Textbooks in Statistical Science

Göran Kauermann · Giacomo De Nicola ·
Cornelius Fritz · David Rügamer

Statistics—Coping with Uncertainty

A Modern Introduction to Statistics in Times of Machine Learning

Springer

Textbooks in Statistical Science

Textbooks in Statistical Science aims to equip learners with both foundational and advanced statistical concepts. The book series welcomes theoretically oriented books that deepen understanding of statistical principles, as well as applied texts that integrate software implementations and real-world case studies. We particularly encourage lecturers to transform their well-tested course materials and lecture notes into textbooks, enabling them to reach a broader audience and share effective teaching approaches developed in the classroom.

Topics of particular interest include—but are not limited to—data collection, machine learning, big data analytics, interpretability, statistical inference, regression and multivariate methods, Bayesian statistics, experimental design, computational statistics, and emerging areas at the interface of statistics, data science, and artificial intelligence.

Göran Kauermann, Giacomo De Nicola, Cornelius Fritz and
David Rügamer

Statistics—Coping with Uncertainty
A Modern Introduction to Statistics in Times of Machine Learning



Göran Kauermann
Department of Statistics, LMU Munich, Munich, Germany

Giacomo De Nicola
Harvard T.H. Chan School of Public Health, Harvard University, Boston, MA, USA

Cornelius Fritz
School for Computer Science and Statistics, Trinity College Dublin, Dublin, Ireland

David Rügamer
Munich Uncertainty Quantification AI Lab, LMU Munich, Munich, Germany

ISSN 3120-6824 e-ISSN 3120-6832
Textbooks in Statistical Science
ISBN 978-3-032-33657-6 e-ISBN 978-3-032-33658-3
https://doi.org/10.1007/978-3-032-33658-3

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2026

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. Unless otherwise expressly agreed by the Publisher in writing, the Publisher

does not authorize use or reproduction of all or any part of this work in any manner for the purpose of training, fine-tuning, evaluating, or developing artificial intelligence models, technologies, or systems. In accordance with Article 4(3) of the EU Directive 2019/790 on copyright and related rights in the Digital Single Market, the Publisher expressly reserves all rights with respect to this work under the text and data mining exception of the directive.

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.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

To Ursula, Niclas, and Linus.
Thanks to all of you!
(G.K.)

Preface

With this book, we aim to provide a new, modern approach to statistics without neglecting its many methodological and mathematical foundations. We choose an information-theoretic perspective and, at numerous points, attempt to draw parallels to methods from machine learning. In this sense, we hope that the book will not only be found helpful by aspiring statisticians but also by readers who see themselves more in the fields of computer science and machine learning.

A common thread throughout the book is uncertainty and how to deal with it. We present statistics as the scientific discipline that provides concepts for handling uncertainty. This naturally includes the quantification of uncertainty, but more generally also an understanding of what uncertainty is and the fact that we are exposed to uncertainty in our everyday lives.

Statistics is traditionally viewed as the science of data analysis. From our perspective, this view is too narrow, especially when considering the tremendous developments in machine learning in recent years. Data are also processed there, leading to complex models and generative AI. Statistics cannot and should not compete here. Some methods in machine learning are often superior to statistical procedures. However, on the one hand, these machine learning methods usually require massive datasets, and on the other hand, the models are often more like black boxes. The latter in particular makes it difficult to understand how and where uncertainty exists. This is where statistics comes into play.

The book therefore follows less of a “how to do data analysis” style and instead focuses on conveying concepts. What characterizes statistics? How is uncertainty quantified in the frequentist sense? How does this differ from Bayesian approaches? What role do simulations play, and why is the likelihood approach so central? These are some of the topics of the book, which, as mentioned, are presented conceptually.

Our aim is to present the versatility of statistical ideas, concepts, and questions in a form that is easy to understand and digest. This does not mean that we are superficial, quite the contrary. Mathematical derivations are part of the material, but they are not the main focus. We hope that our intention will be recognized by the readers.

Finally, we would like to thank a few people. First of all, Eugen Gorich and Vanessa Kleisch, who took care of text and code. We are also grateful to Jan Anders and Sergio Buttazzo for many suggestions and their continuous help getting the book better in shape. We also thank the many students of the course who found one typo or another in working versions of the book. And, in a way, we would also like to thank our statistical teachers, who ignited the fire of statistics in us.

Competing Interests The authors have no competing interests to declare that are relevant to the content of this manuscript.

Göran Kauermann
Giacomo De Nicola
Cornelius Fritz
David Rügamer
Munich, Germany
Boston, MA, USA
Dublin, Ireland
Munich, Germany
April 2026

Contents

1 Introduction

1.1 Statistics and Uncertainty

1.2 Notation and Technical Requirements

2 Uncertainty

2.1 What is Uncertainty

2.2 Sources of Uncertainty

2.3 Classical Statistical Paradigm

2.4 Data-Generating Process

2.5 Statistical Modeling

2.6 Exercises

3 Learning and Estimating

3.1 Kullback-Leibler Divergence

3.2 Squared Loss and Likelihood

3.3 Properties of Estimators

3.4 Estimation and Uncertainty

3.5 Likelihood and Bayes

3.6 Exercises

4 Maximum Likelihood Estimation

4.1 Properties of the Likelihood

4.2 Consistency and Asymptotic Normality

4.3 Cramér-Rao Bound

4.4 Likelihood Ratio

4.5 Parameter Transformation
4.6 Misspecified Models
4.7 Numerical Calculation of the Estimate
4.8 The Likelihood Principle
4.9 Exercises

5 Data-Driven Decisions
5.1 Statistical Evidence and the *p*-Value
5.2 Statistical Hypothesis Tests
5.3 Classical Statistical Parameter Tests
5.4 Power of a Test
5.5 Neyman-Pearson Lemma
5.6 Testing and Confidence Intervals
5.7 Multiple Testing
5.8 Universal Inference
5.9 Sensitivity, Specificity, Accuracy
5.10 Exercises

6 Regression
6.1 Simple Linear Model
6.2 Multiple Linear Regression
6.3 Logistic Regression (Classification)
6.4 Generalized Linear Models
6.5 Exercises

7 Model Selection
7.1 Modeling Dependence

7.2 Training and Test Data
7.3 AIC and BIC
7.4 AIC, BIC, and Hypothesis Testing
7.5 Forward Selection and Small Data Adjustment
7.6 Model Selection and Dimension of the Model
7.7 Exercises

8 Simulating Random Variables and Simulation-Based Inference
8.1 Simulating Uniformly Distributed Random Variables
8.2 Simulating from a Distribution Function and a Density
8.3 Simulating Proportional to a Density (MCMC)
8.4 Simulating from a Multivariate Normal Distribution
8.5 Gibbs Sampling and Markov Fields
8.6 Uncertain Input Parameters
8.7 Bootstrapping
8.8 Permutation
8.9 Exercises

9 Bayesian Inference
9.1 The Bayesian Principle
9.2 Prior and Posterior
9.3 Hyperparameters and Empirical Bayes
9.4 Inference Based on MCMC
9.5 Approximate Bayes Computation (ABC)
9.6 Variational Bayes Reasoning
9.7 Bayes and Regularization

9.8 Exercises

10 Inference in Extreme Data

10.1 Rare Events

10.2 Extreme Data

10.3 Exercises

11 Multivariate Data

11.1 Models for Contingency Tables

11.2 Graphical Models

11.3 Multivariate Normal Distribution

11.4 Exercises

12 Non-i.i.d. Data

12.1 Time Series Models

12.2 Spatial Data

12.3 Network Data

12.4 Functional Data

12.5 Exercises

13 Relating Machine Learning and Statistics

13.1 Primer on Machine Learning and Neural Networks

13.2 Machine Learning from a Statistics Perspective

13.3 Regularization, Penalization, and Bayes

13.4 Uncertainty in Machine Learning

13.5 Exercises

A Background in Probability Theory

A.1 Defining Probability
A.2 Conditional Probability and Bayes' Theorem
A.3 Real Random Variables
A.4 Characterizing the Distribution of a Random Variable
A.5 Expected Values
A.6 Stochastic Convergence
A.7 Probabilistic Inequalities
B Parametric Distribution Families
B.1 Frequently Used Distributions
B.2 The Exponential Family
B.3 Fisher Regularity
References
Index

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
G. Kauermann et al., *Statistics—Coping with Uncertainty*, Textbooks in Statistical Science
https://doi.org/10.1007/978-3-032-33658-3_1

1. Introduction

Göran Kauermann[1] ✉, Giacomo De Nicola[2], Cornelius Fritz[3] and David Rügamer[4]
(1) Department of Statistics, LMU Munich, Munich, Germany
(2) Harvard T.H. Chan School of Public Health, Harvard University, Boston, MA, USA
(3) School for Computer Science and Statistics, Trinity College Dublin, Dublin, Ireland
(4) Munich Uncertainty Quantification AI Lab, LMU Munich, Munich, Germany

1.1 Statistics and Uncertainty

Predicting any future event, whether it is tomorrow's weather, the safety assessment of a technical system, or the outcome of rolling a die, is invariably subject to uncertainty. We cannot generally predict the outcome, at least not with absolute certainty. This does, however, not apply to all scenarios. There are indeed physical laws that state that if "A" is done then "B" happens. Imagine a stone that you are holding in your hand. If you let this stone go, it falls to the ground due to the force of gravity. Natural laws, therefore, apply to deterministic systems. But even the example with the stone can be altered in such a way that making a prediction becomes difficult. Assume that you hold a light feather in your hand and let it fall. In vacuum conditions, the feather falls like the stone to the ground. However, in non-laboratory conditions, there might also be wind, which could affect the feather such that it does not fall but even flies away. Hence, the trajectory of the

feather will be affected by a combination of multiple laws that are only understood up to some precision. This implies that even if physical laws determine the world, additional factors beyond our control influence the setting in many situations where we are faced with uncertainty. In this context, uncertainty indicates that the relationships are too complex to be understood or even described with mathematical or physical models.

In the era of massive data and tremendous numerical power, a novel way of understanding the world has emerged: understanding the world through data. This approach has seen considerable success in recent years, ranging from computer vision, natural language processing, and generative artificial intelligence, utilizing tools like machine learning and deep learning. With massive databases and computational power, one might be tempted to declare that uncertainty diminishes due to increasing data volume. Hence, one might be inclined to put the topic of uncertainty in the background in times of massive data and complex machine learning algorithms. This is, however, a fallacy given that quantifying uncertainty remains as relevant as ever. In the foreseeable future, we will not be able to understand all complex structures in the world, leaving us with uncertainty.

Uncertainty, on the one hand, and ignorance or lack of knowledge, on the other hand, are closely tied together. This point demonstrates that uncertainty has an individual component. Experts know more than laypeople. Hence, the question arises of how to quantify knowledge or, conversely, how to quantify missing knowledge. While answering this question is certainly more in the realm of philosophy, we aim to get more concrete in this book by relating the question of missing knowledge to data. Put differently, we aim to quantify the knowledge—or, equivalently, the reduction of uncertainty—that results from data. Thus, the main objective is not to assume that data can explain the world if we just had enough of it. To the contrary, we should accept that data is inherently insufficient and cannot explain but provide information on the world, even if the amount of data is small.

This book focuses on the quantification of uncertainty, and we claim that uncertainty and its quantification lie within the realm of statistics and statistical reasoning. In fact, we combine coping with uncertainty and statistics in this book. Relying on probability theory and its calculus, we will demonstrate that this is a practical approach. We will introduce the classical concepts of statistics, namely, maximum likelihood as well as Bayesian reasoning, but unlike classical statistics textbooks, we continuously link the ideas to machine learning concepts. Our focus will always be on coping with uncertainty through the rich toolbox that one and a half centuries of statistical developments offer.

We bring this introduction to an end by providing a formal definition of statistics. It was Sir David Cox, certainly one of the most influential personalities in statistics, who defined statistics as "the discipline concerned with the study of variability, with the study of uncertainty and with the study of decision-making in the face of uncertainty." (see Lindsay et al., 2004). Cox continues: "Whereas these are issues that are crucial throughout the sciences and engineering, statistics is inherently an interdisciplinary science." We believe statistics cannot be described more succinctly. Regarding the mentioned interdisciplinarity, in this book, we want to emphasize links to the field of machine learning.

Companion Materials Datasets required for selected exercises and additional materials are available on the companion webpage: https://coping-with-uncertainty.github.io/.

1.2 Notation and Technical Requirements

1.2.1 Random Variables

Before we get into the material, we need to clarify our notation in this book. Uncertainty is closely related to probability theory; therefore, we expect some basic knowledge in this field. In Appendices A and B, we list the main requirements, such as frequently used distribution models and central results from the field of probability. Inexperienced readers

are advised to familiarize themselves with this material first. We emphasize that this book aims to remain on a technical low level, but still, one or another, notation or result is essential. Thus, we detail in this chapter the core results, allowing the reader to get familiar with the notation.

We define with $F(\cdot)$ the **distribution function** or **cumulative distribution function (cdf)** of a random variable Y, which is assumed to take values in the real numbers $\mathbb{R}$,

$$P(Y \leq y) = F(y),$$

for $y \in \mathbb{R}$. We will call this function "distribution function" throughout the book. Generally, we abide by the convention to write random variables in capital letters, while lowercase letters are outcomes of the random variable. A continuous random variable can take any value, and the stochastic behavior of a continuous random variable is described by its **density function** or **probability density function (pdf)**. We use the terminology "density function" in this book. The density function $f(\cdot)$ is a nonnegative function, such that

$$P(Y \in [a, b]) = \int_a^b f(y) \mathrm{d}y.$$

The density results from $\partial F(y) \ / \ \partial y = f(y)$. If the random variable Y is discrete-valued, then $F(\cdot)$ is a step function. It can be burdensome to always distinguish between discrete-valued and continuous random variables, so we use a general notation throughout this book. We make this clear by the example of the mean value of a random variable. If Y is continuous, we have

$$E(Y) = \int_{-\infty}^{\infty} y f(y) \mathrm{d}y.$$

If Y is discrete-valued with outcomes $-\infty < a_1 < a_2 < \ldots < a_K < \infty$, the expected value is given by

$$E(Y) = \sum_{k=1}^{K} a_k \ P(Y = a_k)$$

where $P(Y = k) = F(a_k) - F(a_{k-1})$. To avoid such a distinction, we write the expectation in a general form using the integral notation

$$E(Y) = \int_{-\infty}^{\infty} y \, \mathrm{d} \, F(y).$$

We typically abbreviate the expectation with the parameter $\mu = E(Y)$ and the variance with the parameter $\sigma^2 = E(\{Y - E(Y)\}^2)$ and work with standard distributions such as the normal distribution, χ^2 distribution, binomial distribution, or Poisson distribution, to name but a few. We give details on these distributions in Appendix B.

1.2.2 Asymptotic Results

Numerous central derivations are based on a few asymptotic statements from probability theory, which we want to list here. These are the central limit theorem, the law of large numbers, and the Glivenko-Cantelli theorem, which we list below. As the name suggests, the most central result is the central limit theorem. It states that the distribution of the sum of any independent and identically distributed random variables, which is a setup typically called i.i.d. setting that we define in detail later, converges to a normal distribution.

Theorem (Central Limit Theorem) *Let $Y_1, Y_2, Y_3, \ldots$ be independent and identically distributed random variables, with mean zero and variance σ^2. The distribution of $Z_n = \sum_{i=1}^{n} Y_i \, / \, \sqrt{n}$ converges in distribution to a normal distribution with mean zero and variance σ^2, which is denoted by*

$$Z_n \xrightarrow{d} N\big(0, \sigma^2\big).$$

The second important convergence statement refers to the empirical distribution function. Given data $(y_1, y_2, \ldots, y_n) \in \mathbb{R}^n$, the empirical distribution at $y \in \mathbb{R}$ is

$$F_n(y) = \tfrac{1}{n} \sum_{i=1}^{n} 1_{\{Y_i \leq y\}}.$$

According to the Glivenko-Cantelli theorem, this distribution converges with increasing *n* to the true distribution function.

Theorem (Glivenko-Cantelli Theorem) *Let* $Y_1, Y_2, Y_3, \ldots$ *be independent and identically distributed random variables drawn from the (possibly unknown) distribution function* $F(\cdot)$*. Then,*

$$\lim_{n\to\infty} F_n(y) = F(y) \quad \text{for all } y \in \mathbb{R}. \tag{1.1}$$

The final asymptotic statement highlighted here is the law of large numbers, which states that the arithmetic mean of random variables converges to the true mean value.

Theorem (Law of Large Numbers) *Let* $Y_1, Y_2, Y_3, \ldots$ *be independent and identically distributed random variables drawn from some distribution function with mean* μ*. Then, for any small but positive value* δ*, we get for* $Y = \sum_{i=1}^{n} Y_i \Big/ n$*:*

$$\lim_{n\to\infty} P(|Y - \mu| > \delta) = 0 \tag{1.2}$$

Proposition 1.1 (Weak Law of Large Numbers) *Let* $X_1, \ldots, X_n$ *be independent and identically distributed random variables with* $E(X_i) = \mu$ *and* $\mathrm{Var}(X_i) = \sigma^2 < \infty$ *for* $i = 1, \ldots, n$*. Then,*

$$\overline{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{P} \mu.$$

In other words, the sample mean converges in probability to the expected value μ*.*

Note that we here only stated the weak law of large numbers, while a strong law of large numbers also exists, which is, however, less central to the arguments in this book. We will use these three statements throughout the book on various occasions. Readers unfamiliar with these asymptotic results are encouraged to consult Dekking et al. (2006) to gain a better understanding of them.

References

Dekking, F. M., Kraaikamp, C., Lopuhaä, H. P., & Meester, L. E. (2006). *A modern introduction to probability and statistics: Understanding why and how*. Springer.

Lindsay, B. G., Kettenring, J., & Siegmund, D. O. (2004). A report on the future of statistics. [https://doi.org/10.1214/088342304000000404]

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
G. Kauermann et al., *Statistics—Coping with Uncertainty*, Textbooks in Statistical Science
https://doi.org/10.1007/978-3-032-33658-3_2

2. Uncertainty

Göran Kauermann[1] ✉, Giacomo De Nicola[2], Cornelius Fritz[3] and David Rügamer[4]
(1) Department of Statistics, LMU Munich, Munich, Germany
(2) Harvard T.H. Chan School of Public Health, Harvard University, Boston, MA, USA
(3) School for Computer Science and Statistics, Trinity College Dublin, Dublin, Ireland
(4) Munich Uncertainty Quantification AI Lab, LMU Munich, Munich, Germany

This chapter introduces the idea of taking a statistical view on data. We will start with defining different sources of uncertainty, which we label as *aleatoric* and *epistemic uncertainty*. A disentanglement of the two is not always possible nor advisable, but the consideration of uncertainty is. We will introduce what we call the "data-generating process," which is and remains unknown to us. In any data analysis, our primary goal is to approximate this unknown data-generating process using suitable "statistical models." We will refer to the classical statistical paradigm, building on drawing a sample from a population to infer something about a quantity of interest.

2.1 What is Uncertainty

Before we can cope with uncertainty, we need to define it, which is no easy endeavor. To motivate, let's start with a simple thought experiment showing that uncertainty can be modeled mathematically with

probability distributions. Probability is typically expressed with some kind of random experiment. Imagine a dice cup with a (fair) die. If we shake the dice cup and then place it on the table, the probability that the die will fall with the number 6 upward should be 1/6. The throw of the die is called a random experiment, since the outcome is random. In this case, a probability model (such as a binomial, multinomial, or geometric distribution introduced in Appendix B) is used to calculate the probability of certain events, for example, one can calculate how likely it is to roll a double, i.e., two identical numbers in two subsequent rolls of the die or, equivalently, two equal numbers with two dice in the dice cup. In fact, what sounds well understood and simple today has been a major part of probability theory, originating with the letter exchange by Fermat and Pascal (see Devlin, 2010).

Recall the simple experiment of rolling a die in the dice cup, but modify it as follows: we shake the dice cup and put it on the table without lifting it. The thrown die is therefore invisible to us under the cup. Again, we ask ourselves what the probability is that the die shows the number 6. This is now no longer a random experiment because the die has been thrown and it either shows a 6 or not. But yet, as long as the cup is not lifted, our state of knowledge has not changed. In other words, both before the die is rolled (i.e., before the random experiment) and after the die is rolled with the cup covered (i.e., after the random experiment), our knowledge about the outcome is the same. This suggests quantifying the probability that the die under the cup—invisible to us—has the number 6 upward is 1/6, although this is not the probability of a random experiment. This thought experiment reveals two important aspects: First, uncertainty can often be represented in the form of probabilities. Second, a concrete random experiment is not always necessary to quantify uncertainty. We will utilize probability theory to quantify uncertainty even when no formal random experiment is carried out. Generally, we will make use of probability models, which will allow us to quantify knowledge and, conversely, the remaining uncertainty. It can even be argued, as Spiegelhalter (2024) suggests, that using probability to express our

ignorance and uncertainty always represents an “individual” probability rather than a universal one. In any case, probability theory allows for the quantification of uncertainty.

The idea of representing uncertainty through probability distributions is also reflected in Bayesian statistics, which will be treated in Sect. 3.5 and Chap. 9. But, for now, we do not conceptually distinguish between the Bayesian and what is later called the frequentist’s point of view.

We may also look at the die experiment from a different angle and question whether rolling a die is, in fact, a random experiment at all. If we shake the dice cup, this is a physical process. If we repeat shaking the cup at the same angle, velocity, acceleration, etc., the result would be the same, assuming the outcome is determined solely by the physical input parameters. This means that if we knew the physical relations and had access to all physical quantities, we could determine the side the die lands on. Suppose we could equip the dice cup with hundreds of sensors that record its motion. Additionally, we could install video cameras to record the person shaking the cup. We would then ask the person to repeat shaking the dice cup thousands of times while recording the sensor signals and taking the videos. By protocoling the results (as well as the initial position of the dice before shaking the dice cup), we could train a machine learning model with the data. This now becomes a classical supervised learning framework, with sensor data and video sequences as input and the number of points on the die as outcome. Potentially, we would be able to reduce randomness and uncertainty and would be able to predict the side the die lands on with reduced (or even no) error. This is, however, a very complex setup. But we could build a robot that randomly shakes the dice cup and train a machine learner on the hundred thousand collected dice throws. By doing so, we replace the random experiment with a prediction problem of the number of points on the die based on all collected data. In this case, the random experiment of rolling a die is no longer random but becomes deterministic, which raises an even more philosophical question: Is there randomness at all, or is everything (pre)determined?

Such a view leads us to determinism, which, however, does not directly address the question of quantifying uncertainty. Put differently, whether the world is deterministic or stochastic, our knowledge of what is happening in the world or in the near future is limited, and ignorance remains inevitable. That, in turn, means we face uncertainty, which should be quantified. With respect to the dice cup, it means that unless we can record and process all physical quantities relevant to the side the dice falls on, we are in the circumstance of ignorance, which we will quantify with the help of probability theory.

We can motivate the idea of uncertainty also from a different angle by looking at data. Infinite data do not generate infinite information but sometimes the opposite. Consider the example where we take an image of a horse with a very modern camera that can capture images at arbitrary resolutions. We could look at a coarse version of the image, which is low resolution, and see the horse. We could zoom in and see every single hair of the horse, but the more we zoom in, the more we lose the possibility of detecting the horse, since we focus too much on small things. Therefore, we prefer an image resolution that allows for uncertainties while still conveying the information we want to see.

Quantifying uncertainty is, in practice, much more complicated, and simply giving numbers that express uncertainty does not make them helpful or valid. We have all experienced situations where the weather forecast predicted rain with $100\,\%$ probability, but in fact no rain fell. While a dry day rather than a rainy one is tolerable for most people, this shows that even rainfall probabilities are approximate and not a deterministic forecast. In Fig. 2.1, we show a different example demonstrating that the quantification of uncertainties can be very wrong. The idea traces back to Goodfellow et al. (2015) and demonstrates the so-called adversarial effects or attacks. In machine learning with complex models, properly quantifying uncertainties can be quite complicated. Still, based on the architecture of a neural net, one can re-express the weights of the last layer as probabilities. In the jargon of machine learning, this is denoted as *softmax*, which is then interpreted as a multinomial distribution or, putting it differently,

which is used to express certainty (or vice versa uncertainty). This phenomenon is illustrated in Fig. 2.1 where animal pictures are classified with respect to the animal shown in the picture. A cat is detected, not with a high probability but with a certainty of about $57\,\%$. If now "adversarial noise" is superimposed on the image, that is, noise on the image so that classification of the image goes toward a different animal without visually changing the picture, we obtain the right-hand side picture. The latter is the sum of the first two pictures, which the employed algorithm classifies as a tarantula with a probability of about $99\,\%$. This is an illustration of a pitfall in complex neural networks. We include the example here to motivate the point that "valid" uncertainties are difficult to calculate and quantifying and coping with uncertainty are one of the principal challenges in real data-driven reasoning, such as machine learning. We also refer to Hüllermeier and Waegeman (2021) for a general discussion of uncertainty in machine learning. Many of the ideas proposed there are based on concepts derived from this book.

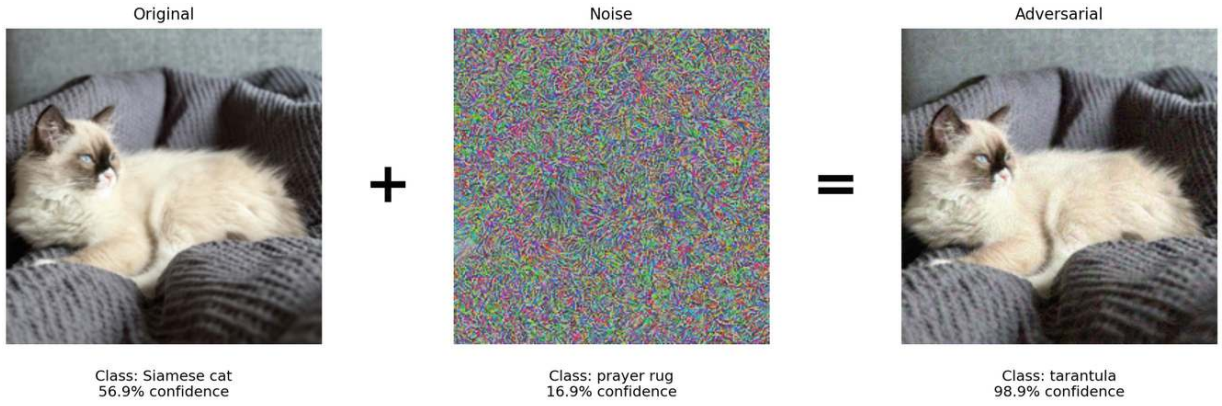


Fig. 2.1 Example for adversarial effects. Adding some "adversarial" noise to the data can lead to wrong predictions

2.2 Sources of Uncertainty

As shown above, uncertainties are multifaceted. In machine learning, uncertainties are typically classified as *aleatoric* and *epistemic* uncertainty. A clear distinction between the two is difficult, but we

want to spend some effort on disentangling them. The terms “aleatory” and “epistemology” were already used in 1975 to describe different historical perspectives on the concept of probability. These terms then developed into the nowadays standard terms “aleatoric” and “epistemic.” Following Hacking (1975), the aleatoric view considers probability as the stable frequency of an event, while the epistemic notion of probability relates to the degree of belief. In more recent years, the terms are used to describe uncertainty in modeling (Hora, 1996; Kiureghian & Ditlevsen, 2009), with the generally established understanding that the total uncertainty decomposes into aleatoric and epistemic uncertainty not necessarily in a mathematical additive way. Aleatoric uncertaintyis thereby defined as uncertainty arising from the inherent randomness of an event. In contrast, epistemic uncertainty expresses uncertainty due to a lack of knowledge, which in principle can be reduced by increasing the amount of data or by conducting an excessive number of experiments. We again refer to Hüllermeier and Waegeman (2021) or Gruber et al. (2026) for a deeper discussion.

Conceptually, the decomposition into aleatoric and epistemic uncertainty is still not unique, and we therefore propose to disentangle the two types of uncertainties as follows: Assume we have some input variable x and some output variable y. This can be a classical regression setup, where, in statistics, x are called covariates or independent variables, while y is the response variable or dependent variable. Nevertheless, this is the classical setup in supervised machine learning, where a neural network is trained to predict y from input data x. The input data itself can be complex and/or unstructured, such as images or text. Aleatoric uncertainty is then uncertainty originating from a stochastic/nondeterministic relationship between the input variables x and the output y. This can be written by a probability model for y, conditional on x, that is

$$Y \sim G(\cdot \mid x) \tag{2.1}$$

where $G(\cdot \mid x)$ is an arbitrary distribution function which depends on x. The dependence of (2.1) on x can be of any type, and we do not impose any constraints or restrictions at this point. We will later call

(2.1) the data-generating process. As long as the conditional distribution (2.1) is not degenerated, we cannot predict *y* from *x* without error. Nondegeneracy means in this context that the distribution is not concentrated at a single value y_0. Assuming that the variance exists, aleatoric uncertainty refers to a strictly positive variance $\mathrm{Var}(Y \mid x)$. The probability model in Eq. (2.1) serves as a definition for aleatoric uncertainty.

With this definition of aleatoric uncertainty, we can then also define epistemic uncertainty as all remaining uncertainty that is not aleatoric. In practice, this is uncertainty due to the fact that we typically do not know $G(\cdot \mid x)$ in (2.1) and replace or rather approximate it with some numerical model, be it a regression model or a neural network. Epistemic uncertainty also refers to the phenomenon that we have only a limited amount of data. Putting it differently, if an infinite amount of data were available, our prediction models could get arbitrarily complex to approximate any function $G(\cdot \mid x)$ up to negligible error and uncertainty due to finite data vanishes, meaning that epistemic uncertainty decreases to zero. But aleatoric uncertainty remains since there is always a probabilistic connection between *x* and *y*.

It is important to note that the situation is even more complex due to what is known as estimation uncertainty. This refers to the fact that any model we train or estimate using data will inherently carry uncertainty because of the finite nature of the data. While we will discuss this topic in greater detail later in the book, we refer readers to Li and Meng (2021) for a broader discussion of this more philosophical issue. As of the time of writing, understanding and quantification of uncertainty, as well as its integration into modern machine learning tools, remain an active area of research with significant advancements, as reviewed by Gawlikowski et al. (2023).

Let's demonstrate the two sources of uncertainty in the dice cup example from above. If no additional variables are gathered, it is not possible to further reduce the uncertainty concerning the rolling of a die, and all uncertainty is of aleatoric nature. If, however, all relevant physical quantities are measured (assuming they would be

measurable), the process of rolling a die becomes deterministic, and no aleatoric uncertainty is present. A clear distinction between aleatoric and epistemic uncertainty depends, therefore, on the available data. If the dice cup is not equipped with hundreds of sensors, aleatoric uncertainty remains. This also means that aleatoric uncertainty is determined by the state of general knowledge and technical developments available, for instance, while rain probabilities some years ago were quite unreliable, they are much more accurate today with increased knowledge of meteorology, extended numerical power for simulation models, and, just recently, with the use of machine learning (see, for example, Bochenek & Ustrnul, 2022). For the dice cup, this holds as well. If we are at some future point in time able to record all movements of the dice cup, then the set of input variables x will be excessive, and aleatoric uncertainty will be small, if not zero. Meaning that with sufficient data, we can predict the outcome. But, for now, dice cups are simple nontechnical objects, and shaking them is a nondeterministic event, so we are faced with aleatoric uncertainty, and collecting data with a standard dice cup will not reduce uncertainty, meaning that epistemic uncertainty can be ignored at this point.

Finally, the terms aleatoric and epistemic uncertainty are not yet familiar in the statistical literature, even though they translate quite easily into statistical jargon. Aleatoric uncertainty arises due to some true underlying data-generating process, being the nature that determines y from x with additional randomness superimposed. If y deterministically follows from x and we can observe all x, then there is no uncertainty, which is, however, not the case in most real-life applications. Aleatoric uncertainty is irreducible and follows the probability model (2.1). Epistemic uncertainty, on the other hand, translates to model uncertainty and model misspecification, as well as estimation uncertainty using the statistical jargon, all of which are covered in Chap. 3.

In some machine learning applications, aleatoric uncertainty is of secondary interest. Consider, for instance, classifying images of cars and airplanes into one of the two categories. It is reasonable to assume

that this can be done unambiguously, at least by humans, except for blurred images. Hence, given an image x, there is no or at least only an ignorable amount of uncertainty, whether the object shown in the image is a car or an airplane. However, such clear discrimination does not apply to all image classification tasks, for instance, the analysis and interpretation of medical images are more likely to involve uncertainty. Consider an X-ray or ultrasound image shown to different physicians. Different experts might reach different conclusions about what is seen. In such cases, statistical concepts help us cope with uncertainty.

2.3 Classical Statistical Paradigm

Traditional inferential statistics aims to make a statement about a population based on data typically drawn from a sample; see Fig. 2.2 for a sketch of this setup. We have a population of individuals for which we want information about a quantity. For simplicity, age is the quantity of interest. We are interested in the average age of the population and, therefore, draw a sample. We obtain this sample by randomly selecting individuals from the population, typically without replacement, so that each individual is drawn only once. Based on the information from the sample, we calculate the quantity of interest, here the average age of the sampled individuals. The average age in the sample is not equal to the average age in the population, unless the population is completely homogeneous or unless we have included each individual in the sample. The latter case is also called a census, which can be regarded as a special case of a random sample. Typically, however, uncertainty from the sampling process on the average age of the population is inevitable. Therefore, we need to assess uncertainty next. Note, however, that the sample also yields information on the average age and the heterogeneity of sizes in the population, as measured by the variance of the recorded ages in the sample. Based on these two quantities, we can calculate a so-called confidence interval, which we will cover in Sect. 3.4. For now, consider the confidence interval as the range of values within which the true but unknown average age of the population lies

with a given degree of certainty, based on repeatedly resampling the data. Aleatoric uncertainty in this example arises from heterogeneity in the population's age distribution. However, because we draw without replacement, we can, in fact, reduce the remaining uncertainty to zero by including all individuals in the sample. For the last point, a central requirement is that the population is finite.

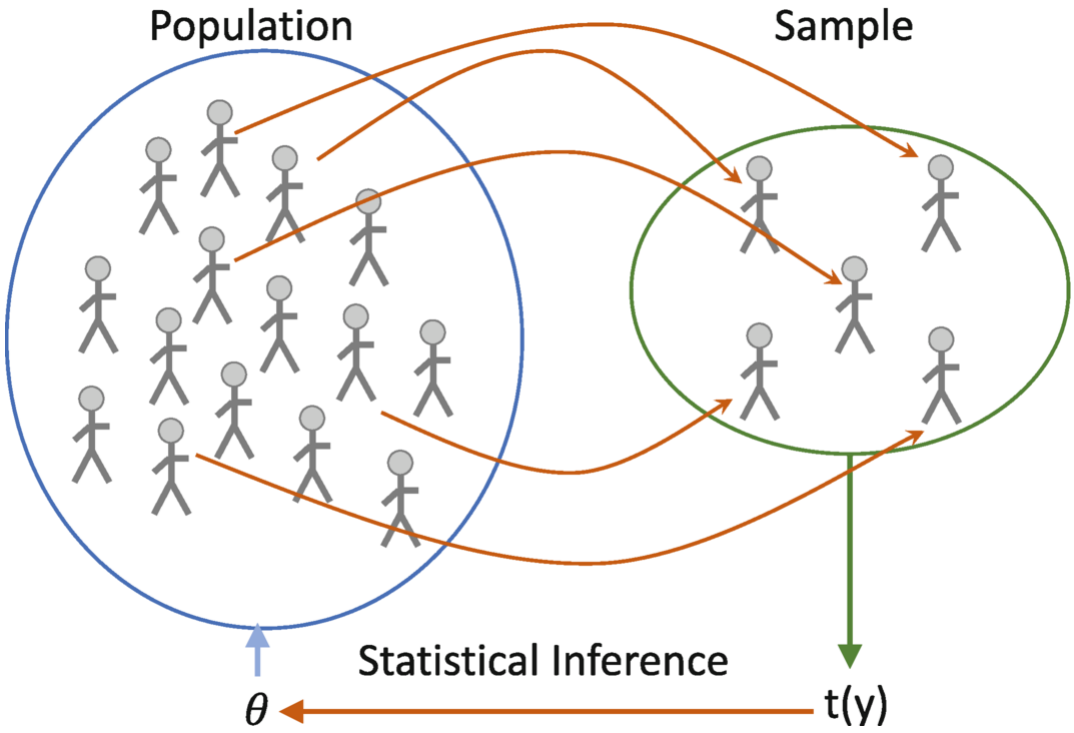


Fig. 2.2 Classical statistical paradigm. A sample is drawn from a population. The sample statistics allow us to draw inference about a population parameter. The quantity $t(y)$ is calculated from the sample and denoted as a statistic

Sampling and survey design are central areas of statistics, and we refer to Thompson (2010) for more details. Our discussions in this book do not focus on survey methods, which encompass methods to take more informative or less intrusive samples. In other words, we aim to deal with situations where data do not trace back to the classical survey sample in Fig. 2.2. Still, we can utilize the main ideas derived but need to extend these to infinite populations or, more generally, to general data-generating processes such as (2.1), which will be appropriately defined in the subsequent chapter.

2.4 Data-Generating Process

We now extend the finite population setup to a general setting. For this, we assume that some random process generates the data. For simplicity, we ignore the input variables x and consider only Y. We assume Y to be drawn from some unknown distribution given by

$$Y \sim G(\cdot), \tag{2.2}$$

where $G(\cdot)$ is an unknown distribution function. From now on, we call (2.2) the **data-generating process**, i.e., a natural law. Ignoring the input variables x for the moment, there are apparent similarities between (2.2) and the definition of aleatoric uncertainty in (2.1). In fact, aleatoric uncertainty is the irreducible uncertainty of Y, which can best be comprehended as a random distribution given in (2.2). This setup will be central throughout the book, and we want to highlight from the start that (2.2) has a very general structure. We assume neither that $G(\cdot)$ is known nor that it can be learned with infinite data, just that it exists.

The statistical paradigm shown in Fig. 2.2 is extended as follows: first, we assume that the distribution $G(\cdot)$ is unknown. We are, however, able to obtain random draws from $G(\cdot)$ which yield the observed data. Following the conventions of probability theory, we denote the data with lowercase letters $y = (y_1, \ldots y_n)$ and assume that these are *n***independent and identically distributed** (in short i.i.d.) according to $G(\cdot)$, i.e.,

$$Y_i \sim G(\cdot) \; i.i.d. \text{ for } i = 1, \ldots, n. \tag{2.3}$$

We also call this that Y_i are independently "drawn" from $G(\cdot)$. Based on the data $(y_1, \ldots, y_n)$, we intend to draw inferences about general properties of the unknown distribution $G(\cdot)$. These could be quantities like the mean or the variance, which are given by

$$\mu = \int y \mathrm{d}\, G(y) \quad \text{or} \quad \sigma^2 = \int (y - \mu)^2 \mathrm{d}\, G(y),$$

but it can also be the entire distribution. From the Glivenko-Cantelli theorem (see Sect. 1.2.2), we know that in setting (2.3) the empirical

distribution function of y_i converges to the true distribution function $G(\cdot)$ for increasing sample size n. We also know that for any sufficiently regular function $t(y)$ we obtain with the law of large numbers (see again Sect. 1.2.2) that

$$\frac{1}{n}\sum_{i=1}^{n} t(Y_i) \rightarrow \int t(y)\mathrm{d}\, G(y)$$

with increasing data size n, regardless of whether $G(\cdot)$ is known or not.

Note that the data-generating process (2.2) is simplified by omitting the input variables x, even though our primary focus is on the more general setup (2.1) with input variables included. It is, however, notationally simpler to omit the input variables x for now, which allows us to introduce the main ideas of statistical inference. After that, we will extend setup (2.2) to the more general case (2.1), where variables x are included as well.

2.5 Statistical Modeling

Next, we introduce concepts to approximate $G(\cdot)$ through a parametric distribution function $F(\cdot;\theta)$, which is fully characterized by parameter vector $\theta = (\theta_1, \ldots, \theta_p)^\top$, for instance, we may assume a continuous distribution for $G(\cdot)$ that we approximate through a normal distribution, with mean μ and variance σ^2. In this case, the parameters are given by $\theta = \left(\mu, \sigma^2\right)^\top$. Generally, we define a *statistical model* as a set of distribution functions,

$$\mathscr{F} = \{F(\cdot;\theta), \theta \in \Theta\}, \tag{2.4}$$

where Θ is the set of admissible values for θ. We assume (or pretend) that the data are generated by an element of the set $\mathscr{F}$ corresponding to θ_0, i.e., $F(\cdot;\theta_0)$. Thus, the modeled data-generating process is

$$Y \sim F(\cdot;\theta_0), \tag{2.5}$$

where θ_0 can be seen as the "TRUE" parameter (see Chap. 3 for a more in-depth discussion).

In machine learning jargon, the choice of $\mathscr{F}$ relates to choosing a loss function, as will be discussed in Sect. 13.2. The probability model,

in fact, determines the assumed aleatoric uncertainty. We will revisit this point later, but it is important to emphasize now that the choice of $\mathscr{F}$ is both critical and inherently hypothetical. Throughout the book, we will develop tools to evaluate how well the statistical model approximates the unknown data-generating process $G(\cdot)$.

Example 2.1 Assume we have data $y = (y_1, \ldots, y_n)$, each drawn from a normal distribution with mean μ and variance σ^2. This results in the following probability model,

$$\mathscr{F} = \left\{F(.; (\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}^+\right\},$$

where $F(\cdot; \mu, \sigma^2)$ is the distribution function of a normal distribution with the following density:

$$f(y, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\ \sigma^2}}\ \exp\left(-\tfrac{1}{2}\left(\tfrac{y-\mu}{\sigma}\right)^2\right).$$

Through the shape of the normal distribution, this model relates to square losses, as will become apparent in Chap. 3. Hence, machine learning based on minimizing the squared loss is equivalent to choosing the set $\mathscr{F}$ to consist of normal distributions. ▷

Having defined the class of hypothetical data-generating processes, we want to choose the parameter θ in an optimal, principled way. Bear in mind that $G(.)$ is (and may remain) unknown. The aim is to set the parameter θ data-driven to some value $\hat{\theta}$ with the help of the observed data y. Thus, $\hat{\theta}$ can be perceived as a function of the data $y_1, \ldots y_n$:

$$\hat{\theta} = t(y_1, \ldots, y_n). \tag{2.6}$$

We call $t : \mathbb{R}^n \to \Theta$ a **statistic** which is a function of the data. The statistic in (2.6) is also called **estimator**, and we will discuss properties of estimators in the next chapter. The function $t(\cdot)$ is only in simple cases analytically available, and in fact in most cases $\hat{\theta}$ is the result of an estimation algorithm $t(\cdot)$, in machine learning typically called *learning algorithm*. It is, therefore, more plausible to consider $t(\cdot)$ not as a

function but as an algorithm applied to data $y_1, \ldots y_n$ generating an estimator $\hat{\theta}$.

