999580	0.999590	0.999599	0.999608	0.999617	0.999626	0.999635	0.999643
4.3	0.999652	0.999660	0.999667	0.999675	0.999683	0.999690	0.999697	0.999704	0.999711	0.999718
4.4	0.999724	0.999731	0.999737	0.999743	0.999749	0.999755	0.999760	0.999766	0.999771	0.999776
4.5	0.999782	0.999787	0.999792	0.999796	0.999801	0.999806	0.999810	0.999815	0.999819	0.999823
4.6	0.999827	0.999831	0.999835	0.999839	0.999842	0.999846	0.999850	0.999853	0.999856	0.999860
4.7	0.999863	0.999866	0.999869	0.999872	0.999875	0.999878	0.999881	0.999884	0.999886	0.999889
4.8	0.999891	0.999894	0.999896	0.999899	0.999901	0.999903	0.999906	0.999908	0.999910	0.999912
4.9	0.999914	0.999916	0.999918	0.999920	0.999922	0.999923	0.999925	0.999927	0.999928	0.999930

Table A.22 Integral of Student's function ($f = 50$), or probability p , as function of critical value T_{crit}

	0	1	2	3	4	5	6	7	8	9
0.0	0.000000	0.007939	0.015877	0.023814	0.031748	0.039678	0.047605	0.055527	0.063443	0.071353
0.1	0.079256	0.087150	0.095036	0.102912	0.110778	0.118632	0.126474	0.134304	0.142120	0.149922
0.2	0.157708	0.165479	0.173233	0.180970	0.188689	0.196388	0.204069	0.211729	0.219367	0.226985
0.3	0.234579	0.242151	0.249698	0.257221	0.264719	0.272191	0.279637	0.287055	0.294445	0.301807
0.4	0.309140	0.316443	0.323715	0.330957	0.338167	0.345345	0.352490	0.359602	0.366680	0.373723
0.5	0.380732	0.387705	0.394642	0.401542	0.408406	0.415232	0.422020	0.428770	0.435481	0.442152
0.6	0.448784	0.455376	0.461927	0.468437	0.474906	0.481333	0.487717	0.494060	0.500359	0.506616
0.7	0.512829	0.518998	0.525123	0.531204	0.537240	0.543231	0.549177	0.555077	0.560932	0.566742
0.8	0.572505	0.578222	0.583892	0.589516	0.595093	0.600623	0.606106	0.611542	0.616931	0.622272
0.9	0.627566	0.632812	0.638010	0.643161	0.648263	0.653318	0.658326	0.663325	0.668196	0.673059
1.0	0.677875	0.682642	0.687362	0.692033	0.696657	0.701233	0.705762	0.710243	0.714676	0.719062
1.1	0.723400	0.727691	0.731935	0.736132	0.740282	0.744386	0.748442	0.752452	0.756416	0.760334
1.2	0.764206	0.768032	0.771812	0.775547	0.779237	0.782882	0.786482	0.790037	0.793548	0.797015
1.3	0.800438	0.803818	0.807154	0.810447	0.813697	0.816904	0.820070	0.823193	0.826274	0.829314
1.4	0.832312	0.835270	0.838187	0.841064	0.843901	0.846698	0.849455	0.852174	0.854853	0.857495
1.5	0.860098	0.862663	0.865190	0.867681	0.870135	0.872552	0.874933	0.877278	0.879587	0.881862
1.6	0.884101	0.886306	0.888477	0.890614	0.892718	0.894788	0.896826	0.898831	0.900805	0.902746
1.7	0.904656	0.906535	0.908383	0.910201	0.911989	0.913747	0.915476	0.917176	0.918847	0.920490
1.8	0.922105	0.923693	0.925253	0.926786	0.928293	0.929773	0.931227	0.932656	0.934060	0.935439
1.9	0.936793	0.938123	0.939429	0.940711	0.941970	0.943206	0.944420	0.945611	0.946780	0.947927
2.0	0.949053	0.950158	0.951243	0.952306	0.953350	0.954374	0.955378	0.956363	0.957329	0.958276
2.1	0.959205	0.960116	0.961009	0.961884	0.962742	0.963584	0.964408	0.965216	0.966008	0.966784
2.2	0.967544	0.968289	0.969019	0.969734	0.970434	0.971120	0.971791	0.972449	0.973093	0.973724
2.3	0.974341	0.974946	0.975538	0.976117	0.976684	0.977239	0.977782	0.978313	0.978833	0.979342
2.4	0.979840	0.980327	0.980803	0.981269	0.981725	0.982171	0.982607	0.983033	0.983450	0.983857

(continued)

Table A.22 (continued)

	0	1	2	3	4	5	6	7	8	9
2.5	0.984255	0.984645	0.985026	0.985398	0.985761	0.986117	0.986464	0.986804	0.987135	0.987459
2.6	0.987776	0.988085	0.988387	0.988683	0.988971	0.989252	0.989527	0.989796	0.990058	0.990314
2.7	0.990564	0.990807	0.991046	0.991278	0.991505	0.991726	0.991942	0.992153	0.992359	0.992560
2.8	0.992756	0.992947	0.993333	0.993493	0.993666	0.993835	0.993999	0.994160	0.994316	
2.9	0.994469	0.994618	0.994763	0.994904	0.995042	0.995176	0.995307	0.995435	0.995559	0.995681
3.0	0.995799	0.995914	0.996026	0.996135	0.996242	0.996345	0.996447	0.996545	0.996641	0.996734
3.1	0.996825	0.996914	0.997000	0.997084	0.997165	0.997245	0.997323	0.997398	0.997471	0.997543
3.2	0.997612	0.997680	0.997746	0.997810	0.997872	0.997933	0.997992	0.998050	0.998106	0.998160
3.3	0.998213	0.998264	0.998314	0.998363	0.998411	0.998457	0.998501	0.998545	0.998587	0.998628
3.4	0.998669	0.998707	0.998745	0.998782	0.998818	0.998853	0.998886	0.998919	0.998951	0.998982
3.5	0.999012	0.999042	0.999070	0.999098	0.999125	0.999151	0.999176	0.999201	0.999225	0.999248
3.6	0.999270	0.999292	0.999314	0.999334	0.999354	0.999374	0.999393	0.999411	0.999429	0.999447
3.7	0.999463	0.999480	0.999496	0.999511	0.999526	0.999540	0.999554	0.999568	0.999581	0.999594
3.8	0.999607	0.999619	0.999631	0.999642	0.999653	0.999664	0.999674	0.999684	0.999694	0.999704
3.9	0.999713	0.999722	0.999731	0.999739	0.999747	0.999755	0.999763	0.999770	0.999777	0.999784
4.0	0.999791	0.999798	0.999804	0.999810	0.999816	0.999822	0.999828	0.999833	0.999839	0.999844
4.1	0.999849	0.999854	0.999858	0.999863	0.999867	0.999871	0.999875	0.999879	0.999883	0.999887
4.2	0.999891	0.999894	0.999898	0.999901	0.999904	0.999907	0.999910	0.999913	0.999916	0.999919
4.3	0.999921	0.999924	0.999926	0.999929	0.999931	0.999933	0.999936	0.999938	0.999940	0.999942
4.4	0.999944	0.999945	0.999947	0.999949	0.999951	0.999952	0.999954	0.999955	0.999957	0.999958
4.5	0.999960	0.999961	0.999962	0.999964	0.999965	0.999966	0.999967	0.999968	0.999969	0.999970
4.6	0.999971	0.999972	0.999973	0.999974	0.999975	0.999976	0.999977	0.999978	0.999979	
4.7	0.999980	0.999980	0.999981	0.999982	0.999982	0.999983	0.999983	0.999984	0.999985	
4.8	0.999986	0.999986	0.999987	0.999987	0.999988	0.999988	0.999988	0.999989	0.999990	
4.9	0.999990	0.999990	0.999991	0.999991	0.999991	0.999992	0.999992	0.999992	0.999993	

Table A.23 Comparison of integrals of Student's function at different critical values

<i>f</i>	Critical value				
	<i>T</i> = 1	<i>T</i> = 2	<i>T</i> = 3	<i>T</i> = 4	<i>T</i> = 5
1	0.500000	0.704833	0.795168	0.844042	0.874334
2	0.577351	0.816497	0.904534	0.942809	0.962251
3	0.608998	0.860674	0.942332	0.971992	0.984608
4	0.626099	0.883884	0.960058	0.983870	0.992510
5	0.636783	0.898061	0.969901	0.989677	0.995896
6	0.644083	0.907574	0.975992	0.992881	0.997548
7	0.649384	0.914381	0.980058	0.994811	0.998435
8	0.653407	0.919484	0.982929	0.996051	0.998948
9	0.656564	0.923448	0.985044	0.996890	0.999261
10	0.659107	0.926612	0.986657	0.997482	0.999463
11	0.661200	0.929196	0.987921	0.997914	0.999598
12	0.662951	0.931345	0.988934	0.998239	0.999691
13	0.664439	0.933160	0.989762	0.998488	0.999757
14	0.665718	0.934712	0.990449	0.998684	0.999806
15	0.666830	0.936055	0.991028	0.998841	0.999842
16	0.667805	0.937228	0.991521	0.998968	0.999870
17	0.668668	0.938262	0.991946	0.999073	0.999891
18	0.669435	0.939179	0.992315	0.999161	0.999908
19	0.670123	0.939998	0.992639	0.999234	0.999921
20	0.670744	0.940735	0.992925	0.999297	0.999932
21	0.671306	0.941400	0.993179	0.999351	0.999940
22	0.671817	0.942005	0.993406	0.999397	0.999948
23	0.672284	0.942556	0.993610	0.999438	0.999954
24	0.672713	0.943061	0.993795	0.999474	0.999959
25	0.673108	0.943524	0.993962	0.999505	0.999963
26	0.673473	0.943952	0.994115	0.999533	0.999967
27	0.673811	0.944348	0.994255	0.999558	0.999970
28	0.674126	0.944715	0.994383	0.999580	0.999973
29	0.674418	0.945057	0.994501	0.999600	0.999975
30	0.674692	0.945375	0.994610	0.999619	0.999977
31	0.674948	0.945673	0.994712	0.999635	0.999979
32	0.675188	0.945952	0.994806	0.999650	0.999981
33	0.675413	0.946214	0.994893	0.999664	0.999982
34	0.675626	0.946461	0.994975	0.999677	0.999983
35	0.675826	0.946693	0.995052	0.999688	0.999984
36	0.676015	0.946912	0.995123	0.999699	0.999985
37	0.676194	0.947119	0.995191	0.999709	0.999986
38	0.676364	0.947315	0.995254	0.999718	0.999987
39	0.676525	0.947501	0.995314	0.999727	0.999988
40	0.676678	0.947678	0.995370	0.999735	0.999989

(continued)

Table A.23 (continued)

f	Critical value				
	T = 1	T = 2	T = 3	T = 4	T = 5
41	0.676824	0.947846	0.995424	0.999742	0.999989
42	0.676963	0.948006	0.995474	0.999749	0.999990
43	0.677095	0.948158	0.995522	0.999755	0.999990
44	0.677222	0.948304	0.995568	0.999761	0.999991
45	0.677343	0.948443	0.995611	0.999767	0.999991
46	0.677458	0.948576	0.995652	0.999773	0.999992
47	0.677569	0.948703	0.995691	0.999778	0.999992
48	0.677675	0.948824	0.995729	0.999782	0.999992
49	0.677777	0.948941	0.995765	0.999787	0.999993
50	0.677875	0.949053	0.995799	0.999791	0.999993
∞	0.682690	0.954500	0.997301	0.999937	1.000000

A.7 The Linear Correlation Coefficient r

The linear correlation coefficient r defined in (14.2) is equal to the square root of the product bb' , where b is the best-fit slope of the linear regression of Y on X , and b' is the slope of the linear regression of X on Y . The probability distribution function of r , under the hypothesis that the variables X and Y are not correlated, is given by (14.3),

$$f_r(r) = \frac{1}{B(1/2, f/2)} (1 - r^2)^{f/2 - 1},$$

where N is the size of the sample, and $f = N - 2$ is the effective number of degrees of freedom of the dataset. This distribution is known as the *symmetric beta distribution* or r -distribution, see Appendix A.3.

Table A.24 reports the two-sided critical values of r calculated according to

$$1 - p = \int_{-r_{crit}}^{r_{crit}} f_r(r) dr, \quad (\text{A.11})$$

where p is the probability for a given value of the correlation coefficient to exceed, in absolute value, the critical value r_{crit} . The critical values are function of the number of degrees of freedom and of the probability p .

To evaluate the probability distribution function in the case of large f , a convenient approximation can be given using the asymptotic expansion for the Gamma function (see [1]):

$$\Gamma(az + b) \simeq \sqrt{2\pi} e^{-az} (az)^{az+b-1/2}. \quad (\text{A.12})$$

Table A.24 Critical values of the linear correlation coefficient

f	Probability p to have an absolute value of r below the critical value						
	0.50	0.60	0.70	0.80	0.90	0.95	0.99
2	0.500	0.600	0.700	0.800	0.900	0.950	0.990
3	0.404	0.492	0.585	0.687	0.805	0.878	0.959
4	0.347	0.426	0.511	0.608	0.729	0.811	0.917
5	0.309	0.380	0.459	0.551	0.669	0.754	0.875
6	0.281	0.347	0.420	0.507	0.621	0.707	0.834
7	0.260	0.321	0.390	0.472	0.582	0.666	0.798
8	0.242	0.300	0.365	0.443	0.549	0.632	0.765
9	0.228	0.282	0.344	0.419	0.521	0.602	0.735
10	0.216	0.268	0.327	0.398	0.497	0.576	0.708
20	0.152	0.189	0.231	0.284	0.360	0.423	0.537
30	0.124	0.154	0.189	0.233	0.296	0.349	0.449
40	0.107	0.133	0.164	0.202	0.257	0.304	0.393
50	0.096	0.119	0.147	0.181	0.231	0.273	0.354
60	0.087	0.109	0.134	0.165	0.211	0.250	0.325
70	0.081	0.101	0.124	0.153	0.195	0.232	0.302
80	0.076	0.094	0.116	0.143	0.183	0.217	0.283
90	0.071	0.089	0.109	0.135	0.173	0.205	0.267
100	0.068	0.084	0.104	0.128	0.164	0.195	0.254
200	0.048	0.060	0.073	0.091	0.116	0.138	0.181
300	0.039	0.049	0.060	0.074	0.095	0.113	0.148
500	0.030	0.038	0.046	0.057	0.073	0.087	0.114
1000	0.021	0.027	0.033	0.041	0.052	0.062	0.081