We want to note here the verbal distinctions between **estimation**, **estimator**, and **estimate**. *Estimation* is the general term for generating information from data to fit or calibrate a model. An *estimator*, as already mentioned, is the computational rule or algorithm that is applied to the data. The result of this algorithm is then a concrete outcome, one or more specific values, which we call the *estimate*. A sharp distinction between the estimator (the algorithm) and the estimate (the result) is not always helpful or feasible. However, as the book progresses, it will quickly become clear what we are referring to.

We also want to highlight that classical statistics textbooks start their development from a more restricted model by replacing the data-generating process (2.3) through

$$Y_i \sim F(\cdot; \theta_0) \; i.i.d. \text{ for } i = 1, \ldots, n. \tag{2.7}$$

In other words, one assumes that the true data-generating process comes from $\mathscr{F}$. If the model (2.7) holds, we hope that $\hat{\theta}$ is close to θ_0, the true parameter. Ideally, if the data size n increases, we want $\hat{\theta} \to \theta_0$. We will call this consistency, which is a reasonable requirement for data analytic procedures. The open question is, though, how to construct or find the algorithm $t(\cdot)$ such that consistency is guaranteed. We propose to find $\hat{\theta}$ by optimizing some target function. In statistics, the likelihood function, which is the density of the data according to (2.5) but viewed as a function of θ, not the data, is often used as a target function and thus is of central importance. We will introduce the likelihood function and will discuss the properties of its maximum in depth in Chap. 4.

However, the true data-generating process (2.3) and its approximation (2.7) may differ. If they coincide, we have $G(\cdot) \in \mathscr{F}$, and in this case, we can state $G(\cdot)$ as

$$G(\cdot) = F(\cdot, \theta_0). \tag{2.8}$$

As mentioned above, this setup underlies most classical statistics textbooks. We remain more general here and emphasize that $\mathscr{F}$ is just a model class approximating reality. To see this, assume we want to

analyze the height of male humans in a particular area. Assuming that the height of a person is normally distributed, we estimate the mean μ and variance σ^2 based on some recorded heights forming our data. Though it might be reasonable to assume that the heights of people are normally distributed, the natural law behind the distribution of human sizes is unknown. Still, the normal distribution might serve as a good proxy, since the hypothesis that height is normally distributed is a reasonable starting point for the analysis. In this matter, we can refer to the general quote: "All models are wrong, but some are useful" (Box, 1976). However, if (2.8) holds, we can derive mathematical properties for the estimators. Importantly, some of these properties extend to scenarios where not (2.8) but the general setup (2.3) holds.

In summary, we here envision statistical modeling as selecting an appropriate distributional model to account for aleatoric uncertainty, analogous to choosing an appropriate loss function in machine learning. A key question is whether the chosen model is suitable, i.e., whether the approximation (2.8) for (2.3) is justifiable. While this is an important consideration, it is not a primary focus of this book.

2.6 Exercises

Exercise 1 In image-based object classification, the visual appearance of objects determines which class they are assigned to. You are tasked with building a classifier that, given an image of a car, is able to classify it into one of three categories (A, "small"; B, "medium"; C, "large." What kind of data would be needed to train the model? Which potential sources of uncertainty are at play?

Exercise 2 A scientist decides to feed a seagull every time it comes to his apartment for a period of 100 days and counts how many times per day the seagull comes. The results are contained in the dataset `seagull.csv` on the online supplementary material.

(a) After loading the dataset in R, compute mean, variance, minimum, and maximum, and adequately plot the data. Among the

distributions you know, which one do you think would fit well here and why?

(b) Fit the distribution you picked at point (a) to the data. Overlay the plot of the fitted distribution to the plot of the raw data you produced at point (a).

(c) Do you think your model is a reasonable choice for this application? Does it produce a good fit? How could you quantify how good your model is for the data at hand?

References

Bochenek, B., & Z. Ustrnul (2022). Machine learning in weather prediction and climate analyses—applications and perspectives. *Atmosphere,13*(2), 180. [https://doi.org/10.3390/atmos13020180]

Box, G. E. (1976). Science and statistics. *Journal of the American Statistical Association,71*(356), 791–799. [MathSciNet][https://doi.org/10.1080/01621459.1976.10480949]

Devlin, K. (2010). The pascal-fermat correspondence: How mathematics is really done. *The Mathematics Teacher,103*(8), 579–582. [https://doi.org/10.5951/MT.103.8.0579]

Gawlikowski, J., Tassi, C. R. N., Ali, M., Lee, J., Humt, M., Feng, J., Kruspe, A., Triebel, R., Jung, P., Roscher, R., et al. (2023). A survey of uncertainty in deep neural networks. *Artificial Intelligence Review,56*(Suppl 1), 1513–1589. [https://doi.org/10.1007/s10462-023-10562-9]

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. In *ICLR proceedings*.

Gruber, C., Schenk, P. O., Schierholz, M., Kreuter, F., & Kauermann, G. (2026). Sources of uncertainty in machine learning—a statisticians' view. *Statistical Science*, *41*(3), 671–696. [MathSciNet][https://doi.org/10.1214/25-STS1003]

Hacking, I. (1975). *The emergence of probability: A philosophical study of early ideas about probability, induction and statistical inference*. Cambridge University Press.

Hora, S. C. (1996). Aleatory and epistemic uncertainty in probability elicitation with an example from hazardous waste management. *Reliability Engineering & System Safety,54*(2), 217–223. Treatment of Aleatory and Epistemic Uncertainty.

Hüllermeier, E., & Waegeman, W. (2021). Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning,110*(3), 457–506. [MathSciNet][https://doi.org/10.1007/s10994-021-05946-3]

Kiureghian, A. D., & Ditlevsen, O. (2009). Aleatory or epistemic? Does it matter? *Structural Safety,31*(2), 105–112. [https://doi.org/10.1016/j.strusafe.2008.06.020]

Li, X., & Meng, X.-L. (2021). A multi-resolution theory for approximating infinite-p-zero-n: Transitional inference, individualized predictions, and a world without bias-variance tradeoff. *Journal of the American Statistical Association,116*(533), 353–367. [MathSciNet][https://doi.org/10.1080/01621459.2020.1844210]

Spiegelhalter, D. (2024). Does probability exist? *Nature,636*, 19. [https://doi.org/10.1038/d41586-024-03807-2]

Thompson, S. K. (2010). *The unfinished game: Pascal, fermat, and the seventeenth-century letter that made the world modern*. Basic Books.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
G. Kauermann et al., *Statistics—Coping with Uncertainty*, Textbooks in Statistical Science
https://doi.org/10.1007/978-3-032-33658-3_3

3. Learning and Estimating

Göran Kauermann[1] ✉, Giacomo De Nicola[2], Cornelius Fritz[3] and David Rügamer[4]
(1) Department of Statistics, LMU Munich, Munich, Germany
(2) Harvard T.H. Chan School of Public Health, Harvard University, Boston, MA, USA
(3) School for Computer Science and Statistics, Trinity College Dublin, Dublin, Ireland
(4) Munich Uncertainty Quantification AI Lab, LMU Munich, Munich, Germany

Transitioning from the concept of uncertainty to formal statistical principles, we now adopt an information-theoretic perspective that emphasizes a general data-generating process. A key tool in this approach is the Kullback-Leibler divergence, which naturally leads to the likelihood concept. Then, we explore traditional statistical results for estimated parameters, such as consistency, sufficiency, and efficiency. With these foundations in place, uncertainty and confidence intervals are formally defined. As such, this chapter gives an overview of classical statistical properties.

3.1 Kullback-Leibler Divergence

In Chap. 2, we defined the data-generating process, implying that the data have been generated through

$$Y_i \sim G(y), \text{ i.i.d. with } i = 1, \dots n. \tag{3.1}$$

We assume that $G(\cdot) \in \mathscr{G}$, where $\mathscr{G}$ refers to the set of all possible distributions for Y. Thus, the outcome Y_i is not deterministic and hence considered as random or, as we phrased it before, uncertain to us. Therefore, we model the distribution G via a statistical model

$$\mathscr{F} = \{F(\cdot;\theta), \theta \in \Theta\}.$$

This scenario is sketched in the upper-left plot of Fig. 3.1. The true data-generating process $G(\cdot)$ is visualized as a point in the set of all possible data-generating processes. We model this data-generating process using a set of distributions $\mathscr{F}$, visualized as a subset that need not contain $G(\cdot)$. In a special case, it is possible that $G(\cdot) \in \mathscr{F}$, indicating that the data-generating process is an element of $\mathscr{F}$. The aim is to find the best approximation of the true but unknown data-generating process $G(\cdot)$ within the set $\mathscr{F}$. Therefore, we search for the element in $\mathscr{F}$ which has the smallest "distance" to $G(\cdot)$. We cautiously used the phrase "distance" here, which correctly refers to a measure for the discrepancy between two distributions. Hence, "distance" is not meant as a rigorous metric in mathematical terms but as a discrepancy measure.

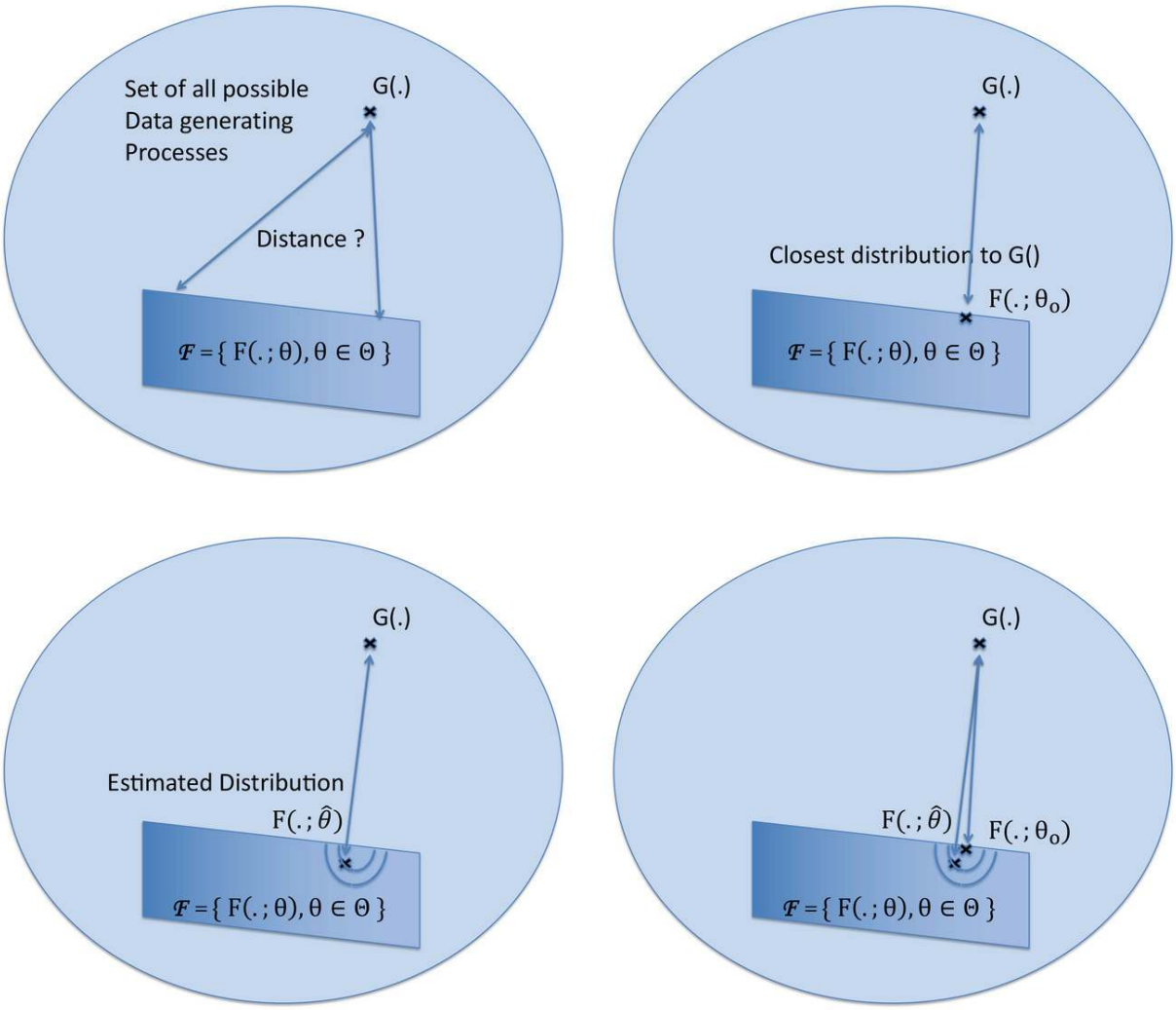


Fig. 3.1 Visualization of the true data-generating process and its approximation through statistical models. The top-left plot shows the true data-generating process $G(\cdot)$ and a statistical model $\mathscr{F}$. The top-right plot shows the optimal parameter value, θ_0. The bottom left shows the estimator $\hat{\theta}$, and the difference from θ_0 is sketched in the bottom right plot

To do so, we need to tackle several questions, which we cover step by step:

Step1:
: Define a distance that describes how far $G(\cdot)$ is away from a distribution $F(\cdot;\theta) \in \mathscr{F}$.

Step2:
: Use this measure for two distributions to find a distribution $F(\cdot;\theta) \in \mathscr{F}$ closest to $G(\cdot)$ via an optimization problem over $\theta \in \Theta$.

Step3:
: Solve this problem involving an intractable expected value via an empirical approximation.

We start with the first step: defining a measure that quantifies the "distance" between two distribution functions. A practical and mathematically useful measure is the Kullback-Leibler divergence.

Definition 3.1 *The Kullback-Leibler divergence for two distribution functions$G(\cdot)$and$F(\cdot)$is defined as*

$$\begin{aligned} KL(G,F) &= \int \log \tfrac{g(y)}{f(y)} g(y) d\, y \\ &= \int \log g(y) g(y) d\, y - \int \log f(y) g(y) d\, y, \end{aligned}$$

where$g(y)$and$f(y)$are the densities or probability functions of G and F, respectively.

The term

$$\int \log g(y) g(y) d\, y$$

is also called entropy.

If y is discrete-valued, the integrals above are summations. In this case, y takes $a_0 = -\infty < a_1 < a_2 < a_3 < \ldots < \infty$, such that $\pi_k := P(Y = a_k) = G(a_k) - G(a_{k-1})$. The entropy can then be written as

$$\sum_k \log\,(\pi_k)\pi_k,$$

which is positive unless the distribution has a point mass of 1 at a single point.

We write $KL(G(\cdot), F(\cdot))$ when we explicitly want to underline in the notation that $G(\cdot)$ and $F(\cdot)$ are, in fact, distribution functions. Both notations are used interchangeably. Note also that if the distribution functions $G(\cdot)$ and $F(\cdot)$ have densities $g(\cdot)$ and $f(\cdot)$, respectively, it is also common to use the densities as arguments and write the Kullback-Leibler divergence as $KL(g(\cdot), f(\cdot))$.

The Kullback-Leibler divergence is always positive and takes the value 0 if and only if $G(\cdot) \equiv F(\cdot)$. For continuous variables, we can show this using $-\log\,(y) \geq 1 - y$ with equality for $y \equiv 1$:

$$\begin{aligned} KL(G, F) &= \int \log\,\tfrac{g(y)}{f(y)} g(y) \mathrm{d}\,y \\ &= -\int \log\,\tfrac{f(y)}{g(y)} g(y) \mathrm{d}\,y \\ &\geq \int \left(1 - \tfrac{f(y)}{g(y)}\right) g(y) \mathrm{d}\,y \\ &= \int g(y) \mathrm{d}\,y - \int f(y) \mathrm{d}\,y \\ &= 1 - 1 = 0. \end{aligned}$$

The Kullback-Leibler divergence is not a distance measure since it is not symmetric, i.e.,

$$KL(G, F) = KL(F, G)$$

does not generally hold, nor does it fulfill the triangle inequality. However, we can omit symmetry and use the true data-generating process as an anchor point for calculating the Kullback-Leibler divergence.

With this measure, we can compare different distributions, as the upper-right plot in Fig. 3.2 illustrates. In the top plot, we show two exemplary distributions, $g(y)$ and $f(y)$, that differ by some amount. The second plot shows the log ratio $\log\,(g(y)\,/\,f(y))$ of the two distributions, and we see that this log ratio takes large values for both small and large values of y. The density $g(y)$ for large and small values of y is, however, small, and by weighting the log ratio in the second plot with the corresponding density, we

obtain the third plot in Fig. 3.2. The discrepancies at the tails of *y* are now down-weighted, and the smaller discrepancies in the center part of the distribution are becoming dominant. Integrating out the function, as sketched in the bottom plot, provides the Kullback-Leibler divergence.

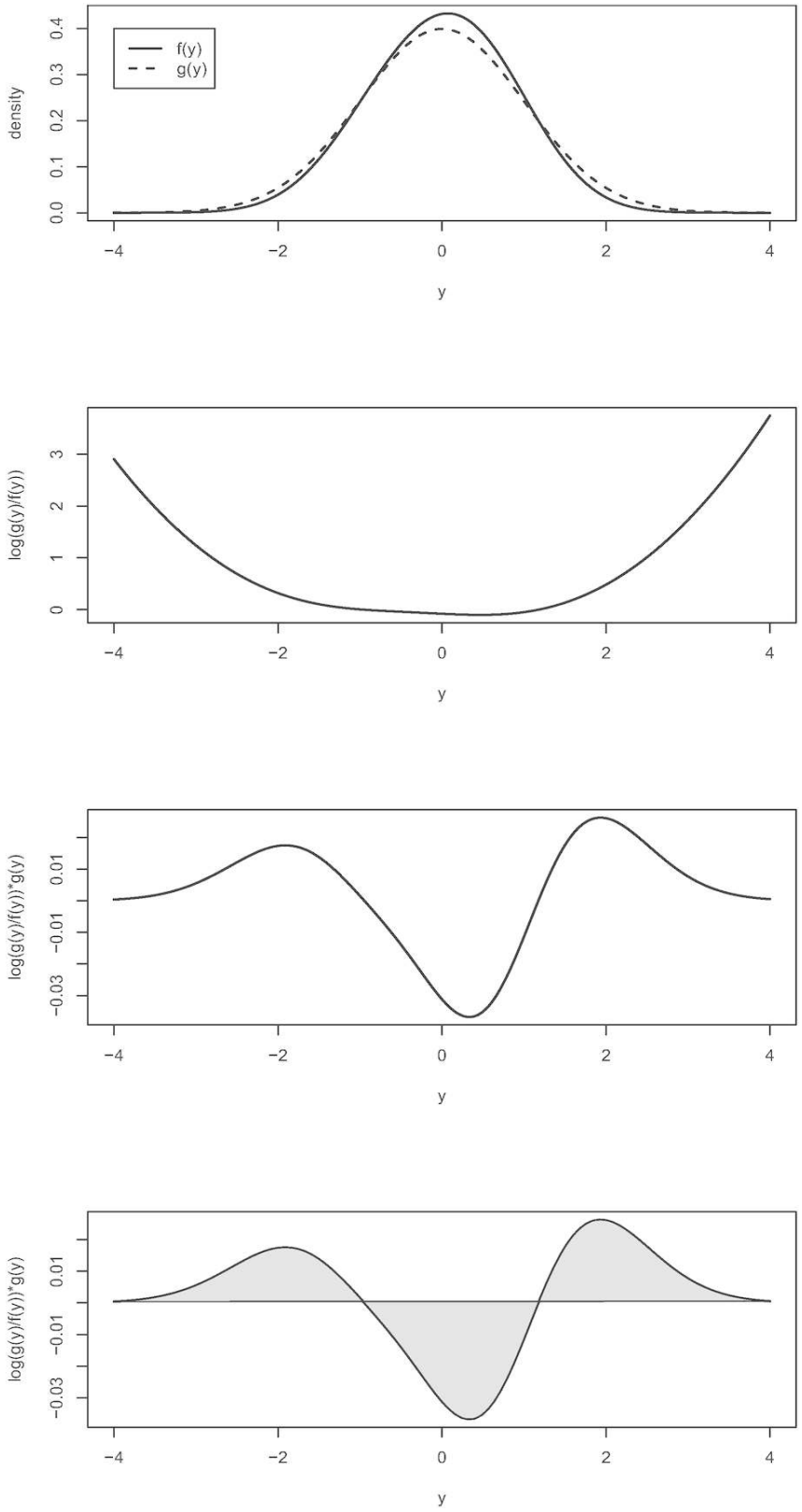


Fig. 3.2 Visualization of Kullback-Leibler divergence. The top row shows two densities $g(\cdot)$ and $f(\cdot)$. The middle row gives the log ratio $\log\,(g(\cdot)\,/\,f(\cdot))$, and the bottom row shows the ratio weighted with

$g(\cdot)$. Integration of the bottom plot curve gives the Kullback-Leibler divergence $KL(g(\cdot), f(\cdot))$

With this distance measure, we can calculate the distance between two distributions. This implies that for each element F in $\mathscr{F}$, we can calculate

$$KL(G(\cdot), F(\cdot;\theta)) = \int \log \tfrac{g(y)}{f(y;\theta)} \mathrm{d}\, G(y).$$

To address the second step, we, however, need a distance measure between the set $\mathscr{F}$ and $G(\cdot)$. How can we define a distance between a point, $G(\cdot)$, and a set, $\mathscr{F}$? We solve this problem with the minimum Kullback-Leibler divergence

$$KL(G(\cdot), \mathscr{F}) := \min_{F(\cdot;\theta)\in\mathscr{F}} KL(G(\cdot), F(\cdot;\theta)). \tag{3.2}$$

If we assume that integration and differentiation are interchangeable (technically, this means that $\mathscr{F}$ is a set of Fisher-regular distributions), we can obtain the minimum through differentiation, that is, we set

$$\begin{aligned} 0 &= \tfrac{\partial}{\partial\theta} \int \log \tfrac{g(y)}{f(y;\theta)} \mathrm{d}\, G(y) \\ &= -\int \tfrac{\partial}{\partial\theta} \log f(y;\theta) \mathrm{d}\, G(y). \end{aligned}$$

The parameter θ_0, which minimizes the distance, is now implicitly defined through the equation

$$0 = \int \tfrac{\partial \log f(y;\theta_0)}{\partial\theta} \mathrm{d}\, G(y) = E\Big(\tfrac{\partial \log f(Y;\theta_0)}{\partial\theta}\Big). \tag{3.3}$$

In this setting, we understood θ_0 as the "closest" parameter $\theta \in \Theta$ to the true data-generating process $G(\cdot)$, which is shown in the top-right plot of Fig. 3.1. In practice, the above integrals cannot be calculated, since the distribution $G(\cdot)$ is unknown. Consequently, it appears that as long as we do not know the true data-generating process, we are unable to choose the optimal model $F(\cdot;\theta_0)$ corresponding to the smallest distance to $G(\cdot)$. For every real-world data example, this will always be the case, so the proposed approach appears to be a cul-de-sac. However, we have data $y_1, \ldots, y_n$ which were generated according to $G(\cdot)$. For any measurable function (statistic) $t(Y)$, this means

$$E\Big(t(Y)\Big) = \int t(y) \mathrm{d}\, G(y).$$

Based on the law of large numbers (see Sect. 1.2.2), we know that the mean can be estimated through the corresponding empirical version, that is,

$$\frac{1}{n}\sum_{i=1}^{n} t(y_i) \xrightarrow{n\to\infty} E\Big(t(Y)\Big).$$

This means that the Kullback-Leibler distance can be estimated through

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\log\frac{g(y_i)}{f(y_i;\theta)}\\
&=\frac{1}{n}\sum_{i=1}^{n}\log g(y_i)-\frac{1}{n}\sum_{i=1}^{n}\log f(y_i;\theta)\\
&\xrightarrow{n\to\infty}\int \log g(y)g(y)\mathrm{d}\,y-\int \log f(y;\theta)g(y)\mathrm{d}\,y.
\end{aligned}$$

The first component is still not estimable, since we do not know $g(\cdot)$ and hence cannot calculate $\frac{1}{n}\sum_{i=1}^{n}\log g(y_i)$. This means we are unable to estimate the entropy of the data-generating process. But we are able to estimate the **negative** second component through

$$\frac{1}{n}\sum_{i=1}^{n}\log f(y_i;\theta)\to\int \log f(y;\theta)g(y)\mathrm{d}\,y.$$

We want to maximize this term, which minimizes the Kullback-Leibler divergence. Omitting the $1/n$ term, which does not depend on θ, in the sum above becomes the *log-likelihood function*.

Definition 3.2 *The***log-likelihood function***is defined as*

$$l(\theta)=\sum_{i=1}^{n}\log f(y_i;\theta).$$

The parameters of the log-likelihood function are the parameters, although it depends on the observations$y_1,\ldots,y_n$, which is suppressed in the notation. Formally, however, we have$l(\theta)=l(\theta;y_1,\ldots,y_n)$.

The log-likelihood function is of central importance in statistics, which we will discuss in more depth in the next chapter.

Note that θ_0 is defined in (3.3) through

$$0=\int\frac{\partial\log f(y;\theta_0)}{\partial\theta}\mathrm{d}\,G(y)$$

which we replace with the empirical version

$$0 = \sum_{i=1}^{n} \frac{\partial \log f(y_i;\hat{\theta})}{\partial \theta} \tag{3.4}$$

where $\hat{\theta}$ solves (3.4). The logarithm is a monotonically increasing function, so $\hat{\theta}$ is maximizing not only the log likelihood but also the likelihood itself. Therefore, we call $\hat{\theta}$ the maximum likelihood estimator. At this point, it might be unclear where the phrase "maximum" comes from since we are "minimizing" the Kullback-Leibler divergence. We agree that this might appear obscure, but bear in mind that we are maximizing the log-likelihood function. We will develop a thorough theoretical framework for maximum likelihood estimation in Chap. 4, but the motivation for why this is a useful concept is based on the minimum Kullback-Leibler approach. We visualize this in Fig. 3.1 in the bottom left. Still, $\hat{\theta}$ is not necessarily equal to θ_0 so that the distance between $F(\cdot;\hat{\theta})$ and $G(\cdot)$ is usually larger than the minimal distance between $F(\cdot;\theta_0)$ and $G(\cdot)$ (see the bottom-right plot in Fig. 3.1). This is visualized through the half-circles around θ_0. These circles aim to visualize uncertainty, to which we will come back in the next chapter. We will show explicitly how to quantify the distance between $\hat{\theta}$ and θ_0. For now, we can comprehend this increased distance as uncertainty due to the fact that $\hat{\theta}$ is calculated from the data, which is the basis for approximating $E(\partial \log f(Y;\theta_0) / (\partial\theta))$ by its empirical pendant. We will show that with increasing sample size we obtain that

$$F(\cdot;\hat{\theta}) \to F(\cdot;\theta_0)$$

or more precisely

$$\hat{\theta} \to \theta_0,$$

but this requires some more technical derivations, which we postpone for now.

3.2 Squared Loss and Likelihood

The likelihood approach allows us to minimize the Kullback-Leibler distance between the data-generating process $G(\cdot)$ and a statistical model $\mathscr{F} = \{F(\cdot;\theta), \theta \in \Theta\}$. It leads to an estimation principle by finding the root of the log-likelihood function. Assuming that the log likelihood is well-behaved in that it is concave, we can achieve this by setting its first-order

derivative to zero. We want to elaborate on this further and connect it to the familiar squared-loss function approach. We initiated this discussion already in Example 2.1, but we keep it simpler here. Assume we are interested in the mean of data $y_1, y_2, \ldots, y_n$. A reasonable approach to find the mean is to minimize the squared losses between observations y_i and the mean μ. This leads to the squared-loss function approach

$$\min_{\mu \in \mathbb{R}} \sum_{i=1}^{n} (y_i - \mu)^2. \tag{3.5}$$

We obtain the same loss function by assuming a normal distribution, i.e., $\mathscr{F}$ is the set of all normal distributions with mean μ and variance σ^2, which, in turn, implies the following log-likelihood function:

$$l(\mu, \sigma^2) = \sum_{i=1}^{n} \Big(-\log(\sigma) - \tfrac{1}{2} \tfrac{(y_i - \mu)^2}{\sigma^2} \Big), \tag{3.6}$$

If we are interested in μ only, i.e., omitting the variance σ^2 for the moment, we observe that maximizing (3.6) with respect to μ is equivalent to minimizing (3.5):

$$\begin{aligned} \max_{\mu \in \mathbb{R}} l(\mu, \sigma^2) &= \max_{\mu \in \mathbb{R}} \sum_{i=1}^{n} -\tfrac{1}{2} \tfrac{(y_i - \mu)^2}{\sigma^2} \\ &= \max_{\mu \in \mathbb{R}} -\sum_{i=1}^{n} (y_i - \mu)^2 \\ &= \min_{\mu \in \mathbb{R}} \sum_{i=1}^{n} (y_i - \mu)^2 \end{aligned}$$

Therefore, a squared loss is equivalent to assuming a normal distribution (with constant variance) for the data $y_1, \ldots, y_n$.

3.3 Properties of Estimators

3.3.1 Consistency

This section is devoted to classical concepts of (mathematical) statistics. In particular, we will introduce concepts like consistency, biasedness, and sufficiency, which are properties of estimators in a statistical model $\mathscr{F} = \{F(\cdot; \theta), \theta \in \Theta\}$. Let us start with consistency. Informally, an estimator $\hat{\theta}$ is consistent if for increasing sample size the difference between $\hat{\theta}$ and the target parameter θ_0 gets smaller in probability. Target parameter means, here, that it is either the true parameter if the data-generating process $G(\cdot)$ is within $\mathscr{F}$, that is, $G(\cdot)$ can be written as $G(\cdot) = F(\cdot; \theta_0)$, or if $G(\cdot) \neq \mathscr{F}$, the target parameter θ_0 gives the closest distribution in $\mathscr{F}$ based on the Kullback-Leibler divergence, as described in Sect. 3.2.

We can formulate consistency utilizing the Kullback-Leibler idea. With increasing data size, we postulate

$$\int \log \frac{f(y;\hat{\theta})}{f(y;\theta_0)} \mathrm{d}\, G(y) \xrightarrow[n\to\infty]{} 0. \tag{3.7}$$

In principle, we need to be more precise here since $\hat{\theta}$ is based on a given database, but we aim to develop a statement for an arbitrary sample drawn from the data-generating process. In other words, we should formulate (3.7) as a probability statement. We remain somewhat sloppy for the moment and get more precise later. Let us look, therefore, at (3.7) more closely. A simple calculation leads to

$$\begin{aligned}\int \log \tfrac{f(y;\hat{\theta})}{f(y;\theta_0)} \mathrm{d}\, G(y) &= \int \log f(y;\hat{\theta}) \mathrm{d}\, G(y) - \int \log f(y;\theta_0) \mathrm{d}\, G(y) \\ &= KL(G(\cdot), F(\cdot;\theta_0)) - KL(G(\cdot), F(\cdot;\hat{\theta})).\end{aligned}$$

Hence, the Kullback-Leibler divergences need to converge to the same constant as the number of data points n increases. Applying a second-order Taylor series expansion around θ_0, which is defined in (3.3), to the estimated log density yields

$$\begin{aligned}\log f(y;\hat{\theta}) \approx\ & \log f(y;\theta_0) + \tfrac{\partial \log f(y;\theta_0)}{\partial \theta} (\hat{\theta} - \theta_0) \\ & + \tfrac{1}{2} (\hat{\theta} - \theta_0)^\top \tfrac{\partial^2 \log f(y;\theta_0)}{\partial \theta \partial \theta^\top} (\hat{\theta} - \theta_0).\end{aligned}$$

Assuming that differentiation and integration are exchangeable, we get

$$\int \log \frac{f(y;\hat{\theta})}{f(y;\theta_0)} \mathrm{d}\, G(y) \approx \frac{1}{2} (\hat{\theta} - \theta_0)^\top \underbrace{E\left(- \frac{\partial^2 \log f(y;\theta_0)}{\partial \theta \partial \theta^\top} \right)}_{:= I(\theta_0)} (\hat{\theta} - \theta_0), \tag{3.8}$$

where the expectation is taken with respect to the data-generating process. We call the matrix $I(\theta_0)$ the Fisher information, or Fisher matrix, which plays an important role in maximum likelihood estimation, which is covered in Chap. 4. As remarked above, (3.7) is a statement based on data, and hence $\hat{\theta}$ is a concrete value calculated through some algorithm. Hence, we can denote $\hat{\theta}$ as $\hat{\theta} = t(y_1, \ldots, y_n)$. Based on the aleatoric uncertainty given in the data-generating process (2.3), the data are random, and hence the outcome

of the algorithm is considered random. If we replace statement (3.7) by taking expectation over all possible samples, we obtain

$$E\left(\int \log \frac{f(y;\hat{\theta})}{f(y;\theta_0)} \mathrm{d}\, G(y)\right) \xrightarrow[n\to\infty]{} 0.$$

Note that the expectation is taken over the data $Y_1, \ldots, Y_n$, and formally the above expectation equals

$$\underbrace{\int \cdots \int}_{n \text{ integrals}} \int \log \frac{f(y; t(y_1, \ldots, y_n))}{f(y;\theta_0)} \, \mathrm{d}\, G(y) \mathrm{d}\, G(y_1) \ldots G(y_n).$$

Utilizing the Taylor series approximation from above, this in turn simplifies to

$$E\Big((\hat{\theta} - \theta_0)^\top I(\theta_0)(\hat{\theta} - \theta_0)\Big) \xrightarrow[n\to\infty]{} 0,$$

where the expectation is taken over the data $Y_1, \ldots, Y_n$. To simplify further, we restrict the deviations, without loss of generality, to univariate parameters, thereby reducing the notation. This allows us to rewrite the above to

$$\begin{aligned} E\Big(I(\theta_0)(\hat{\theta} - \theta_0)^2\Big) &\propto E\Big((\hat{\theta} - \theta_0)^2\Big) \\ &= E\Big((\hat{\theta} - E(\hat{\theta}) + E(\hat{\theta}) - \theta_0)^2\Big) \\ &= \underbrace{E\Big((\hat{\theta} - E(\hat{\theta}))^2\Big)}_{= \mathrm{Var}(\hat{\theta})} + \underbrace{\Big(E(\hat{\theta}) - \theta_0\Big)^2}_{=: \mathrm{bias}^2(\hat{\theta}, \theta)}, \end{aligned} \tag{3.9}$$

where $\propto$ stands for proportionality, which allows us to omit components not depending on $\hat{\theta}$. The first term is the variance of the estimator, while the second is called bias. The bias thereby quantifies the systematic difference between the estimator's expectation and the target parameter. Decomposition (3.9) refers to the **bias-variance-trade-off**, as we aim to minimize their sum, which is defined as mean squared error (MSE).

Definition 3.3 *The mean squared error of an estimator $\hat{\theta}$ is defined as*

$$\begin{aligned} MSE(\hat{\theta}, \theta_0) &= E\Big((\hat{\theta} - \theta_0)^2\Big) \\ &= Var(\hat{\theta}) + bias^2(\hat{\theta}, \theta_0), \end{aligned}$$

where

$$bias(\hat{\theta}, \theta_0) = E(\hat{\theta}) - \theta_0$$

defines the bias. We call an estimator unbiased if

$$bias(\hat{\theta}, \theta_0) = 0$$

and asymptotically unbiased if

$$\text{bias}(\hat{\theta}, \theta_0) \xrightarrow[n \to \infty]{} 0.$$

We sketch the idea of the MSE in Fig. 3.3 when shooting toward the bull's eye. We want both small variance and small bias. The abovementioned bias-variance trade-off is easily described with this figure. The best marksman is the one whose shots score the most points, which can be quantified by the reciprocal of the squared distances from the bull's-eye. Here, both dispersion (variance) and bias matter.

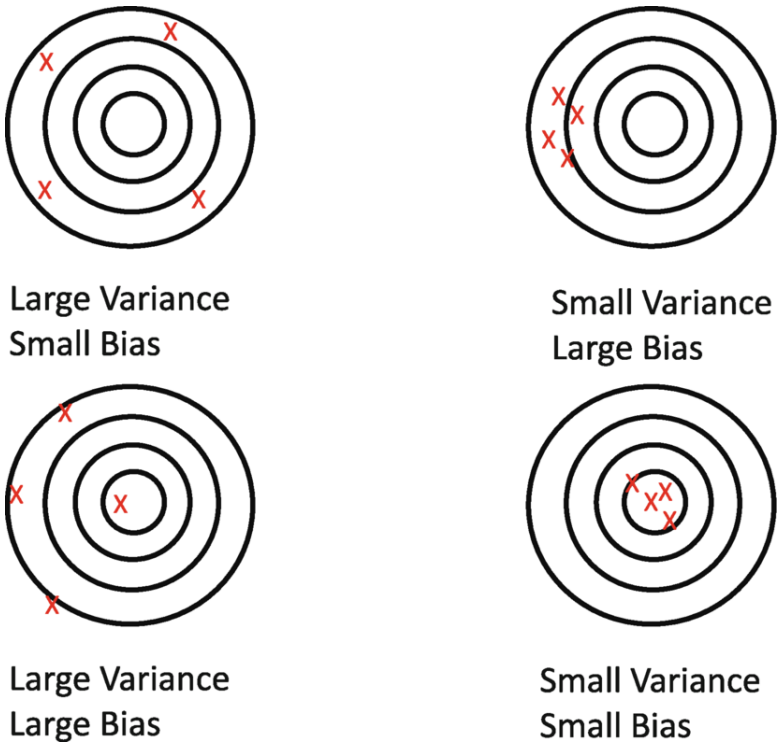


Fig. 3.3 Sketch of bias and variance when shooting at the bull's eye

With these requirements, we can now write down the original idea (3.7) more formally to define consistency.

Definition 3.4 *The estimator $\hat{\theta}$ is MSE consistent for θ_0 (or simply consistent) if, for increasing data size n with data drawn from the data-generating process (2.3), it holds*

$$\mathrm{MSE}(\hat{\theta}, \theta_0) \xrightarrow[n\to\infty]{} 0,$$

where n is the data size used to derive the estimate $\hat{\theta}$.

Note that there are various alternative definitions of consistency. The most common type of consistency is weak consistency.

Definition 3.5 *The estimator $\hat{\theta}$ is weakly consistent for θ_0 if for every $\delta > 0$*

$$P(|\hat{\theta} - \theta_0| \geq \delta) \xrightarrow[n\to\infty]{} 0, \tag{3.10}$$

where n is the data size used to derive the estimate $\hat{\theta}$ and the stochasticity in the probability stems from sampling $Y_1, \ldots, Y_n$ drawn from some data-generating process.

We can relate the definitions (3.4) and (3.5) to one another using the Chebyshev's inequality (see Sect. A.7), which is one of the most important inequalities surveyed in Appendix A.7. This inequality states for any random variable Y with mean μ and variance $\sigma^2 > 0$ the following inequality holds for any $k > 0$:

$$P\big(|X - \mu| \geq k\,\sigma\big) \leq \tfrac{1}{k^2}$$

Taking now $\hat{\theta}$ as random variable with mean θ and variance $\mathrm{Var}(\hat{\theta}) = E((\hat{\theta} - \theta)^2)$ if $\hat{\theta} = t(Y_1, \ldots, Y_n)$ and assuming sufficient regularity conditions, Chebyshev's inequality allows to obtain for arbitrary $\delta > 0$

$$\begin{aligned} P\big(|\hat{\theta} - \theta_0| \geq \delta\big) &= P\big((\hat{\theta} - \theta_0)^2 \geq \delta^2\big) \\ &\leq \tfrac{E\left((\hat{\theta}-\theta_0)^2\right)}{\delta^2} = \tfrac{\mathrm{MSE}(\hat{\theta},\theta_0)}{\delta^2}. \end{aligned}$$

We see that in this case (3.10) is equivalent to MSE consistency, since if $\mathrm{MSE}(\hat{\theta}, \theta_0) \to 0$ for $n \to \infty$ we obtain (3.10).

3.3.2 Sufficiency

We now define sufficiency, the concept of how to condense data so that no relevant information is lost. This compressed data should then encapsulate all the necessary information to estimate the parameter. To motivate the idea, suppose a binary variable $Y_i \sim \mathrm{Binom}(1, \pi)$ is observed and we want to estimate parameter π via its maximum likelihood estimate, the arithmetic mean $\overline{Y} = \sum_{i=1}^{n} Y_i \Big/ n$. If we distill the data to the mean and the sample size, we have successfully extracted all the essential information needed for calculating the maximum likelihood estimate. To demonstrate this, look at the following possible samples of (y_1, y_2, y_3, y_4) under $n = 4$:

- $y = (1, 0, 0, 0)$
- $y = (0, 1, 0, 0)$
- $y = (0, 0, 1, 0)$
- $y = (0, 0, 0, 1)$

Since the arithmetic mean for all four samples is identical, they also yield the same maximum likelihood estimate. Evidently, the concrete order of the observations does not provide any relevant information about π, and the only quantity that gives information on π is $\overline{y} = 1 \,/\, 4$. We want to formalize this via the concept of **sufficiency**.

If $t(y_1, \ldots, y_n)$ gives sufficient information about parameter θ, all the information concerning θ is in $t(y_1, \ldots, y_n)$, and knowing $t(y_1, \ldots, y_n)$ is enough to estimate the parameter. The concept of sufficiency allows us to delete our data entirely if we have calculated a sufficient estimator. This can be useful particularly in the era of big data and given the emergence of fields such as federated learning and learning under privacy constraints. Sufficiency states that one only needs to calculate and store $t(y_1, \ldots, y_n)$ to get all information from our sample relevant to our parameter.

Definition 3.6 *A statistic $(Y_1, \ldots, Y_n)$ is called* **sufficient** *for θ if the conditional distribution $P(Y_1 = y_1, \ldots, Y_n = y_n \mid \; t(Y_1, \ldots, Y_n) = t_0;\; \theta)$ does not depend on θ.*

This definition looks confusing at first glance, so an example might help to clarify.

Example 3.1 Let $Y_1, \ldots, Y_n$ be Bernoulli variables with values 0 or 1 and $P(Y_i = 1) = \pi$. Our statistic is the arithmetic mean $t(\cdot) = t(Y_1, \ldots, Y_n) = \sum_{i=1}^{n} Y_i \,/\, n$. We next show that this statistic is sufficient for π. The statistic takes values in the set $\{0, \frac{1}{n}, \frac{2}{n}, \ldots, \frac{n}{n}\}$, and we denote the resulting value calculated from the data $(y_1, \ldots, y_n)$ with t_0. Correspondingly, $n_0 = nt_0$ is the sum of all of our variables. Then, using the Bayes rule (see Appendix), we can calculate the probability of the sample conditional on the sufficient statistic as follows:

$$
\begin{aligned}
P(Y_1 = y_1, \quad & \ldots, Y_n = y_n \mid \; t(Y_1, \ldots, Y_n) = t_0; \; \pi) \\
&= \frac{P\left(Y_1 = y_1, \ldots, Y_n = y_n, \sum_{i=1}^{n} Y_i = n_0; \, \pi\right)}{P\left(\sum_{i=1}^{n} Y_i = n_0; \, \pi\right)} \\
&= \begin{cases} \dfrac{\prod_{i=1}^{n} \pi^{y_i} (1-\pi)^{(1-y_i)}}{\binom{n}{n_0} \pi^{n_0} (1-\pi)^{(n-n_0)}} & \text{for } \sum_{i=1}^{n} y_i = n_0 \\ 0 & \text{otherwise} \end{cases} \\
&= \begin{cases} \dfrac{\pi^{\sum_{i=1}^{n} y_i} (1-\pi)^{(\sum_{i=1}^{n} 1-y_i)}}{\binom{n}{n_0} \pi^{n_0} (1-\pi)^{(n-n_0)}} & \text{for } \sum_{i=1}^{n} y_i = n_0 \\ 0 & \text{otherwise} \end{cases} \\
&= \begin{cases} \dfrac{\pi^{n_0} (1-\pi)^{(n-n_0)}}{\binom{n}{n_0} \pi^{n_0} (1-\pi)^{(n-n_0)}} & \text{for } \sum_{i=1}^{n} y_i = n_0 \\ 0 & \text{otherwise} \end{cases} \\
&= \begin{cases} \dfrac{1}{\binom{n}{n_0}} & \text{for } \sum_{i=1}^{n} y_i = n_0 \\ 0 & \text{otherwise} \end{cases}
\end{aligned}
$$

Apparently, this distribution does not depend on π.▷

More generally, it is challenging to prove sufficiency in the form defined above. In such settings, the Fisher-Neyman factorization is a valuable tool to prove the sufficiency of a statistic.