For large values of f , the ratio of the Gamma functions that appear in the r -distribution can therefore be approximated as

$$\frac{\Gamma\left(\frac{f+1}{2}\right)}{\Gamma\left(\frac{f}{2}\right)} \simeq \sqrt{\frac{f}{2}}.$$

The equivalence of hypothesis testing with the r -distribution and with the standard beta distribution for r^2 was established in Sect. 14.4. It is therefore not necessary to provide separate tables for the critical values of the r^2 statistic.

A.8 The Kolmogorov–Smirnov Statistics

The one-sample Kolmogorov–Smirnov statistic D_N is defined in (19.7) as

$$D_N = \max_x |F_N(x) - F(x)|,$$

where $F(x)$ is the parent distribution, and $F_N(x)$ the sample distribution.

The cumulative distribution of the test statistic can be approximated by

$$P(D_N < z/(\sqrt{N} + 0.12 + 0.11/\sqrt{N})) \simeq \Phi(z),$$

where

$$\Phi(z) = \sum_{r=-\infty}^{\infty} (-1)^r e^{-2r^2z^2}.$$

and it is independent of the form of the parent distribution $F(x)$. For large values of N , we can use the asymptotic equation

$$P(D_N < z/\sqrt{N}) = \Phi(z).$$

In Table A.25 are listed the critical values of $\sqrt{N} D_N$ for various levels of probability. Values of the Kolmogorov–Smirnov statistic above the critical value indicate a rejection of the null hypothesis that the data are drawn from the parent model.

The two-sample Kolmogorov–Smirnov statistic is

$$D_{NM} = \max_x |F_M(x) - G_N(x)|,$$

where $F_M(x)$ and $G_N(x)$ are the sample cumulative distribution of two independent sets of observations of size M and N . This statistic has the same distribution as the one-sample Kolmogorov–Smirnov statistic, with the substitution of $MN/(M + N)$ in place of N , and in the limit of large M and N , (19.12).

Table A.25 Critical values of the Kolmogorov–Smirnov statistic D_N

N	Probability p to have $D_N \times \sqrt{N}$ below the critical value						
	0.50	0.60	0.70	0.80	0.90	0.95	0.99
1	0.750	0.800	0.850	0.900	0.950	0.975	0.995
2	0.707	0.782	0.866	0.967	1.098	1.191	1.314
3	0.753	0.819	0.891	0.978	1.102	1.226	1.436
4	0.762	0.824	0.894	0.985	1.130	1.248	1.468
5	0.765	0.827	0.902	0.999	1.139	1.260	1.495
6	0.767	0.833	0.910	1.005	1.146	1.272	1.510
7	0.772	0.838	0.914	1.009	1.154	1.279	1.523
8	0.776	0.842	0.917	1.013	1.159	1.285	1.532
9	0.779	0.844	0.920	1.017	1.162	1.290	1.540
10	0.781	0.846	0.923	1.020	1.166	1.294	1.546
15	0.788	0.855	0.932	1.030	1.177	1.308	1.565
20	0.793	0.860	0.937	1.035	1.184	1.315	1.576
25	0.796	0.863	0.941	1.039	1.188	1.320	1.583
30	0.799	0.866	0.943	1.042	1.192	1.324	1.588
35	0.801	0.868	0.946	1.045	1.194	1.327	1.591
40	0.803	0.869	0.947	1.046	1.196	1.329	1.594
45	0.804	0.871	0.949	1.048	1.198	1.331	1.596
50	0.805	0.872	0.950	1.049	1.199	1.332	1.598
60	0.807	0.874	0.952	1.051	1.201	1.335	1.601
70	0.808	0.875	0.953	1.053	1.203	1.337	1.604
80	0.810	0.877	0.955	1.054	1.205	1.338	1.605
90	0.811	0.878	0.956	1.055	1.206	1.339	1.607
100	0.811	0.879	0.957	1.056	1.207	1.340	1.608
200	0.816	0.883	0.961	1.061	1.212	1.346	1.614
300	0.818	0.885	0.964	1.063	1.214	1.348	1.617
500	0.820	0.887	0.966	1.065	1.216	1.350	1.620
1000	0.822	0.890	0.968	1.067	1.218	1.353	1.622
∞	0.828	0.895	0.973	1.073	1.224	1.358	1.628

References

1. Abramowitz, M., Stegun, I.: *Handbook of Mathematical Functions: With Formulas, Graphs, and Mathematical Tables*. Dover, New York (1970)
2. Akritas, M.G., Bershady, M.A.: Linear Regression for astronomical data with measurement errors and intrinsic scatter. *Astrophys. J.* **470**, 706 (1996)
3. Anderson, E.: The irises of the Gaspé peninsula. *Bull. Am. Iris Soc.* **59**, 2–5 (1935)
4. Aristoteles: *Analytica Posteriora*, Canon (collected posthumously) (384–322 BC)
5. Arnold, B., Balakrishnan, N., Nagaraja, H.: *A First Course in Order Statistics*. Society for Industrial and Applied Mathematics, Philadelphia (2008)
6. Baker, S., Cousins, R.D.: Clarification of the use of chi-square and likelihood functions in fits to histograms. *Nuclear Instrum. Methods Phys. Res.* **221**, 437–442 (1984)
7. Barnard, G.A.: Significance tests for 2x2 tables. *Biometrika* **34**(1–2), 123–138 (1947)
8. Bayes, T., Price, R.: An essay towards solving a problem in the doctrine of chances. *Philos. Trans. R. Soc. London* **53** (1763)
9. Bernstein, S.: On the axiomatic foundation of the theory of probability (in Russian). *Commun. Kharkiv Math. Soc.* **15**, 209–274 (1917)
10. Bonamente, M.: Distribution of the C statistic with applications to the sample mean of Poisson data. *J. Appl. Stat.* **47**(11), 2044–2065 (2020)
11. Bonamente, M., Hasler, N., Bulbul, E., Carlstrom, J.E., Culverhouse, T.L., Gralla, M., Greer, C., Hawkins, D., Hennessy, R., Joy, M., Kolodziejczak, J., Lamb, J.W., Landry, D., Leitch, E.M., Marrone, D.P., Miller, A., Mroczkowski, T., Muchovej, S., Plagge, T., Pryke, C., Sharp, M., Woody, D.: Comparison of pressure profiles of massive relaxed galaxy clusters using the Sunyaev-Zel'dovich and X-ray data. *New J. Phys.* **14**(2), 025010 (2012)
12. Bonamente, M., Joy, M.K., Carlstrom, J.E., Reese, E.D., LaRoque, S.J.: Markov chain Monte Carlo joint analysis of Chandra X-ray imaging spectroscopy and Sunyaev-Zel'dovich Effect data. *Astrophys J.* **614**, 194 (2004)
13. Bonamente, M., Joy, M., LaRoque, S.J., Carlstrom, J.E., Nagai, D., Marrone, D.P.: Scaling relations from Sunyaev-Zel'dovich effect and chandra x-ray measurements of high-redshift galaxy clusters. *Astrophys. J.* **675**, 106–114 (2008)
14. Bonamente, M., Spence, D.: A semi-analytical solution to the maximum-likelihood fit of Poisson data to a linear model using the Cash statistic. *J. Appl. Stat.* **49**(3), 522–552 (2022)
15. Brooks, S.P., Gelman, A.: General methods for monitoring convergence of iterative simulations. *J. Comput. Graphical Stat.* **7**(4), 434–455 (1998)
16. Bulmer, M.: *Principles of Statistics*. Dover, New York (1967)
17. Carlin, B., Gelfand, A., Smith, A.: Hierarchical Bayesian analysis for changepoint problems. *Appl. Stat.* **41**, 389–405 (1992)

18. Cash, W.: Parameter estimation in astronomy through application of the likelihood ratio. *Astrophys. J.* **228**, 939 (1979)
19. Conway, T.R., Maxwell, W.L.: A queuing model with state dependent service rates. *J. Ind. Eng.* **12**, 132–136 (1962)
20. Cox, D., Miller, H.: *The Theory of Stochastic Processes*. Chapman and Hall, London (1965)
21. Cramer, H.: *Mathematical Methods of Statistics*. Princeton University Press, Princeton (1946)
22. David, H.: *Order Statistics*, 2nd edn. Wiley, New York (1981)
23. Doob, J.L.: Heuristic approach to the Kolmogorov-Smirnov theorems. *Ann. Math. Stat.* **20**(3), 393–403 (1949)
24. Duns, J.S.: *Commentaria oxoniensia ad IV libros magistri Sententiarum* (ca. 1298–1299)
25. Efron, B.: Bootstrap methods: another look at the jackknife. *Ann. Stat.* **7**(1), 1–26 (1979)
26. Emslie, A., Massone, A.: Bayesian confidence limits of electron spectra obtained through regularized inversion of solar hard X-ray spectra. *Astrophys. J.* **759**, 122 (2012)
27. Feller, W.: On the Kolmogorov-Smirnov limit theorems for empirical distributions. *Ann. Math. Stat.* **19**(2), 177–189 (1948)
28. Fisher, R.: On the interpretation of χ^2 from contingency tables, and the calculation of p. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **85**, 87–94 (1922)
29. Fisher, R.: On a distribution yielding the error functions of several well known statistics. *Proc. Int. Congr. Math.* **2**, 805–813 (1924)
30. Fisher, R.: *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh (1925)
31. Fisher, R.: *Statistical Methods for Research Workers*, 5th edn. Oliver and Boyd, Edinburgh (1934)
32. Fisher, R.: *The Design of Experiments*. Oliver and Boyd, Edinburgh (1935)
33. Fisher, R.: The fiducial argument in statistical inference. *Ann. Eugenics* **6**(4), 391–398 (1935)
34. Fisher, R.: The use of multiple measurements in taxonomic problems. *Ann. Eugenics* **7**, 179–188 (1936)
35. Flegal, J.M., Jones, G.L.: Batch means and spectral variance estimators in Markov chains Monte Carlo. *Ann. Stat.* **38**(2), 1034–1070 (2010)
36. Galton, F.: *Natural Inheritance*. Macmillan, London (1889)
37. Gamerman, D.: *Markov Chain Monte Carlo*. Chapman and Hall CRC, Boca Raton (1997)
38. Gehrels, N.: Confidence limits for small numbers of events in astrophysical data. *Astrophys. J.* **303**, 336–346 (1986)
39. Gelman, A., Rubin, D.: Inference from iterative simulation using multiple sequences. *Stat. Sci.* **7**, 457–511 (1992)
40. Geman, S., Geman, D.: Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.* **PAMI-6**, 721–741 (1984)
41. Geweke, J.: Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In: Bernardo, J.M., Berger, J.O., Dawid, A.P., Smith, A.F.M. (eds.) *Bayesian Statistics 4*. Clarendon Press, Oxford, UK (1992)
42. Ghent, A.W.: A method for exact testing of 2x2, 2x3, 3x3, and other contingency tables, employing binomial coefficients. *Am. Midl. Nat.* **88**(1), 15–27 (1972)
43. Gilks, W., Richardson, S., Spiegelhalter, D.: *Markov Chain Monte Carlo in Practice*. Chapman and Hall CRC, Boca Raton (1996)
44. Glynn, P.W., Whitt, W.: Estimating the asymptotic variance with batch means. *Oper. Res. Lett.* **10**, 431–435 (1991)
45. Gosset, W.S.: The probable error of a mean. *Biometrika* **6**, 1–25 (1908)
46. Greenwood, M., Yule, G.U.: The statistics of anti-typhoid and anti-cholera inoculations, and the interpretation of such statistics in general. In: *Proceedings of the Royal Society of Medicine, Sect. Epidemiol. State Med.* **8**, 113–94 (1915)
47. Hammersley, J., Handscomb, D.: *Monte Carlo Methods*. Springer, Dordrecht (1964)
48. Hastings, W.K.: Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**(1), 97–109 (1970)
49. Heisenberg, W.: Über den anschaulichen Inhalt der quantentheoretischen Kinematik und Mechanik. *Zeitschrift für Physik* **43**(3–4), 172–198 (1927)