Theorem (Fisher-Neyman Factorization) *A statistic* $(Y_1, \ldots, Y_n)$ *is sufficient for* θ *if and only if the density or probability function decomposes to*

$$f(y_1, \ldots, y_n; \theta) = h(y_1, \ldots, y_n) g\Big(t(y_1, \ldots, y_n); \theta\Big), \tag{3.11}$$

where $h(\cdot)$ *is a function of the data independent of* θ *and* $g(\cdot)$ *is a function of* θ *that depends on the data only through the sufficient statistic* $(y_1, \ldots, y_n)$.

The proof of this statement is given at the end of this section.

Example 3.2 Let $Y_1, \ldots, Y_n$ be independent and identically distributed variables from a Gamma distribution with the following density,

$$f(y; \alpha, \beta) = \tfrac{\beta^\alpha}{\Gamma(\alpha)} y^{\alpha-1} \exp\,(-\beta\, y)\, \mathbb{I}(y > 0)$$

with shape α, rate β, and Γ function given by $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} \exp\,(-t)\, \mathrm{d}\, t$ that is a continuous generalization of factorials. The goal is to show that the two-dimensional statistic $t(y_1, \ldots, y_n) = (\prod_{i=1}^n y_i, \sum_{i=1}^n y_i)$ is sufficient for (α, β). Since the distribution of $Z = \prod_{i=1}^n Y_i$ with Y_i being i.i.d. Gamma variables has no closed form, it is difficult to calculate the conditional probability necessary for showing sufficiency according to Definition 3.6. We can, however, write down the joint distribution needed for the Fisher-Neyman factorization:

$$\begin{aligned}
&f(y_1, \ldots, y_n; \alpha, \beta) \\
&= \prod_{i=1}^n \frac{\beta^\alpha}{\Gamma(\alpha)} y_i^{\alpha-1} \exp(-\beta\, y_i)\, \mathbb{I}(y_i > 0) \\
&= \underbrace{\left(\prod_{i=1}^n \mathbb{I}(y_i > 0)\right)}_{h(y_1, \ldots, y_n)} \underbrace{\left(\frac{\beta^\alpha}{\Gamma(\alpha)}\right)^n \left(\prod_{i=1}^n y_i\right)^{\alpha-1} \exp\left(-\beta \left(\sum_{i=1}^n y_i\right)\right)}_{g\left(t(y_1, \ldots, y_n); \theta\right)}
\end{aligned}$$

The claim follows from the factorization shown in the final line. ▷

Note that sufficiency itself is a relatively weak statement, as the original sample $(y_1, \ldots, y_n)$ itself is already sufficient. We therefore also require some concept of minimality defined as follows:

Definition 3.7 *The statistic $t(y_1, \ldots, y_n)$ is minimal sufficient for θ if $t(\cdot)$ is sufficient and for any other sufficient statistic $\tilde{t}(y_1, \ldots, y_n)$ there exists a function $m(\cdot)$ such that $t(y_1, \ldots, y_n) = m(\tilde{t}(y_1, \ldots, y_n))$.*

The definition states that if there exists a minimal sufficient statistic, then it can be calculated directly from any other sufficient statistic. Hence, we may reduce the data $y_1, \ldots, y_n$ to the value of the minimal sufficient statistic, but we may not reduce it further without losing information about the parameter θ.

Proof Assuming that $t(\cdot)$ is sufficient, then $f(y_1, \ldots, y_n \mid t(y) = t; \theta)$ does not depend on θ. Because $t(\cdot)$ is calculated from $y_1, \ldots, y_n$, with the basic definition of conditional probabilities, we get

$$f(y_1, \ldots, y_n; \theta) = \underbrace{f(y_1, \ldots, y_n \mid t(y_1, \ldots, y_n) = t; \theta)}_{h(y_1, \ldots, y_n)} \underbrace{f_t(t \mid y_1, \ldots, y_n; \theta)}_{g(t(y_1, \ldots, y_n); \theta)},$$

where the first component does not depend on θ and the second component is the distribution of $t(\cdot)$, which by construction depends only on $t(y_1, \ldots, y_n)$ and θ. Let us assume now that the density $f(y_1, \ldots, y_n)$ is factorized as in (3.11). The marginal density for $t(y_1, \ldots, y_n)$ is

$$\begin{aligned} f_t(t; \theta) &= \int_{t(y_1, \ldots, y_n) = t} f(y_1, \ldots, y_n; \theta) \mathrm{d}\, y_1 \ldots \mathrm{d}\, y_n \\ &= \int_{t(y_1, \ldots, y_n) = t} h(y_1, \ldots, y_n) g(t; \theta) \mathrm{d}\, y_1 \ldots \mathrm{d}\, y_n \\ &= g(t; \theta) \int_{t(y_1, \ldots, y_n) = t} h(y_1, \ldots, y_n) \mathrm{d}\, y_1 \ldots \mathrm{d}\, y_n. \end{aligned}$$

The conditional distribution can then be written as

$$\begin{aligned} & f(y_1, \ldots, y_n \mid t(y_1, \ldots, y_n) = t; \theta) \\ &= \frac{f(y_1, \ldots, y_n, t(y_1, \ldots, y_n) = t; \theta)}{f_t(t; \theta)} \\ &= \begin{cases} \frac{h(y_1, \ldots, y_n) g(t; \theta)}{g(t; \theta)} & \text{for } t(y_1, \ldots, y_n) = t \\ 0 & \text{otherwise,} \end{cases} \end{aligned}$$

which does not depend on θ because $g(t; \theta)$ cancels out. □

3.3.3 Efficiency

Lastly, we discuss the concept of efficiency, which addresses the question of which of the two possible estimators performs better. In this context, "better" refers to having a smaller mean squared error introduced in Sect. 3.2.

Imagine you have one estimator, denoted as $\hat{\theta}_1 = t_1(Y_1, \ldots, Y_n)$, and another one, denoted as $\hat{\theta}_2 = t_2(Y_1, \ldots, Y_n)$, which provide an estimate for θ_0. The comparison between the two boils down to evaluating their performance using mean squared error (MSE). If the MSE of $\hat{\theta}_1$ is less than the MSE of $\hat{\theta}_2$, then $\hat{\theta}_1$ is considered superior to $\hat{\theta}_2$. This concept is largely theoretical, as calculating both MSEs is not feasible because they depend on the unknown data-generating process $G(\cdot)$.

Example 3.3 Assume we have i.i.d. normally distributed random variables $Y_i \sim \mathrm{N}(\mu, \sigma^2)$. We want to estimate the mean via the statistic

$$t(y) = \sum\nolimits_{i=1}^{n} w_i y_i$$

for some weights w_i for $i = 1, \ldots, n$. For unbiasedness from Definition 3.3, observe that $\sum_{i=1}^{n} w_i = 1$ needs to hold for

$$E\Big(t(Y)\Big) = \sum\nolimits_{i=1}^{n} w_i \mu.$$

Since $\mathrm{bias}(t(y)) = 0$, the MSE depends only on the variance of the statistic. Therefore, we need to choose the weights so that the variance of $t(Y)$ is minimized, thereby finding an estimator with minimal MSE. Note that

$$\mathrm{Var}\Big(t(Y)\Big) = \sum\nolimits_{i=1}^{n} w_i^2 \sigma^2.$$

Without loss of generality, we can reparametrize the weights and set $w_i = \frac{1}{n} + d_i$, where $\sum_{i=1}^{n} d_i = 0$ holds for unbiasedness:

$$\begin{aligned}\mathrm{Var}\Big(t(Y)\Big) &= \sum_{i=1}^{n}\left(\frac{1}{n}+d_i\right)^2\sigma^2 \\ &= \sum_{i=1}^{n}\left(\frac{1}{n^2}+d_i^2\right)\sigma^2 \\ &= \frac{\sigma^2}{n^2} + \underbrace{\sum_{i=1}^{n} d_i^2\sigma^2}_{\geq 0}\end{aligned}$$

Because $\sum_{i=1}^{n}\sigma^2\ d_i^2 \geq 0$ unless $d_i = 0$, we see that the arithmetic mean, which results when we set the weights w_i equal to $\frac{1}{n}$, has the smallest variance and hence the estimator is MSE-efficient. ▷

Definition 3.8 *The statistic* $(y_1, \ldots, y_n)$ *is* **efficient** *for* θ *if it has the smallest MSE among all possible statistics.*

We will show later that maximum likelihood estimation is (approximately) efficient. This means that estimators based on optimizing the likelihood function are asymptotically unbiased and have the smallest variance.

3.4 Estimation and Uncertainty

So far, we have developed estimation strategies and highlighted the importance of the Kullback-Leibler divergence. Estimation, or learning, can be thought of as a process or algorithm that generates a value $\hat{\theta}$ for observed data $(y_1, \ldots, y_n)$. While the estimation process itself is treated as an algorithm, producing the same output for the same input, the input itself is random and drawn from $Y_i \sim G(\cdot)$ as generally defined (2.3). This introduces randomness, meaning that the estimated value $\hat{\theta}$ has variability due to the randomness of the data. If we used different datasets, we would therefore get different estimates. This variability is known as the *estimation variance*, and next, we introduce ways to quantify it.

To tackle this challenge, we introduce *confidence intervals*. Instead of relying on a single estimate $\hat{\theta}$, we construct a range or interval within which the true parameter value is likely to lie. This interval is bounded to the left by

$t_l(y_1, \ldots, y_n)$ and to the right by $t_r(y_1, \ldots, y_n)$, representing the lower and upper bounds of the interval. For simplicity, we here only cover the special case of an univariate parameter. The interval limits are functions of the data and, as such, can be understood as the output of an algorithm applied to the observed data. We need $t_l(Y_1, \ldots, Y_n) < t_r(Y_1, \ldots, Y_n)$ for all values of $Y_1, \ldots, Y_n$ to obtain a sensical interval. These statistics can now be defined as follows:

Definition 3.9 *The interval* $CI = \big[t_l(Y_1, \ldots, Y_n), t_r(Y_1, \ldots, Y_n)\big]$ *is called a* **confidence interval** *for* θ *with confidence level* $(1 - \alpha)$ *if*

$$P\Big(t_l(Y_1, \ldots, Y_n) \leq \theta \leq t_r(Y_1, \ldots, Y_n)\Big) \geq 1 - \alpha. \tag{3.12}$$

The probability $P(\cdot)$ *is allowed to depend on the parameter* θ*, and usually it does, which is omitted in the notation for simplicity. The value* $(1 - \alpha)$ *is called the confidence level, and* α *is chosen to be a small value, e.g.,* $\alpha = 0.01$ *or* $\alpha = 0.05$.

We can equivalently state (3.12) as

$$P_\theta\Big(\theta \in [t_l(Y_1, \ldots, Y_n), t_r(Y_1, \ldots, Y_n)]\Big) \geq 1 - \alpha$$

to make the concept of an interval more visible. It is thereby important to bear in mind that $t_l(Y_1, \ldots, Y_n)$ and $t_r(Y_1, \ldots, Y_n)$ are both random due to the assumed data-generating process. In other words, the interval itself is random.

Defining $t_l(Y_1, \ldots, Y_n) = -\infty$ and $t_r(Y_1, \ldots, Y_n) = \infty$ always gives a $(1 - \alpha)$ confidence interval, as $P(-\infty \leq \theta \leq \infty) = 1$ for any $\theta \in \mathbb{R}$. This interval is, however, useless; hence, we want to find an interval with minimal length still fulfilling the definition of a confidence interval from (3.12). In other words, we prefer "$=$" in (3.12) as opposed to "$\geq$."

The construction of such a confidence interval is straightforward if a pivotal statistic exists.

Definition 3.10 *A quantity* $g\big(t(Y_1, \ldots, Y_n); \theta\big)$ *is called a* **pivotal statistic** *if its distribution does not depend on* θ*. The distribution of* $g\big(t(Y_1, \ldots, Y_n); \theta\big)$ *is also called a pivotal distribution.*

Exact pivotal quantities are rare. However, approximate pivotal distributions are quite common due to the central limit theorem. In fact, if the sample size n is large, the estimator $\hat{\theta} = t(Y_1, \ldots, Y_n)$ in many cases follows approximately a normal distribution, which is denoted by

$$\hat{\theta} = t(Y_1, \ldots, Y_n) \overset{a}{\sim} \mathrm{N}\big(\theta, \mathrm{Var}\big(\hat{\theta}\big)\big). \tag{3.13}$$

In this context, the notation $\overset{a}{\sim}$ stands for "asymptotically distributed", put differently, this distribution solely holds for sufficiently large n. With (3.13), we can construct an approximate pivotal statistic through

$$g\big(t(Y_1, \ldots, Y_n); \theta\big) = \frac{t(Y_1, \ldots, Y_n) - \theta}{\sqrt{\mathrm{Var}(\hat{\theta})}} = \frac{\hat{\theta} - \theta}{\sqrt{\mathrm{Var}(\hat{\theta})}} \overset{a}{\sim} \mathrm{N}(0, 1). \tag{3.14}$$

Since $\mathrm{N}(0, 1)$ does not depend on θ, the statistic is pivotal. The pivotal distribution can be used to construct a confidence interval as follows. With (3.14), we have

$$\begin{aligned} 1 - \alpha &\approx P\left(z_{\alpha/2} \leq \frac{\hat{\theta} - \theta}{\sqrt{\mathrm{Var}(\hat{\theta})}} \leq z_{1-\alpha/2} \right) \\ \Leftrightarrow 1 - \alpha &\approx P\left(\hat{\theta} - z_{1-\alpha/2}\sqrt{\mathrm{Var}\big(\hat{\theta}\big)} \leq \theta \leq \hat{\theta} - z_{\alpha/2}\sqrt{\mathrm{Var}\big(\hat{\theta}\big)} \right), \end{aligned}$$

where z_α is the α quantile of a $\mathrm{N}(0, 1)$ distribution for any $\alpha \in [0, 1]$. Because $z_\alpha = -z_{1-\alpha}$ holds with $\alpha \in [0, 1]$ for any symmetric distribution, such as the normal, we obtain the confidence interval as

$$CI = \left[\underbrace{\hat{\theta} - z_{1-\alpha/2}\sqrt{\mathrm{Var}\big(\hat{\theta}\big)}}_{t_l(Y_1, \ldots, Y_n)}, \underbrace{\hat{\theta} + z_{1-\alpha/2}\sqrt{\mathrm{Var}\big(\hat{\theta}\big)}}_{t_r(Y_1, \ldots, Y_n)} \right]. \tag{3.15}$$

Under a confidence level of 0.95, the corresponding value $z_{1-\alpha/2}$ is 1.96:

$$CI \approx \left[\hat{\theta} \pm 1.96\sqrt{\mathrm{Var}\big(\hat{\theta}\big)} \right] \tag{3.16}$$

Figure 3.4 visualizes the construction of the above confidence interval (3.15). In the top-left plot, we show the asymptotic normal distribution of the estimator $\hat{\theta}$ and its variance. Standardizing the estimator through

$$\frac{\hat{\theta} - \theta}{\sqrt{\mathrm{Var}(\hat{\theta})}}$$

provides the standard normal distribution as a pivotal distribution (top-right plot). From this distribution, we can derive the confidence interval (bottom-left plot), which is then transferred to the distribution of $\hat{\theta}$ (bottom-right plot).

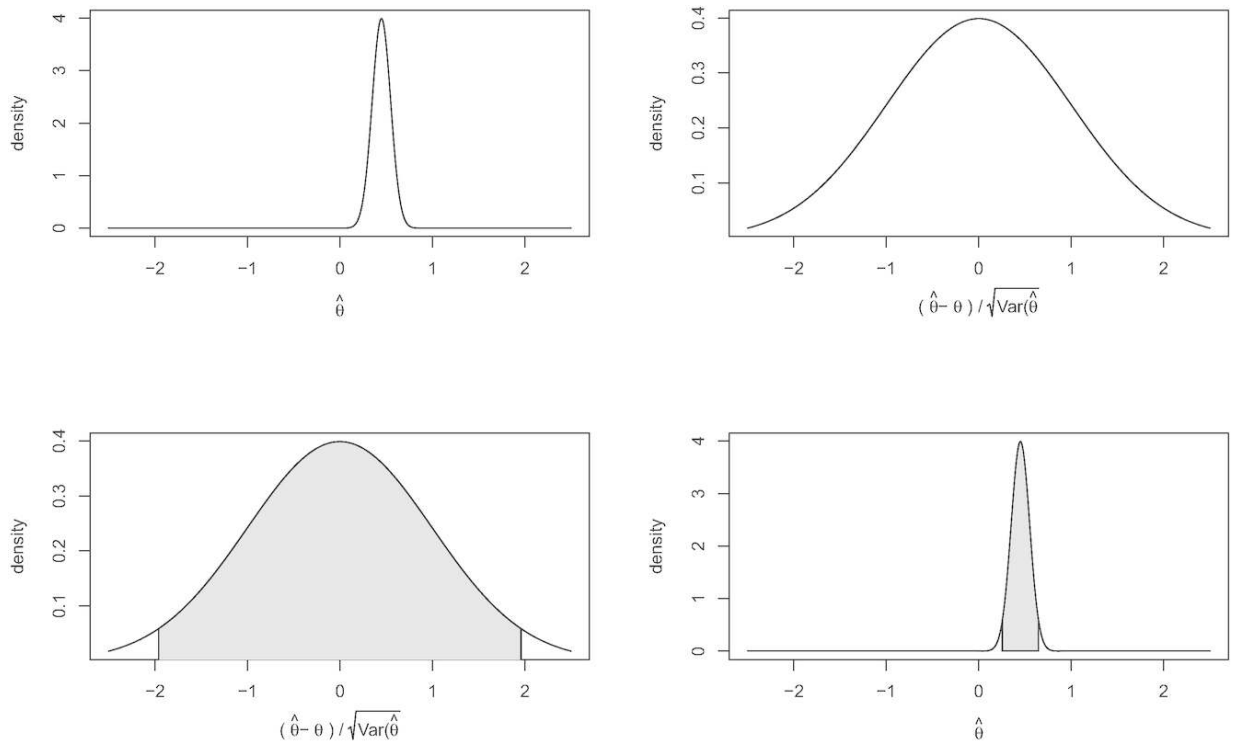


Fig. 3.4 Visualization of constructing a pivotal distribution by standardizing estimate $\hat{\theta}$ (top-left plot) through $(\hat{\theta} - \theta_0) \Big/ \sqrt{\mathrm{Var}(\hat{\theta})}$ (top-right plot). This gives a standard normal distribution as pivotal from which confidence intervals can be calculated (bottom left) and then transformed back to the level of the estimate (bottom right)

Note that, in many cases, the variance $\mathrm{Var}\left(\hat{\theta}\right)$ may itself either depend on the unknown parameter θ or depend on some other unknown parameters. In the first case, it is reasonable to replace θ with its estimate $\hat{\theta}$, which is called a plug-in estimate. If it depends on additional parameters, we need to estimate these. In particular, in situations with small data sizes, this can lead to different distributions.

A historical problem in statistics is the construction of a confidence interval for the mean of a normal distribution. We assume that data $Y_1, \ldots, Y_n$ are normally distributed with

$$Y_i \sim \mathrm{N}(\mu, \sigma^2) \; i.i.d. \tag{3.17}$$

The mean is estimated through $\overline{Y}$, and we obtain for the estimate

$$\widehat{\mu} = \overline{Y} \sim \mathrm{N}(\mu, \sigma^2 \,/\, n).$$

A confidence interval is then easily constructed through

$$CI = \left[\bar{y} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{y} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]. \tag{3.18}$$

In real-life applications, though, we do not know the variance σ^2 and hence need to estimate this. The estimate is again easy to derive through

$$\widehat{\sigma}^2 := S^2 = \tfrac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2.$$

The term $n-1$ makes the estimator unbiased, which we leave as an exercise to show. Note that

$$\tfrac{\overline{Y}-\mu}{\sigma/\sqrt{n}} \sim \mathrm{N}(0,1)$$

is a pivotal statistic with a known distribution. But if we replace the variance with its estimate S^2, the resulting distribution is no longer a normal distribution:

$$\tfrac{\overline{Y}-\mu}{\widehat{\sigma}/\sqrt{n}} \nsim \mathrm{N}(0,1) \tag{3.19}$$

The estimation of the variance leads to additional uncertainty.

This phenomenon was identified and quantified by William Sealy Gosset (1876–1937). Gosset is now a legend in the field of statistics, not only for his seminal scientific contributions but also for the circumstances of his discoveries. During his professional life, he was the head brewer at Guinness, the renowned Irish beer company, where he was hired to study data originating from all parts of their brewing process (Boland, 1984). Faced with often very small sample sizes where asymptotic expressions such as 3.13 no longer hold, he solved the problem shown in (3.19). Since he was at that time an employee at the Guinness brewery, he was unable to publish this because of the possible risk of revealing a trade secret and had to share his findings using the pseudonym Student (1908), see also Ziliak (2008). Gosset's work led to the derivation of the distribution, which is today known as the (Student's) t-distribution and the correct distribution of the left-hand side of (3.19),

$$\tfrac{\overline{Y}-\mu}{\widehat{\sigma}/\sqrt{n}} \sim t_{(n-1)}, \tag{3.20}$$

where the index $n-1$ is the only parameter characterizing the distribution, called the degree of freedom. The proof is given in Appendix B. Note that

(3.20) is again a pivotal statistic as it does neither depend on μ nor σ^2. One can then construct confidence intervals through

$$CI = \left[\overline{y} - t_{(n-1),1-\alpha/2}\frac{S_n}{\sqrt{n}}, \overline{y} + t_{(n-1),1-\alpha/2}\frac{S_n}{\sqrt{n}}\right], \tag{3.21}$$

where $t_{(n-1),\alpha}$ is the α quantile of the t distribution with $n-1$ degrees of freedom with $\alpha \in [0, 1]$. Since the estimated variance σ^2 adds uncertainty to estimating μ, we have that $t_{(n-1),1-\alpha/2} > z_{1-\alpha/2}$ for $0 < \alpha < 0.5$ making the confidence interval (3.21) with estimated variance wider than the one with given variance (3.18).

The proof of (3.20) is somewhat clumsy. Therefore, we want to provide a graphical motivation for why estimating the variance actually leads to increased variation. Let us therefore look at Fig. 3.5. We show the standard normal distribution in the top-left plot and zoom into the region of the $z_{1-\alpha/2}$ quantile for $\alpha = 0.05$ in the top-right plot. If the variance σ^2 is estimated by S_n^2, we induce additional uncertainty. We visualize this in Fig. 3.5 by a varying variance which takes a value (roughly) symmetrically to the left and right of the true $z_{1-\alpha/2}$ quantile (see the left plot in the middle row). One can see that the areas to the left and right are unequal, with the left area larger than the right, as evident in the distribution function calculated at the two symmetric points around the quantile (middle-right plot). Apparently, while there is symmetry on the horizontal axis, there is no symmetry on the vertical axis. If we instead postulate symmetry on the vertical axis, that is, for the coverage probability (bottom left plot), this leads to asymmetric quantiles, taking larger values to the right. Hence, we need to increase the $z_{1-\alpha/2}$ quantile to achieve $1-\alpha$ coverage probability. This motivates the use of a *t* distribution, as shown in the bottom-right plot. Note that for increasing *n* we get the standard normal distribution as the limit.

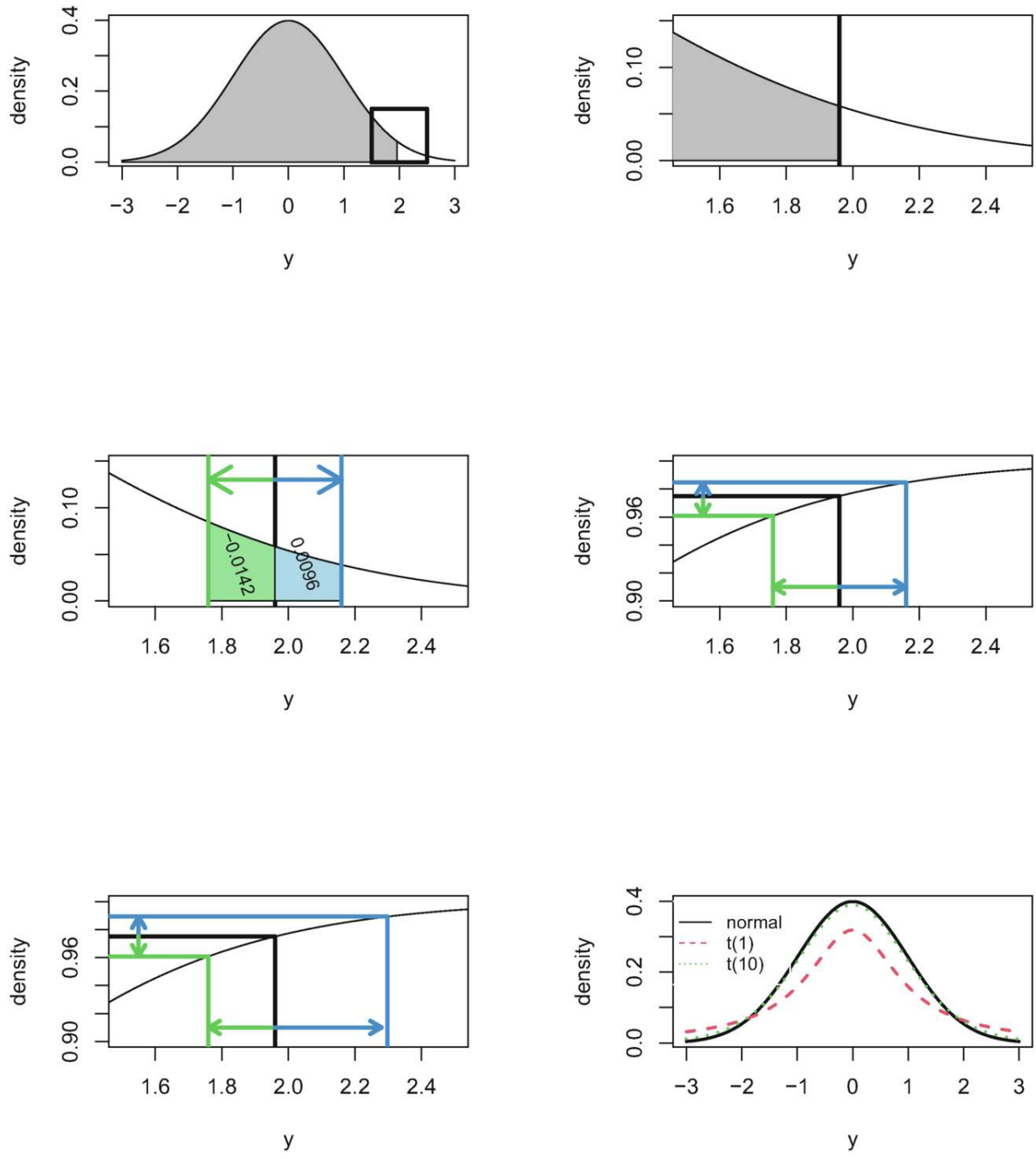


Fig. 3.5 Standard normal distribution with 95% quantile (top left) and zoom-in version (top right). If the variance is estimated unbiasedly, the quantile takes realized values roughly symmetric to the left and the right of the $1 - \alpha$ quantile of the normal distribution (middle left). This leads to asymmetric coverage probabilities (middle right). Postulating symmetry in the coverage probabilities (bottom left) induces asymmetric quantiles, with larger values occurring to the right. This motivates the role of the t distribution leading to larger quantiles (bottom right)

Before we close this section, we will discuss the interpretation of confidence intervals. Assume that we have calculated a confidence interval

with values $[t_l(y_1, \ldots, y_n), t_r(y_1, \ldots, y_n)]$. One might now assume that the confidence interval can be interpreted as the probability that the true unknown parameter is within the confidence interval:

$$P(\theta \in [t_l(y_1, \ldots, y_n), t_r(y_1, \ldots, y_n)]) = ? \tag{3.22}$$

To make a concrete example, assume that we have collected normally distributed data $y_1, \ldots y_n$ with $\overline{y} = 3.2$ and $S_n^2 = 4.2$ with $n = 20$. Based on the derivations from above, we obtain the 0.95 confidence interval through the $1 - \alpha \,/\, 2$ quantile of the *t* distribution with 19 degrees of freedom, which equals 2.093, so that the (rounded) confidence interval results to

$$[3.2 - \sqrt{4.2 \,/\, 20} \times 2.093, 3.2 + \sqrt{4.2 \,/\, 20} \times 2.093] \quad = [2.24, 4.16].$$

To interpret this, we might formulate (3.22) to $P(\mu \in [2.24, 4.16]) = ?$. Formally, this is not a question based on a probability model, since the data $y_1, \ldots, y_n$ are given and there is no random experiment. Put differently, the unknown parameter is either in the interval or not, and the probability statement (3.22) is not about a random process. The formal statement is therefore that we have a confidence level of $(1 - \alpha)$ 100 percent that the parameter is in the interval. Hence, the term “probability” is to be replaced by “confidence.”

Definition 3.11 (Confidence Principle) *If we repeat the calculation with new data of the same size multiple times, then in* $\alpha \times 100\,\%$ *of the cases the parameter would* **not** *be within the interval, and consequently, in* $(1 - \alpha) \times 100\,\%$ *of the cases, it would be.*

Though this is formally a correct description, it might seem unusual since it does not state anything about the concrete sample. Therefore, let us look at this from the perspective of uncertainty and remember the example from Sect. 2.1 with the dice cup not lifted. The situation here is very similar; the data are given, and we can quantify our uncertainty about the parameter θ. To do so, we may utilize probabilities. In this case, we could formulate our knowledge as follows: We believe the true parameter lies in the confidence interval with a probability of $(1 - \alpha) \times 100$ percent, bearing in mind that “probability” here does not refer to a proper random experiment but to a quantification of our knowledge. This is what is meant by “confidence,” that we have $(1 - \alpha) \times 100$ percent confidence.

3.5 Likelihood and Bayes

Now that we have covered properties of estimators, let us return to the likelihood function. The original work by Fisher (1912) introduced the concept of likelihood and catalyzed the adoption of likelihood-based approaches using Bayesian principles (see also Aldrich, 1997 or Edwards, 1997). However, the term "Bayesian statistics" had not yet been coined in that era. In fact, it took nearly 50 more years for this terminology to emerge, which was formerly known as "inverse probability," as motivated by Fienberg (2006). In essence, inverse probability involves a reversal of the typical reasoning process. Given a parameter θ, we construct a probability model $F(\cdot;\theta)$ for the variable Y. Conversely, the shift in perspective involves deriving the probability model for θ given that we have observed y. These are two different perspectives on uncertainty. In the first case, the data are treated as random, while the distribution's parameter is treated as fixed. Data are used to obtain information about this unknown parameter, leading to estimation. This approach is generally referred to as *frequentist inference*. In the inverse perspective, knowledge (or lack thereof) about the parameter is modeled as probability distributions. The data are used to update the knowledge about the parameter. This approach is known as *Bayesian inference*. Both approaches coexist in the field of statistics and are, at least in principle, conceptually different. However, we will see that both approaches, at least for large samples, lead to the same result.

Here, we consider the Bayesian perspective in more details. We argued before that probability can be used to express uncertainty, but it can also express knowledge. We proceed with the presumption that prior information concerning the parameter θ is available. The goal is then to effectively incorporate this into our analysis. This prior information expresses our preliminary beliefs regarding the parameter θ, and we quantify it through a probability distribution, called **prior distribution**, denoted as $p(\theta)$. The structure of this distribution may encompass additional parameters, which we ignore for the moment. Under the i.i.d. setup with data $y = (y_1, \ldots, y_n)$, we can update the prior distribution to another distribution, called the **posterior distribution**. We want to answer what information we gained from the data about our parameter θ. We achieve this by "inverting" the

distributions. Assuming a continuous random variable Y, this posterior distribution is defined as follows,

$$\begin{aligned} p(\theta|y) &= \frac{f(y;\theta)p(\theta)}{f(y)} \\ &= \frac{\prod_{i=1}^{n} f(y_i;\theta)p(\theta)}{f(y)} \\ &= \frac{\exp(l(\theta))p(\theta)}{f(y)}, \end{aligned} \tag{3.23}$$

where the first component in the numerator is the likelihood, that is, $\exp(\cdot)$ of the log-likelihood, $p(\theta)$ is the density of the prior, and $f(y)$ is the normalization constant

$$f(y) = \int_{\theta \in \Theta} \prod_{i=1}^{n} f(y_i;\theta)p(\theta)d\theta.$$

The latter is usually numerically intractable, which will be considered in Chap. 9 again. Since (3.23) defines an entire distribution, one can be interested in numerous quantities to characterize it, for instance, its mode.

Definition 3.12 *The***posterior mode***is defined through*

$$\hat{\theta}_{postmode} = \underset{\theta \in \Theta}{arg\,max}\ p(\theta|y).$$

The posterior mode only depends on the numerator in (3.23) and raises an optimization problem similar to what we covered on maximum likelihood estimation. In fact, choosing $p(\theta)$ to be a constant, implying a so-called flat prior, the posterior mode matches the maximum likelihood estimator, that is,

$$\underset{\theta \in \Theta}{\text{argmax}}\ p(\theta|y) = \underset{\theta \in \Theta}{\text{argmax}}\ l(\theta). \tag{3.24}$$

If the maximum likelihood estimator is, however, different than the mode of the prior, this is not the case.

Instead of calculating a single value for the parameter, like the posterior mode, we can derive intervals comparable to confidence intervals defined above. These can be constructed using a similar interpretation, but through Bayesian reasoning. In this case, though, confidence intervals are formally introduced as *credibility intervals* in the literature. In Bayesian statistics, we consider the parameter θ to be random so that we can explicitly look at the posterior distribution.

Definition 3.13 *A* **credibility interval** *for parameter θ is defined as follows. Our knowledge about the parameter θ is given by the posterior distribution $p(\theta \mid y_1, \ldots, y_n)$. The credibility interval $[t_l(y), t_r(y)]$ is implicitly defined as*

$$P\Big(\theta \in \big[t_l(y_1, \ldots, y_n), t_r(y_1, \ldots, y_n)\big] \mid (y_1, \ldots, y_n)\Big)$$
$$= \int_{t_l(y_1,\ldots,y_n)}^{t_r(y_1,\ldots,y_n)} p(\theta \mid y_1, \ldots, y_n) d\theta \geq 1 - \alpha.$$

A natural choice is to set $t_l()$ and $t_r()$ such that

$$\int_{-\infty}^{t_l(y_1,\ldots,y_n)} p(\theta \mid y_1, \ldots, y_n) d\theta = \int_{t_r(y_1,\ldots,y_n)}^{\infty} p(\theta \mid y_1, \ldots, y_n) \mathrm{d}\theta = \frac{\alpha}{2}, \quad (3.25)$$

that is, we cut off a probability mass of $\alpha \,/\, 2$ in the left and right tails of the posterior probability. We visualize such a construction in the top plot of Fig. 3.6. While the construction is simple, the credibility interval is not optimal. If we look at the left-hand side of the interval in Fig. 3.6, we see that the density at the boundary is higher than the density at the right-hand side of the interval. Hence, we include values with lower density in the interval and exclude values with higher density. More formally, it occurs that for $\theta_1 \notin [t_l(y_1, \ldots, y_n), t_r(y_1, \ldots, y_n)]$ and for $\theta_2 \in [t_l(y_1, \ldots, y_n), t_r(y_1, \ldots, y_n)]$ one has

$$p(\theta_1 \mid y_1, \ldots, y_n) > p(\theta_2 \mid y_1, \ldots, y_n).$$

This drawback can be avoided by using a **highest posterior density** (HDI) credibility interval or, in short, the highest density interval. The construction principle is shown in Fig. 3.6. We lower the horizontal line until the interval obtained through cut points with the density has the postulated credibility $(1 - \alpha)$. Mathematically, this is written down as

$$HDI(y_1, \ldots, y_n) = \{\theta; p(\theta \mid y_1, \ldots, y_n) \geq c\},$$

where *c* is chosen such that

$$\int_{\theta \in HDI(y_1,\ldots,y_n)} p\big(\theta \mid (y_1, \ldots, y_n)\big) \mathrm{d}\ \theta = 1 - \alpha.$$

That is, we choose an interval where values are greater than a threshold density *c*, such that the region integrates to $1 - \alpha$.

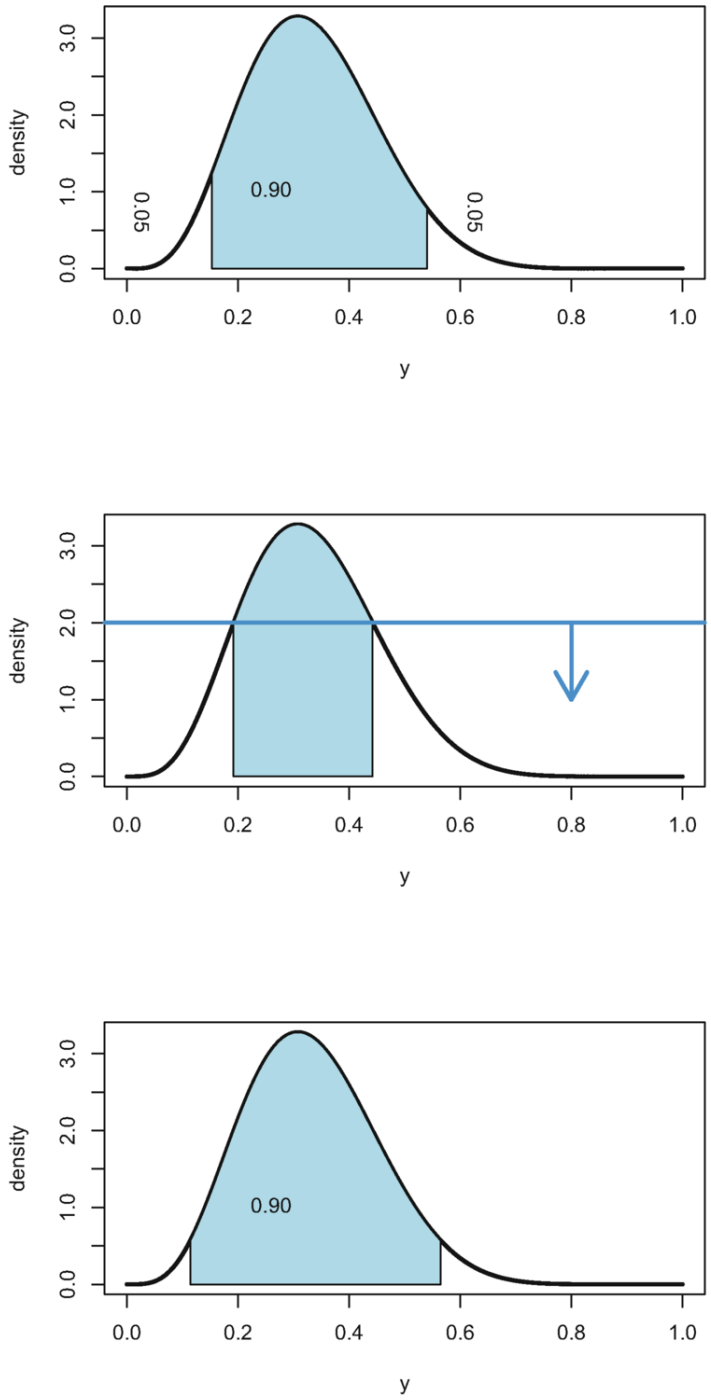


Fig. 3.6 Credibility interval, where we cut off to the left and the highest-density credibility interval for binomial data with $n = 5$ and $y = 0$ (left) and $n = 5$ and $y = 2$ (right)

Besides deriving the posterior mode or credibility intervals, one can also look directly at the posterior distribution (3.23) itself. This is numerically burdensome due to the component $f(y)$, which is typically numerically intractable. We will extensively explore how to circumvent this issue in Chap. 9.

3.6 Exercises

Exercise 1 For independently identically distributed random variables Y_i with mean value μ and variance σ^2, show that

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2$$

is unbiased. Hint: Rewrite the summands to $(Y_i - \mu) - (\overline{Y} - \mu)$.

Exercise 2

(a) Give the definition of the Kullback-Leibler divergence, and show why $KL(g(\cdot), f(\cdot)) \neq KL(f(\cdot), g(\cdot))$, except when $f(\cdot) = g(\cdot)$.

(b) Under real-world conditions, the true data-generating process $G(\cdot)$ is rarely known. Assume we have a sample $x = (2, 2, 3, 2, 3, 0, 2, 1, 4, 3)^\top$, but we don't know the underlying data-generating process, so we make the assumption $X_i \sim \text{Bin}(4, 0.5)$. Calculate the Kullback-Leibler divergence between the empirical distribution of the sample, $g(\cdot)$, and the assumed distribution, $f(\cdot)$.

Exercise 3 Let $X = (X_1, X_2, \ldots, X_n)^\top$ be an i.i.d. sample. The X_i are Poisson distributed with unknown parameter $\lambda > 0$ for $i = 1, \ldots, n$:

(a) Show that the statistic $t(X) = \sum_{i=1}^{n} X_i$ is sufficient for λ by directly computing the conditional density of X given $t(X) = t_0$.

(b) Show that the statistic $t(X) = \sum_{i=1}^{n} X_i$ is sufficient for λ by using Neyman factorization.

(c) Are the statistics $\overline{X}$, the arithmetic mean, and $\overline{X}^{-1}$ sufficient for λ as

well?

(d) Are the statistics t and $\overline{X}$ unbiased for λ? Are they consistent for λ?

Exercise 4 A company ships tea in wooden crates, each containing 10 tea packages. The weight of the individual tea packages is normally distributed with $\mu = 6$ kg and $\sigma = 0.06$ kg. The weight of the empty wooden crate is normally distributed with $\mu = 5$ kg and $\sigma = 0.05$ kg.

(a) Under assuming independence of the individual delivery components, specify an interval symmetrical to the expected value in which the gross weight of the ready-to-dispatch wooden crate lies in 95% of the cases.

(b) A customer of the above tea company checks whether the tea packages really comply with the specified target value of 6 kg. For this purpose, he takes a sample of $n = 16$ tea packages from a delivery. The arithmetic mean was found to be 5.95 kg. Calculate a 95% confidence interval for the average weight of the individual tea packages in the delivery, and interpret the result.