50. Helmert, F.: Die genauigkeit der formel von peters zur berechnung des wahrscheinlichen fehlers director beobachtungen gleicher genauigkeit. *Astron. Nachr.* **88**, 192–218 (1876)
51. Hubble, E., Humason, M.: The velocity-distance relation among extra-galactic nebulae. *Astrophys. J.* **74**, 43 (1931)
52. Humphrey, P.J., Liu, W., Buote, D.A.: χ^2 and Poissonian data: biases even in the high-count regime and how to avoid them. *Astrophys. J.* **693**, 822–829 (2009)
53. Isobe, T., Feigelson, E.D., Akritas, M.G., Babu, G.J.: Linear regression in astronomy. I. *Astrophys. J.* **364**, 104 (1990)
54. Jaynes, E.T.: Information theory and statistical mechanics. *Phys. Rev.* **106**(4), 620–630 (1957)
55. Jaynes, E.T.: Information theory and statistical mechanics. Brandeis University Summer Institute Lectures in Theoretical Physics, pp. 182–218 (1963)
56. Jeffreys, H.: Theory of Probability, 2nd edn. Oxford University Press, London (1948)
57. Jones, G.L.: On the Markov chain central limit theorem. *Probab. Surv.* **1** (2004)
58. Jones, G.L., Haran, M., Caffo, B.S., Neath, R.: Fixed-width output analysis for Markov Chain Monte Carlo. *J. Am. Stat. Assoc.* **101**(476), 1537–1547 (2006)
59. Kastra, J.S.: On the use of C-stat in testing models for X-ray spectra. *Astron. Astrophys.* **605**, A51 (2017)
60. Kelly, B.C.: Some aspects of measurement error in linear regression of astronomical data. *Astrophys. J.* **665**(2), 1489–1506 (2007)
61. Knoll, M., Wonodi, C.: Oxford-AstraZeneca COVID-19 vaccine efficacy. *The Lancet* **397**, 72–74 (2021). This paper is a comment to the clinical study by Voysey M. et al. on behalf of the Oxford COVID Vaccine Trial Group, *The Lancet*, **397**, 99–111 (2021)
62. Kolmogorov, A.: Sulla determinazione empirica di una legge di distribuzione. *Giornale dell’Istituto Italiano degli Attuari* **4**, 1–11 (1933)
63. Kolmogorov, A.: Foundations of the Theory of Probability. Chelsea, New York (1950)
64. Lampton, M., Margon, B., Bowyer, S.: Parameter estimation in X-ray astronomy. *Astrophys. J.* **208**, 177–190 (1976)
65. Le Clerc de Buffon, G.: Géométrie, Histoire de l’Academie Royale des Sciences, pp. 43–45. Paris (1733)
66. Lutz, M.: Programming Python. O’Reilly (1996)
67. Mandl, F.: Statistical Physics, 2nd edn. Wiley, Chichester (1988)
68. Marsaglia, G., Tsang, W., Wang, J.: Evaluating Kolmogorov’s distribution. *J. Stat. Softw.* **8**, 1–4 (2003)
69. Mendel, G.: Versuche über pflanzenhybriden, Verhandlungen des naturforschenden Vereines in Brünn, p. 3047. IV p, Bd (1865)
70. Metropolis, N.: The Beginning of the Monte Carlo Method. Los Alamos Science, Special Issue (1987)
71. Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., Teller, E.: Equation of state calculations by fast computing machines. *J. Chem. Phys.* **21**, 1087–1092 (1953)
72. Metropolis, N., Ulam, S.: The Monte Carlo method. *J. Am. Stat. Assoc.* **44**(247), 335–341 (1949). PMID: 18139350
73. Miller, R.G.: The jackknife-a review. *Biometrika* **61**(1), 1–15 (1974)
74. Nemmen, R.S., Georganopoulos, M., Guiriec, S., Meyer, E.T., Gehrels, N., Sambruna, R.M.: A universal scaling for the energetics of relativistic jets from black hole systems. *Science* **338**(6113), 1445 (2012)
75. Occam, W.: Summa totius logicae (ca. 1323)
76. Pearson, K.: On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. London, Edinburgh, Dublin Philos. Mag. J. Sci. **50**(302), 157–175 (1900)
77. Pearson, K.: On the χ^2 test of goodness of fit. *Biometrika* **14**(1–2), 186–191 (1922)
78. Pearson, K., Lee, A.: On the laws on inheritance in men. *Biometrika* **2**, 357–462 (1903)

79. Polack, F.P., Thomas, S.J., Kitchin, N., Absalon, J., Gurtman, A., Lockhart, S., Perez, J.L., Pérez Marc, G., Moreira, E.D., Zerbini, C., Bailey, R., Swanson, K.A., Roychoudhury, S., Koury, K., Li, P., Kalina, W.V., Cooper, D., Frenck, R.W., Hammitt, L.L., Tureci, O., Nell, H., Schaefer, A., Unal, S., Tresnan, D.B., Mather, S., Dormitzer, P.R., Şahin, U., Jansen, K.U., Gruber, W.C.: Safety and efficacy of the BNT162b2 mRNA Covid-19 vaccine. *New England J. Med.* **383**(27), 2603–2615 (2020). PMID: 33301246
80. Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P.: Numerical Recipes 3rd Edition: The Art of Scientific Computing, 3 edn. Cambridge University Press (2007)
81. Protassov, R., van Dyk, D.A., Connors, A., Kashyap, V.L., Siemiginowska, A.: Statistics, handle with care: detecting multiple model components with the likelihood ratio test. *Astrophys. J.* **571**(1), 545–559 (2002)
82. Quenouille, M.H.: Approximate tests of correlation in time-series. *J. R. Stat. Soc. Series B (Methodol.)* **11**(1), 68–84 (1949)
83. Quenouille, M.H.: Notes on bias in estimation. *Biometrika* **43**(3/4), 353–360 (1956)
84. Raftery, A., Lewis, S.: How many iterations in the Gibbs sampler? *Bayesian Stat.* **4**, 763–773 (1992)
85. Ross, S.: Stochastic Processes. Wiley, New York (1995)
86. Ross, S.: Introduction to Probability Models. Academic, San Diego (2003)
87. Scargle, J.D., Norris, J.P., Jackson, B., Chiang, J.: Studies in astronomical time series analysis. vi. Bayesian block representations. *Astrophys. J.* **764**(2), 167 (2013)
88. Sellers, K.F., Shmueli, G.: A flexible regression model for count data. *Ann. Appl. Stat.* **4**(2), 943–961 (2010)
89. Serfling, R.: Approximation Theorems of Mathematical Statistics. Wiley, New York (1980)
90. Shannon, C.E.: A mathematical theory of communication. *Bell Syst. Tech. J.* **27**(3), 379–423 (1948)
91. Shmueli, G., Minka, T.P., Kadane, J.B., Borle, S., Boatwright, P.: A useful distribution for fitting discrete data: revival of the Conway-Maxwell-Poisson distribution. *J. R. Stat. Soc. Series C (Appl. Stat.)* **54**(1), 127–142 (2005)
92. Siegrist, K.: Random. Random Services (1997–2021). <https://www.randomservices.org/random>
93. Simard, R., L'Ecuyer, P.: Computing the two-sided Kolmogorov-Smirnov distribution. *J. Stat. Softw. Articles* **39**(11), 1–18 (2011)
94. Smirnov, N.V.: On the estimation of the discrepancy between empirical curves of distribution for two independent samples. *Bull. Math. Univ. Moscou* **2**, 2 (1939)
95. Snedecor, G.: Statistical Methods. The Iowa State University Press, Ames, Iowa (1937)
96. Stephens, M.A.: EDF statistics for goodness of fit and some comparisons. *J. Am. Stat. Assoc.* **69**(347), 730–737 (1974)
97. Thomson, J.: Cathode rays. *Philos. Mag.* **44**, 293 (1897)
98. Tremaine, S., Gebhardt, K., Bender, R., Bower, G., Dressler, A., Faber, S.M., Filippenko, A.V., Green, R., Grillmair, C., Ho, L.C., Kormendy, J., Lauer, T.R., Magorrian, J., Pinkney, J., Richstone, D.: The slope of the black hole mass versus velocity dispersion correlation. *Astrophys. J.* **574**(2), 740–753 (2002)
99. Tukey, J.: Bias and confidence in non-quite large samples (abstract). *Ann. Math. Stat.* **29**(2), 614 (1958)
100. Von Eye, A., Schuster, C.: Regression Analysis for Social Sciences. Academic, New York (1998)
101. Wasserstein, R.L., Lazar, N.A.: The ASA's statement on p-values: context, process, and purpose. *Am. Stat.* **70**(2), 129–133 (2016)
102. Wilks, S.: Mathematical Statistics. Princeton University Press, Princeton (1943)
103. Wilks, S.S.: The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math. Stat.* **9**(1), 60–62 (1938)
104. Williamson, E., Walker, A., Bhaskaran, K., et al.: Factors associated with COVID-19-related death using OpenSAFELY. *Nature* **584** (2020)

105. Wu, C.F.J.: Jackknife, bootstrap and other resampling methods in regression analysis. *Ann. Stat.* **14**(4), 1261–1295 (1986)
106. Yates, F.: Contingency tables involving small numbers and the χ^2 test. *Suppl. J. R. Stat. Soc.* **1**(2), 217–235 (1934)
107. Yates, F.: Tests of significance for 2x2 contingency tables. *J. R. Stat. Soc. Series A (Gen.)* **147**(3), 426–449 (1984)
108. Zhou, X.-H., McClish, D.K., Obuchowski, N.A.: Statistical Methods in Diagnostic Medicine, vol. 569. Wiley (2009)

Index

Symbols

- $1/\sqrt{N}$ factor, 66
- D_N statistic, 340
 - approximation, 341
- D_{NM} statistic, 343
 - approximation, 343
- ΔC statistic, 303, 307
- $\Delta\chi^2$ statistic, 239
- χ^2 distribution, 155
 - degrees of freedom, 155
 - hypothesis testing, 158
 - mean, 157
 - moment generating function, 156
 - probability density function, 155
 - reduced χ^2 , 157
 - table of critical values, 443
 - use of bivariate errors, 329
 - variance, 157
- \hat{R} statistic, 400
- π
 - estimate with Monte Carlo integration, 354
 - simulation with hit-or-miss method, 356
- χ^2_{\min} statistic, 216, 233, 333
 - bias with Poisson data, 309
 - for 2×2 contingency tables, 185
- python , 414, 415
 - codes, 420

A

- acceptability of goodness of fit, 315
- acceptable region, 148
- acceptance probability, 388
- acceptance rate, 389
- accessible state, 373
- Akritas, M., 324

- alleles, 14
- analytical solutions, 413
- Anderson, E., 247
- Aristotle, 339
- association in contingency tables, 183
- asymptotic variance, 380
- auxiliary distribution, 387
- average, 133
 - linear, 134
 - logarithmic, 137
 - relative-error weighted, 142
 - weighted, 134

B

- Baker, S., 279
- Barnard, G.A., 190
- batch mean, 382
- Bayes' theorem, 17
 - for predictive values, 198
- Bayes, T., 17
- Bayesian method of inference, 126
- Bayesian method of probability, 6
- BCEs estimators, 324
- Bernoulli variable, 43
- Bernstein, S., 4
- Bershady, M., 324
- beta distribution, 269, 437
 - standard, 269
 - symmetric, 267
- Beta function, 163, 165, 437
 - alternative definition, 438
- biased statistic
 - see unbiased statistic, 358
- bin
 - for Markov chains, 370
- bin of data, 38, 293

- non uniform, 299
 - binary experiment, 43
 - binary Markov chain, 371
 - asymptotic variance, 380
 - Raftery Lewis diagnostic, 403
 - stationary distribution, 377
 - binomial coefficient, 45
 - binomial distribution, 43
 - Ehrenfest chain, 379
 - exact test for contingency tables, 191
 - for bootstrap method, 364
 - mean, 46
 - moments, 46
 - Monte Carlo hit-or-miss method, 355
 - probability mass function, 45
 - variance, 47
 - biometric characteristic experiment by K. Pearson, 38
 - bisector model, 327
 - bivariate data
 - use of χ^2 , 329
 - bivariate errors, 323
 - bootstrap method, 359, 362
 - Bowyer, S., 238
 - Brooks, S., 400
 - Buffon coin drop, 351
 - Bulmer, M.G., 118, 162
 - burn-in period, 395
- C**
- C statistic
 - see* Cash statistic, 278
 - candidate of MCMC, 387
 - Carlin, B., 394
 - Cartesian coordinates, 75
 - transformation to polar, 94
 - Cash statistic, 278
 - analytical expressions for expectation and variance, 282
 - approximation with χ^2 , 279
 - discretization in low-count regime, 308
 - expectation and variance, 281
 - hypothesis testing, 284, 301
 - Cash, W., 279
 - Cauchy distribution, 73, 172, 173
 - Cauchy–Schwartz inequality, 33, 82
 - central limit theorem, 69
 - for stationary stochastic processes, 380
 - limitations, 73
 - central moments, 28
 - ceteris paribus, 339
 - change of variables, method, 74
- chi squared distribution, *see* χ^2 distribution
 - classical method, 6
 - coefficient of determination, 259
 - distribution function, 272
 - coefficient of multiple determination
 - see* coefficient of determination, 259
 - coin toss experiment, 6
 - combinations, 44
 - conditional probability, 8
 - confidence intervals, 51, 113
 - central, 114
 - Gaussian variables, 116
 - model parameters, 238, 244
 - one-sided, 114
 - Poisson data, 307
 - Poisson mean, 121
 - reduced number of parameters, 241
 - several fit parameters, 239
 - significance, 113
 - consistent statistic, 148
 - contingency table, 181
 - χ^2 test, 184
 - 2×2 tables, 182
 - χ^2_{\min} , 185
 - Fisher's exact test, 187
 - higher dimensions, 194
 - margins, 182
 - convergence tests of MCMC, 395
 - convolution of two distribution, 77
 - Conway–Maxwell–Poisson distribution, 278
 - coordinate transformation
 - Cartesian to polar, 75
 - Jacobian, 75
 - correlation, 30
 - counting process, 55
 - Cousins, R.D., 279
 - covariance, 30
 - covariance matrix, 217, 221
 - Covid-19 statistics from OpenSAFELY, 194
 - Covid-19 vaccine efficacy, 205
 - Cox, D.R., 369, 381
 - Cramér's theorem on χ^2_{\min} , 234
 - use in F test, 337
 - Cramér, H., 234
 - cumulative distribution function, 23
- D**
- debiased model variance, 317
 - degree of belief, 6
 - degrees of freedom
 - χ^2 distribution, 155