References

Aldrich, J. (1997). R.A. Fisher and the making of maximum likelihood 1912–1922. *Statistical Science,12*(3), 162–176.
[MathSciNet][https://doi.org/10.1214/ss/1030037906]

Boland, P. J. (1984). A biographical glimpse of William Sealy Gosset. *The American Statistician,38*(3), 179–183.
[https://doi.org/10.1080/00031305.1984.10483195]

Edwards, A. W. (1997). What did fisher mean by "inverse probability" in 1912–1922? *Statistical Science,12*(3), 177–184.
[MathSciNet][https://doi.org/10.1214/ss/1030037907]

Fienberg, S. E. (2006). When did Bayesian inference become "Bayesian"? *Bayesian Analysis,1*(1), 1–40.
[MathSciNet][https://doi.org/10.1214/06-BA101]

Fisher, R. A. (1912). On an absolute criterion for fitting frequency curves. *Messenger of Mathematics,41*, 155–156.

Student (1908). The probable error of a mean. *Biometrika,6*(1), 1–25. [https://doi.org/10.2307/2331554]

Ziliak, S. T. (2008). Retrospectives: Guinnessometrics: the economic foundation of "student's" t. *Journal of Economic Perspectives,22*(4), 199–216. [https://doi.org/10.1257/jep.22.4.199]

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
G. Kauermann et al., *Statistics—Coping with Uncertainty*, Textbooks in Statistical Science
https://doi.org/10.1007/978-3-032-33658-3_4

4. Maximum Likelihood Estimation

Göran Kauermann[1] ✉, Giacomo De Nicola[2], Cornelius Fritz[3] and David Rügamer[4]
(1) Department of Statistics, LMU Munich, Munich, Germany
(2) Harvard T.H. Chan School of Public Health, Harvard University, Boston, MA, USA
(3) School for Computer Science and Statistics, Trinity College Dublin, Dublin, Ireland
(4) Munich Uncertainty Quantification AI Lab, LMU Munich, Munich, Germany

Building on the introduction likelihood estimation in Chap. 3, we here present key mathematical results related to maximum likelihood estimation. Under a data-generating process that characterizes the stochastic variation of some outcome of interest via a distributional assumption and some parameters θ, we next want to find an estimate of θ for given observations $y_1, \ldots, y_n$. There are multiple ways to achieve this goal, for instance, maximum likelihood estimation, the method of moments estimator, or general M estimation. Given these options, let's first state desirable properties of estimators: One would prefer an estimator that is correct in expectation, has known uncertainty, follows a known distribution so that confidence intervals can be derived, and is as robust as possible. In this context, robustness means that the estimator should vary as little as possible under repeated observations of the data.

In this chapter, we focus on maximum likelihood estimation and show in Sects. 4.1 to 4.3 that it enjoys three fundamental properties that relate to the desiderata mentioned above and underscore the widespread use of maximum likelihood estimators in applied statistics: First, we show that the error between the maximum likelihood estimator and the true data-generating parameter converges to zero as the number of observations increases. Second, we derive that under repeated sampling the distribution of the maximum likelihood estimator converges to a Gaussian distribution. The Fisher information gives the variance of the estimator. Third, we introduce the Cramér-Rao bound as the minimal achievable variance of an unbiased estimator and demonstrate that maximum likelihood

estimators attain it. In summary, we show that the maximum likelihood estimator is consistent, asymptotically normal, and optimal under unbiased estimators.

In Sect. 4.4, we examine the likelihood ratio and its distribution, a concept revisited in Chap. 5 that helps specify formal hypothesis tests. The likelihood function is derived from a distributional assumption, but what happens if this assumption does not hold? In such settings, the specified distributional model does not include the true, unknown data-generating process. As discussed in Chap. 3, the Kullback-Leibler divergence provides a framework for deriving more general results, showing that maximum likelihood estimation remains a robust approach even when models are misspecified. We highlight this characteristic further in Sect. 4.6. We conclude with Sect. 4.8 by emphasizing the likelihood principle, asserting that all information about the parameter is contained in the data, a principle that holds even when data are only partially observed.

4.1 Properties of the Likelihood

Going beyond the material covered in Sect. 3.1, we here derive useful properties of the maximum likelihood estimator. Like above, we thereby assume the data-generating process (3.1), that is,

$$Y_i \sim G(y), \text{ i.i.d. with } i = 1, \dots n. \tag{4.1}$$

We aim to approximate this data-generating process with the parametric model

$$\mathscr{F} = \{F(\cdot; \theta), \theta \in \Theta\}.$$

We assume that some basic mathematical operations are valid on $\mathscr{F}$ and require that the statistical model $\mathscr{F}$ is Fisher regular as formally defined in Appendix B.3. Fisher regularity implies, among other things, that integration with respect to the random variable and differentiation with respect to the parameter are exchangeable.

We have derived in Sect. 3.1 that the closest distribution in $\mathscr{F}$ to the true data-generating process can be obtained through minimizing the Kullback-Leibler divergence. Thereby, we define the "true" parameter θ_0 implicitly through

$$E\left(\frac{\partial \log f(y;\theta_0)}{\partial \theta}\right) = \int \frac{\partial \log f(y;\theta_0)}{\partial \theta} \mathrm{d}\, G(y) = 0. \tag{4.2}$$

Note that $\frac{\partial \log f(y;\theta_0)}{\partial \theta}$ is short for $\frac{\partial \log f(y;\theta)}{\partial \theta}\big|_{\theta=\theta_0}$. We will rewrite this as a property of the log likelihood function, which we define as

$$l(\theta) = \sum_{i=1}^{n} \log f(y_i \, \theta).$$

Sometimes, we write $l(\theta; y)$ to stress that the likelihood depends on the data. Treating the data as random implies that the likelihood function itself can be

considered random. In this case, we have $l(\theta) = \sum_{i=1}^{n} \log f(Y_i; \theta)$, where the observed data y_i are replaced by the respective random variables Y_i for $i = 1, \ldots, m$. We do not differentiate between the two versions in notation, but, conceptually, these are different. With this in mind, we can reformulate (4.2) to

$$E\left(\frac{\partial l(\theta_0)}{\partial \theta}\right) = 0. \tag{4.3}$$

We term the first-order derivative of the likelihood function the **score function**, and (4.3) says that the expected value of the score function is zero for the "true" parameter θ_0.

The implicit definition for θ_0 stems from the general data-generating process, wherein the expectation in Eq. (4.3) is determined by using the distribution $G(\cdot)$. If we consider the scenario where $G(\cdot)$ belongs to $\mathscr{F}$, i.e., assume that the actual process generating the data is one among the set of distributions $\mathscr{F}$, interesting simplifications arise. In this case, there exists a unique parameter θ_0 such that $G(\cdot) \equiv F(\cdot; \theta_0)$. As before, the parameter θ_0 denotes the true parameter. Identity (4.3) has a different justification in this case, which we want to motivate next. Note that for all parameters $\theta \in \Theta$ we have that $F(\cdot; \theta)$ is a proper distribution, which means that the density integrates out to one:

$$1 = \int f(y; \theta) \mathrm{d}\, y \tag{4.4}$$

This holds for all elements in $\mathscr{F}$, that is, for all $\theta \in \Theta$. Differentiating both sides of Eq. (4.4) with respect to θ and making use of the fact that integration and differentiation are exchangeable gives

$$0 = \frac{\partial 1}{\partial \theta} = \int \frac{\partial f(y; \theta)}{\partial \theta} \mathrm{d}\, y = \int \frac{\partial f(y; \theta)}{\partial \theta} \frac{f(y; \theta)}{f(y; \theta)} \mathrm{d}\, y$$

$$[2ex] = \int \underbrace{\frac{\partial \log f(y; \theta)}{\partial \theta}}_{s(\theta; y)} f(y; \theta) \mathrm{d}\, y,$$

where $s(\theta; y) = \dfrac{\partial}{\partial \theta} \log f(y; \theta)$ is the score function and the result $\frac{\partial \log(f(x))}{\partial x} = \frac{\partial f(x)}{\partial x} \frac{1}{f(x)}$ is used for the equality in the second line. This shows that the score function has mean zero for the true parameter, i.e.,

$$E\Big(s\big(\theta; Y\big)\Big) = 0, \tag{4.5}$$

where we dropped the index 0 at θ to simplify notation. Note that the expectation is now calculated with respect to density $f(\cdot; \theta)$, that is, the parameter in the density and the parameter in the score are the same. The statement in Eq. (4.5) is called the **first Bartlett identity**. The interpretation of (4.5) is that although the

position of the maximum of the likelihood function may vary, on average, there is a peak in the likelihood function at the true parameter, since the expected slope is zero.

Let us further differentiate both sides of Eq. (4.5) with respect to θ. Allowing thereby that θ is a multivariate (column)vector, we obtain by incorporating the transpose operator

$$
\begin{aligned}
0 = \frac{\partial 0}{\partial \theta^\top} &= \frac{\partial}{\partial \theta^\top} \int \frac{\partial}{\partial \theta} \log f(y;\theta) f(y;\theta) \mathrm{d}\, y \\
[2ex] &= \int \left(\frac{\partial^2}{\partial \theta \partial \theta^\top} \log f(y;\theta) \right) f(y;\theta) \mathrm{d}\, y + \int \frac{\partial \log f(y;\theta)}{\partial \theta} \frac{\partial f(y;\theta)}{\partial \theta^\top} \mathrm{d}\, y \\
&= E\left(\frac{\partial^2}{\partial \theta \partial \theta^\top} \log f(Y;\theta) \right) + \int \frac{\partial \log f(y;\theta)}{\partial \theta} \frac{\partial \log f(y;\theta)}{\partial \theta^\top} f(y;\theta) \mathrm{d}\, y,
\end{aligned}
$$

which is equivalent to

$$
\begin{aligned}
E\Big(s(\theta;Y)(s(\theta;Y))^\top\Big) &= E\Big(-\tfrac{\partial^2}{\partial\theta\partial\theta^\top} \log f(Y;\theta)\Big) \\
E\Big(s(\theta;Y)(s(\theta;Y))^\top\Big) - E\Big(s(\theta;Y)\Big)^2 &= E\Big(-\tfrac{\partial^2}{\partial\theta\partial\theta^\top} \log f(Y;\theta)\Big) \qquad (4.6) \\
\mathrm{Var}\Big(s(\theta;Y)\Big) &= E\Big(-\tfrac{\partial^2}{\partial\theta\partial\theta^\top} \log f(Y;\theta)\Big),
\end{aligned}
$$

where the second row holds due to $E\Big(s(\theta;Y)\Big) = 0$ and the third row is called the **second Bartlett identity**. Note that this is a matrix-valued component if θ is vector-valued. As before, we point out that the expectation is taken over the random variables $Y_i \sim f(\cdot;\theta)$ for $i = 1, \ldots, n$ so that the parameter in the density and the parameter in the arguments of the expectation and variance in (4.6) are the same. We define the expectation on the right-hand side of Eq. (4.6) as the Fisher information matrix, which is sometimes also called Fisher matrix or Fisher information.

As its name suggests, the maximum likelihood estimate maximizes the (log) likelihood function. Because the score function has a nonnegative variance, it follows from (4.6) that the second-order derivative is, in expectation, negative. This, in turn, guarantees that maximizing the log likelihood with respect to the coefficient is a well-defined optimization problem, since the likelihood function is concave and hence has a unique maximum. This property holds in expectation only. For a concrete sample, we might encounter local maxima. The Fisher information plays a fundamental role, as it indicates whether the maximization problem is well-defined.

The term “information” can thereby be understood literally, and we can show that the Fisher information grows in the same order as the sample size. We want to demonstrate this with a simple example. Assume that data y_i are normally distributed, and assume for simplicity that the variance is known and equal to one. In Fig. 4.1, we plot the resulting likelihood for three samples, one with sample size $n = 5$ (black outer line), one with sample size $n = 50$ (blue middle line), and one with sample size $n = 500$ (purple inner line). We see that the maximum gets more exposed as the sample size grows. Mathematically, we can describe this with the curvature at the peak. The curvature is thereby the second-order derivative, which in turn is (in expectation) the Fisher information. The larger the database, the larger the information. We combine all the above in the following definitions.

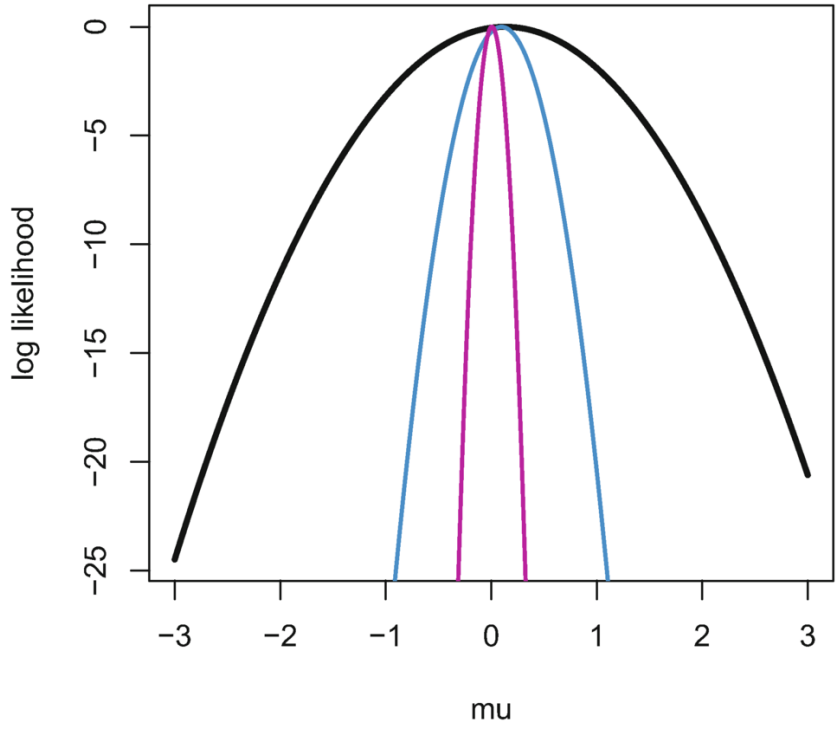


Fig. 4.1 Log-likelihood function for μ for normally distributed data with sample sizes $n = 5$ (black outer line), $n = 50$ (blue middle line), and $n = 500$ (purple inner line)

Definition 4.1 *The first derivative of the log-likelihood function is called the***score function**

$$s\big(\theta; y_1, \ldots, y_n\big) = \frac{\partial l(\theta; y_1, \ldots, y_n)}{\partial \theta}. \tag{4.7}$$

*Differentiating the score function and taking negative expectation yield the***Fisher information (matrix)**

$$I(\theta) = -E\left(\frac{\partial s(\theta; Y_1, \ldots, Y_n)}{\partial \theta^\top}\right) = -E\left(\frac{\partial^2 l(\theta; Y_1, \ldots, Y_n)}{\partial \theta \partial \theta^\top}\right).$$

If the data-generating process is within the model class, we obtain

$$Var\Big(s(\theta;Y)\Big) = I(\theta). \tag{4.8}$$

With these prerequisites, we can give a formal definition of the maximum likelihood estimator.

Definition 4.2 *For a random sample* $y_1, \dots y_n$*, the* **maximum likelihood** *estimator is defined by*

$$\hat{\theta} = \text{argmax } l(\theta) = \text{argmax } l(\theta; y_1, \dots, y_n),$$

which, for Fisher regular distributions, occurs when

$$s\Big(\hat{\theta}\Big) = s\Big(\hat{\theta}; y_1, \dots, y_n\Big) = 0.$$

For distributions that are not Fisher regular, the maximum likelihood estimator is not necessarily estimable through differentiation. The following example shows such a case.

Example 4.1 Assume that Y_i are $i.i.d.$ uniformly distributed between 0 and θ, that is, $Y_i \sim \text{Uniform}[0;\theta]$ with the following density,

$$f(y_i;\theta) = \begin{cases} \frac{1}{\theta} & \text{for } 0 \leq y_i \leq \theta \\ 0 & \text{otherwise} \end{cases},$$

and likelihood function

$$l(\theta) = \begin{cases} \left(\frac{1}{\theta}\right)^n & \text{for } \theta \geq \max(y_1, \dots, y_n) \\ 0 & \text{otherwise.} \end{cases}.$$

The log likelihood apparently results from

$$l(\theta) = \begin{cases} n \log\left(\frac{1}{\theta}\right) & \text{for } \theta \geq max(y_1, \dots, y_n) \\ -\infty & \text{otherwise.} \end{cases}.$$

We see that the likelihood function is not continuous and in fact takes the form as shown in Fig. 4.2. The maximum is achieved at the value $y_{max} = \max(y_1, \dots, y_n)$.

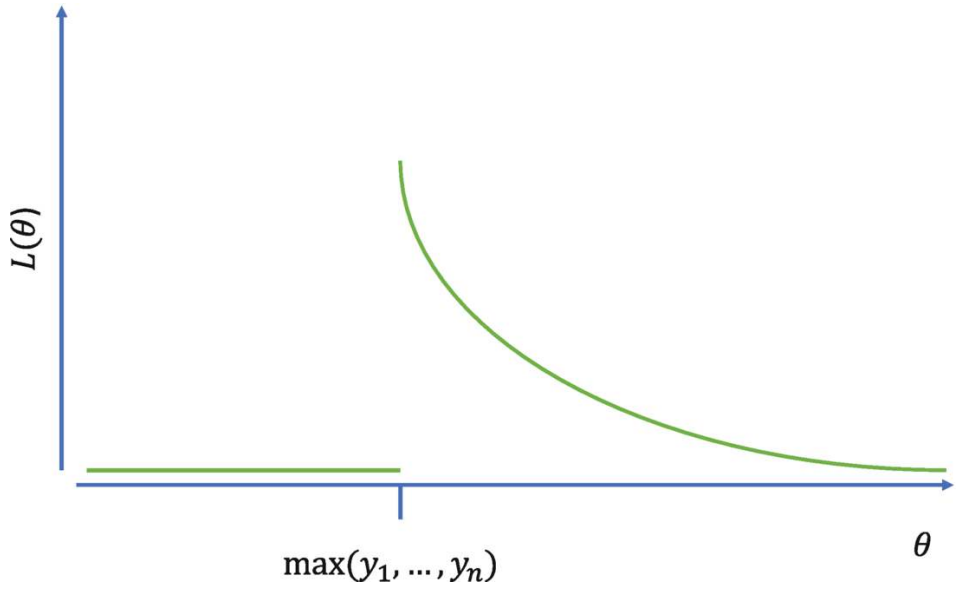


Fig. 4.2 Likelihood function for a uniform distribution on $[0, \theta]$

We can further investigate a potential bias in the estimator, that is, we aim to calculate $E(Y_{max})$, where Y_{max} is the random variable corresponding to y_{max}. Note that

$$P(Y_{max} \leq y) = P((Y_1 \leq y) \wedge (Y_2 \leq y) \wedge \ldots \wedge (Y_n \leq y)) = (F(y))^n$$

holds since every random variable $Y_1, \ldots, Y_n$ has to be smaller than some *y* for the corresponding maximum to be smaller than *y*. Hence, the density of the random variable $Y_{max} = \max\{Y_1, \ldots, Y_n\}$ for any *i.i.d.* random variables $Y_1, \ldots, Y_n$ is

$$f_{Y_{max}}(y) = \frac{\partial (F(y))^n}{\partial y} = n\ F^{n-1}(y) f(y).$$

For $Y_i \sim \text{Uniform}(0; \theta_0)$ with θ_0 as the true parameter, the density turns out to be

$$f_{Y_{max}}(y) = \begin{cases} n\left(\frac{y}{\theta_0}\right)^{n-1} \frac{1}{\theta_0} & \text{for } y \leq \theta_0 \\ 0 & \text{otherwise} \end{cases}.$$

Calculating the expected value leads to

$$E(Y_{max}) = \int_0^{\theta_0} y\ n \frac{y^{n-1}}{\theta_0^n}\ \mathrm{d}\ y = \int_0^{\theta_0} n \frac{y^n}{\theta_0^n}\ \mathrm{d}\ y = \frac{n}{n+1} \frac{y^{n+1}}{\theta_0^n}\Big|_0^{\theta_0} = \frac{n}{n+1}\theta_0.$$

We see that the estimator is biased downward and we have $E(Y_{max}) < \theta_0$. ▷

4.2 Consistency and Asymptotic Normality

The likelihood function itself is random, as it is a function of realizations of a stochastic data-generating process. To illustrate, consider a simple example.

Example 4.2 Assuming $Y_i \sim \mathrm{B}(1, \pi)$ $i.i.d.$, $i = 1, \ldots, n$ with unknown parameter $\pi = P(Y_i = 1)$, we set to $\pi_0 = 0.3$ in simulations. In this setup, it is straightforward to derive that the maximum likelihood estimator for a sample $y_1, \ldots, y_n$ is given by $\widehat{\pi} = \overline{y} = 1 \Big/ n \sum_{i=1}^{n}$ (see Example 4.4). We vary the number of observations *n* between 5 and 150, repeat the sampling and estimation under a particular configuration 1000 times, and then plot the absolute error of the estimators $\widehat{\pi}$ in Fig. 4.3. As the number of observations *n* grows, the absolute error vanishes. ▷

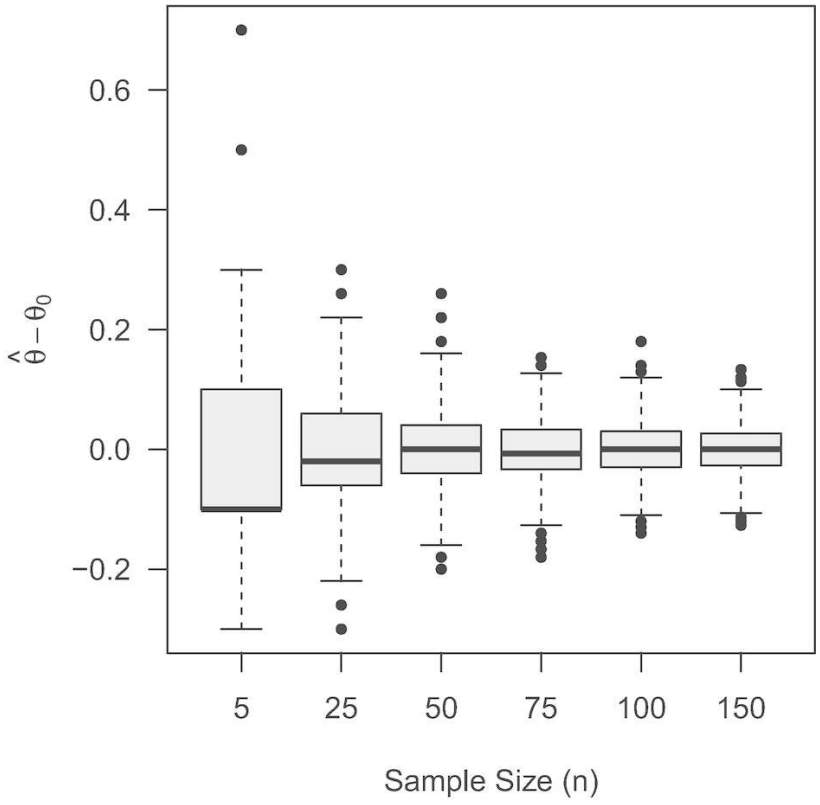


Fig. 4.3 Boxplots showing the absolute error of maximum likelihood estimators for a binomial distribution with $\pi_0 = 0.3$ for $n = 5, \ldots, 150$ under 1000 repetitions

This finding suggests that this maximum likelihood estimator is weakly consistent (see Definition 3.5),

$$P(|\hat{\pi} - \pi| \geq \delta) \xrightarrow[n \to \infty]{} 0,$$

since the absolute error gets arbitrarily small for large *n*. This phenomenon is not particular to Example 4.2 but holds generally for maximum likelihood estimators.

Theorem 4.1 *The maximum estimator $\hat{\theta}$ is weakly consistent for θ_0 if the data size used for the estimate goes to ∞.*

$$\hat{\theta} \xrightarrow{p} \theta_0 \quad \text{as } n \to \infty$$

We sketch a proof of Theorem 4.1 at the end of this section.

Example 4.3 Continuing Example 4.2, we show a sample of possibly observed log-likelihood functions in Fig. 4.4 for $n = 10$ (left-hand side) and $n = 100$ (right-hand side). For $n = 10$, there are 11 possible log-likelihood functions according to the possible outcome of $Y = \sum_{i=1}^{10} Y_i \in \{0, 1, \ldots, 10\}$. The width of the functions plotted in Fig. 4.4 is proportional to their occurrence probability. The same holds for $n = 100$ on the right-hand side, but now there are 101 possible log-likelihood functions. The top plots show the possible maxima of the log-likelihood functions and the probability of each. The true value $\pi_0 = 0.3$ is indicated as a vertical line. Apparently, the log-likelihood function randomly varies under resampling the data, supporting that it is a random variable. The maximum value of the log-likelihood function also occurs in the vicinity of the true parameter π_0. Finally, with increasing sample size, the maximum likelihood estimator appears to follow a normal distribution. This observation can, in fact, be formally proved, and we conclude with the following result. ▷

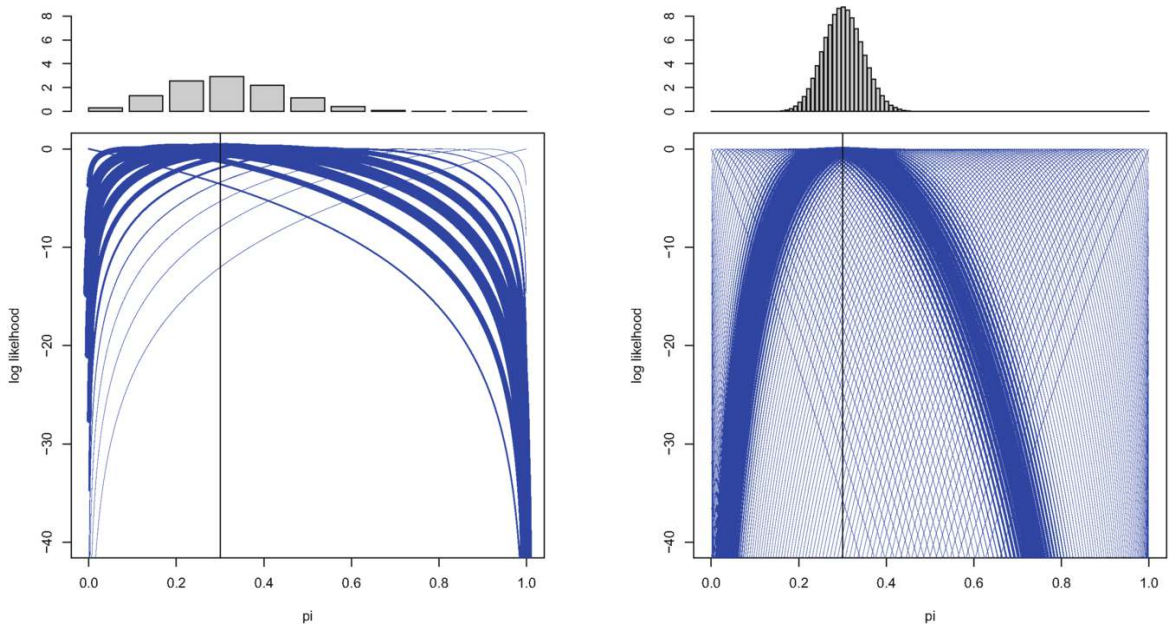


Fig. 4.4 Log-likelihood functions for a binomial distribution ($\pi_0 = 0.3$). The top row represents $n = 10$, and the bottom row represents $n = 100$. Lines are weighted by their probability. Note the concentration of the maximum likelihood estimates $\widehat{\pi}$ around the true parameter as *n* increases, illustrating weak consistency

Theorem 4.2 *Assume a Fisher regular distribution with parameter θ_0 from which an i.i.d. sample $Y_1, \ldots, Y_n$ is drawn. The maximum likelihood estimator is asymptotically normally distributed with*

$$\hat{\theta} \overset{a}{\sim} N\big(\theta_0, I^{-1}(\theta_0)\big). \tag{4.9}$$

A formal proof of this statement is lengthy; we, therefore, only sketch the main ideas at the end of this section. Before doing so, we emphasize that (4.9) is *the* central result in maximum likelihood theory and thus of fundamental importance for inferring properties of $\hat{\theta}$ under resampling. Knowing the behavior of the estimator $\hat{\theta}$, let us derive confidence intervals for the maximum likelihood estimator for a reasonably large sample size n, since a pivotal distribution is directly available from (4.9). This holds since the Fisher matrix $I(\theta)$ can usually be calculated quite easily or at least be approximated through simulations. We will get back to the latter point in Chap. 8.

Apparently, the Fisher matrix depends on the true parameter, which is unknown in practice. However, this difficulty can be overcome by substituting the parameter with its estimator, yielding $I\big(\hat{\theta}\big)$. Provided that the estimate $\hat{\theta}$ concentrates around θ_0 for large n (see Theorem 4.1), this difficulty can be overcome by substituting the parameter with its estimator, yielding $I\big(\hat{\theta}\big)$. The inverse $I^{-1}\big(\hat{\theta}\big)$ is then taken as an estimate for the variance matrix of the maximum likelihood estimator, that is, $\widehat{Var}\big(\hat{\theta}\big) = I^{-1}\big(\hat{\theta}\big)$. Writing $\hat{\theta} = (\hat{\theta}_k)$ and $\theta_0 = (\theta_{0,k})$ and denoting the kth diagonal entry of any matrix $X \in \mathbb{R}^{K \times K}$ with $K > k$ by $X_{k,k}$, this in turn means that for the k-th component of $\hat{\theta}$, we obtain

$$\frac{(\hat{\theta}_k - \theta_{0,k})}{\sqrt{\widehat{\mathrm{Var}}(\hat{\theta})_{k,k}}} \overset{a}{\sim} \mathrm{N}(0,1). \tag{4.10}$$

The following example demonstrates the above results.

Example 4.4 Assume that $Y \sim \mathrm{B}(n, \pi)$. The log likelihood results to

$$l(\pi) = y \log(\pi) + (n - y) \log(1 - \pi),$$

and the derivative leads to the score function

$$\frac{\partial l(\pi)}{\partial \pi} = \frac{y}{\pi} - \frac{n-y}{1-\pi} = 0. \tag{4.11}$$

Taking expectation, we see

$$E\left(\frac{\partial l(\pi)}{\partial \pi}\right) = \frac{n\pi}{\pi} - \frac{n - n\pi}{1-\pi} = \frac{(1-\pi)n\pi - n\pi + n\pi^2}{\pi(1-\pi)} = 0.$$

Setting the score (4.11) to zero provides the maximum likelihood estimate $\widehat{\pi} = \dfrac{y}{n}$.

Taking the second-order derivative gives

$$\frac{\partial^2 l(\pi)}{\partial \pi \partial \pi} = -\frac{y}{\pi^2} + \frac{n-y}{(1-\pi)^2}.$$

Calculating the expectation of the term above leads to the Fisher information

$$\begin{aligned} I(\pi) &= E\left(-\frac{\partial^2 l(\pi)}{\partial \pi \partial \pi}\right) \\ &= \frac{n\pi}{\pi^2} + \frac{n-n\pi}{(1-\pi)^2} \\ &= \frac{n\pi(1-\pi)^2 + n\pi^2(1-\pi)}{\pi^2(1-\pi)^2} \\ &= \frac{n(1-\pi)+n\pi}{\pi(1-\pi)} \\ &= \frac{n}{\pi(1-\pi)}. \end{aligned}$$

Hence, the maximum likelihood estimator $\widehat{\pi} = \dfrac{y}{n}$ follows a normal distribution with mean π_0 and variance $\dfrac{\widehat{\pi}(1-\widehat{\pi})}{n}$. ▷

We now provide the proofs for the statements above: First, we sketch the proof of Theorem 4.1.

Proof We denote the entire sample $(y_1, \ldots, y_n)$ with the corresponding random variables $(Y_1, \ldots, Y_n)$ by y and Y, respectively. By definition, the maximum likelihood estimator $\hat{\theta}$ is the argument maximizing the likelihood $l(\theta; y)$,

$$\hat{\theta} = \arg\max_{\theta \in \Theta} l(\theta; y) = \arg\max_{\theta \in \Theta} \tfrac{1}{n} l(\theta; y),$$

meaning that $\hat{\theta}$ also maximizes the likelihood scaled with $1 \,/\, n$. Using the law of large numbers, it becomes evident that

$$\tfrac{1}{n} l(\theta; y) \to E(l(\theta; Y)). \tag{4.12}$$

Under regularity assumptions,[1] the parameter maximizing the right-hand side of (4.12) then also converges in probability the parameter maximizing the left-hand side. Finally, we need to show that the maximizer of the right-hand side is the true parameter θ_0. Noting $l(\theta; y_1, \ldots, y_n) = \log f(y_1, \ldots, y_n; \theta)$, we can write

$$E(\log f(Y; \theta)) - E(\log f(Y; \theta_0)) = E\left(\log\left(\tfrac{f(Y;\theta)}{f(Y;\theta_0)}\right)\right)$$

for any $\theta \in \Theta$. Since the logarithm is a concave function, applying Jensen's inequality yields

$$\begin{aligned} E\left(\log\left(\tfrac{f(Y;\theta)}{f(Y;\theta_0)}\right)\right) \leq \log E\left(\tfrac{f(Y;\theta)}{f(Y;\theta_0)}\right) &= \log \int \tfrac{f(y;\theta)}{f(y;\theta_0)} f(y;\theta_0) \mathrm{d}\, y \\ &= \log \int f(y;\theta) \mathrm{d}\, y = \log(1) = 0, \end{aligned} \tag{4.13}$$

where the first inequality uses that the true data-generating parameter of Y is $f(y;\theta_0)$ and the second equality follows since $f(y;\theta)$ is a density for all $\theta \in \Theta$, which integrates to 1 by the probability axioms of Kolmogorov. Finally, we can rearrange (4.13) for any $\theta \in \Theta$ as follows,

$$E(l(\theta, Y)) \leq E(l(\theta_0, Y)),$$

showing that θ_0 is the argument maximizing $E(l(\theta, Y))$ over $\theta \in \Theta$. We have thus shown that $\hat{\theta} \xrightarrow{d} \theta_0$, proving the weak consistency of maximum likelihood estimators. □

Second, the proof of Theorem 4.2:

Proof We now want to sketch the proof of the asymptotic normality of the maximum likelihood estimator. Note that the maximum likelihood estimator is defined through

$$0 = s(\hat{\theta}).$$

Applying a zeroth-order Taylor series expansion around θ_0 and writing the rest in the Lagrange form lead to

$$0 = s(\hat{\theta}) = s(\theta_0) + \frac{\partial s(\tilde{\theta})}{\partial \theta^\top}(\hat{\theta} - \theta_0), \tag{4.14}$$

where $\tilde{\theta}$ lies on the line between $\hat{\theta}$ and θ_0. Solving (4.14) for $\hat{\theta} - \theta_0$ gives

$$\hat{\theta} - \theta_0 = -\left[\frac{\partial s(\tilde{\theta})}{\partial \theta^\top}\right]^{-1} s(\theta_0). \tag{4.15}$$

Note that the score can be written as

$$s(\theta_0) = \sum_{i=1}^{n} \underbrace{\frac{\partial \log f(Y_i;\theta_0)}{\partial \theta}}_{=:s_i(\theta_0)},$$

where $s_i(\theta_0)$ are independent identically distributed random variables with mean value 0. The central limit theorem and (4.8) yield

$$s(\theta_0) \overset{a}{\sim} \mathrm{N}(0, I(\theta_0)). \tag{4.16}$$

Moreover, bear in mind that

$$\frac{\partial s(\theta)}{\partial \theta^\top} = \sum\nolimits_{i=1}^{n} \frac{\partial s_i(\theta)}{\partial \theta^\top},$$

meaning that the second-order derivative of the log likelihood grows with order n. This, in turn, means that its inverse decreases with order $1 \,/\, n$. Looking at (4.15), we can conclude that with increasing sample size we obtain $\hat{\theta} \to \theta_0$ and therefore $\tilde{\theta} \to \theta_0$. Consequently, for increasing n, formula (4.15) can be approximated through

$$\hat{\theta} - \theta_0 \approx -\left[\frac{\partial s(\theta_0)}{\partial \theta^\top}\right]^{-1} s(\theta_0). \tag{4.17}$$

The matrix term itself can be written as

$$\frac{\partial s(\theta_0)}{\partial \theta^\top} = \underbrace{E\left\{\frac{\partial s(\theta_0)}{\partial \theta^\top}\right\}}_{=-I(\theta_0)} + \underbrace{\left\{\frac{\partial s(\theta_0)}{\partial \theta^\top} - E\left(\frac{\partial s(\theta_0)}{\partial \theta^\top}\right)\right\}}_{=:U(\theta_0)}.$$

Note that $U(\theta_0)$ decomposes to

$$U(\theta_0) = \sum_{i=1}^{n} \underbrace{\frac{\partial s_i(\theta_0)}{\partial \theta^\top} - E\left(\frac{\partial s_i(\theta_0)}{\partial \theta^\top}\right)}_{U_i(\theta_0)},$$

where the $U_i(\theta_0)$ are independent matrix valued random variables so that $U(\theta_0)$ has standard deviation of order $\sqrt{n}$, while $I(\theta_0)$ grows with order n. This in turn allows to simplify (4.17) further by dropping $U(\theta_0)$ so that we end up with

$$(\hat{\theta} - \theta_0) \approx I^{-1}(\theta_0)\ s(\theta_0). \tag{4.18}$$

Utilizing (4.3) directly proves (4.9). □

4.3 Cramér-Rao Bound

The concept of asymptotic normality enables the construction of confidence intervals. We will now establish that the maximum likelihood estimator is not only consistent but also efficient, meaning that it holds the smallest possible variance among unbiased estimators (see Definition 3.3). To achieve this, we will examine a central mathematical concept known as the Cramér-Rao bound. It refers to the independent contributions derived by Cramér (1946) and Rao (1945). However, it is important to note that several other authors also reached the same conclusion, with one of the earliest references being Fréchet (1943).[2] The result is as follows:

Theorem 4.3 *Let $\hat{\theta} = t(Y_1, \ldots, Y_n)$ be an estimator of θ based on the data $Y_1, \ldots, Y_n$ with $Y_i \sim F(\cdot; \theta_0)$, i.i.d., where $F(\cdot; \theta)$ is a Fisher regular distribution. The mean squared error of $\hat{\theta}$ is always larger than the lower* **Cramér-Rao bound**:

$$MSE(\hat{\theta}, \theta_0) \geq Bias^2(\hat{\theta}, \theta_0) + \frac{\left(1 + \frac{\partial Bias(\hat{\theta}, \theta_0)}{\partial \theta}\right)^2}{I(\theta_0)} = CR(\hat{\theta}, \theta_0)$$

If the estimator is unbiased, we get

$$MSE(\hat{\theta}, \theta) \geq \frac{1}{I(\theta)}. \tag{4.19}$$

A proof sketch is given at the end of this section. For simplicity of notation, we omitted any vector or matrix notation and assumed (without loss of generality, though) that θ is a scalar.

The Cramér-Rao bound demonstrates why the inverse Fisher information plays a central role. In fact, for unbiased estimators or at least for asymptotically unbiased estimators like the maximum likelihood estimator, it proves that the inverse Fisher information is the smallest variance achievable. In other words, it shows that maximum likelihood estimation is efficient. One could literally reformulate this to "the maximum likelihood approach makes the most efficient use of the data."

Proof Here, we prove the validity of the Cramér-Rao bound. Note that for unbiased estimators we have

$$\theta = E(\hat{\theta}) = \int t(y) f(y;\theta) \mathrm{d}\ y.$$

Differentiating both sides with respect to θ yields

$$\begin{aligned} 1 &= \int t(y) \tfrac{\partial f(y;\theta)}{\partial \theta} \mathrm{d}\ y \\ &= \int t(y) \tfrac{\partial \log f(y;\theta)}{\partial \theta} f(y;\theta) \mathrm{d}\ y \\ &= \int t(y) s(y;\theta) f(y;\theta) \mathrm{d}\ y \end{aligned}$$

where $s(\theta; y) = \frac{\partial \log f(y;\theta)}{\partial \theta}$. The score function has mean zero,

$$E\Big(s(\theta;y)\Big) = \int s(\theta;y) f(y;\theta) \mathrm{d}\ y = 0,$$

which implies

$$\begin{aligned} 1 &= \int t(y) s(\theta;y) f(y;\theta) \mathrm{d}\ y = \int \Big(t(y) - \theta\Big)\Big(s(\theta;y) - 0\Big) f(y;\theta \mathrm{d}\ y a) \mathrm{d}\ y \\ &= \mathrm{Cov}\Big(t(Y); s(\theta;Y)\Big). \end{aligned}$$

Invoking the Cauchy-Schwarz inequality (A.8), we obtain

$$1 = \mathrm{Cov}\Big(t(Y); s(\theta;Y)\Big) \leq \sqrt{\mathrm{Var}_\theta\big(t(Y)\big)} \sqrt{\mathrm{Var}_\theta\big(s(\theta;Y)\big)}.$$

Note that

$$
\begin{aligned}
\mathrm{Var}_\theta\Big(s(\theta;Y)\Big) &= \int \Big(s(\theta;y) - 0\Big)^2 f(y,\theta) \mathrm{d}\, y \\
&= \int s^2(\theta;y) f(y;\theta) \mathrm{d}\, y \\
&= \int \tfrac{\partial \log f(y;\theta)}{\partial \theta} \tfrac{\partial \log f(y;\theta)}{\partial \theta} f(y;\theta) \mathrm{d}\, y \\
&= I(\theta),
\end{aligned}
$$

using (4.6). According to Definition 3.3, we have $\mathrm{Var}_\theta\big(t(Y)\big) = \mathrm{MSE}(t(Y),\theta)$ for unbiased estimators. So the following statement holds

$$
\begin{aligned}
1 &\leq \sqrt{\mathrm{Var}_\theta(t(Y))}\sqrt{\mathrm{Var}_\theta(s(\theta;Y))} \\
\tfrac{1}{\sqrt{\mathrm{Var}_\theta(s(\theta;Y))}} &\leq \sqrt{\mathrm{Var}_\theta(t(Y))} \\
\tfrac{1}{I(\theta)} &\leq \mathrm{MSE}(t(Y),\theta),
\end{aligned}
$$

which concludes the proof. For simplicity of notation, the final lines were written down only for univariate statistics and parameters, respectively. □

4.4 Likelihood Ratio

After investigating the mathematical properties of the maximum likelihood estimator $\hat{\theta}$, let us look at the likelihood function itself, which is sketched in Fig. 4.5. The asymptotic statements derived above relate to the horizontal axis, i.e., statements concerning $\hat{\theta} - \theta_0$. However, statements about the likelihood function $l(\theta)$ refer to the vertical axis. For the latter, we define the log-likelihood ratio.