- sampling distribution of variance, 162
Student's t-distribution, 171
design matrix, 252
detailed balance, 390
deviation, 27
diagnostic test, 195
likelihood ratios, 197
predictive values, 197
sensitivity and specificity, 196
distribution function, 21, 22
properties, 23
full conditional, 393
Duns, John (Scotus), 339
- E**
efficient statistic, 148
Ehrenfest chain, 372
transition probability, 372
stationary distribution, 377, 384
electron discovery experiment by J.J. Thomson, 28
empirical method of probability, 6
Emslie, A.G., 129
entropy, 105
ergodic average, 379
ergodic Markov chain, 376
ergodic theorem, 379
use for MCMC analysis, 407
error function, 428
error matrix, 217, 221, 251
for multi-variable regression, 252
error propagation, 83
error propagation formula, 84
exponential of variable, 88
logarithm of variable, 88
power of variable, 88
product and division, 86
sum of constant, 85
table of common functions, 89
weighted sum of two variables, 85
event, 4
expectation, 24
expectation of random variable, 24
experiment, 3
explained variance, 257
exponential distribution, 23
cumulative distribution function, 23
simulation, 92, 357
- F**
F-distribution, 163
mean, 165
variance, 165
F statistic, 163
comparison of sample variances, 168
distribution function, 163
hypothesis testing, 165
tables of critical values, 444
F test, 333
for nested model component, 336
multi-variable linear regression, 256
degrees of freedom, 334
two independent χ^2 measurements, 334
factorial function
Stirling's approximation, 59
Feller, W., 341
fiducial inference, 114
Gaussian distribution, 118
Student t-distribution, 174
Fisher's theorem on distribution of χ^2_{\min} , 185, 194
Fisher, R.A., 97, 115, 148, 155, 163, 185, 234, 247
fit statistic, 214
fractional errors, 137
frequentist method, 6
full conditional distribution, 393, 394
full-width at half maximum (FWHM), 52
function of random variables
mean, 81
variance, 83
- G**
Galton, F., 214
Gamerman, D., 380, 394
gamma distribution, 155, 394, 436
moment generating function, 157
shape and rate parameters, 436
Gamma function, 128, 165, 437
Legendre duplication formula, 439
asymptotic expansion, 470
Gamma function, incomplete, 128
Gaussian distribution, 47
confidence intervals, 116
cumulative distribution function, 52
mean, 48
moments, 50
probability density function, 48
simulation, 92, 358
upper and lower limits, 118
variance, 48
tables, 427
Gehrels approximation, 123
tables of upper and lower limits, 434

Gehrels, N., 123
 Gelman Rubin statistic, 395, 399
 between-chain variance, 399
 within-chain variance, 399
 Gelman, A., 400
 Geman, D., 393
 Geman, S., 393
 genes, 14
 genotype, 14
 Geweke z-score, 395, 398
 Gibbs distribution, 393
 Gibbs sampler, 393
 Gilks, W.R., 386
 golden standard, 196
 goodness of fit
 χ^2_{\min} statistic, 233
 Cash statistic, 301
 Gaussian data, 216
 Gosset, W.S. (Student), 171
 Greenwood, M., 181

H

half-width at half maximum (HWHM), 52
 Hammersley, J.M., 356
 Handscomb, D.C., 356
 Hastings, W. K., 387
 Heisenberg's uncertainty principle, 22
 Helmert, F.R., 162
 Hubble experiment on galaxy distances, 225
 Hubble's law, 225
 Hubble, E., 225
 hypergeometric distribution, 188
 hypothesis testing, 147
 χ^2 distribution, 158
 and confidence level, 148
 Cash statistic, 301
 F-statistic, 165
 rejection region, 149
 sampling distribution of variance, 162
 Student's t-distribution, 172
 rejection and acceptability, 149

I

iid, independent and identically distributed variables, 36, 66
 immunization statistics for COVID-19 vaccine, 205
 impossible event, 4
 independence chain, 405
 independent assortment principle, 14
 inoculation statistics experiment by M. Greenwood and U. Yule, 181

interesting and uninteresting parameters, 241
 intrinsic scatter, 316
 alternative method using χ^2_{red} , 317
 direct calculation, 316
 parameter estimation, 320
 intrinsic variance
 see intrinsic scatter, 315
 iris data by Fisher and Anderson, 248
 irreducible aperiodic Markov chains, 376
 Isobe, T., 327

J

jackknife method, 359, 360
 pseudo-values, 360
 resampled dataset, 360
 second order, 361
 Jaynes, E.T., 98, 105
 Jeffreys priors, 388
 Jeffreys, H., 18, 127, 388
 Jensen's inequalities, 82
 joint distribution function, 31
 jones, G., 380

K

Knoll, M., 205
 Kolmogorov axioms, 5
 Kolmogorov Smirnov test, 339
 D_N statistic, 340
 D_{NM} statistic, 343
 approximation, 341
 comparison of data with model, 339
 non-parametric nature, 339
 table of critical values, 472
 two-sample test, 342
 Kolmogorov, A., 4, 339

L

Lagrange's multipliers, 107
 Lampton, M., 238
 law of large numbers, 26, 353, 380
 Le Clerc, G.L. *comte de Buffon*, 351
 least-squares method, 216
 Lewis, S., 402, 405
 likelihood, 17, 126, 386
 Gaussian data, 99
 Poisson data, 278
 two-dimensional Gaussian data, 215
 limiting probability, 375
 linear average, 134
 linear correlation coefficient, 266

- distribution function, 267
distribution of r^2 , 272
table of critical values, 471
- linear regression, 216
 multi-variable, 249
 multi-variable with uniform variance, 251
multiple, 219
Poisson data, 292
with factorized linear model, 293
with uniform variance, 222
model sample variance, 237
- log-normal distribution, 138
 standard, 138
- logarithmic average, 137
 weighted, 140
- Lutz, M., 414, 415
- M**
- marginalization of random variables, 31, 34, 38
marginalization of uninteresting parameters, 242
- Margon, B., 238
- Markov chain Monte Carlo, 385
 acceptance probability, 388
 augmentation, 407
 candidates, 387
 convergence tests, 395
 posterior distribution, 386
 prior distribution, 387
 regression analysis, 386
 burn-in period, 395, 398
 correlation of links, 397
 mixing, 397
 posterior distribution, 388
 stopping time, 395
- Markov chains, 369
 accessible states, 373
 binary, 371
 dependence of samples, 386
 ergodicity, 376
 Markovian property, 371
 recurrent and transient states, 372
 short memory property, 371
 state of system, 370
 communication of states, 376
 detailed balance, 390
 irreducible aperiodic, 376
 limiting probability, 375
 periodicity, 376
 recurrent and transient states, 375
- stationary distribution, 375
time reversible, 390
- Markovian property, 371
- Massone, A.M., 129
- maximum entropy method, 105
- maximum likelihood method, 98
 non-linear regression, 228
bivariate data, 324
fit to two-dimensional data, 214
Gaussian data, 99
other distributions, 102
Poisson data, 278
Poisson distribution, 102
- MCMC
 see Markov chain Monte Carlo, 385
- mean, 25
 weighted, 134
Bayesian expectation for Poisson mean, 127
- median, 135
 insensitivity to outliers, 136
 sample, 135
 sample median theorem, 136
- Mendel, G., 11
- method of moments, 103
- Metropolis Hastings algorithm, 387
 acceptance rate, 389
 uniform priors and proposals, 389
- Metropolis, N., 352, 387
- Miller, H.D., 369, 381
- Miller, R., 361
- mixing properties of MCMC, 397
- model fitting, 214
- model sample variance, 237
 for multi-variable regression, 254
- moment generating function, 66
 Gaussian distribution, 68
 uniform distribution, 69
 Poisson distribution, 68
 sum of two uniform variables, 72
 sum of uniform variables, 71
- moments of distribution function, 24
- Monte Carlo, 351, 352
 efficiency, 356
 function evaluation, 355
 traditional integration, 352
 Hit-or-miss, 355
 multi-dimensional integration, 353
 simulation of variables, 357
- multi-variable data, 247
- multi-variable linear regression, 249
 coefficient of determination, 259
 design matrix, 252

error matrix, 251
 F test, 256
 iris data, 252, 256
 t test, 254
 tests for significance, 254
 multiple linear regression, 219
 best-fit parameters, 220
 multiplicative errors, 137