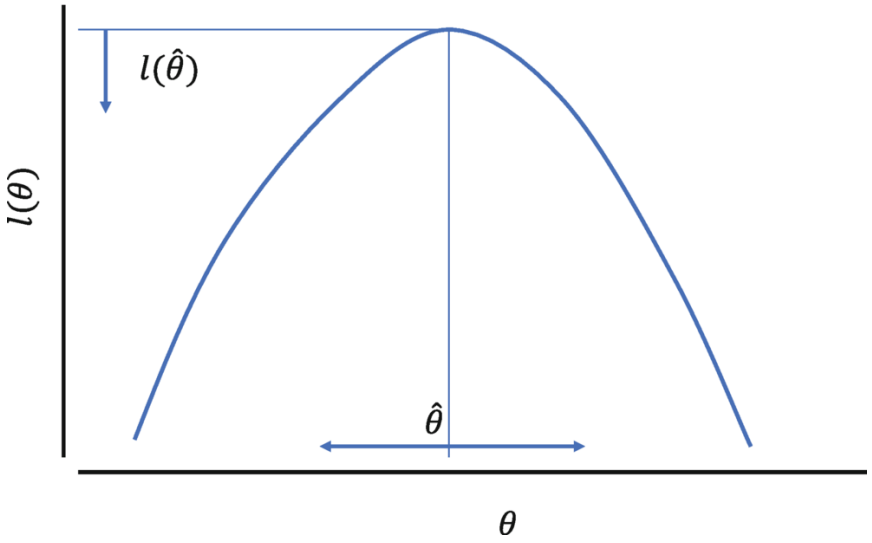


Fig. 4.5 Sketch of log-likelihood function. Asymptotic properties of $\hat{\theta}$ refer to "horizontal" statements, while properties of $l(\hat{\theta})$ look at the "vertical" perspective

Definition 4.3 *The* **log-likelihood ratio** *is defined as*

$$lr(\hat{\theta};\theta) = 2\{l(\hat{\theta}) - l(\theta)\} = 2\log\frac{L(\hat{\theta})}{L(\theta)} \tag{4.20}$$

with $L(\theta) = \exp(l(\theta))$ *as likelihood function.*

The role of factor 2 above will become apparent in the subsequent property. By definition, the log-likelihood ratio is always nonnegative since $l(\hat{\theta})$ defines the maximum of the log-likelihood function. Considering the estimate $\hat{\theta}$ as a random variable turns the likelihood ratio into a random variable as well, for which we can derive its asymptotic distribution. In other words, besides the asymptotic distribution on the horizontal axis, we can derive a related statement for the vertical axis.

Property (Likelihood-Ratio) *The likelihood ratio* (4.20) *for a Fisher regular distribution converges for sample size n increasing to a Chi-squared distribution* $\mathcal{X}_p^2$ *with p degrees of freedom, that is,*

$$2\{l(\hat{\theta}) - l(\theta_0)\} \overset{a}{\sim} \mathcal{X}_p^2, \tag{4.21}$$

where p is the dimension of the parameter θ. □

We sketch the proof below. We emphasize that this statement will play a central role for statistical tests and will be reused in Chap. 5.

Proof Proof of asymptotic distribution of the log-likelihood ratio. Applying Taylor series expansion, we get

$$l(\theta_0) \approx l(\hat{\theta}) + \underbrace{\frac{\partial l(\hat{\theta})}{\partial \theta^\top}}_{=0}(\theta_0 - \hat{\theta}) + \frac{1}{2}(\theta_0 - \hat{\theta})^\top \left(\frac{\partial^2 l(\hat{\theta})}{\partial\theta\,\partial\theta^\top}\right)(\theta_0 - \hat{\theta}) \tag{4.22}$$

$$\Leftrightarrow 2\Big\{l(\hat{\theta}) - l(\theta_0)\Big\} \approx -(\theta_0 - \hat{\theta})^\top \left(\frac{\partial^2 l(\hat{\theta})}{\partial\theta\partial\theta^\top}\right)(\theta_0 - \hat{\theta}). \tag{4.23}$$

Using rigorous asymptotic arguments, which we simplify here, one can show that the second-order derivative above can be approximated through its expectation calculated at the parameter θ_0, $I(\theta_0)$. Therefore, Eq. (4.23) can be simplified to

$$2\Big\{l(\hat{\theta}) - l(\theta_0)\Big\} \approx -(\theta_0 - \hat{\theta})^\top I(\theta_0)(\theta_0 - \hat{\theta}). \tag{4.24}$$

Using Eq. (4.18), we can approximate $\hat{\theta} - \theta_0$ with $I^{-1}(\theta_0)s(\theta_0)$ which allows to simplify (4.24) further to

$$2\Big\{l(\hat{\theta}) - l(\theta_0)\Big\} \approx -s(\theta_0)^\top I^{-1}(\theta_0)s(\theta_0). \tag{4.25}$$

Given (4.16), we can utilize (B.5) to show (4.21). □

4.5 Parameter Transformation

Often, the parameter of a model is restricted, for instance, in a Binomial model, the parameter π is restricted such that $0 \leq \pi \leq 1$. Maximization in the presence of restricted parameters can be numerically difficult, and it is therefore often helpful to transform the parameter into an unrestricted space, estimate this transformed parameter, and, finally, transform it back to its original scale.

Example 4.5 For a binomial model, instead of maximizing with respect to $\pi \in [0, 1]$, one can maximize with respect to

$$\gamma = \log\left(\frac{\pi}{1-\pi}\right) \in \mathbb{R} \quad \Leftrightarrow \quad \pi = \frac{\exp(\gamma)}{1+\exp(\gamma)} \in [0, 1].$$

▷

Generally, let θ be the parameter of a model and define the log likelihood as $l_\theta(\theta)$, where we include the subscript to demonstrate the parameterization. Let $\gamma = h(\theta)$ be some invertible transformation, so that $\theta = h^{-1}(\gamma)$. The log likelihood for γ is then

$$l_\gamma(\gamma) = l_\theta\big(h^{-1}(\gamma)\big).$$

Differentiation leads to

$$\frac{\partial l_\gamma(\gamma)}{\partial \gamma} = \frac{\partial l_\theta(\overbrace{h^{-1}(\gamma)}^{=\theta})}{\partial \theta} \; \frac{\partial \overbrace{h^{-1}(\gamma)}^{=\theta}}{\partial \gamma}.$$

Setting $\widehat{\gamma} = h(\widehat{\theta})$ leads to

$$\frac{\partial l_\gamma\left(\hat{\gamma}\right)}{\partial \hat{\gamma}} = \underbrace{\frac{\partial l_\theta\left(\hat{\theta}\right)}{\partial \theta}}_{=0} \; \frac{\partial \theta}{\partial \gamma} = 0,$$

so that $\widehat{\gamma} = h(\hat{\theta})$ results as maximum likelihood estimator of the transformed parameter. Hence, we obtain the maximum likelihood estimate of the transformed parameter just by transforming the maximum likelihood estimate of the original parameter. Taking the second-order derivative (assuming for simplicity a scalar-valued parameter) yields

$$\begin{aligned}\frac{\partial^2 l_\gamma(\gamma)}{\partial\gamma\,\partial\gamma} &= \frac{\partial}{\partial\gamma}\left(\frac{\partial l_\theta(\theta)}{\partial\theta}\,\frac{\partial\theta}{\partial\gamma}\right)\\ &= \frac{\partial^2 l_\theta(\theta)}{\partial\gamma\,\partial\theta}\,\frac{\partial\theta}{\partial\gamma} + \frac{\partial l_\theta(\theta)}{\partial\theta}\,\frac{\partial^2\theta}{\partial\gamma\,\partial\gamma}\\ &= \frac{\partial\theta}{\partial\gamma}\,\frac{\partial^2 l_\theta(\theta)}{\partial\theta\,\partial\theta}\,\frac{\partial\theta}{\partial\gamma} + \frac{\partial l_\theta(\theta)}{\partial\theta}\,\frac{\partial\theta}{\partial\gamma}.\end{aligned}$$

Taking expectation leads to

$$\begin{aligned}I_\gamma(\gamma) &= E\left(-\frac{\partial^2 l_\gamma(\gamma)}{\partial\gamma\,\partial\gamma}\right)\\ &= \frac{\partial\theta}{\partial\gamma}\,\underbrace{I_\theta(\theta)}_{=E\left(-\frac{\partial^2 l_\theta(\theta)}{\partial\theta\,\partial\theta}\right)}\,\frac{\partial\theta}{\partial\gamma},\end{aligned}$$

since $E\left(\frac{\partial l_\theta(\theta)}{\partial\theta}\right) = 0$. In other words, the Fisher information for γ easily results from the Fisher information for θ. This in turn leads to

$$\begin{aligned}\widehat{\gamma} &\overset{a}{\sim} \mathrm{N}\left(\gamma, I_\gamma^{-1}(\gamma)\right)\\ \widehat{\gamma} &\overset{a}{\sim} \mathrm{N}\left(\gamma, \frac{\partial\gamma}{\partial\theta} I_\theta^{-1}(\theta)\frac{\partial\gamma}{\partial\theta}\right).\end{aligned}$$

We can conclude that parameter transformations can be handled easily in maximum likelihood estimation and the asymptotic results extend directly to the transformed parameter.

4.6 Misspecified Models

In Chap. 3, we initially approached maximum likelihood estimation with the perspective that the data followed the general data-generating process (3.1). The motivation behind maximum likelihood estimation stemmed from seeking the distribution within the statistical model $\mathscr{F} = \{F(\cdot;\theta); \theta \in \Theta\}$ that best approximates $G(\cdot)$. This alignment was measured using the Kullback-Leibler divergence as a nonsymmetric measure of distance.

In the current chapter, we have so far focused on cases where $G(\cdot) \in \mathscr{F}$. However, we are now moving to broaden our scope. Let us explicitly assume that $G(\cdot) \notin \mathscr{F}$. The distribution in $\mathscr{F}$ which is closest to $G(\cdot)$ was implicitly defined through

$$\int \frac{\partial l(\theta_0)}{\partial\theta}\, dG(y) = E\left(s(\theta_0)\right) = 0, \tag{4.26}$$

so that the first Bartlett identity (4.5) remains valid, even though through different reasoning as derived in Eq. (3.3). However, the second Bartlett identity does not

hold. Instead, we obtain the Fisher matrix through

$$I(\theta_0) := -\int \frac{\partial^2 l(\theta_0)}{\partial\theta\partial\theta^\top} dG(y), \tag{4.27}$$

while the variance of the score equals

$$V(\theta_0) := \int s(\theta)s^\top(\theta) dG(y) = \mathrm{Var}\big(s(\theta_0)\big). \tag{4.28}$$

In general, we have for $G(\cdot) \notin \mathscr{F}$ that $V(\theta_0) \neq I(\theta_0)$. This makes statistical inference slightly more complicated and Theorem 4.2 extends as follows.

Theorem 4.4 *Assume a model $\mathscr{F}$ of Fisher regular distributions and an $i.i.d.$ sample $Y_1, \ldots, Y_n$ drawn from $Y_i \sim G(\cdot)$. The maximum likelihood estimator is asymptotically normally distributed with*

$$\hat{\theta} \overset{a}{\sim} N\big(\theta_0, I^{-1}(\theta_0)V(\theta_0)I^{-1}(\theta_0)\big), \tag{4.29}$$

where θ_0 is the parameter that minimizes the Kullback-Leibler divergence $KL(G(\cdot), F(\cdot,\theta))$ for $\theta \in \Theta$. The Fisher information $I(\theta_0)$ is derived through (4.27), and the variance $V(\theta_0)$ equals (4.28).

The proof of the statement is rather straightforward and given at the end of this section.

It should be noted that since $G(\cdot)$ is unknown, neither the Fisher matrix nor the variance can be calculated analytically. However, empirical estimates are readily available. To obtain these, one just replaces the integral by its arithmetic version based on the data. This leads to the estimates

$$\begin{aligned} \hat{I}(\theta_0) &= \frac{1}{n}\sum_{i=1}^n -\frac{\partial^2 \log f(y_i;\hat{\theta})}{\partial\theta\partial\theta^\top} \\ \widehat{V}(\theta_0) &= \frac{1}{n}\sum_{i=1}^n \frac{\partial \log f(y_i;\hat{\theta})}{\partial\theta}\frac{\partial \log f(y_i;\hat{\theta})}{\partial\theta^\top}, \end{aligned}$$

which can then be inserted in (4.29) to obtain a pivotal distribution.

The proposed estimates thereby have a long history since inference in misspecified models is of central importance in real-world data examples. As sketched in Chap. 3, the true data-generating process remains unknown in nearly all cases. Hence, it is important to obtain inference on the parameter estimators $\hat{\theta}$ even if $G(\cdot) \notin \mathscr{F}$. This was focused in particular in regression models in the presence of heteroskedasticity, where Eicker (1967) and Huber (1967) proposed so-called robust variance estimates (see also White, 1980). Their estimates are the same or similar to (4.29).

Proof It remains to prove Theorem 4.4. Taking (4.26), we expand the score around θ_0 which gives

$$
\begin{aligned}
0 &= s(\hat{\theta}; y_1, \ldots, y_n) \\
&\approx s(\theta_0; y_1, \ldots, y_n) - J(\theta_0; y_1, \ldots, y_n)(\hat{\theta} - \theta_0) \\
\Leftrightarrow (\hat{\theta} - \theta_0) &\approx J^{-1}(\theta_0; y_1, \ldots, y_n) s(\theta_0; y_1, \ldots, y_n),
\end{aligned} \tag{4.30}
$$

where $J(\theta)$ is the observed Fisher information defined through

$$
J(\theta; y_1, \ldots, y_n) = -\sum_{i=1}^{n} \frac{\partial^2 \log f(y_i;\theta)}{\partial\theta\partial\theta^\top}.
$$

The observed Fisher information can asymptotically be replaced by its expected value given in (4.27) so that (4.30) can be written in short as

$$
(\hat{\theta} - \theta_0) \approx I^{-1}(\theta_0) s(\theta_0). \tag{4.31}
$$

Reflecting that the score is a sum of independent random variables proves normality following from the central limit theorem, where the mean and the variance directly follow from (4.26), (4.27), and (4.28). □

4.7 Numerical Calculation of the Estimate

In today's landscape, optimizing a function numerically is no longer a technical hurdle due to the abundance of algorithms integrated into various software packages. Consequently, we can keep this section brief. However, the properties we derived for the maximum likelihood estimator synergize nicely with a simple method called the Newton-Raphson procedure.

To reiterate, let us outline the Newton-Raphson approach using Fig. 4.6. Our goal is to locate the root of the score function, illustrated in the top row. By initializing the parameter at a starting value $\theta_{(1)}$, we can approximate the score function linearly. This linear approximation is depicted in the middle plot of Fig. 4.6. Notably, for linear functions, the root can be found analytically, leading to the parameter update $\theta_{(2)}$. Continuing iteratively, as demonstrated in the bottom plot of the figure, yields the update $\theta_{(3)}$ and so forth. This iterative process ultimately converges to the maximum likelihood estimate $\hat{\theta}$.

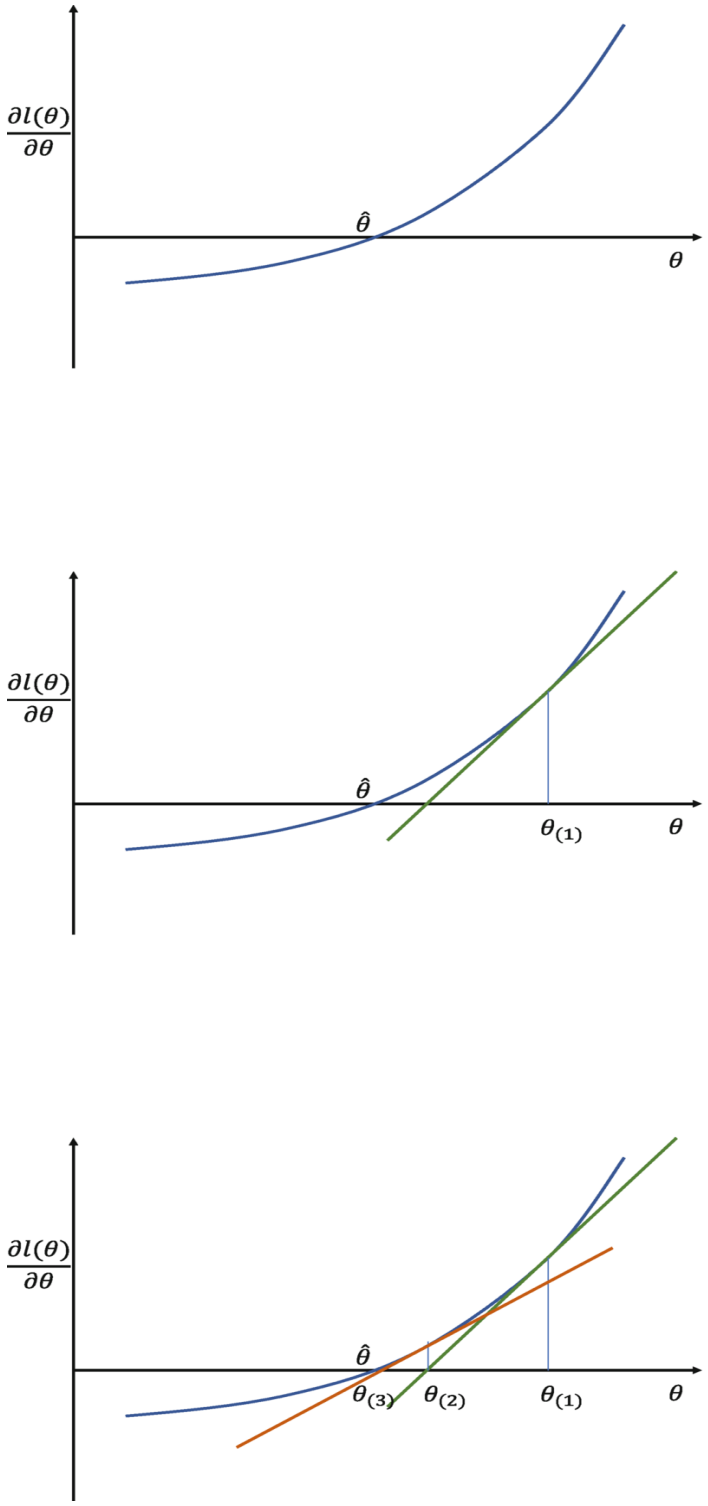


Fig. 4.6 Sketch of Newton-Raphson procedure

We apply the same idea to the calculation of the maximum likelihood estimate. The Newton-Raphson approach can here be simplified by replacing the derivative of the score function, which is the second-order derivative of the log-likelihood

function, by its expectation, that is, the (negative) Fisher matrix. This is called Fisher scoring.

Note that in first-order approximation, we have

$$0 = s(\hat{\theta}) \approx s(\theta_0) - I(\theta_0)(\hat{\theta} - \theta_0).$$

Rewriting this gives

$$\hat{\theta} = \theta_0 + I^{-1}(\theta_0)s(\theta_0).$$

The simple iteration scheme can now easily be derived from this formula. Starting with some value $\theta_{(0)}$ and setting $t = 0$, we iterate as follows:

1. Calculate $\theta_{(t+1)} := \theta_{(t)} + I^{-1}(\theta_{(t)})s(\theta_{(t)})$.
2. Repeat step 1 until $\|\theta_{(t+1)} - \theta_{(t)}\| < d$.
3. Set $\hat{\theta} = \hat{\theta}_{(t+1)}$.

As with any Newton-Raphson procedure, the process may fail if the starting value $\theta_{(0)}$ is too far away from the target value $\hat{\theta}$. In this case, it can help to work with a reduced step size. Hence, one can add some $0 < \delta < 1$:

$$\theta_{(t+1)} = \theta_{(t)} + \delta I^{-1}(\theta_{(t)})s(\theta_{(t)})$$

The step size can even be chosen adaptively based on the current step *t*, i.e., $\delta(t)$, which traces back to Robbins and Monro (1951).

4.8 The Likelihood Principle

Using the likelihood has proven to be quite advantageous. This concept can be taken a step further by adhering to the likelihood principle. This principle operates in two main steps: First, one picks a probability model. Second, one then derives the (log-)likelihood function which in fact contains all information about the parameters of the model.

Definition 4.4 *The likelihood principle states that for inference about a parameterθall relevant information is contained in the likelihood function for the data at hand*

To illustrate the likelihood principle, let us consider a simple example. Imagine we are measuring air pollution, specifically, the concentration of airborne particles denoted as Y. To measure Y, we are using a device with a predefined and known level of precision, represented by c. If the air pollution value Y falls below this

precision level, the device is not capable of providing an accurate measurement. Any values below this threshold *c* are essentially unrecordable due to the technical limitations of the device. When such measurements occur, they are noted as *c* in the data. This notation signifies that the actual value Y is less than or equal to *c*. These instances are referred to as “censored” observations. We demonstrate the idea with a simple example. In Fig. 4.7, we show some simulated data. The measurements *y* are thereby generated through $y_i = \exp(z_i)$ with $z_i \sim \mathrm{N}(-1, 1)$. Hence, the measurements follow a log-normal distribution with $\mu = -1$ and $\sigma^2 = 1$. The horizontal line is the precision threshold, with the true values below this threshold shown in gray. Note that these values remain unobservable and they are shifted to the threshold level, which in this case is $c = 0.5$.

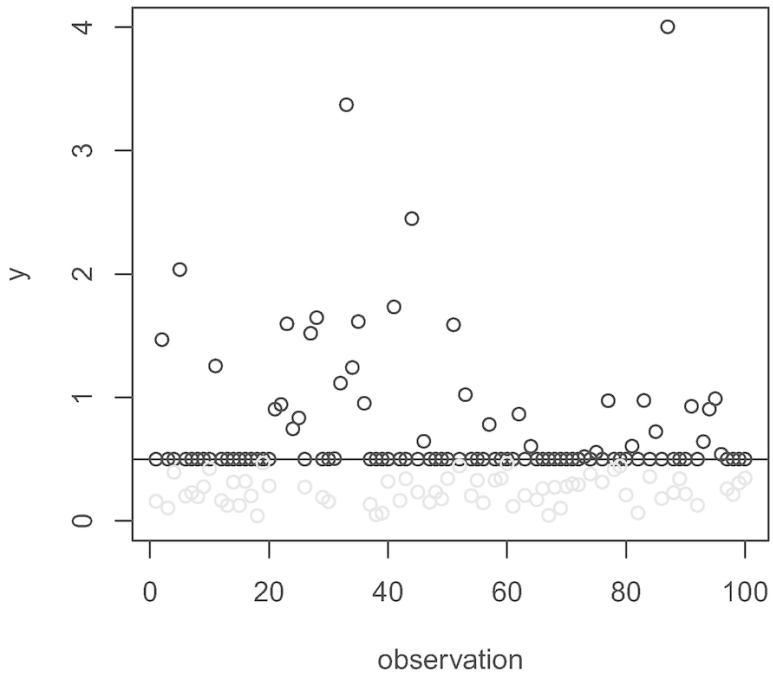


Fig. 4.7 Simulated data with a precision threshold. Values below this threshold are set to the threshold. The true simulated values below the threshold are shown as gray points

In our scenario, we take a log-normal distribution for the measurements or, equivalently, a normal distribution for the logarithm of the measurements. Our interest lies in the parameters μ and σ^2. Taking the recorded data and just calculating the mean value of the logarithmic measurement would lead to the optimization problem

$$\sum_{i=1}^{n} \big(\log(y_i) - \mu\big)^2 \to \min.$$

In the data example shown in Fig. 4.7, we end up with the mean estimate $\widehat{\mu} = -0.42$, which is clearly biased. Hence, the resulting mean estimator is biased

upward and hence takes too large values, which is obvious since observations with value $y_i = c$ are censored and just represent the precision level of the measurement device and not the true air pollution.

Instead of taking the mean of the observed data, the likelihood principle would lead to a more elaborate extraction of information. For observations with $y_i > c$, we obtain from the density of a normal distribution. However, for observations $y_i = c$, we know that the true air pollution value is below c, leading to the likelihood contribution $P(Y_i \leq c)$. Putting this together gives the log likelihood

$$l(\mu, \sigma^2) = \sum_{i=1}^{n} \mathbb{I}\{Y_i > c\} \left[-\tfrac{1}{2} \log \sigma^2 - \tfrac{1}{2} \tfrac{(\log(y_i) - \mu)^2}{\sigma^2} \right] \tag{4.32}$$

$$+ \sum_{i=1}^{n} \mathbb{I}\{Y_i \leq c\} \Phi\left(\tfrac{\log(c) - \mu}{\sigma} \right), \tag{4.33}$$

where $\Phi(\cdot)$ is the distribution function of a standard normal distribution and the latter terms express the probabilities

$$P(Y_i \leq c; \mu, \sigma^2) = P(\underbrace{\log(Y_i)}_{\sim \mathrm{N}(\mu, \sigma^2)} \leq \log(c); \mu, \sigma^2) \tag{4.34}$$

$$= P\left(\underbrace{\frac{\log(Y_i) - \mu}{\sigma}}_{\sim \mathrm{N}(0,1)} \leq \frac{\log(c) - \mu}{\sigma} \right) \tag{4.35}$$

$$= \Phi\left(\tfrac{\log(c) - \mu}{\sigma} \right). \tag{4.36}$$

Applying this to the data shown in Fig. 4.7, we get $\widehat{\mu} = -1.04$ and $\widehat{\sigma}^2 = 1.04$, which is very close to the simulation values.

In summary, the likelihood stands as a foundational cornerstone in statistics, and it will be a recurring concept throughout the rest of this book. Its applicability extends directly to more complex models than those discussed here. With the assurance of efficiency, we can be certain that we are making optimal use of the available data.

4.9 Exercises

Exercise 1 Consider two samples from a Bernoulli distribution with parameter p. The first one has 2 successes over 10 trials, and the second one has 20 successes over 100 trials.

(a) Write down the likelihood function and compute it for both samples.

(b) Can you find the value of p which makes the likelihood function take the maximum value possible for those two samples? What about the general case

of a sample of *n* trials from any Bernoulli distribution with parameter *p*?

(c) Draw the two likelihood functions and compare them. Which one of the two samples, in your opinion, contains more information? And why?

Exercise 2 Consider the geometric probability distribution

$$f(x \mid p) = p(1-p)^{x-1} \; ; \; p \in (0,1) \; , \; x \in \{1,2,3,\ldots\}.$$

This distribution, sometimes also called the inverse Bernoulli distribution, counts the number of independent Bernoulli trials that take place until a success is realized.

(a) Suppose that we draw an independent sample $(X_1 \ldots . X_n)$ from this distribution. Write the log-likelihood function of the sample.

(b) Find the maximum likelihood estimator (MLE) $\widehat{p}$ for p, and prove that it is indeed the MLE.

(c) By using the asymptotic properties of the MLE, derive the asymptotic distribution of the ML estimator $\widehat{p}$.

Exercise 3 You are given an i.i.d. sample $X = (X_1, \ldots, X_n)$ of continuous random variables X_i with the following density:

$$f(x_i;\theta) = \frac{1}{\sqrt{2\pi}}\theta^{-\frac{3}{2}}\; x_i^2 \; \exp\left(-\frac{\frac{2}{x_i}}{2\,\theta}\right), \quad \theta > 0, \quad i = 1, \ldots, n \tag{4.37}$$

Further, $\mathbb{E}(X_i) = 0$ and $\mathrm{Var}(X_i) = 3\theta$ hold.

(a) Determine the maximum likelihood estimator for θ. **Note:** You don't have to show that the MLE is indeed a maximum.

(b) The provided distribution fulfils the criteria for Fisher regularity. Show that the expected Fisher information is $I_X(\theta) = \dfrac{1.5\; n}{\theta^2}$.

(c) Determine the approximate distribution of the following statistic:

$$T(X) = \frac{\sum_{i=1}^{n} X_i^2}{2\theta^2} - \frac{3n}{2\theta}$$

(d) A standard approach to the construction of the $(1-\alpha)$ confidence interval for a parameter is the Wald interval. In case of the MLE, we make use of its asymptotic variance so that the interval looks like this:

$$\left[\hat{\theta}_{ML} + z_{\frac{\alpha}{2}}\sqrt{I_n^{-1}(\hat{\theta}_{ML})}\right]$$

$$\left[\hat{\theta}_{ML} \pm z_{1-\alpha/2} \sqrt{I_X^{-1}(\hat{\theta}_{ML})}\right]$$

Calculate an approximate $(1-\alpha)$ confidence interval ($\alpha = 0.05$) for θ using the following sample:

$$x = (-2, 1, 0, -0.5, 2, 1, 0, 0.5, 2, -1, 0, 0.5, 2, 1, 0, -0.5)$$

References

Cramér, H. (1946). *Mathematical methods of statistics*. Princeton University Press.

Eicker, F. (1967). Limit theorems for regressions with unequal and dependent errors. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, Berkeley, CA, 1965/66 (Vol. I: Statistics, pp. 59–82). University California Press.

Fréchet, M. (1943). Sur l'extension de certaines évaluations statistiques au cas de petits échantillons. *Revue de l'Institut International de Statistique,11*, 182–205.
[https://doi.org/10.2307/1401114]

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, Berkeley, CA, 1965/66 (Vol. I: Statistics, pp. 221–233). University California Press.

Rao, C. R. (1945). Information and the accuracy attainable in the estimation of statistical parameters. *Bulletin of the Calcutta Mathematical Society,37*, 81–89.
[MathSciNet]

Robbins, H., & Monro, S. (1951). A stochastic approximation method. *The Annals of Mathematical Statistics, 22*, 400–407.
[MathSciNet][https://doi.org/10.1214/aoms/1177729586]

Stigler, S. M. (1980). Stigler's law of eponymy. *Transactions of the New York Academy of Sciences,39*(1), 147–157.
[https://doi.org/10.1111/j.2164-0947.1980.tb02775.x]

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica,48*, 817–838.
[MathSciNet][https://doi.org/10.2307/1912934]

Footnotes

1 Namely, that the maximizer of $E(l(\theta; Y_1, \ldots, Y_n))$ is unique and the convergence is uniform, i.e., holds for any θ, and the function $E(l(\theta; Y))$ continuous.

2 See Stigler (1980) for a more general discussion on the phenomenon of concepts not being named after the first inventors itself called Stigler's law of eponymy.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
G. Kauermann et al., *Statistics—Coping with Uncertainty*, Textbooks in Statistical Science
https://doi.org/10.1007/978-3-032-33658-3_5

5. Data-Driven Decisions

Göran Kauermann[1] ✉, Giacomo De Nicola[2], Cornelius Fritz[3] and David Rügamer[4]
(1) Department of Statistics, LMU Munich, Munich, Germany
(2) Harvard T.H. Chan School of Public Health, Harvard University, Boston, MA, USA
(3) School for Computer Science and Statistics, Trinity College Dublin, Dublin, Ireland
(4) Munich Uncertainty Quantification AI Lab, LMU Munich, Munich, Germany

Statistical inference is commonly understood as the process of decision-making based on statistical tests. It is a central pillar of statistical reasoning and ultimately revolves around hypothesis tests and *p*-values. We will provide a gentle introduction to the topic that follows the chronological development of the different ideas. We therefore begin with the concept of a *p*-value before formally introducing hypothesis tests. Our main goal is not to present a full catalog of different test procedures but rather to motivate the (simple) core ideas of statistical tests. From there, different test strategies naturally follow. We also present the key mathematical results that underpin optimal tests. Furthermore, we show that hypothesis tests and confidence intervals, as discussed in Sect. 3.4 , are closely related and can be viewed as two sides of the same coin.

5.1 Statistical Evidence and the *p*-Value

We begin by focusing on data-driven decision-making, starting with the assessment of a specific hypothesis. One should therefore bear in mind that we can rarely prove that a hypothesis holds; we can, however, collect evidence against it. When such evidence becomes sufficiently strong, it calls the hypothesis into question and may lead us to reject it. To illustrate this idea, consider the hypothesis that the Earth is flat. One might adhere to this belief until enough evidence is accumulated to contradict it, such as satellite imagery or measurements of the Earth's curvature. In the context of statistical inference and uncertainty quantification, our goal is therefore to construct a measure that captures the strength of evidence against a given hypothesis.

We motivate the idea with a historical example. Pierre-Simon Laplace (1749–1827) was one of the most influential mathematicians of his time. He not only proved the central limit theorem but also made numerous lasting contributions to mathematics, many of which remain foundational today. He can also be regarded as an early pioneer of statistical thinking, having addressed precisely the question central to our discussion: How much evidence is required to reject a hypothesis? Laplace turned his attention to

the gender ratio of newborns, a phenomenon already noted in his time due to the consistently higher number of male births. This raised the question of whether the observed imbalance was merely the result of chance or instead reflected a natural law governing human reproduction, implying a systematically higher probability of male births. At the time, such natural laws were often associated with religious interpretations and viewed as manifestations of divine providence, a connection Laplace explicitly rejected. Remarkably, Laplace collected data from various regions across Europe and found strikingly consistent patterns. Even so, he refrained from asserting the existence of a universal law until a sufficiently large body of data—by the standards of the time—had been assembled. Methods for formally quantifying evidence against a hypothesis were not yet available, and numerical calculations were laborious. These limitations led Laplace to develop a series of approximations to address the problem.

To make this idea concrete, let us consider the question using contemporary data. Suppose we entertain the hypothesis that boys and girls are born in equal proportions. In 2021, however, 408,490 boys were born in Germany, compared to 387,027 girls. How much evidence do these data provide against the hypothesis? Figure 5.1 illustrates this question: the left panel shows the number of newborns by gender, while the right panel displays the proportion of male births over recent years. As Laplace did in his time, we may ask whether the observed imbalance provides sufficient evidence to reject the hypothesis of equal gender proportions. At the same time, we can formulate a related question: Is the gender ratio constant over time? The right-hand panel shows some year-to-year variation, with a noticeable peak in 2020. Is this peak merely the result of random fluctuation, or does it indicate genuine temporal variation? Notably, 2020 coincides with the onset of the COVID-19 pandemic. This raises the question of whether the pandemic may have influenced the gender ratio of newborns.

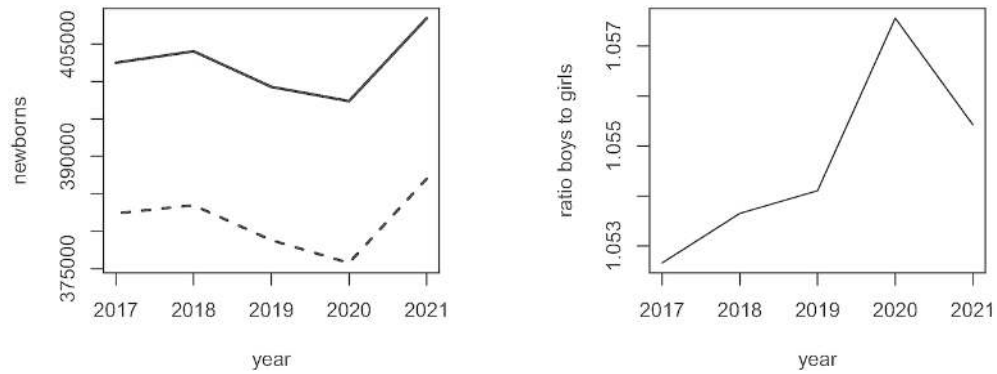


Fig. 5.1 Left-hand side: number of male (solid line) and female (dashed line) newborns. Right-hand side: proportion of males in the last years

To address these questions more formally, we begin by defining the hypothesis of interest.

Definition 5.1 *We denote the hypothesis as* H_0 *and call it the* **null hypothesis**.

In statistical significance testing, which we develop in the following sections, hypotheses can often be formulated as statements about model parameters. As an example, we revisit the gender distribution of newborns and model it using a Bernoulli distribution. Let Y_i denote the coded gender of the *i*-th newborn, where $Y_i = 1$ represents a male child and $Y_i = 0$ a female child (for simplicity, we ignore nonbinary categories here). The number of newborn male babies in a year then results through $Y = \sum_{i=1}^{n} Y_i$, where *n* equals the number of newborns in the considered year. Assuming independence—at least approximately—it follows that

$$Y := \sum_{i=1}^{n} Y_i \sim \mathrm{B}(n, \pi).$$

This allows us to formulate the hypothesis as

$$H_0 : \pi = \tfrac{1}{2}$$

Next, we need to construct evidence against the hypothesis. It is not difficult to see that $\widehat{\pi} = Y \,/\, n$ is the maximum likelihood estimate, and the further away $\widehat{\pi}$ is from 1/2, the more evidence there is against the hypothesis. While this intuition is straightforward, the real challenge lies in quantifying this evidence. This can be done with the so-called $p-$value, a concept which traces back to the abovementioned analyses of Laplace but which was formally introduced by Pearson (1900) and heavily promoted by Fisher (see, for example, Fisher, 1955).

Definition 5.2 *The p-value expresses the probability of observing data that contradict the hypothesis even more than the observed data, assuming the hypothesis to hold.*

This definition needs some motivation. If the hypothesis holds and males and females are born in equal proportions, then we could calculate the probability of observing even more male babies than the number observed in a particular year. To be concrete, in 2021, we observed $y_{obs} = 408.490$ male babies out of $n = 795.478$ newborns. We denote the observed data with the subscript *obs*. If we define our H_0 to be an equal gender share and we assume that to be true, we obtain the *p*-value through

$$p\text{-value} \;:= P(Y > y_{obs} \mid H_0),$$

where conditioning on the hypothesis means that we assume

$$Y \sim \mathrm{B}(795.478, \pi = 0.5).$$

This formulation focuses specifically on deviations in the direction of excess male births. An exact calculation would be computationally demanding, but the central limit theorem provides a powerful simplification. Indeed, once Laplace derived the central limit theorem, he could have calculated the *p*-value for the gender example at that time. Note that under H_0 the variance of *Y* is given by $n \cdot \pi(1 - \pi)$, which in our example reduces to $n \cdot 0.5 \cdot 0.5$. A straightforward calculation therefore yields

$$
\begin{aligned}
p\text{-value} &= P(Y > y_{obs} \mid H_0) \\
&= P\Big(\underbrace{\frac{Y - n \cdot 0.5}{\sqrt{n \cdot 0.5 \cdot 0.5}}}_{\overset{a}{\sim}\mathrm{N}(0,1)} > \frac{y_{obs} - n \cdot 0.5}{\sqrt{n \cdot 0.5 \cdot 0.5}} \Big) \\
&= 1 - \Phi\Big(\underbrace{\frac{408490 - 795478/2}{\sqrt{795478 \cdot 0.25}}}_{=24.1082} \Big) \\
&\approx 0
\end{aligned}
$$

where $\Phi(\cdot)$ is the distribution function of the standard normal distribution. The *p*-value is effectively zero. Hence, the probability of observing data that contradict the hypothesis even more strongly than the observed data is negligibly small. Consequently, the numbers that we observed carry strong evidence against the hypothesis. We can therefore reject the hypothesis and, in this case, conclude that males are more frequent among newborns than females.

Fisher argued that the smaller the *p*-value, the stronger the evidence against the null hypothesis H_0, and he proposed the following thresholds:

$$
\begin{aligned}
p\text{-value} &\leq 0.1 \quad \Leftrightarrow \quad \text{weak evidence against } H_0 \\
p\text{-value} &\leq 0.05 \quad \Leftrightarrow \quad \text{increased evidence against } H_0 \\
p\text{-value} &\leq 0.01 \quad \Leftrightarrow \quad \text{strong evidence against } H_0
\end{aligned}
$$

Let us continue with the birth examples but now question whether there is evidence that the gender proportion changes over time. To do so, we look at the right-hand side of Fig. 5.1. Let Y_t be observed male newborns in year *t* and $\overline{Y}_t = Y_t \,/\, n_t$ the corresponding observed proportion. The hypothesis is now that

$$
H_0 : Y_t \sim \mathrm{B}(n_t, \pi) \text{ for } t = 2017, \ldots, 2021, \tag{5.1}
$$

where π is explicitly unspecified—assumed to be unequal to $1/2$—but constant over time. There are multiple ways of checking this hypothesis, but to keep it simple, we look at pairwise comparisons of the years. For 2 years *t* and *t′*, we look at the quantity

$$
D_{t,t\prime} := \overline{Y}_t - \overline{Y}_{t\prime}. \tag{5.2}
$$

The larger the absolute difference, that is, the larger the absolute value of $D_{t,t\prime}$, the more evidence there is against the hypothesis that the gender proportion is time constant. We can again rely on asymptotic normality, and we assume the events in different years to be independent. A *p*-value for the 2 years *t* and *t′* can then be calculated as follows. Let $d_{t,t\prime,obs}$ be the observed values, for instance, comparing years 2020 and 2021, we have $\overline{y}_{21} = 0.5134825$ and $\overline{y}_{20} = 0.5139873$ so that $d_{21,20,obs} = -0.0005048$. Note that if H_0 holds, then $D_{t,t\prime}$ should have zero mean since π is the same in both years. This can be utilized and the *p*-value results through

$$
\begin{aligned}
p\text{-value} &= P(|D_{t,t'}| > |d_{21,20,obs}| \mid H_0) \\
&= P\Big(\Big| \underbrace{\frac{D_{t,t'}}{\sqrt{\mathrm{Var}(D_{t,t'})}}}_{\overset{a}{\sim}\mathrm{N}(0,1)} \Big| > \Big| \frac{d_{t,t',obs}}{\sqrt{\mathrm{Var}(D_{t,t'})}} \Big| \Big).
\end{aligned}
$$

The variance can be estimated through the binomial distribution in the form

$$
\widehat{Var}(D_{t,t\prime}) = \frac{\overline{Y}_t * (1-\overline{Y}_t)}{n_t} + \frac{\overline{Y}_{t\prime} * (1-\overline{Y}_{t\prime})}{n_{t\prime}},
$$

where the sum follows due to assumed independence between different years. In Fig. 5.2, we show the resulting standardized values of $d_{t,t\prime,obs}$ with an over-imposed standard normal distribution for the ten pairwise comparisons of the years 2017 to 2021. The resulting *p*-values are shown on the right-hand side. We see that the variation visible in Fig. 5.1 can be explained by pure random fluctuation, and there is no evidence that the proportion of male newborns varies over the years. Putting it differently, the data give no evidence that 2020 was special, and the peak visible in Fig. 5.1 for that year is fully explainable through randomness.

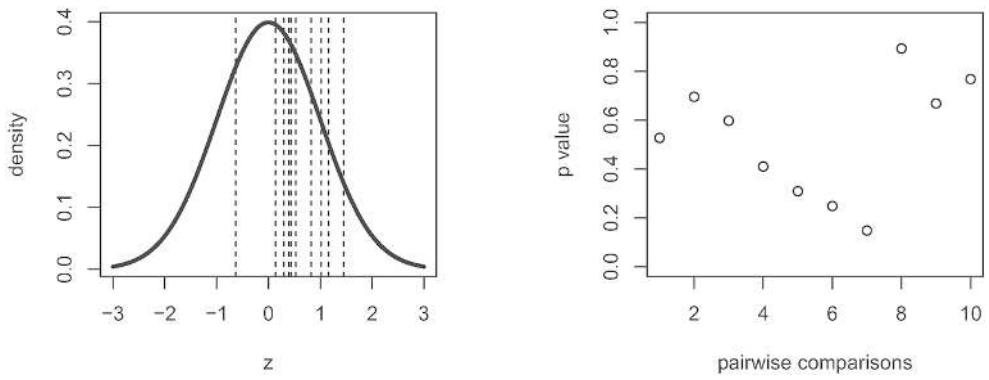


Fig. 5.2 Left-hand side: standard normal distribution and standardized values of $d_{t,t\prime,obs}$. Right-hand side: resulting *p*-values

The *p*-value is calculated from data, and hence, considering the data to come from some data-generating process, one can consider the *p*-value itself as a random variable. This is an interesting perspective that we want to elaborate further. To do so, we look at an artificial setup, which allows us to demonstrate the relevant mathematical features. We consider a simple parametric test and assume that $Y_i \sim \mathrm{N}(\mu, \sigma^2)$, $i.i.d.$, where for simplicity σ^2 is known. Our hypothesis is expressed on the parameter μ

$$
H_0 : \mu \leq \mu_0.
$$

To quantify the evidence against the hypothesis, we make use of $\overline{Y} = \sum_{i=1}^{n} Y_i \Big/ n$, where apparently large values speak against the hypothesis. Let $\overline{y}_{obs}$ be the observed arithmetic mean. The *p*-value then results through

$$
p\text{-value} = P(\overline{Y} > \overline{y}_{obs} \mid H_0).
$$

This number can be calculated explicitly. Note first that the above probability gets maximal for $\mu = \mu_0$. This allows us to calculate

$$\begin{aligned} p\text{-value} &= P(\overline{Y} > \overline{y}_{obs} \mid \mu = \mu_0) \\ &= P\Big(\underbrace{\frac{\overline{Y} - \mu_0}{\sqrt{\sigma^2/n}}}_{\sim \mathrm{N}(0,1)} > \frac{\overline{y}_{obs} - \mu_0}{\sqrt{\sigma^2/n}} \Big) \\ &= 1 - \Phi\Big(\underbrace{\frac{\overline{y}_{obs} - \mu_0}{\sqrt{\sigma^2/n}}}_{:= z_{obs}} \Big). \end{aligned}$$

Note that z_{obs} is based on the data, which can itself be treated as random. In this case, $Z_{obs} \sim \mathrm{N}(0,1)$ with z_{obs} as its concrete realization. Treating now Z_{obs} as random makes the *p*-value random as well, and we may formulate the distribution

$$\begin{aligned} P(p\text{-value} \leq p) &= P\big((1 - \Phi(Z_{obs})) \leq p\big) \\ &= P\big(\Phi(Z_{obs}) > 1 - p\big) \\ &= 1 - P\big(\Phi(Z_{obs}) \leq 1 - p\big) \\ &= 1 - P\big(\Phi^{-1}(\Phi(Z_{obs})) \leq \Phi^{-1}(1 - p)\big) \\ &= 1 - P\big(Z_{obs} \leq \Phi^{-1}(1 - p)\big) \\ &= 1 - \Phi\big(\Phi^{-1}(1 - p)\big) \\ &= 1 - (1 - p) = p. \end{aligned}$$

Hence, we find that under the hypothesis, the *p*-value has a uniform distribution. This is an important property and a requirement for the *p*-value to be unbiased. We summarize this statement as follows.

Property *If the hypothesis H_0 holds, then the p-value follows a uniform distribution, that is,*

$$p - value \sim Uniform[0,1].$$

But what happens if the hypothesis does not hold? We can carry out analogous calculations. Let μ be the true mean value and assume now that $\mu > \mu_0$. In other words, the null hypothesis is false. In this case, $Z_{obs} \nsim \mathrm{N}(0,1)$. But we can still calculate the distribution of the *p*-value. Note that

$$\underbrace{\frac{\overline{Y} - \mu}{\sqrt{\sigma^2/n}}}_{\sim \mathrm{N}(0,1)} = \underbrace{\frac{\overline{Y} - \mu_0}{\sqrt{\sigma^2/n}}}_{\sim \mathrm{N}\left(\frac{\mu - \mu_0}{\sqrt{\sigma^2/n}}, 1\right)} - \frac{\mu - \mu_0}{\sqrt{\sigma^2/n}}.$$

The distribution of the *p*-value then results through the following, where the first steps are like above,

$$\begin{aligned} P(p\text{-value} \leq p) &= P\big((1-\Phi(Z_{obs})) \leq p\big) \\ &= P\big(\Phi(Z_{obs}) > 1-p\big) \\ &= 1 - P\big(\Phi(Z_{obs}) \leq 1-p\big) \\ &= 1 - P\big(\Phi^{-1}(\Phi(Z_{obs})) \leq \Phi^{-1}(1-p)\big) \\ &= 1 - P\big(Z_{obs} \leq \Phi^{-1}(1-p)\big) \\ &= 1 - \Phi\big(z_{1-p} - c\big), \end{aligned}$$

where $c = (\mu - \mu_0) \,/\, \sqrt{\sigma^2 \,/\, n}$ and z_{1-p} is the $1-p$ quantile of the $\mathrm{N}(0,1)$ distribution. We show the distribution of the *p*-value in Fig. 5.3 for different values of *c*. The larger *c*, the larger the probability that the *p*-value takes small values. This makes sense, since *c* represents the distance in the mean values and hence mirrors how far the true model is away from the hypothetical model. Indeed, this is what we intend to see: small *p*-values are more likely to occur when the null hypothesis is false.

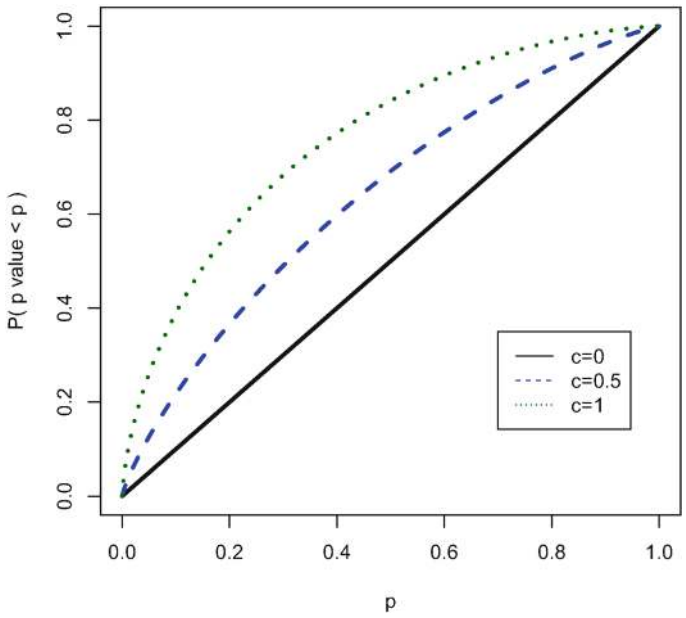


Fig. 5.3 Distribution of *p*-value for different setting of *c*

Looking at the above derivations and properties of the *p*-value, we see that its interpretation needs sufficient understanding of the concept, and it is not surprising that numerous misconceptions of the approach exist. Goodman (2008) lists 12 common misunderstandings, for example, "if the *p*-value is less than 0.05, there is only a 5% chance that the hypothesis is true." This statement is incorrect, since the *p*-value does not assign a probability to the validity of the hypothesis. A similarly common error is "if the *p*-value exceeds 0.05, the hypothesis is confirmed". This is also incorrect. A hypothesis can never be proven true; it can only fail to be rejected if there is insufficient

evidence against it. Hence, it is important to keep the correct definition of the *p*-value in mind in order to interpret it properly. In general, the *p*-value is a concept to quantify evidence against the hypothesis—nothing more but also nothing less.

Example 5.1 We conclude our discussion on the *p*-value by looking at an early instance where the concept of a *p*-value was put into practice. Let's turn the clock back to 1865 when Gregor Mendel (1822–1884) released experimental findings aimed at proving the principle of inheritance, now recognized as Mendelian inheritance (Mendel, 1865). Mendel, an Augustinian friar and amateur biologist, stands out as one of the pioneering figures in genetics and the study of inheritance theory.

His research revolved around peas, a choice that allowed him to uncover the rule of dominance through the crossbreeding of peas with distinct colors and shapes. To be specific, he mated purebred yellow and round peas with purebred green and wrinkled peas. Both the yellow color and round shape held dominance, manifesting in the traits of the first-generation offspring. This initial generation comprised hybrids, which Mendel crossed once more. Based on his assumption of random gene inheritance, he predicted that the second generation would exhibit color and shape in predetermined proportions on average. He then conducted experiments and compared the observed numbers to the numbers that would result if his assumed inheritance laws applied. The original numbers of the experiment that he reported are given in Table 5.1

Table 5.1 Mendel's data

Mendel argued that the observed numbers were close to the proportions predicted by his hypothesis, which he interpreted as support for his theory of inheritance. While a hypothesis cannot be proven true, we can quantify the evidence against it using a *p*-value. Later in the chapter, we derive this formally using the Chi-squared test. For the moment, we proceed via the log-likelihood function and compute the corresponding likelihood ratio. We index the four possible outcomes by 1, 2, 3, and 4, and let $Y_i \in \{1, 2, 3, 4\}$ denote the outcome of the *i*-th pea, with $n = 556$ observations in total. We assume that Y_i are i.i.d. with a multinomial distribution and cell probabilities $\pi_0 = (9 \,/\, 16, 3 \,/\, 16, 3 \,/\, 16, 1 \,/\, 16)$ under the null hypothesis. We can take the log-likelihood function

$$l(\pi) = \sum\nolimits_{i=1}^{n} \sum\nolimits_{k=1}^{4} 1\{Y_i = k\} \log\left(\pi_k\right)$$

with π_k as the *k*-th element of vector π and $1\{Y_i = k\}$ as the indicator vector taking value 1 if $Y_i = k$ and 0 otherwise. It is not difficult to show that the maximum likelihood estimate for π_k is

$$\widehat{\pi}_k = \frac{\sum_{i=1}^{n} 1\{Y_i = k\}}{n},$$

so that the log-likelihood ratio is calculated through $2\{l(\widehat{\pi}) - l(\pi_0)\}$. The larger this value, the more evidence exists against the hypothesis. We have derived the asymptotic

distribution of the likelihood ratio in Sect. 4.4 leading to a Chi-squared distribution with p degrees of freedom, where p is the dimension of the parameter. In this example, $p = 3$, since the sum of all π_k is equal to one; hence, knowing π_1, π_2, and π_3 determines π_4. We calculate the value of the log-likelihood ratio for Mendel's data, which is 0.238, and with it, we obtain p-value

$$p\text{-value} = P(\mathcal{X}^2_p > 0.238) = 0.971,$$

which is also shown in Fig. 5.4. The p-value is very large, indicating essentially no evidence against the hypothesis. It is worth noting that, at the time Mendel conducted his experiments, there was little understanding of how much random variation could be expected around theoretical proportions. Given the unusually close agreement between theory and data, it's possible (and perhaps likely) that the reported experiment may have been one among several conducted, with results chosen for their strong concordance with the theoretical prediction. ▷

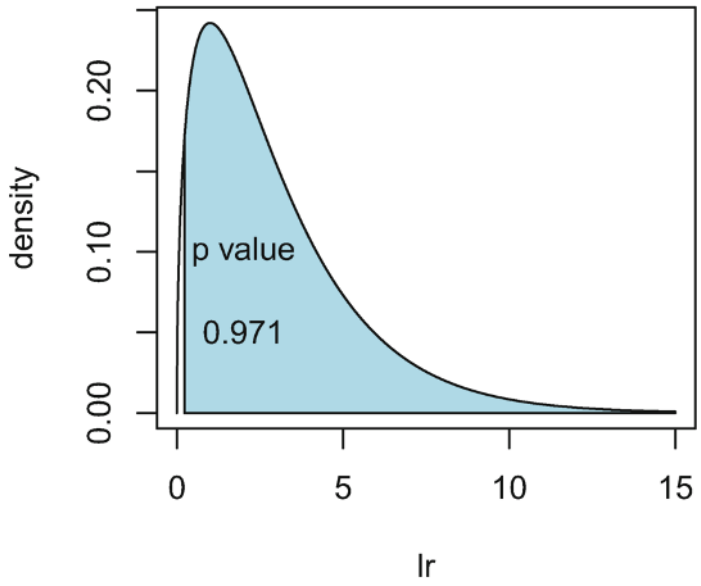


Fig. 5.4 Likelihood ratio of Mendel data with corresponding p-value

5.2 Statistical Hypothesis Tests

So far, we have examined the validity of a hypothesis and quantified the evidence against it using the p-value. We now adopt a different perspective and consider the problem from a decision-theoretic standpoint. In this framework, we must decide either in favor of or against the hypothesis, which requires specifying a competing alternative.

Definition 5.3 *The null hypothesis is denoted by H_0 and the alternative hypothesis by H_1. We assume that exactly one of these two conditions holds.*

A data-driven decision process requires choosing between two possibilities based on the observed data: either we retain the null hypothesis (H_0) or we select the alternative

hypothesis (H_1). We denote these decisions by "H_0" and "H_1", respectively.

As in any decision problem, errors are possible. If the null hypothesis is true but the decision is "H_1", a mistake has occurred. Conversely, if the alternative hypothesis is true but the decision is "H_0", an error has also been made. These outcomes can be summarized in a decision matrix, shown in Table 5.2. The two types of errors are known as type I and type II errors or equivalently as α- and β-errors.

Table 5.2 Error matrix in significance tests

It is impossible to make both type I and type II error probabilities equal to zero simultaneously. For example, if we attempt to eliminate the type I error by always deciding in favor of "H_0", we will inevitably incur a type II error whenever the alternative hypothesis holds. While this would be the correct decision if the hypothesis holds, we will always run into the type II error if the alternative is valid. Completely avoiding an error is therefore not the right approach, but limiting the probability that an error occurs seems more useful. To do so, we need to understand that the test decision is based on data, which itself can be considered as a random variable. Consequently, the decision based on data needs to be viewed as random as well. Therefore, we may apply probability theory to quantify the probability that we make one or another error. This is the idea of statistical hypothesis tests.

Definition 5.4 *A* **statistical hypothesis test** *assumes two states* H_0 *and* H_1 *and two possible decisions "*H_0*" and "*H_1*". The decision rule is thereby constructed such that*

$$P\big(\text{“}H_1\text{”}\big|H_0\big) \leq \alpha \tag{5.3}$$

for a fixed small value of α*. The term* α *is* **called significance level**.

This definition is sufficient to specify a statistical test; the remaining elements follow conceptually. To make this explicit, we describe in general terms how a statistical hypothesis test is constructed.

1. Construct or select a test statistic $t(Y_1, \ldots, Y_n)$ that discriminates between H_0 and H_1, for example, large values of $t(Y_1, \ldots, Y_n)$ may provide evidence in favor of H_1, while small values support H_0. For definiteness, we assume in what follows that large values favor H_1.
2. Determine the distribution of $t(Y_1, \ldots, Y_n)$ under H_0. This step is crucial, as it allows us to derive a decision rule. In the following, we assume that this distribution is known.
3. Define the decision rule: reject the null hypothesis if $t(Y_1, \ldots, Y_n)$ exceeds a threshold value c, that is,

$$\text{“}\mathrm{H}_1\text{”} \quad \text{if} \quad t(Y_1, \ldots, Y_n) > c.$$

Choose the critical value c such that

4. Choose the critical value c such that
$$P(\text{“H}_1\text{”} \mid H_0) = P\big(t(Y_1, \ldots, Y_n) > c \mid H_0\big) \leq \alpha,$$
where α denotes the significance level.
5. Compute the observed value $t_{obs} = t(y_1, \ldots, y_n)$ and compare it with c. If $t_{obs} \leq c$, the decision is “H_0”; otherwise, the decision is “H_1”.

The construction of a hypothesis test is similar to the derivation of a p-value from above. However, while the p-value gives a measurement for the evidence against the hypothesis, the result of a hypothesis test is binary in the sense that the outcome is “H_0” or “H_1”. Still, the two concepts can be linked as follows.

Property *Assuming a statistical hypothesis test with significance level α, then it holds*
$$p\text{-}value \leq \alpha \Leftrightarrow \text{“}H_1\text{”}. \tag{5.4}$$

The proof of this statement is rather simple, which is why we give it here.

Proof Assume we have a test statistic $t()$ such that we decide in favor of H_1 if $t(y_{obs}) > c$, where y_{obs} are the observed data and the critical value c is determined such that
$$P(t(Y) > c \mid H_0) \leq \alpha.$$
Because
$$p\text{-value } = P(t(Y) > t(y_{obs}) \mid H_0),$$
it is directly clear that if $t(y_{obs}) > c$ then the p-value $\leq \alpha$ and vice versa. □

Property (5.4) gives a very simple and practical rule of thumb. Looking at the p-value, we can easily come up with a test decision. However, the p-values provide more information than the test decision itself, since it also determines how much evidence there is against the hypothesis. This speaks in favor of using the p-value, but in practice, statistical significance tests are dominant. Efron (2010) writes with respect to this issue “Fisher's famous $\alpha = 0.05$ direction for ‘significance’ has been overused, but has served a crucial purpose nevertheless in bringing order to scientific reporting.”

5.3 Classical Statistical Parameter Tests

Many tests have been developed in statistics for particular purposes, and we cannot provide a complete list here. There are, however, some tests which are commonly used. These will be discussed in the following. We thereby consider the following setup. We assume that the data are generated through a Fisher regular distribution
$$Y_i \sim f(y; \theta) \quad i.i.d.. \tag{5.5}$$
The decision problem equals
$$H_0 : \theta \in \Theta_0 \text{ and } H_1 : \theta \in \Theta_1,$$

where Θ_0 and Θ_1 are a disjoint decomposition of the parameter space Θ, i.e., $\Theta = \Theta_0 \cup \Theta_1$ and $\Theta_0 \cap \Theta_1 = \emptyset$. We restrict the subsequent presentation to the case where $\Theta_0 = \{\theta_0\}$ and $\Theta_1 = \Theta \setminus \{\theta_0\}$ or $\Theta_0 = \{\theta : \theta \leq \theta_0\}$ and $\Theta_1 = \Theta \setminus \Theta_0$, respectively. We call the first a **two-sided test**, i.e., $H_0 : \theta = \theta_0$ and $H_1 : \theta \neq \theta_0$, and the latter a **one-sided test**, i.e., $H_0 : \theta \leq \theta_0$ and $H_1 : \theta > \theta_0$.

5.3.1 Gauß Test and *t*-Test

We begin with the most frequently used tests, the Gauß test and the t-test. Both tests are based on the normal distribution and focus on the mean value.

We consider normal random variables

$$Y_i \sim \mathrm{N}(\mu, \sigma^2) \quad i.i.d., i = 1, \ldots, n$$

,and we want to test a hypothesis on μ. Exemplarily, we consider the one-sided hypothesis $H_0 : \mu \leq \mu_0$ versus $H_1 : \mu > \mu_0$. We take $\overline{Y}$ as the test statistic and use the decision rule

$$\text{“H}_1\text{”} \Leftrightarrow \overline{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i > c, \tag{5.6}$$

where the critical value c needs to be determined such that (5.3) holds. It is not difficult to see that $c = \mu_0 + z_{1-\alpha}\sigma \,/\, \sqrt{n}$. This follows since

$$\begin{aligned}
& P(\text{“H}_1\text{”} \mid H_0) \leq \alpha \\
\Leftrightarrow \quad & \underbrace{P(\overline{Y} > c \mid \mu \leq \mu_0)}_{\geq P(\overline{Y} > c \mid \mu = \mu_0)} \leq \alpha \\
\Leftarrow \quad & P(\overline{Y} > c \mid \mu = \mu_0) = \alpha \\
\Leftrightarrow \quad & P\left(\underbrace{\frac{\overline{Y} - \mu_0}{\sigma/\sqrt{n}}}_{\sim \mathrm{N}(0,1)} > \frac{c - \mu_0}{\sigma/\sqrt{n}} \right) = \alpha \quad .
\end{aligned}$$

Hence, we get

$$\begin{aligned}
z_{1-\alpha} &= \tfrac{c-\mu_0}{\sigma/\sqrt{n}} \\
c &= \mu_0 + z_{1-\alpha}\sigma \,/\, \sqrt{n}.
\end{aligned}$$

Note that to calculate c we need to know the variance σ^2 of the underlying distribution, which in most practical applications is unknown. In this case, we can rely on the *t*-distribution. If we replace σ with its estimate

$$\widehat{\sigma} = \sqrt{\tfrac{1}{n-1} \sum_{i=1}^{n} (Y_i - \overline{Y})^2},$$

we get

$$\tfrac{\overline{Y}-\mu_0}{\widehat{\sigma}/\sqrt{n}} \sim t_{n-1},$$

where t_{n-1} denotes the *t*-distribution with $n-1$ degrees of freedom. Hence, with calculations analogous as above, we obtain the decision rule (5.6) with

$$c = \mu_0 + t_{(n-1),1-\alpha}\widehat{\sigma} \,/\, \sqrt{n},$$

where $t_{(n-1),1-\alpha}$ is the $1-\alpha$ quantile of the corresponding *t* distribution.

In many applications, the goal is to compare the means of two different samples. In such scenarios, we consider two $i.i.d.$ samples:

$$Y_{1i} \sim \mathrm{N}(\mu_1, \sigma_1^2), i = 1, \ldots, n_1 \text{ and } Y_{2j} \sim \mathrm{N}(\mu_2, \sigma_2^2), j = 1, \ldots, n_2$$

For the two variances σ_1^2 and σ_2^2, we can either assume that they are equal, i.e., $\sigma_1^2 = \sigma_2^2$, which is called **variance homogeneity**, or we allow for different variances in the two samples, which is called **variance heterogeneity**. We want to test the hypothesis $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 \neq \mu_2$. This is a two-sided hypothesis. For two independent $i.i.d.$ samples, we make use of the decision rule

$$\text{“}\mathrm{H}_1\text{”} \Leftrightarrow |\overline{y}_1 - \overline{y}_2| > c$$

where the critical value *c* needs to be determined.

Assume first that σ_1^2 and σ_2^2 are both known. Since $\overline{Y}_1 - \overline{Y}_2$ follows a normal distribution with mean $\mu_1 - \mu_2$ and variance $\sigma_1^2 \,/\, n_1 + \sigma_2^2 \,/\, n_2$, the critical value *c* can be easily obtained with similar calculations as above, which leads to

$$c = z_{1-\frac{\alpha}{2}} \cdot \sqrt{\tfrac{\sigma_1^2}{n_1} + \tfrac{\sigma_2^2}{n_2}}.$$

Due to the two-sided test used here, the normal quantile changes to $1-\alpha \,/\, 2$ compared to $1-\alpha$ used above for the one-sided test.

Note that in the presence of variance homogeneity, that is, assuming $\sigma_1^2 = \sigma_2^2$, the formula above simplifies to

$$c = z_{1-\frac{\alpha}{2}}\sigma \cdot \sqrt{\tfrac{1}{n_1} + \tfrac{1}{n_2}}$$

with $\sigma^2 := \sigma_1^2 = \sigma_2^2$.

If the variances are unknown, we need to estimate them. In this case, we limit ourselves to the case of variance homogeneity, and we estimate the variance σ^2 through the so-called pooled variance estimate

$$\widehat{\sigma}^2 = \tfrac{(n_1-1)\widehat{\sigma}_1^2 + (n_2-1)\widehat{\sigma}_2^2}{n_1+n_2-2},$$

with

$$\widehat{\sigma}_1^2 = \tfrac{1}{n_1-1}\sum\nolimits_{i=1}^{n_1}\left(y_{1i} - \overline{y}_1\right)^2 \text{ and } \widehat{\sigma}_2^2 = \tfrac{1}{n_2-1}\sum\nolimits_{i=1}^{n_2}\left(y_{2i} - \overline{y}_2\right)^2,$$

where $\overline{y}_1$ and $\overline{y}_2$ are the arithmetic means of the first and second sample, respectively. The *t* distribution allows us to calculate the threshold *c* leading to the decision rule

$$c = t_{(n_1+n_2-2),1-\frac{\alpha}{2}}\widehat{\sigma} \cdot \sqrt{\tfrac{1}{n_1} + \tfrac{1}{n_2}}$$

where $t_{(n_1+n_2-2),1-\frac{\alpha}{2}}$ is the $1-\alpha\,/\,2$ quantile of the t distribution with n_1+n_2-2 degrees of freedom.

5.3.2 Likelihood-Based Tests

In Chap. 4, we derived asymptotic properties of the maximum likelihood estimator, which we now turn into three testing principles. We illustrate the idea in Fig. 5.5, where the likelihood function attains its maximum at $l(\hat{\theta})$. We consider the hypothesis $H_0 : \theta = \theta_0$ against the alternative $H_1 : \theta \neq \theta_0$. There are three natural quantities that can be examined to assess the validity of the hypothesis: First, we may consider the difference between θ_0 and $\hat{\theta}$. If this difference is large, it provides evidence against the hypothesis. This leads to the **Wald test**. Second, we may examine the vertical difference between the likelihood at its maximum, $l(\hat{\theta})$, and the likelihood evaluated under H_0, namely, $l(\theta_0)$. Again, if this difference is large, it indicates evidence against the hypothesis. This gives rise to the **likelihood-ratio test**. Finally, we may evaluate the slope of the likelihood function at θ_0, that is, the score function $s(\theta_0) = \partial l(\theta_0)\,/\,\partial\theta$. If this quantity is large in absolute value, it suggests evidence against the hypothesis. This defines the **score test**. We now discuss these three tests in turn.

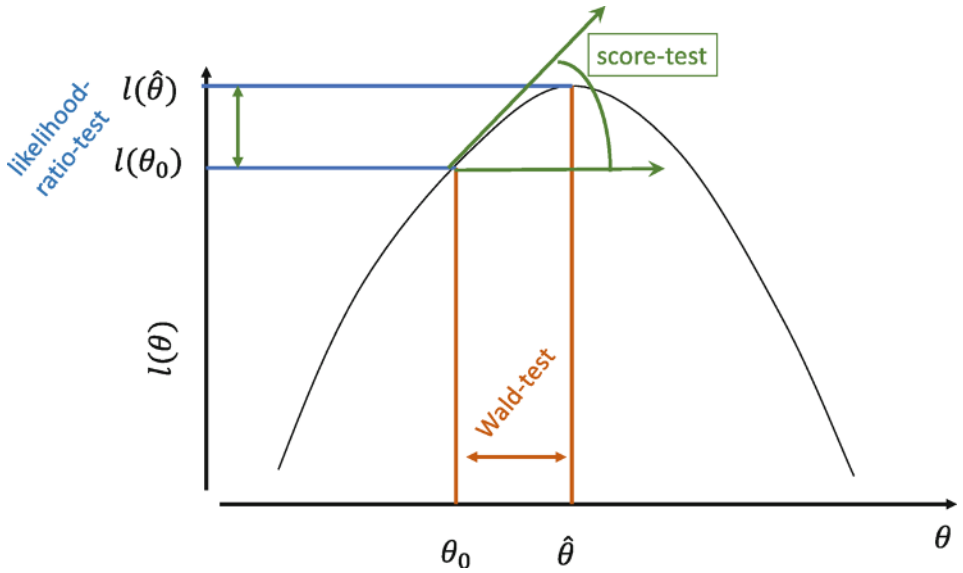


Fig. 5.5 Sketch of Wald, score, and likelihood-ratio test

Note that (under certain regularity conditions) the maximum likelihood estimator is asymptotically normally distributed, i.e.,

$$\hat{\theta} \overset{a}{\sim} \mathrm{N}(\theta, I^{-1}(\theta)).$$

Hence, if n is large, any hypothesis about θ can be treated as a hypothesis about the mean of a normal distribution. With the hypotheses $H_0 : \theta = \theta_0$ and $H_1 : \theta \neq \theta_0$, we obtain, as derived above for the Gauss test, the following decision rule:

$$\text{“H}_1\text{”} \Leftrightarrow \left|\hat{\theta} - \theta_0\right| > c,$$

where c is given by $c = z_{1-\alpha/2}\sqrt{I^{-1}(\theta_0)}$. Tests of this form are known as **Wald tests** and were proposed in Wald (1943). The Wald test is by far the most frequently used testing principle in statistics. It requires that we estimate the parameter as well as its variance.

Often, the variance is computed using the Fisher information evaluated not at the hypothetical value θ_0 but at the estimated value $\hat{\theta}$. In this case, the critical value is given by $c = z_{1-\alpha/2}\sqrt{I^{-1}(\hat{\theta})}$. Since under H_0 the estimator $\hat{\theta}$ converges to θ_0, it is reasonable to use $I(\hat{\theta})$ instead of $I(\theta_0)$. Moreover, standard software packages report $I(\hat{\theta})$, which makes this implementation convenient.

In Sect. 4.4, we derived the asymptotic distribution of the likelihood ratio. This result can now be used to construct a test. The likelihood ratio $lr(\hat{\theta};\theta) = 2\{l(\hat{\theta}) - l(\theta_0)\}$ is asymptotically Chi-squared distributed with p degrees of freedom, where p denotes the dimension of the parameter. Hence, large values of the likelihood ratio $lr(\hat{\theta};\theta)$ evaluated at θ_0 provide evidence against the hypothesis $H_0 : \theta = \theta_0$. This yields the decision rule of the **likelihood-ratio test**

$$\text{“H}_1\text{”} \Leftrightarrow 2\{l(\hat{\theta}) - l(\theta_0)\} > c.$$

The critical value c is obtained from the $1-\alpha$ quantile of the Chi-squared distribution with p degrees of freedom. It will be shown in the next part of the chapter that the likelihood-ratio test is the most powerful test. In other words, it achieves optimal performance in terms of discriminating between the hypotheses. More details are given in Sect. 5.4.

Finally, from (4.16), we can formulate a test based on the score. Since the score has mean value zero at the true parameter, we consider the hypothesis $H_0 : E(s(\theta_0)) = 0$. This leads to the decision rule

$$\text{“H}_1\text{”} \Leftrightarrow \left|s(\theta_0; y_1, \ldots, y_n)\right| > c$$

where $c = z_{1-\alpha/2}\sqrt{I(\theta_0)}$. We explicitly include the observations in the score to emphasize that the test decision is based on the data. The **score test** is similar to the Wald test, but there is an important difference. For the Wald test, the maximum likelihood estimate $\hat{\theta}$ must first be computed in order to assess H_0. This is not required for the score test. Here, it suffices to evaluate the score function at the hypothetical parameter value θ_0. The score test is therefore particularly useful when computing the maximum likelihood estimate is cumbersome or numerically demanding.

5.3.3 Chi-Squared Tests

We now focus on a discrete random variable $V \in \{1, \ldots, K\}$ with probabilities $P(V = k) = \pi_k$, where $\sum_{k=1}^{K} \pi_k = 1$. Assuming an $i.i.d.$ sample $V_1, \ldots, V_n$, the joint distribution is multinomial. Let

$$Y_k = \sum_{i=1}^{n} 1\{V_i = k\},$$

where $1\{.\}$ denotes the indicator function. We define $Y = (Y_1, \ldots, Y_K)$ as the vector of counts in the K categories. Then Y follows the multinomial distribution

$$\begin{aligned} P(Y = y; \pi) &= P((Y_1, \ldots Y_K) = (y_1, \ldots, y_K); \pi) \\ &= \binom{n}{y_1 \quad \ldots \quad y_K} \pi_1^{y_1} \ldots \pi_K^{y_K}. \end{aligned}$$

A typical hypothesis specifies particular values of the probabilities, as in the Mendel example above (see Example 5.1). Let π_0 denote the hypothetical vector and consider the hypothesis $H_0 : \pi = \pi_0$, where $\pi = (\pi_1, \ldots, \pi_K)$ and $\pi_0 = (\pi_{01}, \ldots, \pi_{0K})$. This leads to the so-called Chi-squared statistic

$$X^2 = \sum_{k=1}^{K} \frac{(y_k - e_k)^2}{e_k}, \tag{5.7}$$

where $e_k = n\pi_{0k}$ are the expected counts under H_0. It can be shown that X^2 (under suitable regularity conditions) is asymptotically Chi-squared distributed with $K-1$ degrees of freedom. We sketch the proof at the end of this section. Tests based on the statistic X^2 are known as Chi-squared tests.

The statistic X^2 in (5.7) has an intuitive interpretation. We compare the expected counts e_k, which would arise under H_0, with the observed counts y_k. The squared difference ensures nonnegativity, and dividing by the expected count standardizes the contribution of each cell. A small discrepancy matters more in cells with few expected observations than in cells with many expected observations. Thus, dividing by e_k can be interpreted as a form of standardization.

A central application of Chi-squared tests is the hypothesis of independence. Let $Z_i = (Z_{i1}, Z_{i2})$ be a bivariate discrete random variable with $Z_{i1} \in \{1, \ldots, K_1\}$ and $Z_{i2} \in \{1, \ldots, K_2\}$. We assume that

$$P(Z_{i1} = k, Z_{i2} = l) = \pi_{kl}$$

and test the hypothesis of independence, namely,

$$H_0 : P(Z_{i1} = k, Z_{i2} = l) = P(Z_{i1} = k) \cdot P(Z_{i2} = l).$$

To do so, we apply the Chi-squared test to the observed contingency table:

Table

Here, n_{kl} denotes the number of observations with $Z_{i1} = k$ and $Z_{i2} = l$, that is,

$$n_{kl} = \sum_{i=1}^{n} 1\{Z_{i1} = k, Z_{i2} = l\}.$$

We define $n_{\cdot l} = \sum_k n_{kl}$ as the column totals, $n_{k\cdot}$ as the row totals, and $n_{\cdot\cdot} = n$ as the sample size.

Under independence, the joint probabilities factor as $\pi_{kl} = \pi_{k\cdot}\pi_{\cdot l}$. The maximum likelihood estimates are

$$\widehat{\pi}_{k\cdot} = n_{k\cdot} \,/\, n_{\cdot\cdot} \quad \text{and} \quad \widehat{\pi}_{\cdot l} = n_{\cdot l} \,/\, n_{\cdot\cdot},$$

so that the expected counts are

$$e_{kl} = n_{\cdot\cdot}\widehat{\pi}_{k\cdot}\widehat{\pi}_{\cdot l} = \tfrac{n_{k\cdot}n_{\cdot l}}{n_{\cdot\cdot}}.$$

This leads to the Chi-squared statistic

$$X^2 = \sum\nolimits_{k=1}^{K_1}\sum\nolimits_{l=1}^{K_2} \tfrac{(n_{kl}-e_{kl})^2}{e_{kl}}.$$

The distribution of $\mathscr{X}^2$ can again be approximated by a Chi-squared distribution, but determining the correct number of degrees of freedom requires some care. When the test was introduced by Pearson (1904) and further developed by Fisher (1922), there was considerable debate about this issue; see Baird (1983). According to Fisher's result, the degrees of freedom are given by

df = Number of cells - number of constraints - number of estimated parameters.

For the independence test, there are $K_1 \times K_2$ cells. There is one constraint, since all probabilities must sum to one. We estimate K_1-1 parameters $\widehat{\pi}_{k\cdot}$ and K_2-1 parameters $\widehat{\pi}_{\cdot l}$, giving a total of $K_1 + K_2-2$ estimated parameters. Hence,

$$K_1 \times K_2-1 - (K_1 + K_2-2) = (K_1-1) \times (K_2-1)$$

is the resulting number of degrees of freedom.

Proof We sketch the proof that X^2 defined in (5.7) follows a Chi-squared distribution with $K-1$ degrees of freedom. The hypothesis involves only the first $K-1$ parameters, since $\pi_K = 1 - \pi_1 - \ldots - \pi_{K-1}$. We collect these parameters in the vector θ. A test can then be constructed using the likelihood-based results above. In particular, we consider the statistic

$$2\Big\{l(\hat{\theta}) - l(\theta_0)\Big\}$$

where

$$l(\theta) = \sum\nolimits_{k=1}^{K} y_k \log \pi_k + \text{const.}$$

The log-likelihood ratio can be approximated using a Taylor expansion of the logarithm:

$$\begin{aligned} 2\sum\nolimits_{k=1}^{K} y_k \log\left(\tfrac{\widehat{\pi}_k}{\pi_{0k}}\right) &= -2\sum\nolimits_{k=1}^{K} y_k \log\left(\tfrac{\pi_{0k}}{\widehat{\pi}_k}\right) \\ &= -2\sum\nolimits_{k=1}^{K} y_k \log\left(1 + \tfrac{\pi_{0k}-\widehat{\pi}_k}{\widehat{\pi}_k}\right) \\ &\approx -2\sum\nolimits_{k=1}^{K} y_k\left(\tfrac{\pi_{0k}-\widehat{\pi}_k}{\widehat{\pi}_k}\right) + \sum\nolimits_{k=1}^{K} y_k\left(\tfrac{\pi_{0k}-\widehat{\pi}_k}{\widehat{\pi}_k}\right)^2 \end{aligned}$$

Since $\widehat{\pi}_k = \dfrac{y_k}{n}$ and both π_{0k} and $\widehat{\pi}_k$ sum to one, the first term vanishes. Using $e_k = n\pi_{0k}$, the second term becomes

$$\sum_{k=1}^{K} y_k \left(\frac{\pi_{0k} - \widehat{\pi}_k}{\widehat{\pi}_k} \right)^2 = \sum_{k=1}^{K} \frac{(e_k - y_k)^2}{y_k}.$$

Replacing y_k by its expectation under H_0 yields the Chi-squared statistic (5.7). Thus, X^2 is asymptotically equivalent to the likelihood ratio when π_0 are the true probabilities, and therefore $\mathscr{X}^2$ is asymptotically Chi-squared distributed with $K-1$ degrees of freedom. □

5.3.4 Distribution-Free Tests

A large class of tests are known as distribution-free tests, sometimes also referred to as nonparametric tests. The terminology is not entirely consistent and can be confusing, since nonparametric methods in statistics also include approaches to functional estimation. We therefore use the term distribution-free tests to emphasize that no specific parametric model is assumed for the data. Instead, we rely on a general data-generating process (2.3). The focus is now on hypotheses concerning the distribution $G(\cdot)$, for example, whether the median takes a particular value.

We present the idea in a general two-sample setting, assuming

$$Y_i \sim G_Y(\cdot), i = 1, \ldots n \text{ and } X_j \sim G_X(\cdot), j = 1, \ldots m.$$

We test whether the two distributions coincide, that is,

$$H_0 : G_Y(\cdot) \equiv G_X(\cdot).$$

This hypothesis is quite general. To simplify, we restrict attention to the question of whether the two distributions share the same median, assuming suitable regularity conditions.

To construct a test statistic, we form the pooled sample $z = (y_1, \ldots y_n, x_1, \ldots x_m)$ and order it such that

$$z_1 \leq z_2 \leq z_3 \leq \ldots \leq z_{n+m}$$

where each z_j originates from either the *Y* or the *X* sample. Let $rg(x_j)$ denote the rank of X_j in the pooled and ordered sample *z*, that is, $rg(x_j) = k$ if and only if $z_k = x_j$, assuming there are no ties (i.e., all observations are distinct). Figure 5.6 illustrates the idea. The test statistic is then defined as

$$T = \sum_{j=1}^{m} rg(x_j).$$

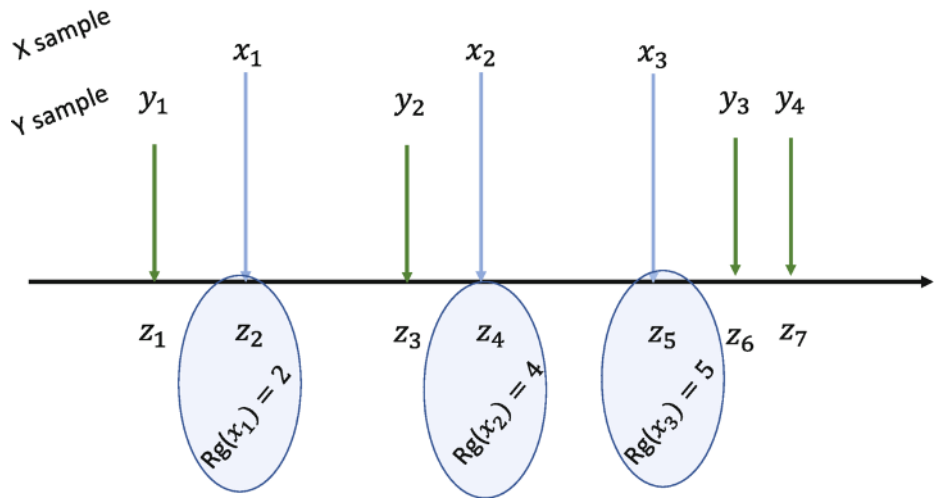


Fig. 5.6 Sketch of Wilcoxon test statistics

Large or small values of T provide evidence against the hypothesis that the median of $G_Y(\cdot)$ coincides with that of $G_X(\cdot)$. Under H_0, the distribution of T is known as the Wilcoxon distribution, and the resulting procedure is called the **Wilcoxon test**, introduced in Wilcoxon (1945). The decision rule is

$$\text{“H}_1\text{”} \Leftrightarrow (T < w_{\alpha/2}) \text{ or } (T > w_{1-\alpha/2}).$$

Here, w_α denotes the α-quantile of the Wilcoxon distribution. The test is implemented in standard statistical software and can therefore be readily applied in practice.

5.3.5 Kolmogorov-Smirnov Test

The Kolmogorov-Smirnov test provides another example of a distribution-free test. It allows for a direct assessment of whether two distributions coincide. In this context, we introduce the test for situations in which the goal is to determine whether the data arise from a specified model. Within the general framework (2.3), we consider the hypothesis

$$H_0 : G(\cdot) = F(\cdot; \theta).$$

For the moment, we assume that θ is known.

The test is based on the difference between the empirical distribution function and the hypothesized distribution. With $F_n(y) = \frac{1}{n} \sum_{i=1}^{n} 1\{Y_i \leq y\}$ denoting the empirical cumulative distribution function, we examine whether the distance between $F(y; \theta)$ and $F_n(y)$ is small, taking the sample size n into account. To this end, we construct a test statistic based on the difference between $F(y; \theta)$ and $F_n(y)$. Specifically, we consider the supremum of the absolute difference

$$D_n := \sup_y |F_n(y) - F(y; \theta)|. \tag{5.8}$$

This quantity is known as the Kolmogorov-Smirnov distance. Figure 5.7 illustrates the construction of D_n with an example. The empirical distribution function of a sample is shown as $F_n(\cdot)$, and the standard normal distribution serves as the hypothetical

distribution $F(\cdot;\theta)$. This is displayed in the top plot. The bottom plot shows the difference $F_n(y) - F(y)$ for all values of *y*, which attains its maximum (in absolute value) at the locations indicated by the vertical line. This determines the value of D_n.