N

Nemmen, R.S., 327
 nested model, 336
 non-detection, 120, 125
 non-informative priors, 127
 non-linear regression, 228
 normal distribution, 69
 see Gaussian distribution, 47
 null hypothesis, 148
 numerical solution, 414

O

Occam's razor, 339
 Occam, William of, 339
 orthonormal transformation, 160
 over-dispersed data, 278
 overbooking probability, 47

P

p-value, 151
 statement of the ASA, 153
 parent distribution, 22
 parent mean
 comparison with sample mean, 169
 parsimony principle, 339
 partition, 4
 partition function, 106
 Pearson, K., 155, 184, 263
 percentile of distribution, 90
 periodicity of Markov chains, 376
 permutations, 44
 plant hybridization experiment by G. Mendel, 11
 point estimate, 115, 133
 Poisson distribution, 52
 Bayesian expectation of mean, 128
 Bayesian upper and lower limits, 129
 likelihood, 278
 mean, 53
 probability mass function, 53
 variance, 54
 Poisson parameter S , 123

Poisson process, 55
 polar coordinates, 75
 transformation to Cartesian, 94
 posterior distribution, 374
 Markov chains, 386
 posterior probability, 18
 Poisson mean, 127
 predictive values, 197
 repeated testing, 202
 Press, W.H., 329
 prevalence of disease, 196
 Price, M., 17
 prior probability, 17, 387
 probability distribution function, 23
 probability mass function, 22
 probability of event, 4
 Venn diagram, 5
 proposal (auxiliary) distribution, 387
 Protassov, R., 338

Q

quantile, 90, 113
 quantile function, 89
 exponential distribution, 90
 uniform distribution, 90
 Quenouille, H.M., 359
 Quenouille–Tukey jackknife
 see jackknife method, 360

R

r distribution, 267
 Raftery Lewis diagnostic, 402, 407
 Raftery, A., 402, 405
 random error, 318
 random variables, 21
 distribution of functions, 74
 functions of, 63
 linear combination, 64
 linear combination of independent measurements, 66
 marginalization, 31
 variance of linear combination, 65
 random walk, 369, 371
 recurrence of states, 374
 transition probability, 371
 recurrent states, 384
 raw data, 38, 83, 236
 Rayleigh distribution, 76
 cumulative distribution, 93
 quantile function, 93
 receiver operating characteristic curve, 204
 reduced χ^2 , 157

- regression, 214
 Gaussian data, 214
 Poisson data, 291
rejection of hypothesis, 149
rejection region, 149
relative uncertainty of random variable, 66
relative-error weighted average, 142
resampling, 359
residual variance, 257, 317
Richardson, S., 386
Roberts, G., 380
Ross, S., 23, 369
Rubin, D., 400
- S**
- sample correlation coefficient, 33
sample covariance, 32
 for binned data, 265
 expectation, 35
sample distribution, 22
sample mean, 25
 comparison of two means, 175
 comparison with parent mean, 169
 distribution function for Gaussian measurements, 71
 weighted, 101
sample median, 135
 variance, 136
sample median theorem, 136
sample space, 4
sample variance, 27
 distribution, 159
 for binned data, 265
 model, 237
 expectation, 35
sampling distribution of mean, 169
sampling distribution of variance
 degrees of freedom, 162
 probability function, 162
Scargle, J.D., 293
second law of thermodynamics, 106
segregation principle, 14
sequence of events, 44
Serfling, R.J., 338
set, 4
Shannon, C.E., 98, 105
Siegrist, K., 23
simulation of number π , 356
simulation of random variables, 91
 exponential, 92
 Gaussian, 92, 357
 Monte Carlo methods, 357
- square of uniform distribution, 92
skewness, 58
Smirnov, N., 339, 343
Snedecor F-distribution
 see F-distribution, 163
Snedecor, G.W., 163
Spiegelhalter, D.J., 386
standard beta distribution
 see beta distribution, 269
standard deviation, 27
standard error, 27
standard Gaussian, 51
stationary distribution, 370
statistical error, 134, 319
statistical independence, 9, 33
statistics, 147
 critical values, 149
 sensitivity, 168
 unbiasedness, consistency and efficiency, 148
Stephens, M.A., 341
Stirling's approximation, 59
stochastic process, 369
stopping time of MCMC, 395
Student
 see Gosset, W.S. (Student), 171
Student's t-distribution, 171
 degrees of freedom, 171
 hypothesis testing, 172
 probability function, 171
 tables, 453
Student's t-statistic, 169, 170
 comparison of two sample means, 175
survival function, 23
systematic error, 134, 319
 additive, 319
 multiplicative, 319
 parameter estimation, 320
- T**
- t-statistic
 see Student's t-statistic, 169, 170
t test for multi-variable linear regression, 254
tea-tasting experiment, 148
Thomson, J.J.
 analysis of experimental data, 347
time reversed chain, 390
total probability theorem, 16
transition kernel, 371
transition probability, 371
triangular distribution, 72, 78
true and false negatives, 196

true and false positives, 196

Tukey, J.W., 359

two-variable dataset, 213

bivariate errors, 323

U

Ulam, S., 352

unbiased estimator, 26

unbiased statistic, 148

uncorrelated variables, 33

under-dispersed data, 278

uniform distribution, 69, 77

probability function, 77

simulation, 92

square, 81

sum of independent variables, 71

sum of two variables, 77

uninteresting variable, 32

upper and lower limits, 114

Bayesian method for Poisson mean, 129

Gaussian variables, 118

Gehrels approximation, 123

Poisson variable, 122, 124

upper limit to non-detection

Poisson, 125

Gaussian, 120

V

vaccine efficacy, 204

van Rossum, G., 414, 415

variance, 27

batch means and spectral estimation, 382

explained, 257

linear combination of variables, 65

residual, 257

weighted sample mean, 102

within and between chain, 399

W

weighted logarithmic average, 140

weighted mean, 134

weighted sample mean, 101

variance, 102

Wilk's theorem on likelihood ratio, 302

Wilks, S.S., 301

Williamson, E.J., 194

Wonodi, C., 205

Wu, C.F.J., 362

Y

Yates continuity correction, 186

Yates, F., 186, 189

Youden's index, 204

Yule, U., 181

Z

z-score, 51

for sample mean test, 169

Geweke test, 398

in χ^2 statistic, 155