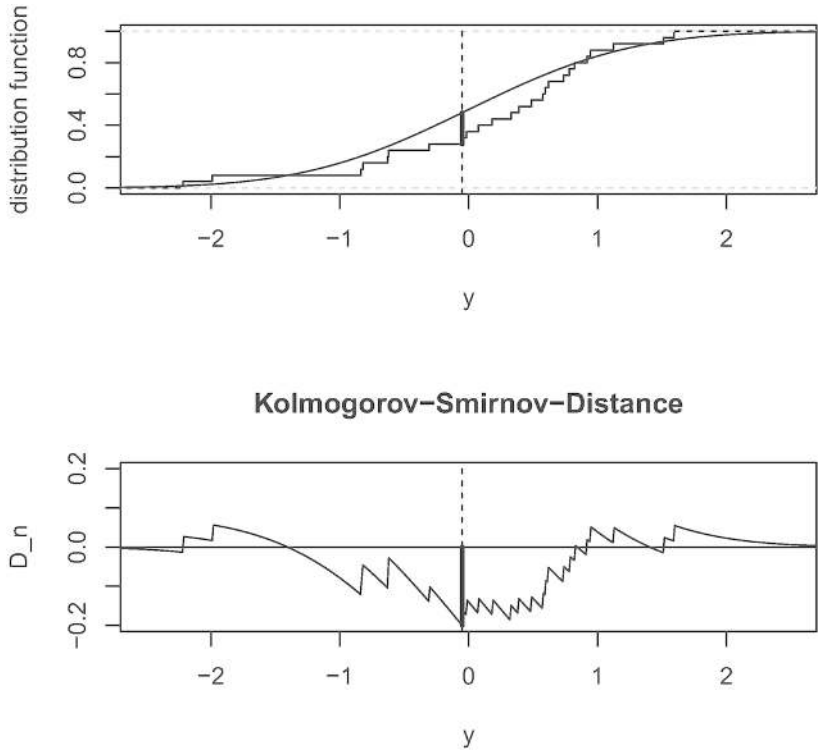


Fig. 5.7 Empirical distribution function and hypothetical distribution (top plot). The difference between these two distributions (bottom plot) with the maximum absolute deviation indicated by a vertical line

The next step is to determine the distribution of D_n under H_0, that is, when $F(y;\theta)$ represents the true distribution. It can be shown that the distribution of D_n does not depend on the specific form of $F(\cdot;\theta)$ but has a universal form. Consequently, D_n follows a pivotal distribution.

Property *The distribution of* D_n *given in* (5.8) *(asymptotically) follows a Kolmogorov-Smirnov distribution. In particular, it is pivotal, meaning that it does not depend on* $F(\cdot;\theta)$. □

The proof is given at the end of this section. The corresponding test decision is

$$\text{“H}_1\text{”} \Leftrightarrow D_n \geq KS_{1-\alpha},$$

where $KS_{1-\alpha}$ denotes the $(1-\alpha)$-quantile of the Kolmogorov-Smirnov distribution.

The test described above assumes that the parameter θ is known, which is rarely the case in practice. In most applications, θ must be estimated, and deriving the exact distribution of the statistic with an estimated parameter,

$$\widehat{D}_n = \sup_y |F_n(y) - F(y; \hat{\theta})|,$$

is considerably more involved. Exact results are available only for specific distributions, such as the normal or exponential distribution. For an unspecified distribution, the Kolmogorov-Smirnov distribution holds only asymptotically for D_n. For further details, see Berger and Zhou (2014).

Proof We now show that the distribution of D_n is pivotal. Let $F^{-1}(y; \theta)$ denote the inverse distribution function; more precisely, for $x \in [0, 1]$,

$$F^{-1}(x; \theta) = \min \{y : F(y; \theta) \geq x\}.$$

Setting $y = F^{-1}(x; \theta)$ yields $F(F^{-1}(x; \theta); \theta) = x$, and therefore

$$P(\sup_y |F_n(y) - F(y; \theta)| \leq t) = P(\sup_x |F_n(F^{-1}(x; \theta)) - x| \leq t) \tag{5.9}$$

where

$$F_n(F^{-1}(x; \theta)) = \frac{1}{n} \sum_{i=1}^n 1\{Y_i \leq F^{-1}(x; \theta)\} = \frac{1}{n} \sum_{i=1}^n 1\{F(Y_i; \theta) \leq x\}.$$

If $F(y; \theta)$ is the true distribution, then $F(Y_i; \theta)$ follows a uniform distribution on $[0, 1]$. Indeed,

$$P(F(Y_i; \theta) \leq x) = P(Y_i \leq F^{-1}(x; \theta)) = F(F^{-1}(x; \theta); \theta) = x.$$

Defining $U_i = F(Y_i; \theta)$, which is uniformly distributed on $[0, 1]$, allows us to rewrite (5.9) as

$$P\left(\sup_x \left| \frac{1}{n} \sum_{i=1}^n 1\{U_i \leq x\} - x \right| \leq t\right).$$

This probability does not depend on the underlying distribution, and thus the distribution of D_n is pivotal. □

5.4 Power of a Test

In Fig. 5.3, we already illustrated how the *p*-value behaves when the hypothesis is false. We now elaborate on this further. So far, we have mainly examined the performance of a test under the hypothesis, that is, assuming that H_0 holds. This perspective led to the *p*-value in Definition 5.2 and to the hypothesis tests introduced in Definition 5.4. Focusing only on behavior under H_0 is restrictive, which can be illustrated by the following pathological test. Consider any hypothesis to be tested using data. Instead of looking at the data at all, we base the test decision on a single uniformly distributed random variable $U \sim \text{Uniform}[0, 1]$. We then reject the hypothesis if $U \leq \alpha$, that is, “H_1” $\Leftrightarrow U \leq \alpha$. Apparently, we have $P(\text{“H}_1\text{”} \mid H_0) = P(\text{“H}_1\text{”}) = P(U \leq \alpha) = \alpha$. Consequently, this decision rule defines a valid statistical hypothesis test according to (5.3). Needless to say, such a test is useless. Hence, condition (5.3) is not sufficient to guarantee that a test is informative. In general, we also need to study the type II error. In particular, a test is useful if it fails to reject H_0 under H_1 only with small probability. To

make this precise, we introduce a few technical definitions. First, we define the power of a test.

Definition 5.5 *The power of a statistical hypothesis test is defined as* $P(\text{“}H_1\text{”}|H_1)$.

The power of a test is the probability of correctly rejecting the null hypothesis when the alternative is true. This is best understood through a simple example. Assume $i.i.d.$ normally distributed data $Y_i \sim \mathrm{N}(\mu, \sigma^2)$ with σ^2 known, and consider the hypothesis $H_0 : \mu \leq \mu_0$. Using the results derived above, we obtain the decision rule:

$$\text{“}\mathrm{H}_1\text{”} \Leftrightarrow \overline{Y} \geq \mu_0 + z_{1-\alpha}\frac{\sigma}{\sqrt{n}}$$

The power of the test can then be computed for any $\mu > \mu_0$ as

$$\begin{aligned} P(\text{“}\mathrm{H}_1\text{”} \mid \mu) &= P\Big(\overline{Y} \geq \mu_0 + z_{1-\alpha}\tfrac{\sigma}{\sqrt{n}} \mid \mu\Big) \\ &= P\Big(\tfrac{\overline{Y}-\mu_0}{\sigma/\sqrt{n}} \geq z_{1-\alpha} \mid \mu\Big) \\ &= P\Big(\tfrac{\overline{Y}-\mu}{\sigma/\sqrt{n}} \geq z_{1-\alpha} + \tfrac{\mu_0-\mu}{\sigma/\sqrt{n}} \mid \mu\Big) \\ &= 1 - \Phi\Big(z_{1-\alpha} + \tfrac{\mu_0-\mu}{\sigma/\sqrt{n}}\Big). \end{aligned}$$

We can plot $P(\text{“}\mathrm{H}_1\text{”} \mid \mu)$ as a function of μ. This is shown in Fig. 5.8 for $\mu_0 = 0$. We consider values $-1 \leq \mu \leq 1$ and two sample sizes, namely, $n = 30$ and $n = 60$. The top plot shows the full function. Note that for values $\mu \leq \mu_0$ the power is less than or equal to α, which follows from (5.3). The maximum over $\mu \leq \mu_0$ is attained at $\mu = \mu_0$. Hence, the left-hand side of the plot is not informative for comparing performance under alternatives: for all $\mu \leq \mu_0$, we have $P(\text{“}\mathrm{H}_1\text{”} \mid \mu) \leq P(\text{“}\mathrm{H}_1\text{”} \mid \mu_0) \leq \alpha$. We visualize this in the middle plot in Fig. 5.8 by suppressing the function values for $\mu \leq \mu_0$.

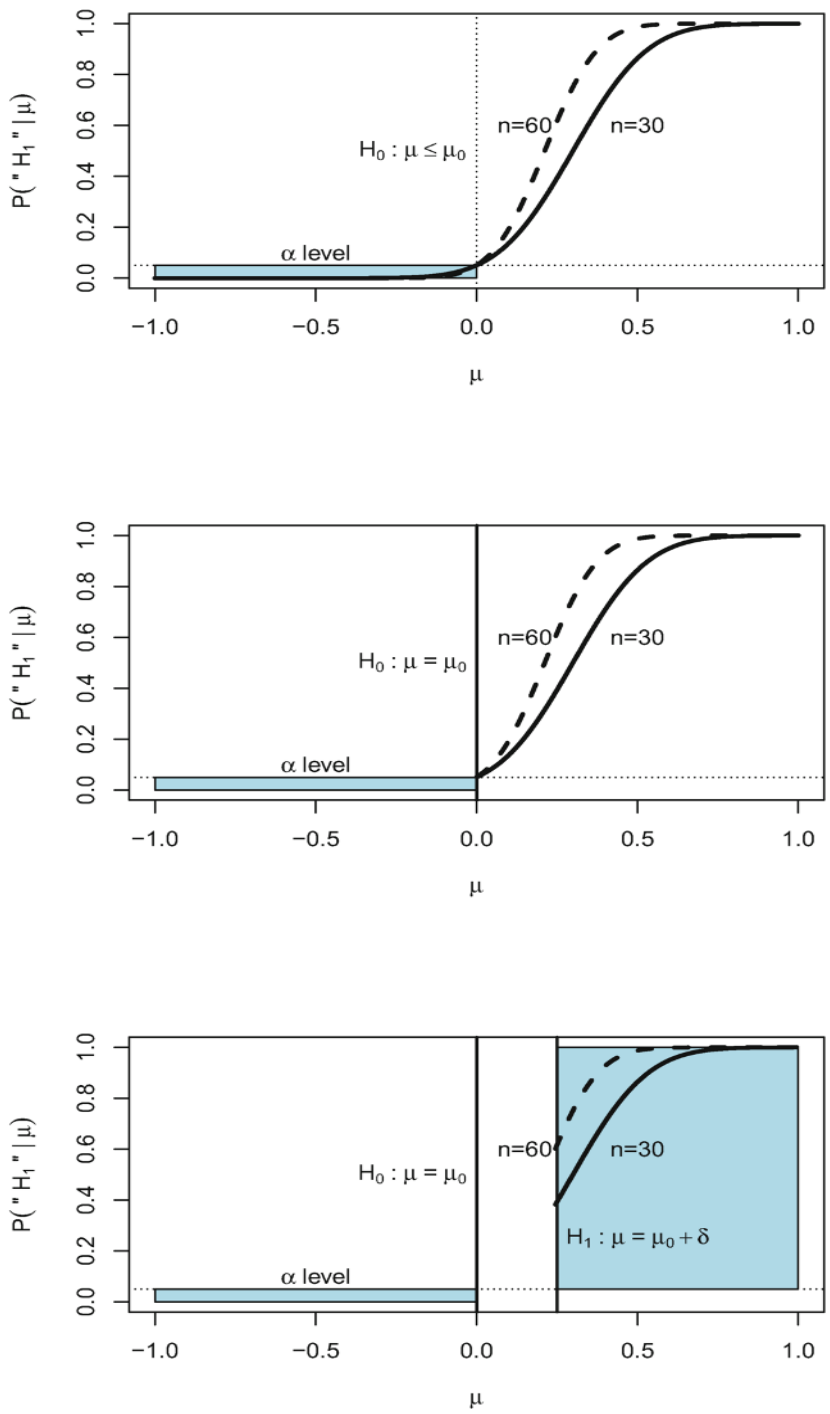


Fig. 5.8 Power of a test on $H_0 : \mu \leq \mu_0$ with $\mu_0 = 0$. The top plot shows $P(\text{“H}_1\text{”} \mid \mu)$ for all values of μ. The middle plot shows $P(\text{“H}_1\text{”} \mid H_1)$ for $\mu > \mu_0$, with the α level indicated for $\mu \leq \mu_0$. The bottom plot highlights the relevant power $P(\text{“H}_1\text{”} \mid \mu = \mu_0 + \delta)$, illustrating the effect of the sample size

For $\mu > \mu_0$, the smallest achievable power is α (equivalently, the largest possible type II error is $1 - \alpha$). This occurs when μ is only infinitesimally larger than μ_0. However, if we consider an alternative bounded away from μ_0, say, $\mu = \mu_0 + \delta$ with $\delta > 0$, then $P(\text{“H}_1\text{”} \mid \mu_0 + \delta)$ increases with increasing sample size. Equivalently, the

type II error probability $P(\text{“H}_0\text{”} \mid H_1)$ decreases. The bottom plot in Fig. 5.8 highlights this idea by focusing on a fixed effect size, for example, $\delta = 0.25$. The parameter δ represents the relevant difference one wants to detect. Once δ is fixed, one can determine the sample size needed to control the type II error. This is particularly important in experimental design, and the corresponding calculations are outlined below for the normal example.

In Fig. 5.9, we focus on the relevant quantities from Fig. 5.8 and show $P(\text{“H}_1\text{”} \mid \mu = \mu_0 + \delta)$ for different sample sizes, namely, $n = 60$, $n = 90$, and $n = 120$. If we require the type II error to be at most $\beta = 0.2$, then the power must be at least $1 - \beta$ at $\mu = \mu_0 + \delta$. For $\beta = 0.2$, this is illustrated in Fig. 5.9 by the blue shaded area in the top-right corner. In this way, we can control both the type I error and the type II error, and we see that the required sample size lies between $n = 90$ and $n = 120$. In fact, we can compute the required sample size exactly by solving the following implicit equation with respect to n:

$$n_0 = \arg\min\{n : 1 - \beta \leq P(\text{“H}_1\text{”} \mid \mu = \mu_0 + \delta)\} \tag{5.10}$$

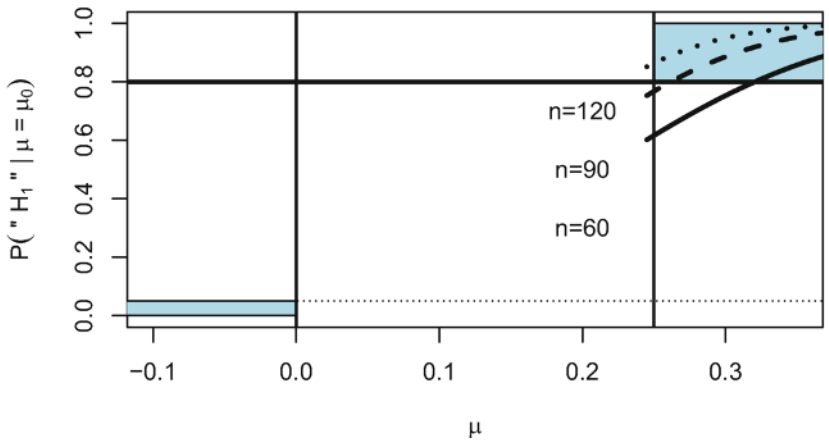


Fig. 5.9 Power of a test on $H_0 : \mu \leq \mu_0$ with $\mu_0 = 0$. Different sample sizes yield different power

$$= \arg\min\left\{n : 1 - \beta \leq 1 - \Phi\left(z_{1-\alpha} + \tfrac{\mu_0 - (\mu_0 + \delta)}{\sigma/\sqrt{n}}\right)\right\} \tag{5.11}$$

$$= \arg\min\left\{n : 1 - \beta \leq 1 - \Phi\left(z_{1-\alpha} + \tfrac{-\delta}{\sigma/\sqrt{n}}\right)\right\} \tag{5.12}$$

In the concrete example, we obtain $n_0 = 99$ for $\beta = 0.2$.

5.5 Neyman-Pearson Lemma

The next question we address is whether an optimal test exists. Referring to Fig. 5.8, we seek a test that satisfies (5.3) while maximizing power, that is, maximizing $P(\text{“H}_1\text{”} \mid \mu = \mu_0 + \delta)$ among all tests of the same level. This is formalized by the following lemma.

Property (Neyman-Pearson-Lemma) *Let* $H_0 : \theta = \theta_0$ *be tested against* $H_1 : \theta = \theta_1$ *with a statistical hypothesis test using level* α*. The most powerful test has the decision rule*

$$\text{“}H_1\text{”} \Leftrightarrow l(\theta_0) - l(\theta_1) \leq c,$$

where c is determined such that $P(\text{“}H_1\text{”} \big| H_0) \leq \alpha$*. This test is called the* **Neyman-Pearson test**. □

The lemma was published by Neyman and Pearson (1933), and the result highlights once more the central role of the likelihood. Utilizing the distribution of the likelihood ratio as in (4.21), we can quantify the threshold value *c* using a Chi-squared distribution. The result also motivates the likelihood-ratio test introduced in Sect. 5.3.2, as it is the most powerful test in the Neyman-Pearson sense. The proof of the lemma is straightforward and is given below.

Proof Let *y* be the data, which gives $l(\theta) = \log f(y;\theta)$ as the log-likelihood. The above decision rule can then be rewritten as

$$l(\theta_0) - l(\theta_1) \leq c \Leftrightarrow \frac{f(y;\theta_0)}{f(y;\theta_1)} \leq k = \exp(c).$$

We define with $\varphi(y)$ the outcome of the Neyman-Pearson test with

$$\varphi(y) = \begin{cases} 1 & \text{if } \frac{f(y;\theta_0)}{f(y;\theta_1)} \leq k \\ 0 & \text{otherwise.} \end{cases}$$

Hence, if $\varphi(y) = 1$, we decide “H_1”, and if $\varphi(y) = 0$, we conclude with decision “H_0”. Now, we take an arbitrary statistical significance test for which we similarly write the test outcome as a function $\psi(y) \in \{0, 1\}$, where $\psi(y) = 1$ means we decide for “H_1” and $\psi(y) = 0$ for “H_0”. Let us now assume that $\theta = \theta_1$, i.e., H_1 holds. We need to prove that

$$P\Big(\psi(Y) = 1; \theta_1\Big) \leq P\Big(\varphi(Y) = 1; \theta_1\Big),$$

that is, if H_1 holds, then the Neyman-Pearson test decides for H_1 with a higher probability than any other arbitrarily chosen test with the same significance level. Note that $P(\psi(Y) = 1; \theta) = E_\theta(\psi(Y))$ and the same holds for the Neyman-Pearson test $\varphi(Y)$, such that

$$P\Big(\varphi(Y) = 1; \theta_1\Big) - P\Big(\psi(Y) = 1; \theta_1\Big) = \int \{\varphi(y) - \psi(y)\} f(y; \theta_1) \mathrm{d}\, y.$$

We have to prove that this integral is greater than or equal to zero. The above integral can be labelled over three regions as

$$\begin{aligned} R_1 &= \left\{y : \tfrac{f(y;\theta_0)}{f(y;\theta_1)} < k\right\} \\ R_2 &= \left\{y : \tfrac{f(y;\theta_0)}{f(y;\theta_1)} > k\right\} \\ R_3 &= \left\{y : \tfrac{f(y;\theta_0)}{f(y;\theta_1)} = k\right\}. \end{aligned}$$

For region R_1, we have $\varphi(y) \equiv 1$ and $f(y; \theta_1) > f(y; \theta_0) \,/\, k$, such that

$$\int_{y\in R_1}\Big[\varphi(y)-\psi(y)\Big]f(y;\theta_1)\mathrm{d}\ y \geq \tfrac{1}{k}\int_{y\in R_1}\Big[\varphi(y)-\psi(y)\Big]f(y;\theta_0)\mathrm{d}\ y.$$

For region R_2, we have $\varphi(y)\equiv 0$ and $-f(y;\theta_1) > -f(y;\theta_0)\ /\ k$, such that

$$\int_{y\in R_2}\Big[\varphi(y)-\psi(y)\Big]f(y;\theta_1)\mathrm{d}\ y = -\int_{y\in R_2}\psi(y)f(y;\theta_1)\mathrm{d}\ y$$

$$\geq -\tfrac{1}{k}\int_{y\in R_2}\psi(y)f(y;\theta_0)\mathrm{d}\ y = \tfrac{1}{k}\int_{y\in R_2}\Big[\varphi(y)-\psi(y)\Big]f(y;\theta_0)\mathrm{d}\ y.$$

And finally for $y\in R_3$, we have $f(y;\theta_1) = f(y;\theta_0)\ /\ k$, such that

$$\int_{y\in R_3}\Big[\varphi(y)-\psi(y)\Big]f(y;\theta_1)\mathrm{d}\ y = \tfrac{1}{k}\int_{y\in R_3}\Big[\varphi(y)-\psi(y)\Big]f(y;\theta_0)\mathrm{d}\ y.$$

Collecting the right-hand sides of the three regions, we can conclude

$$\int\Big[\varphi(y)-\psi(y)\Big]f(y;\theta_1)\mathrm{d}\ y \geq \tfrac{1}{k}\int\Big[\varphi(y)-\psi(y)\Big]f(y;\theta_0)\mathrm{d}\ y. \tag{5.13}$$

As both tests have a significance level of α, one obtains

$$\alpha = P(\text{“H}_1\text{”}\big|H_0) = \int\varphi(y)f(y;\theta_0)\mathrm{d}\ y = \int\psi(y)f(y;\theta_0)\mathrm{d}\ y,$$

such that the right-hand side in (5.13) is equal to zero and the proof is completed. □

5.6 Testing and Confidence Intervals

We have now introduced two key concepts for statistical inference, namely, confidence intervals and statistical hypothesis tests. This raises the question of how the two ideas are related. In fact, we can show that they are not only related but that one can be constructed from the other. In other words, given a confidence interval, one can directly construct a corresponding test, and vice versa.

Let us motivate this relationship and reconsider the definition of a confidence interval given in (3.12) in Definition 3.9. The confidence interval was defined as $[t_l(y), t_r(y)]$ such that

$$P_\theta\big(t_l(Y)\leq\theta\leq t_r(Y)\big)\geq 1-\alpha,$$

where $Y=(Y_1,\ldots,Y_n)$. Given this interval, we can define a statistical hypothesis test for $H_0:\theta=\theta_0$ via the decision rule

$$\text{“H}_1\text{”}\Leftrightarrow\theta_0\notin[t_l(y),t_r(y)].$$

Thus, parameter values outside the confidence interval are rejected, while values inside the interval are not rejected. This construction yields a statistical hypothesis test with significance level α. To see this, define the test decision function

$$\varphi_\theta(Y)=\begin{cases}0 & \text{if } \theta\in[t_l(y),t_r(Y)]\\ 1 & \text{otherwise.}\end{cases}$$

Then,

$$\begin{aligned} P(\text{“H}_1\text{”} \mid H_0) &= 1 - P(\text{“H}_0\text{”} \mid H_0) \\ &= 1 - P(\varphi(Y) = 0 \mid \theta = \theta_0) \\ &= 1 - \underbrace{P(t_l(Y) \leq \theta_0 \leq t_r(Y))}_{\geq 1-\alpha} \leq \alpha. \end{aligned}$$

The above construction takes a confidence interval and builds a hypothesis test from it. We can also proceed in the other direction: starting from a hypothesis test, we can construct a confidence interval based on the test. Assume we have a test defined by

$$\varphi_\theta(Y) = \begin{cases} 0 & \text{if “H}_0\text{”} \\ 1 & \text{if “H}_1\text{”}. \end{cases}$$

The corresponding confidence interval can then be defined as

$$CI(y) = \{\theta : \varphi_\theta(y) = 0\}.$$

Since $P(\text{“H}_1\text{”} \mid H_0) = P(\varphi_\theta(Y) = 1 \mid \theta) \leq \alpha$, we obtain

$$\begin{aligned} P(\theta \in CI(Y) \mid \theta) &= 1 - P(\theta \notin CI(Y) \mid \theta) \\ &= 1 - P(\varphi_\theta(Y) = 1 \mid \theta) \\ &= 1 - P(\text{“H}_1\text{”} \mid H_0) \geq 1 - \alpha. \end{aligned}$$

We thus see that confidence intervals and statistical hypothesis tests are two sides of the same coin.

5.7 Multiple Testing

So far, we have focused on univariate (or at least low-dimensional) parameters and a single hypothesis. But what happens if we have a multivariate parameter $\theta = (\theta_1, \ldots \theta_p)$ and multiple hypotheses of the form

$$H_{0j} : \theta_j = \theta_{0j}$$

with $j \in \mathcal{J} \subset \{1, \ldots, p\}$ and $\theta_0 = (\theta_{01}, \ldots, \theta_{0p})$? Assume that $|\mathcal{J}| = m$, that is, we aim to test m components of θ at the same time. Instead of conducting one joint test on all m components simultaneously, we test each of the m hypotheses separately. Let $p_1, \ldots, p_m$ be the p-values for these m tests, leading to the m decision rules

$$\text{“H}_{1j}\text{”} \Leftrightarrow p_j \leq \alpha \qquad j = 1, \ldots, m.$$

For each individual test, we have a statistical hypothesis test in the sense that

$$P(\text{“H}_{1j}\text{”} \mid H_{0j}) \leq \alpha. \tag{5.14}$$

But what can be said about the probability of at least one false rejection when all tests are considered together?

Let us view the problem from a different angle. For a single test, we have an error matrix as in Table 5.3. Combining all m tests leads to the aggregated error matrix shown in Table 5.3.

Table 5.3 Error matrix for multiple testing

Note that the entries in the table are unknown since we do not know which hypotheses are true and which are false. This motivates the use of capital letters. The same holds for the margins. The only number we know in advance is *m*, the total number of tests, and once the tests are conducted, we also know *r*, the number of rejected hypotheses. Apparently, if $r = m$ or $r = 0$, that is, if all hypotheses are rejected or none are rejected, then we learn more about the entries in the table. In general, however, the interior of the table remains unknown.

Multiple testing focuses on the column "H_{1j}" and asks what can be said about *S* and *V*. To formalize this, we define the family-wise error rate.

Definition 5.6 *The***family-wise error rate***(FWER) is defined as*

$$\alpha_{FWER} := P(V > 0 \mid H_{0j} \text{ for } j = 1, \ldots m), \tag{5.15}$$

*with V as defined in Table*5.3.

We can interpret the FWER as the probability of falsely rejecting at least one of the *m* hypotheses when all hypotheses are true. It is important to relate the FWER to the individual decision rules (5.14). In fact, we will show that $\alpha_{FWER} \geq \alpha$, and even worse, if the tests are not perfectly dependent, we will have $\alpha_{FWER} \to 1$ as *m* increases. To see this, note that

$$\begin{aligned} \alpha_{FWER} &= P(V > 0 \mid H_{0j} \text{ for } j = 1, \ldots m) \\ &= P((p_1 \leq \alpha) \vee \ldots \vee (p_m \leq \alpha) \mid H_{0j} \text{ for } j = 1, \ldots m) \\ &= 1 - P((p_1 > \alpha) \wedge \ldots \wedge (p_m > \alpha) \mid H_{0j} \text{ for } j = 1, \ldots m) \end{aligned} \tag{5.16}$$

where "$\vee$" denotes logical "or" and "$\wedge$" denotes logical "and." If we assume independence of the tests, the last line (5.16) simplifies to

$$1 - \prod_{j=1}^{m} P((p_j > \alpha) \mid H_{0j}) = 1 - (1 - \alpha)^m \tag{5.17}$$

using the fact that the *p*-value is uniformly distributed under the null. It is now easy to see that as *m* increases, the FWER α_{FWER} converges to 1. In particular, we obtain $\alpha_{FWER} > \alpha$, that is, if *m* independent tests are carried out at significance level α, then the overall probability of at least one false rejection, α_{FWER}, is larger than the per-test level α in (5.14).

We may therefore require that the FWER α_{FWER} should not exceed a given significance level and ask how to adjust the individual significance levels so that the overall error remains controlled. It can be shown that α_{FWER} is largest when the tests are independent. Hence, in general,

$$\alpha_{FWER} \geq 1 - (1 - \alpha)^m \quad \Leftrightarrow \quad 1 - \alpha_{FWER} \leq (1 - \alpha)^m. \tag{5.18}$$

It can further be shown that for small α we have $(1 - \alpha)^m > (1 - m\alpha)$. Therefore, if we set the significance level for each individual test to $\alpha_{adjust} = \alpha \,/\, m$, then the resulting FWER is less than or equal to α. This adjustment is called the Bonferroni adjustment and goes back to Bonferroni (1936).

Property (Bonferroni Adjustment) *The FWER α_{FWER} is limited to α if we set the significance level of each individual test in* (5.14) *to $\alpha \,/\, m$, where m is the number of tests. The adjusted significance level is defined by $\alpha_{adjust} = \alpha \,/\, m$.* □

Though the Bonferroni adjustment is simple, it is not always recommended. In particular, if *m* is large, other procedures may be preferable. We refer to Efron (2010) for further details. An alternative is to consider the ratio $V \,/\, r$ in the table above, known as the *false discovery rate*. The idea is that instead of forcing V to be positive only with small probability α, it may be more informative to allow occasional false rejections as long as the proportion $V \,/\, r$ remains small. Approaches along these lines were proposed by Benjamini and Hochberg (1995); see also Storey (2011).

5.8 Universal Inference

The tests and confidence intervals discussed above rely on asymptotic (or exact) distributions. These are not always available, for instance, one may not be able to derive an asymptotic distribution in a given setting. For such situations, Wasserman et al. (2020) proposed a general framework, labeled universal inference. The procedure is designed for large samples and is based on splitting the data into two parts, which, for ease of notation, we take to be of equal size. Let $2n$ denote the total sample size so that the data are $Y_1, \ldots, Y_n, Y_{n+1}, \ldots, Y_{n+n}$. We define the first half as $D_{(0)}$ and the second half as $D_{(1)}$. Let

$$L_{(0)}(\theta) = \prod_{i=1}^{n} \log f(y_i; \theta)$$

denote the corresponding likelihood-based quantity computed from data $D_{(0)}$. Accordingly, we define

$$L_{(1)}(\theta) = \prod_{i=n+1}^{2n} \log f(y_i; \theta)$$

for the second part. The latter is maximized, leading to $\hat{\theta}_{(1)}$ as the maximum likelihood estimator. With this notation, we can define the universal confidence interval.

Definition 5.7 *The universal confidence interval is defined as*

$$CI_{universal} = \left\{ \theta : \frac{L_{(0)}(\hat{\theta}_{(1)})}{L_{(0)}(\theta)} \leq \frac{1}{\alpha} \right\}. \tag{5.19}$$

It can be shown that this confidence interval has a coverage of at least $1 - \alpha$, that is,

$$P(\theta \in CI_{universal}) \geq 1 - \alpha. \tag{5.20}$$

Bear in mind that the quantity $L_{(0)}(\cdot)$ is computed from observations 1 to *n*, while the estimate $\hat{\theta}_{(1)}$ is derived from observations $n+1$ to $2n$. Hence, $\hat{\theta}_{(1)}$ is independent of $Y_1, \ldots, Y_n$. This independence is a key property used in the proof. We sketch the proof at the end of this section.

Using the link between confidence intervals and testing, we can extend the idea to universal testing. We will not provide details here, but we emphasize that the approach extends naturally to more complex confidence sets and more complex testing problems. Assume, for instance, that one has fitted a model $\widehat{f}_{(1)}(y)$ (parametric or nonparametric) using the data $D_{(1)}$. If we want to test whether an alternative model $\widetilde{f}(y)$ is also compatible with the data, we can check whether $\widetilde{f}(y)$ lies in the universal confidence set. To be specific, let $\mathscr{F}$ be a set of densities. The universal confidence interval is then

$$CI_{\text{universal}} = \left\{ f \in \mathscr{F} : \frac{\prod_{i=1}^n \widehat{f}_{(1)}(y_i)}{\prod_{i=1}^n f(y_i)} \leq \frac{1}{\alpha} \right\}. \tag{5.21}$$

Putting it differently, a test of the hypothesis $H_0 : f(\cdot) = \widetilde{f}(\cdot)$ can be carried out by checking

$$\frac{\prod_{i=1}^n \widehat{f}_{(1)}(y_i)}{\prod_{i=1}^n \widetilde{f}(y_i)} \leq \frac{1}{\alpha} \Leftrightarrow \text{“H}_0\text{”}.$$

Let us illustrate universal inference with a simple example. Let $Y_i \sim \mathrm{N}(\mu, 1)$, and we aim to construct a confidence interval for μ. As before, we split the data $Y_1, \ldots, Y_n, Y_{n+1}, \ldots, Y_{2n}$ and use $Y_{n+1}, \ldots, Y_{2n}$ to obtain the mean estimate

$$\widehat{\mu}_{(1)} = \frac{1}{n} \sum_{i=n+1}^{2n} Y_i =: \overline{y}_{(1)}.$$

The likelihood-ratio statistic in (5.19) results to

$$T_n(\mu) := \frac{\exp\left(-\frac{1}{2} \sum_{i=1}^n (y_i - \overline{y}_{(1)})^2 \right)}{\exp\left(-\frac{1}{2} \sum_{i=1}^n (y_i - \mu)^2 \right)}.$$

The universal confidence interval is then defined as

$$CI_{\text{universal}} = \left\{ \mu : \; T_n(\mu) \leq \frac{1}{\alpha} \right\}.$$

Simple calculation gives

$$\begin{aligned} T_n(\mu) \leq \tfrac{1}{\alpha} \quad &\Leftrightarrow - \sum_{i=1}^n (y_i - \overline{y}_{(1)})^2 + \sum_{i=1}^n (y_i - \mu)^2 \leq 2 \log \left(\tfrac{1}{\alpha} \right) \\ &\Leftrightarrow - \sum_{i=1}^n (y_i - \mu + \mu - \overline{y}_{(1)})^2 + \sum_{i=1}^n (y_i - \mu)^2 \leq 2 \log \left(\tfrac{1}{\alpha} \right) \\ &\Leftrightarrow 2n(\overline{y}_{(0)} - \mu)(\mu - \overline{y}_{(1)}) + n(\mu - \overline{y}_{(1)})^2 \geq 2 \log (\alpha) \end{aligned}$$

with $\overline{y}_{(0)} = \dfrac{1}{n} \sum_{i=1}^n y_i$. The universal confidence interval takes the form

$$CI_{\text{universal}} = \left\{ \mu : \; 2n(\overline{y}_{(0)} - \mu)(\overline{y}_{(1)} - \mu) + n(\overline{y}_{(1)} - \mu)^2 \geq 2 \log (\alpha) \right\}.$$

Let us compare this with the exact confidence interval, which in this case can be calculated analytically. Note that

$$\sqrt{n}(\overline{Y}_{(0)} - \mu) \sim \mathrm{N}(0, 1) \text{ and } \sqrt{n}(\overline{Y}_{(1)} - \mu) \sim \mathrm{N}(0, 1).$$

Hence, the confidence interval for $\overline{y}_{(1)}$ results to

$$CI = \left\{ \mu : \; n(\overline{y}_{(1)} - \mu)^2 \leq \mathscr{X}^2_{1,1-\alpha} \right\}.$$

With the exact distribution, we can calculate the coverage of the universal interval. Let Z_0 and Z_1 be independent $\mathrm{N}(0,1)$ variables. Then, the coverage probability of the universal interval equals

$$P(\mu \in CI_{\text{universal}}) = P(2Z_0Z_1 + Z_1^2 \geq 2 \log(\alpha)).$$

We compare the coverage probabilities of the exact interval with its universal counterpart in Fig. 5.10. We see that the coverage is substantially larger than the nominal coverage. Hence, the universal approach is not recommended when an exact distribution is available. The general framework, however, allows universal inference even in settings without known limiting distributions or in more complex testing constellations, such as the general confidence set over distributions in (5.21). The price to pay is increased coverage.

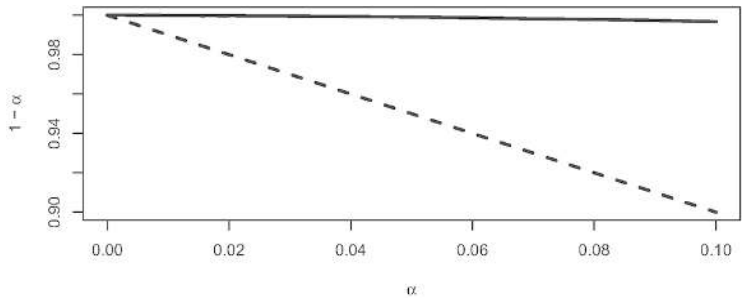


Fig. 5.10 Coverage probability of the universal confidence interval (solid line) compared to the nominal coverage (dashed line)

We can also visualize this loss of information differently. Without data splitting, we can consider $T_n(\mu)$ as a likelihood ratio by replacing $\widehat{\mu}_{(1)}$ by $\widehat{\mu}_{(0)} = \overline{y}_{(0)}$. In this case, the threshold for a confidence interval would follow from the asymptotics of the log-likelihood ratio, and we would obtain

$$CI = \left\{ \mu : \; \frac{\exp\left(-\frac{1}{2}\sum_{i=1}^n (y_i - \overline{y}_{(0)})^2\right)}{\exp\left(-\frac{1}{2}\sum_{i=1}^n (y_i - \mu)^2\right)} \leq \exp\left(\frac{1}{2}\mathscr{X}^2_{1,1-\alpha}\right) \right\}.$$

In Fig. 5.11, we compare the two thresholds: one based on the asymptotic Chi-squared approach and one based on the universal approach. Specifically, we plot the log threshold. The asymptotic threshold (dotted line) is clearly smaller than the universal threshold. The shapes of the two curves are, however, comparable.

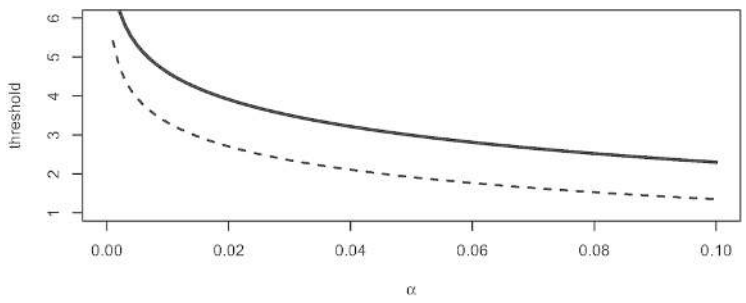


Fig. 5.11 Comparison of threshold based on the correct Chi-squared approximation (dashed line) and the universal threshold (solid line)

Proof To prove the validity of (5.20), we first define with

$$T_n(\theta) = \frac{L_{(0)}(\hat{\theta}_{(1)})}{L_{(0)}(\theta)}$$

the split likelihood-ratio statistics. We can then write

$$P_\theta(\theta \notin CI_{\text{universal}}) \;\; = P_\theta\Big(T_n(\theta) > \tfrac{1}{\alpha}\Big) \leq \alpha E_\theta\big(T_n(\theta)\big),$$

where the inequality results from the Markov inequality (see Sect. A.7). Calculating the expectation is easy since

$$\begin{aligned} E_\theta\big(T_n(\theta)\big) \;\; &= E\Big(\tfrac{L_{(0)}(\hat{\theta}_{(1)})}{L_{(0)}(\theta)}\Big) \\ &= \int \tfrac{\prod_{i=1}^n f(y_i;\hat{\theta}_{(1)})}{\prod_{i=1}^n f(y_i;\theta)} \prod_{i=1}^n f\big(y_i;\theta\big)\mathrm{d}\; y_i \\ &= \int \prod_{i=1}^n f\big(y_i;\hat{\theta}_{(1)}\big)\mathrm{d}\; y_i = 1, \end{aligned}$$

which concludes the proof. □

5.9 Sensitivity, Specificity, Accuracy

Before concluding this chapter, we relate the results to terminology commonly used in classification. This does not introduce new material but presents the same concepts using different wording. To do so, we return to Table 5.3. In classification terminology, a decision "H_1" is called "positive," while a decision "H_0" is called "negative," for instance, consider a medical test: deciding "H_1" corresponds to a positive test result (disease detected), while deciding "H_0" corresponds to a negative test result. With this binary language, we obtain:

- S = true positives (TP)
- U = true negatives (TN)
- V = false positive (FP)
- T = false negative (FN)

With these labels, one defines several measures that characterize test performance:

First, **sensitivity** denotes the proportion of correctly identified positives or, phrased differently, the proportion of correct positive decisions. Common synonyms are *recall* and *true positive rate (TPR)*. Hence,

$$\text{Sensitivity} = \text{Recall} = \text{TPR} = \frac{TP}{P} = \frac{S}{M_1}$$

where $P = FN + TP$ denotes the number of positives (denoted as M_1 in Table 5.3). Roughly speaking, sensitivity corresponds to the power of a test, that is, $P(\text{“H}_1\text{”} \mid H_1)$.

The **specificity** of a test is the proportion of correctly classified negatives. Here, too, common synonyms are *selectivity* and *true negative rate (TNR)*, so that

$$\text{Specificity} = \text{Selectivity} = \text{TNR} = \frac{TN}{N} = \frac{U}{M_0},$$

where $N = TN + FP$ denotes the number of negatives, which is M_0 in Table 5.3. Note that specificity can roughly be understood as $P(\text{“H}_0\text{”} \mid H_0)$.

The **prevalence** is the proportion of positives, that is,

$$\text{Prevalence} = \frac{P}{P+N} = \frac{M_1}{m}.$$

We already mentioned the **false discovery rate (FDR)**, which we repeat here for completeness. It is defined as

$$\text{FDR} = \frac{FP}{FP+TP} = \frac{V}{r},$$

which in turn defines **precision**, also called the *positive predictive value (PPV)*, via

$$\text{Precision} = \text{PPV} = \frac{TP}{FP+TP} = \frac{S}{r} = 1 - FDR.$$

This is the proportion of correctly identified positives among all positive decisions.

In classification, one often combines these measures into a single summary measure. The first is called **accuracy** and is defined as

$$\text{Accuracy} = \frac{TP+TN}{P+N} = \frac{U+S}{m}.$$

This gives the proportion of correct decisions among all tests.

A second commonly used measure is the **F-score**. The idea is to combine sensitivity with precision. Specifically, the F-score is defined as

$$\text{F-score} = 2\frac{\text{sensitivity} \times \text{precision}}{\text{sensitivity} + \text{precision}} = \frac{2TP}{2TP+FP+FN} = \frac{2S}{r+M_1}.$$

The interpretation of these quantities was developed throughout the chapter. The present section is meant as a guide for relating different terms across fields, both in definition and in terminology.

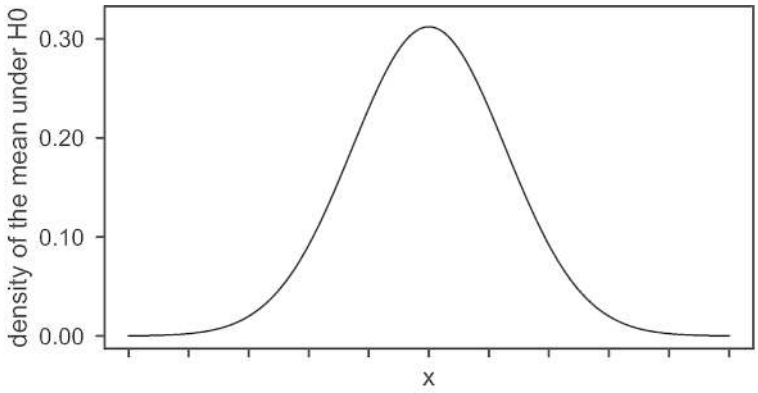
5.10 Exercises

Exercise 1 The best seller of Edumm cheese dairy is their 1000-g ball cheese. The actual weight of the cheese varies during production and can be regarded as normally distributed with known variance $\sigma^2 = 49[g^2]$. The cheesemaker suspects that, in reality, the average cheese weight is even greater than 1000 g, which would have a negative effect on production efficiency. The dairy wants to perform (based on a sample of size n) a statistical test at level $\alpha = 0.05$ on the true mean weight of their ball cheese.

(a) Formulate the hypotheses for this question. Which test would you use and why?

(b) Which test decision do you arrive at for a measured mean cheese weight of 1003 g? Perform the test for $n = 10$ and $n = 30$. Interpret the results. Also, calculate the *p*-values and rejection region of the tests.

(c) The diagram below shows the density of the expected value of the cheese weight under the null hypothesis at a sample size of $n = 30$. Fill in the values on the x-axis, and calculate and mark the rejection region of the null hypothesis. Also, mark the region of the density where the *p*-value is "located."



(d) Calculate a 90% confidence interval for μ at $n = 30$, and also mark it in the diagram from (c).

Exercise 2 We consider an i.i.d. sample $X_1, \ldots, X_n$ from a normal distribution $X_i \sim \mathrm{N}(\mu, \sigma^2)$ and want to construct a statistical test for the hypotheses $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$ at $\alpha = 0.1$.

From a sample of $n = 200$, we got $\widehat{\mu_{ML}} = \frac{1}{n} \sum_{i=1}^{n} x_i = 1.08$. We know that $\sigma = 0.5$. Our hypothesis is $H_0 : \mu = 1$ versus $H_1 : \mu \neq 1$. We will now look at three different tests to come to a test decision.

(a) State the likelihood and log-likelihood function for μ in a normal distribution. What is the idea of the likelihood-ratio test? Derive the likelihood ratio for $\mu = \mu_0$ and a general decision rule for the test problem given in this exercise, where we want to test $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$ with known variance. Then, calculate the test statistic for the concrete values provided in this exercise, and come to a test decision.

(b) Derive the score function $s(\mu; x_1, \ldots, x_n)$. We now want to use the score test to test the same problem. How does the score test work in general, and what is the decision rule in this case? Calculate the test statistic and come to a decision.

(c) State the idea of the Wald test, the general test statistic, and the decision rule for this exercise, and calculate the test statistic for this case (*Hint*: Calculating the Fisher information using the score function you already derived in (b) should result in $I(\mu) = \frac{n}{\sigma^2}$). What is the test decision?

What is an advantage of the score test from (b) over the Wald test?

(d) With the general decision rules you have constructed in (a)–(c), prove that the Wald, score, and likelihood-ratio test are in fact equal (will always come to the same test decision at the same level α) in case of a normally distributed random variable. ($Hint$: The $(1-\alpha)$ quantile of the Chi-squared distribution with $k = 1$ is given by $\mathcal{X}^2_{k,1-\alpha} = z^2_{1-\frac{\alpha}{2}}$.)

Exercise 3 We take a look at the `fev` data set. FEV (forced expiratory volume) is an index of pulmonary function that measures the volume of air expelled after 1 second of constant effort. The data contains FEV on 654 children aged 6–22 who were part of the Childhood Respiratory Disease Study in 1980 in Boston (Tager et al., 1979).

(a) In regression modeling, the normality assumption is often violated. Assume that we want to model `FEV` with a simple model that assumes the outcome to be normally distributed. Use the `shapiro.test` function in `R` to check for normality of the variable `FEV` and interpret the resulting *p*-value.

(b) In many disciplines, the test for normality is a common practice, while in others people have advocated against it. What might be reasons against testing for normality? What might be the effects on later inference, i.e., other tests performed based on the normality test result, keeping in mind the multiple testing problem?

(c) Generalize the general $\mathcal{X}^2$ test to test if two variables in a contingency table are independent, i.e., derive the $\mathcal{X}^2$ test on independence on your own. *Hint:* Derive the distribution of the number in each cell under the independence assumption, and then test if the data as a whole follows this distribution.

(d) Use the $\mathcal{X}^2$ test in `R` with $\alpha = 0.05$ to test if the smoker status has an impact on `FEV`.

Exercise 4 Consider the WDI data set, which contains the CO_2 emission up to 2011 for several large countries along with other information. We want to test whether the mean values of the CO_2 emissions for each continent present significant distances among each other.

(a) Why is it necessary to adjust the *p*-values in the case of multiple testing? What

would otherwise happen if we increased the number of tests?

(b) Use the duality of confidence intervals and tests to discuss the effect of the multiple testing problem on confidence intervals. Use the Bonferroni method to define adjusted confidence intervals.

(c) Test pairwise if the CO_2 emissions per continent are different. Then, do the same, but adjust using the "Bonferroni," "Holm's," and "false discovery rate" methods. *Hint:* Have a look at the documentation of `pairwise.t.test` for the different adjustment methods.

References

Baird, D. (1983). The Fisher/Pearson chi-squared controversy: A turning point for inductive inference. *The British Journal for the Philosophy of Science,34*(2), 105–118.
[MathSciNet][https://doi.org/10.1093/bjps/34.2.105]

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society,57*, 289–300.
[MathSciNet][https://doi.org/10.1111/j.2517-6161.1995.tb02031.x]

Berger, V. W., & Zhou, Y. (2014). *Kolmogorov–Smirnov test: Overview*. Wiley statsref: Statistics reference online.

Bonferroni (1936). *Teoria statistica delle classi e calcolo delle probabilità*. Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze.

Efron, B. (2010). *Large-scale inference: Empirical Bayes methods for estimation, testing, and prediction*. Cambridge University Press.
[https://doi.org/10.1017/CBO9780511761362]

Fisher, R. (1922). On the interpretation of chi-squared from contingency tables, and the calculation of p. *Journal of the Royal Statistical Society, 85*, 87–94.
[https://doi.org/10.2307/2340521]

Fisher, R. (1955). Statistical methods and scientific induction. *Journal of the Royal Statistical Society: Series B (Methodological),17*(1), 69–78.
[MathSciNet][https://doi.org/10.1111/j.2517-6161.1955.tb00180.x]

Goodman, S. (2008). A dirty dozen: Twelve p-value misconceptions. *Semin Hematol, 45*, 235–240.
[https://doi.org/10.1053/j.seminhematol.2008.04.003]

Mendel, G. (1865). Versuche über Pflanzenhybriden. *Verhandlungen des Naturforschenden Vereins Brünn* (4), 3–47.

Neyman, J., & Pearson, E. S. (1933). Ix. on the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character,231*(694–706), 289–337.
[https://doi.org/10.1098/rsta.1933.0009]

Pearson, K. (1904). On the theory of contingency and its relation to association and normal correlation. In *Mathematical contributions to the theory of evolution XIII*. Dulau and Co.

Pearson, K. F. (1900). X. 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. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science,50*(302), 157–175.
[https://doi.org/10.1080/14786440009463897]

Storey, J. D. (2011). False discovery rate. *International Encyclopedia of Statistical Science,1*, 504–508. [https://doi.org/10.1007/978-3-642-04898-2_248]

Tager, I. B., Weiss, S. T., Rosner, B., & Speizer, F. E. (1979). Effect of parental cigarette smoking on the pulmonary function of children. *American Journal of Epidemiology,110*(1), 15–26. [https://doi.org/10.1093/oxfordjournals.aje.a112783]

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transaction of American Mathematical Society,54*, 426–482. [MathSciNet][https://doi.org/10.1090/S0002-9947-1943-0012401-3]

Wasserman, L., Ramdas, A., & Balakrishnan, S. (2020). Universal inference. *Proceedings of the National Academy of Sciences,117*(29), 16880–16890. [MathSciNet][https://doi.org/10.1073/pnas.1922664117]

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin,1*(6), 80–83. [https://doi.org/10.2307/3001968]

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
G. Kauermann et al., *Statistics—Coping with Uncertainty*, Textbooks in Statistical Science
https://doi.org/10.1007/978-3-032-33658-3_6

6. Regression

Göran Kauermann[1] ✉, Giacomo De Nicola[2], Cornelius Fritz[3] and David Rügamer[4]
(1) Department of Statistics, LMU Munich, Munich, Germany
(2) Harvard T.H. Chan School of Public Health, Harvard University, Boston, MA, USA
(3) School for Computer Science and Statistics, Trinity College Dublin, Dublin, Ireland
(4) Munich Uncertainty Quantification AI Lab, LMU Munich, Munich, Germany

The previous chapters introduced the general concepts of statistical inference, focusing on the case where we observe data $y_1, \ldots, y_n$, but in a simplified setting without any additional input variables. Now, we will shift our focus and assume that we observe an outcome variable Y (which may be multivariate) that depends on input variables x, which may be one- or higher-dimensional. In this context, Y is commonly referred to as the "outcome," "target," or "dependent variable," since its distribution depends on x, which are called the "inputs," "features," "independent variables," or "covariates." We observe pairs of values (y_i, x_i) and aim to model the distribution of Y given X. This type of analysis is generally referred to as *regression* in statistics. In this chapter, we introduce the main ideas, acknowledging that the model class has been extensively explored and that it goes far beyond what is discussed here. We limit ourselves to what we deem to be essential and refer to Fahrmeir et al. (2013) for a more in-depth overview of the field.

6.1 Simple Linear Model

The term regression originates from the pioneering work of Sir Francis Galton. While investigating questions about heredity, Galton collected data on the heights of parents and their adult children and noticed a strong relationship between the two. In particular, an approximate linear trend was found, with a slope of 2/3, meaning that children of very tall parents would tend to be taller than the average person while also being shorter than their parents on average. Galton described this tendency toward the population average as *regression* toward the mean and

termed the linear trend line “regression line.” Nowadays, the term regression analysis is widely used to describe statistical methods for estimating relationships between a dependent variable and one or more independent variables. The simplest type of regression model is a linear model, which we introduce now.

The simple linear model aims at quantifying how the mean of the response variable Y depends on a single (independent) variable x. As implied by the name, the relationship is assumed to be linear and takes the following form:

$$Y = \beta_0 + \beta_1 x + \epsilon, \tag{6.1}$$

where β_0 is the intercept and β_1 is the slope of the regression line. The term ϵ is called the error term or **residual** and consists of a random component with mean 0. The expected value of Y is assumed to be a linear function of x, that is,

$$E(Y \mid x) = \beta_0 + \beta_1 x.$$

In other words, we assume (a) a linear relationship between x and Y and (b) that the relationship is only disturbed by the mean-zero residual ϵ. These are the two fundamental assumptions of a linear regression model.

An additional common assumption is that the residuals are independent and normally distributed variables

$$\epsilon_i \sim \mathrm{N}(0, \sigma^2), \quad i.i.d. \text{ for } i = 1, \ldots, n. \tag{6.2}$$

Here, σ^2 is sometimes called the residual variance or error variance, respectively. Then, we can show that conditional on x,

$$Y_i \mid x_i \sim \mathrm{N}(\mu_i, \sigma^2), \tag{6.3}$$

with

$$\mu_i = E(Y_i \mid x_i) = \beta_0 + \beta_1 x_i \text{ and } \mathrm{Var}(Y_i \mid x_i) = \sigma^2$$

The property of constant variance σ^2 is termed *homoscedasticity*, contrasting *heteroscedasticity*, where the variance depends on x as well. We do not discuss the latter here.

Having defined the theoretical aspects of the regression model, the natural question arises: How should we estimate its unknown parameters? In line with the methods introduced in Chaps. 3 and 4, we can solve this problem by maximizing the likelihood function. Under normality and independence of the error terms ϵ_i, the log-likelihood function is given by

$$l(\beta_0, \beta_1, \sigma^2) = \sum\nolimits_{i=1}^{n} \left\{ -\tfrac{1}{2} \log \sigma^2 - \tfrac{1}{2} \tfrac{(y_i - \beta_0 - \beta_1 x_i)^2}{\sigma^2} \right\}. \tag{6.4}$$

Differentiation with respect to the three parameters yields the following maximum likelihood estimates:

$$\widehat{\beta}_0 = \tfrac{1}{n} \sum\nolimits_{i=1}^{n} y_i - \tfrac{1}{n} \sum\nolimits_{i=1}^{n} x_i \widehat{\beta_1} \text{ where } \widehat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \tag{6.5}$$

and

$$\widehat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \widehat{\beta_0} - x_i \widehat{\beta_1}\right)^2.$$

The estimates $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are also known as "least squares estimates," as they can also be obtained through the method of *least squares*. The method, predating maximum likelihood and tracing back to Gauss (1777–1855), consists of finding the parameters that minimize the sum of the squared deviations between predicted and observed values, $\sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_i\right)^2$. The equivalence between the two is apparent from Eq. (6.4), which contains this exact quantity. Graphically, minimizing the least squares criterion can be comprehended as drawing the regression line as close to the observed points as possible, as done, for example, in Fig. 6.1.

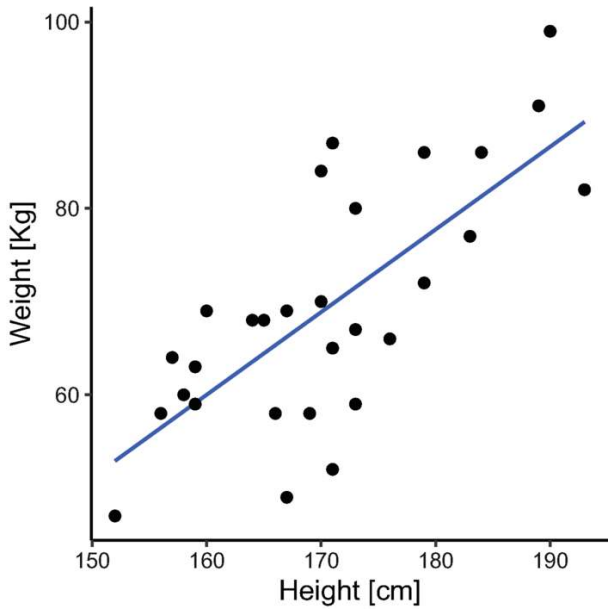


Fig. 6.1 Scatterplot of weight and height data, with regression line

Calculation of the estimates can also be carried out in matrix notation. To do so, let

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \in \mathbb{R}^{nx1} \quad , \quad \boldsymbol{X} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \in \mathbb{R}^{nx2} \quad , \quad \epsilon = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix} \in \mathbb{R}^{nx1}. \tag{6.6}$$

It is important to note that while $\boldsymbol{X}$ is capitalized, it is still assumed to be deterministic. We can then write the regression model (6.1) in the form

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \epsilon \tag{6.7}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1)^\top$ and $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$. The matrix $\boldsymbol{X}$ is often referred to as the design matrix, and by default, it contains a column of 1s to include the intercept (which would otherwise be implicitly assumed to be 0). The matrix equivalent of (6.3) is then

$$\boldsymbol{Y} \mid \boldsymbol{X} \sim \mathrm{N}(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \, \boldsymbol{I}_n), \tag{6.8}$$

where I_n is a diagonal matrix of dimension n. The log likelihood is then

$$l(\boldsymbol{\beta}, \sigma^2) = -\tfrac{n}{2} \, \log \sigma^2 - \tfrac{1}{2\sigma^2} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}). \tag{6.9}$$

Differentiation with respect to $\boldsymbol{\beta}$ and setting the derivative to $\mathbf{0} = (0, 0)^\top$ yield the maximum likelihood estimate

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}, \tag{6.10}$$

which is identical to the estimates in (6.5) but has a more convenient form. One can easily show that $\widehat{\boldsymbol{\beta}}$ is unbiased by combining (6.10) with (6.8):

$$E(\widehat{\boldsymbol{\beta}}) = E((\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y}) = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top E(\boldsymbol{Y} \mid \boldsymbol{X}) = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{\beta}$$

Moreover, as $\mathrm{Var}(\epsilon) = \sigma^2 I_n$, it follows that

$$\mathrm{Var}(\widehat{\boldsymbol{\beta}}) = \sigma^2 (\boldsymbol{X}^\top \boldsymbol{X})^{-1}.$$

It is also easy to show that $\sigma^2 (\boldsymbol{X}^\top \boldsymbol{X})^{-1}$ is equal to the inverse Fisher information matrix, implying that the estimator variance is minimal. Furthermore, $\widehat{\boldsymbol{\beta}}$ is the best linear unbiased estimator (BLUE) for $\boldsymbol{\beta}$, which is also known as the Gauss-Markov property. For more details, we refer to Ravishanker et al. (2021).

Example 6.1 In a class of 30 statistics students, we have information on their height and weight and are interested in learning about the relationship between the two quantities. We consider weight as the response (or dependent) variable and height as the single covariate (or independent variable) to estimate a simple linear regression model of the following form:

$$\mathrm{weight}_i = \beta_0 + \beta_1 \mathrm{height}_i + \epsilon_i.$$

We obtain an estimated slope of $\beta_{\mathrm{height}} = 0.89$, indicating that taller students tend to weigh more. Within our sample, an increase of 1 cm in height is therefore associated with an increase of 0.89 kg in body mass on average. A graphical representation of the data and the corresponding regression line are given in Fig. 6.1. ▷

6.2 Multiple Linear Regression

Until now, we considered the case in which a response variable Y is assumed to (linearly) depend on a single covariate x. In the real world, however, things are often much more complex: Variation in Y is often explained by a multitude of different factors, which can also induce nonlinear dependencies that interact with each other. As a natural next step, we extend single-variable regression covered in Sect. 6.1 to multiple covariates, giving rise to the *multiple linear regression model*,

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon, \tag{6.11}$$

where $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ is a vector of p different covariates assumed to influence the response variable Y. The variables $\beta_1, \beta_2, \ldots, \beta_p$ are regression coefficients weighting each covariate and ϵ_i is the random error term or residual, respectively. As in (6.2), all individual errors ϵ_i are assumed to follow a normal distribution with mean 0 and variance σ^2. This implies that the conditional expectation of Y_i is given by the *structural component* of the model

$$E(Y_i \mid x_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}.$$

We can rewrite the model again in matrix form. Defining the design matrix $\boldsymbol{X}$ by

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ \vdots & \vdots & \vdots & \vdots & \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np} \end{bmatrix} \in \mathbb{R}^{n \times (p+1)}$$

allows to write the regression model as (6.7), with $\boldsymbol{y}$ and ϵ as defined in (6.6) and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^\top$. Accordingly, the maximum likelihood estimate is given by (6.10).

Example 6.2 Considering again the 30 statistics students from Example 6.1, we now have additional information on their gender and age in addition to their height and weight. We are interested in learning how height, gender, and age relate to the students' weight. In the multiple linear regression model, weight will be the response variable, and height, age, and gender are covariates,

$$\text{weight}_i = \beta_0 + \beta_1 \text{height}_i + \beta_2 \text{age}_i + \beta_3 \text{gender}_i + \epsilon_i,$$

where the variable "gender" takes value 1 if the student identifies as male and 0 otherwise. Estimates of β_0, β_1, β_2, and β_3 were obtained via maximum likelihood and are shown in the following table:

Table

In the table, the p-values are obtained by performing a Wald test (see Sect. 5.3.2 for the null hypothesis that the corresponding coefficient equals 0). A high p-value thus means that there is not enough evidence to reject the null hypothesis, implying that we can't statistically rule out that the true coefficient is zero.

Plugging the estimated coefficients into the model equation yields

$$\widehat{\text{weight}} = -7.05 + 0.42 \text{height}_i - 0.06 \text{age}_i + 13.69 \text{gender}_i.$$

Each single coefficients can be interpreted similarly as in the simple regression model, with the additional caveat that the interpretation is valid in a setting where all other coefficients are fixed. This setup is often called *ceteris paribus*, Latin for "all else equal." The coefficient β_k thus describes the "pure" association of the k-th

covariate with the response, "washing out" the effect of the other variables. In our specific case, interpretation can be carried out as follows:

- $\widehat{\beta}_0$: The intercept can usually be interpreted as the expected value of the response when all covariates (i.e., height, age, and gender) take a value of 0. In this case, it does not have a meaningful interpretation, as a height of 0 cm does not make sense from a substantive standpoint.
- $\widehat{\beta}_1$: An increase of 1 cm in height is associated, *ceteris paribus*, with an average increase of 0.42 kg in weight.
- $\widehat{\beta}_2$: Keeping height and gender fixed, an age increase of one year is associated with an average reduction of 0.07 kg in weight. Note that the variance of this coefficient is much higher than the absolute value of the coefficient itself, evident in a high *p*-value. This means that the data at hand does not provide us enough evidence to say that age has any effect on the weight for people comparable to those within our sample.
- $\widehat{\beta}_3$: The average weight for a male student is about 13.69 kg higher than that of a female student with the same height and age. ▷

One may have noticed in Example 6.2 that the variable gender was not continuous or even numeric but rather a binary categorical indicator, taking only values 1 and 0. This is not a problem, as the linear regression model does not make any assumption about the nature of the covariates. In fact, the independent variables can be transformed in many ways, therefore increasing the range of alternative dependence structures at our disposal far beyond simple linear associations. Possibilities include:

- Indicator variables (with any number of categories) to model the association between nonnumeric variables and Y.
- Variable transformations, such as $\log(x)$ or $\frac{1}{x}$, to model nonlinear relationships.
- Using polynomials to model more complex relationships, e.g., including variables x, x^2, and x^3, to model a potential cubic relationship between variable z and the response Y.
- Using products of variables, e.g., including variables x_1, x_2, and their product $x_3 = x_1 x_2$, together in the model equation. In this case, x_3 is named interaction term, as it quantifies how x_1 and x_2 interact with each other relative to their effect on Y.

The fact that nonlinearity can be captured may seem conflicting with the chapter's name: after all, the model class is called "multiple *linear* regression." This confusion can easily be solved by noting that the model class is always linear with

respect to the parameters but not necessarily in the covariates, i.e., the covariates can be transformed nonlinearly.

6.3 Logistic Regression (Classification)

A linear regression model is suitable for response variables that approximately follow a normal distribution. For many applications, however, we are interested in investigating non-normally distributed target variables. Some variables are not even continuous but rather categorical, i.e., taking a finite number of nonnumeric values. In the following, we focus on the case of a binary response variable, i.e., a variable which only takes two possible values. Examples of such variables include the test result for a disease (positive/negative), voting for a certain political party in an election (yes/no), or passing an exam (pass/fail). In machine learning, such problems are usually labelled as *binary classification*, while the statistical terminology refers this as binary regression.

In general, we characterize the two possible outcomes of a binary variable Y, which we label as either a *success*, i.e., $Y = 1$, or *failure*, i.e., $Y = 0$. First, we show that modeling a binary response Y by a linear model as presented in the previous Sect. 6.2, i.e., assuming

$$Y_i = E(Y_i \mid x_{i1}, x_{i2}, \ldots, x_{ip}) + \epsilon_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon_i \quad (6.12)$$

with $\epsilon_i \sim \mathrm{N}(0, \sigma^2)$, isn't feasible. Note that for binary 0 / 1 variables we have $E(Y_i) = P(Y_i = 1)$. Therefore, the linear model breaks down in this case:

- Any probability can only take values between 0 and 1, while the predicted values on the right side of (6.12) are unbounded.
- The errors ϵ_i would not be normally distributed.

We can solve these issues by modeling the conditional expectation $\pi_i := E(Y_i \mid x_{i1}, x_{i2}, \ldots, x_{ip})$ = $P(Y_i = 1 \mid x_{i1}, x_{i2}, \ldots, x_{ip})$, instead of predicting observations Y_i, and by assuming the model,

$$\pi_i = P(Y_i = 1 \mid x_{i1}, x_{i2}, \ldots, x_{ip}) = h(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}),$$

where $h()$ is a function mapping the real axis from $-\infty$ to ∞ uniquely to values in the interval $[0, 1]$. By modeling the expectation, we do not need to make any assumption on the errors, and by using the function *h* to transform the output, we ensure that the predicted probabilities will only take valid values. We term the argument of the *h* function *linear predictor* and usually indicated with the letter η_i:

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}. \quad (6.13)$$

For mathematical reasons, the function $h()$ needs to be invertible, which is guaranteed when choosing it to be monotonically increasing, such that a higher value of the linear predictor η_i implies a higher predicted probability π_i. This

makes cumulative distribution functions of continuous random variables a prime choice, as they are naturally bound to the interval $[0, 1]$ and are always monotonically increasing. A standard pick in this situation is the logistic distribution function,

$$h(\eta) = \frac{\exp(\eta)}{1+\exp(\eta)}, \tag{6.14}$$

which is visualized in Fig. 6.2. The nonlinear S-shape shows any value of η_i implies a conditional probability between 0 and 1 but also visualizes how small changes in the linear predictor have varying effects on the conditional probability π_i depending on from where we look at the change, i.e., the change gets larger the closer η_i is to 0.

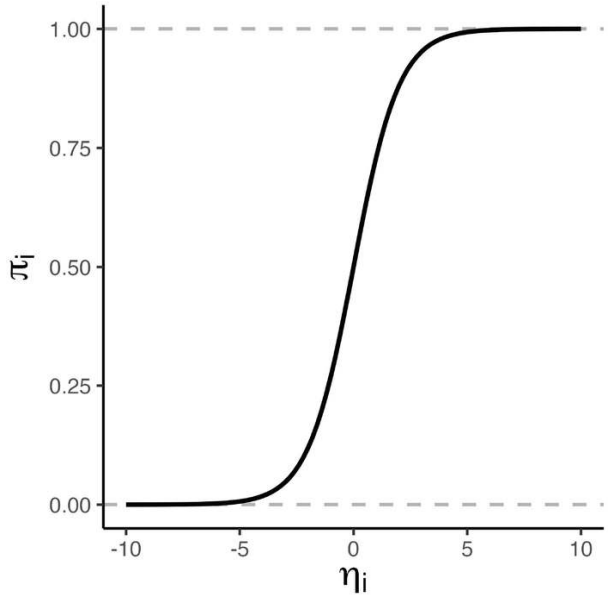


Fig. 6.2 Logistic distribution function, showing the probability $P(Y_i = 1)$, π_i, as a function of the linear predictor η_i

Using the logistic function results in the *logistic regression model* or, more shortly, the *logit* model:

$$\pi_i = P(Y_i = 1 \mid \eta_i) = \frac{\exp(\eta_i)}{1+\exp(\eta_i)}. \tag{6.15}$$

Note that an increase of the linear predictor η to $\eta + 1$ leads to an increase of the predicted probability $P(Y = 1)$ from $h(\eta)$ to $h(\eta + 1)$, which in the logistic case translates to a change from

$$\frac{\exp(\eta)}{1+\exp(\eta)} \text{ to } \frac{\exp(\eta+1)}{1+\exp(\eta+1)}.$$

Solving (6.15) for η_i leads to an alternative representation of the model

$$\eta_i = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \log\left(\frac{P(Y_i=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}. \tag{6.16}$$

This formulation is analogous to that of a linear model, but the quantity that is a linear function of η_i is now $\log\left(\frac{\pi_i}{1-\pi_i}\right)$, also termed *log odds*. This means that we can interpret the coefficients in a similar manner but with respect to the log odds as opposed to the conditional expectation: "an increase of 1 unit in the covariate x_j leads to an increase of β_j in the log odds".

A third useful formulation of the logit model is obtained by taking the exp transformation on both sides of Eq. 6.16:

$$\frac{\pi_i}{1-\pi_i} = \frac{P(Y_i=1)}{P(Y_i=0)} = \exp(\beta_0)\exp(\beta_1 x_{i1})\exp(\beta_2 x_{i2})\cdot\ldots\cdot\exp(\beta_p x_{ip}). \quad (6.17)$$

Through this formulation, one can observe that an increase of 1 unit in covariate x_j will lead to a multiplicative increase (or decrease if the coefficient is negative) by a factor of $\exp(\beta_j)$ in the odds $\frac{\pi_i}{1-\pi_i}$.

Let us now explore the practical usage of logistic regression and its different interpretations through an example.

Example 6.3 A medical doctor gathers data on 47 patients to study the association between age, gender, and the presence of heart disease. In this case, the response variable Y is "disease," taking values

$$Y = \begin{cases} 1 & \text{if heart disease present,} \\ 0 & \text{if heart disease absent.} \end{cases}$$

We can estimate a logistic regression model of the following form:

$$\mathrm{P}(Y_i = 1) = \frac{\exp(\eta_i)}{1+\exp(\eta_i)}.$$

The linear predictor η_i is given by

$$\eta_i = \beta_0 + \beta_1 \mathrm{age}_i + \beta_2 \mathrm{gender}_i,$$

where the variable age is measured in years and the variable "gender" takes the value 1 if the patient identifies as male and 0 otherwise. Parameter estimates for β_0, β_1, and β_2, obtained with maximum likelihood, are given in Table 6.1, together with their related standard errors and *p*-values.

Table 6.1 Coefficients, standard errors, and *p*-values for the regression model in Example 6.3

Plugging the estimated coefficients into the model equation yields

$$\eta_i = \log\left(\frac{\pi_i}{1-\pi_i}\right) = -5.53 + 0.06\mathrm{age}_i + 1.45\mathrm{gender}_i.$$

The single coefficients can be interpreted similarly as in the linear regression model, keeping in mind that the linear interpretation holds for the log odds for the response variable. In our specific case, interpretation can be carried out as follows:

- $\widehat{\beta}_0$: The intercept gives the expected log odds of heart disease for a patient with all covariates set to 0, i.e., for a female patient with age 0. As the value 0 is out of the domain for our variable age, the intercept has no meaningful interpretation.
- $\widehat{\beta}_1$: For a given gender, an age increase of 1 year is associated with an average increase of 0.06 in the log odds of heart disease. Equivalently, we can say that an increase of 1 year in age leads to an increase in the odds of heart disease ($\frac{P(\text{disease}_i=1)}{P(\text{disease}_i=0)}$) by a (multiplicative) factor of $\exp(0.06) = 1.06$.
- $\widehat{\beta}_2$: The log odds of heart disease for male patients are on average 1.45 units higher than those of a female patient of the same age. Alternatively, it's also possible to say that the odds of heart disease for a male patient are higher than those of a female patient of the same age by a (multiplicative) factor of $\exp(1.45) = 4.26$.

The predicted probability of heart disease by age and gender is depicted in Fig. 6.3. From the plot, it is apparent that the risk of heart disease increases for older people and that males have a higher risk than females. ▷

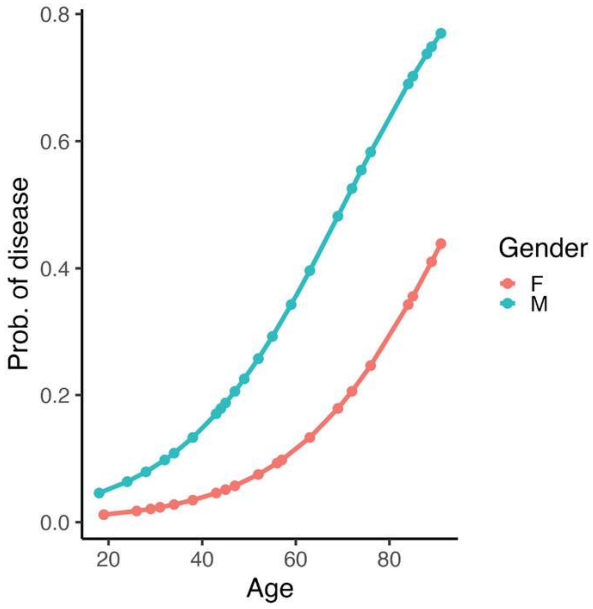


Fig. 6.3 Predicted probability of heart disease by age and gender, estimated via the logistic regression model

6.4 Generalized Linear Models

In the previous section, we showed how the linear modeling framework can be extended to handle binary response variables by focusing on modeling the conditional expectation of the response variable, i.e., $E(Y \mid x)$, and transforming the output of the linear model in a suitable manner. We will now show that this

generalization is possible not only in the binary case but also for a wide variety of different response variable types, such as Poisson, exponential, Gamma, and many more distributions. Such extensions to the linear model make up the class of generalized linear models (GLMs), introduced by Nelder and Wedderburn (1972). More precisely, GLMs possess the following properties:

1. The (conditional) distribution of the response variable Y belongs to the broad class of exponential family distributions. This class contains most of the commonly used distributions (see Sect. B.2 for more details).
2. The expected value of the response value, $E(Y_i \mid x_i)$, is connected to the linear predictor $\eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}$ through a response function $h()$. In other words, we have

$$E(Y_i \mid x_i) = \mu_i = h(\eta_i).$$

The GLM framework is apparently quite general and encompasses many different regression scenarios. A very important role is played by the response function $h(\eta)$, which should be chosen to be compatible with the type of response variable Y that is modeled. In the previous section, we saw the case of a binary variable, in which the response function has the main job of transforming the linear predictor, which is generally unbounded and real, into a probability, which only takes real values between 0 and 1. Another special case is given by a normally distributed response variable: In that scenario, given that a normally distributed random variable can take all values in the real numbers, there is no need to transform the linear predictor η. In other words, the response function h is given by the identity function, i.e., $h(\eta) = \eta$, reverting to the case of the classical (standard) linear model. As said before, many other GLMs are possible: Let us now see an empirical example with a Poisson-distributed response variable.

Example 6.4 An e-commerce website is interested in understanding how review scores, price, and promotions are related to product sales. More in particular, the research team considers total sales over a week for a sample of 200 different products. The following is a snapshot of the data:

Table

Given that we are in a classical count variable setting, it seems reasonable to model sales as the response variable in a Poisson GLM, i.e., a regression model assuming the response variable to follow a Poisson distribution. The latter

assumption is consistent with what is visible from a barplot of sales, which is given in Fig. 6.4.

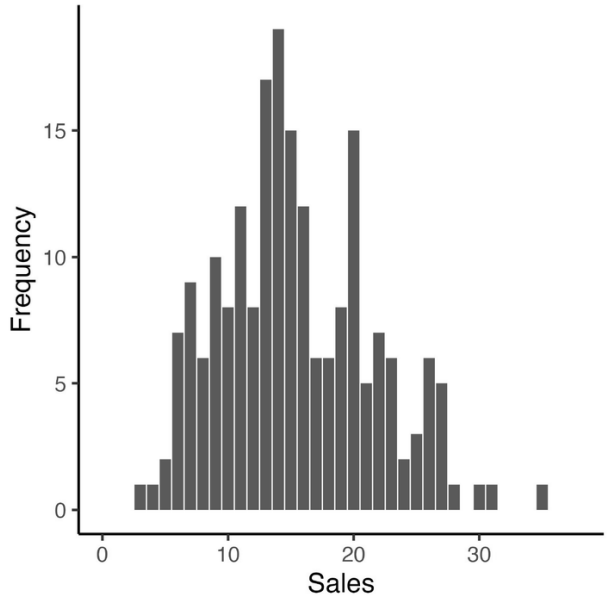


Fig. 6.4 Barplot of "sales," the response variable in our regression model. The shape of the data is consistent with the choice of a Poisson distribution

Once we decided to use a Poisson distributional assumption, all that's left is to choose the response function *h*, which will map the linear predictor η (i.e., a linear combination of the covariate scores, price, and promotion) to the expectation of *Y* (i.e., sales). Given that sales can only take positive values, a reasonable choice is given by the exponential function, which maps real numbers into positive ones. All things considered, our model will thus be

$$\mathrm{P}(Y_i = y \mid x_i) = \exp(\eta_i),$$

with

$$\eta_i = \beta_0 + \beta_1 \mathrm{review_score}_i + \beta_2 \mathrm{price}_i + \beta_3 \mathrm{promotion}_i$$

where the variable "review score" is measured in stars and the variable "price" in Euros and the variable "promotion" takes the value 1 if the price shows a discount on the website and 0 otherwise. Parameter estimates for β_0, β_1, β_2, and β_3, obtained with maximum likelihood, are given in Table 6.2, together with their related standard errors and *p*-values.

Table 6.2 Coefficients, standard errors, and *p*-values for the regression model in Example 6.4

Plugging the estimated coefficients into the model equation yields

$$\eta_i = 0.642 + 0.473 * \mathrm{review_score}_i + 0.001 * \mathrm{price}_i + 0.166 * \mathrm{promotion}_i.$$

The single coefficients can be interpreted similarly as in the binary regression model above, keeping in mind that the linear interpretation holds for the logarithm of the response variable. In our specific case, interpretation can be carried out as follows:

- $\widehat{\beta}_0$: The intercept gives the log expected weekly sales for a product with all covariates set to 0. Once again, as the value 0 is out of the domain for some of our covariates, the intercept has no meaningful interpretation here.
- $\widehat{\beta}_1$: For a given price and promotion, an increase of one star in review is associated with an increase of 0.473 in the logarithm of weekly sales, *ceteris paribus*. Equivalently, we can say that an increase of one star in review is associated with a multiplicative increase in sales by a factor of $e^{0.473} = 1.605$ (*ceteris paribus*), that is, a whopping 60% increase in sales.
- $\widehat{\beta}_2$: The coefficient for price is very small, and its *p*-value is close to 1. From the available data and controlling for review scores and whether or not there is a promotion, there is no evidence of a relationship between item price and sales.
- $\widehat{\beta}_3$: For given review score and price, the presence of a promotion is associated with an increase of 0.166 in the logarithm of weekly sales. Alternatively, we can also say that if a product is on sale we can expect an increase in sales of a factor of $\exp(0.177) = 1.194$, i.e., an increase of 19% in weekly sales, *ceterisparibus*.

▷

As is visible from these simple examples, generalized linear models are powerful tools to measure associations between variables under fairly general conditions. The wide range of distributional families available, together with the choice of the response function, provides the user with remarkable flexibility. This flexibility, however, comes at the cost of putting the choice in the hands of the modeler. Nonetheless, there are several tools to ensure proper model specification, i.e., proper choice of distributional and structural assumptions. For reasons of brevity, we don't discuss these here but instead refer to Fahrmeir et al. (2013) for that as well as for a broader discussion of everything associated with regression.

6.5 Exercises

Exercise 1 The dataset theatre.txt, available in the book online materials, contains data on yearly expenses of theater visits for randomly surveyed citizens of Zurich, Switzerland. The data contains the following variables:

`Age`: Age of the informants in years

`Sex`: Sex of the informants (0, male/1, female)

`Salary`: Yearly salary in thousand Swiss francs (CHF)

`Culture`: Expenses for cultural activities in CHF

`Theater`: Expenses for theater visits in CHF

`Theater_prevYear`: Expenses for theater in the previous year in CHF

Our aim is to understand what variables are associated with theater expenses, and we will use regression models to do so:

(a) Construct a linear model M_0 explaining expenses for theater visits this year by all other covariates, using the function `lm` in `R`. Interpret the estimates obtained for the model coefficients. Which variables have the most impact on theater expenses?

(b) Eliminate all nonsignificant variables in M_0, and fit a new model M_1. Which of the model would you prefer, and why?

(c) Extend M_1 to check whether `Sex_fac` and `Theatre_prevYear` interact.

(d) How does the model fit change if you

- Use the logarithm of the covariates instead of the absolute metrical values?
- Add the square root of the salary to the absolute covariates?

(e) Based on your observations, which model would you choose for this data?

Exercise 2 The dataset `shuttle` describes, for the 23 space shuttle flights previous to the Challenger disaster in 1986, the temperature ($^\circ F$) at takeoff as well as the (non)occurrence of thermal overstress of a certain unit. The data contains the following variables:

Table

Our goal is to understand the relationship between temperature and thermal overstress.

Download the dataset `shuttle` from the online supplementary material, and load the data into `R`. Generate an additional column `tempC`, which gives the temperature in Celsius. *Note: The conversion rate is* $T_F = 1.8 \cdot T_C + 32$.

(a) Graphically compare the temperatures measured at `td` = 1 with those measured at `td` = 0. What is problematic about this way of presenting the data?

(b) Fit, using the function `lm()`, a linear model with temperature in $^\circ F$ as predictor. Generate a plot that shows the observed values of `td` and `temp`, as well as the probabilities estimated by the linear model.

(c) What parameter estimates do we obtain if the temperature is measured in $^\circ C$? What are the consequences if we add `temp` and `tempC` simultaneously to the model?

(d) Consider the `summary` of the linear model estimated at point (b) and interpret the output. Also take a look at the plot of the residuals. Is the linear term significant? Why is the test applied here problematic? What else speaks against the use of the linear model?

(e) How can the slope parameter in the logit model be interpreted?

(f) Calculate the probability of thermal overstress given the temperature of $31\ ^\circ F$ as on the date of the Challenger disaster.

(g) For which temperature does the probability of thermal overstress equal 0.5?

References

Fahrmeir, L., Kneib, T., Lang, S., & Marx, B. (2013). *Regression—models, methods and applications*. Springer. [https://doi.org/10.1007/978-3-642-34333-9]

Nelder, J. A., & Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society Series A: Statistics in Society,135*(3), 370–384. [https://doi.org/10.2307/2344614]

Ravishanker, N., Chi, Z., & Dey, D. K. (2021). *A first course in linear model theory*. Chapman and Hall/CRC. [https://doi.org/10.1201/9781315156651]

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
G. Kauermann et al., *Statistics—Coping with Uncertainty*, Textbooks in Statistical Science
https://doi.org/10.1007/978-3-032-33658-3_7

7. Model Selection

Göran Kauermann[1] ✉, Giacomo De Nicola[2], Cornelius Fritz[3] and David Rügamer[4]
(1) Department of Statistics, LMU Munich, Munich, Germany
(2) Harvard T.H. Chan School of Public Health, Harvard University, Boston, MA, USA
(3) School for Computer Science and Statistics, Trinity College Dublin, Dublin, Ireland
(4) Munich Uncertainty Quantification AI Lab, LMU Munich, Munich, Germany

The choice of the right model is apparently crucial. This was already discussed in Sect. 3.1, and we refer to Fig. 3.1 where we motivated the idea of choosing a (probability) model. We will deepen this view now because the choice of a model is not only about selecting a probability model but also about specifying how input variables x influence the output variable(s) Y. Some ideas in this direction were already visible in the previous chapter on regression, and we want to formalize this further in the subsequent chapter. We will expose numerous links to similar, if not equivalent, questions in the realm of machine learning, but illuminate the issues from a statistical perspective.

7.1 Modeling Dependence

In the previous chapters, we learned about different ways to model random variables by characterizing them via probability distributions. In Chap. 6, we introduced regression models that allow for the inclusion of covariate information. In real-life applications of these models, it is often unclear which covariates should be used in the final model or which distribution best captures the data at hand. To address such issues, we will develop tools for model selection, i.e., selecting an appropriate model for data at hand.

In many, if not most, real-world applications, one is not interested in how quantities behave on their own but rather in the relationships between some variables. For example, when studying quantities relevant to the socioeconomic aspects of a country, such as income and education, it might be interesting to understand how one of these variables influences the other. In general, we may be interested in learning how a variable of interest Y ("outcome," "target," "dependent variable") is influenced by some other variables x ("inputs," "features," "independent variables"). In most settings, the input variable is multivariate. This is exactly the setting of regression, which we covered in the previous chapter. We want to treat the problem from a much broader perspective here. We aim to find a model for Y given x or, more formally, to select a model for the conditional distribution of Y given x, i.e., $F(y|x, \theta)$, where the distribution depends on a set of parameters $\theta = (\theta_1, \theta_2, \ldots, \theta_p)$.

We need to get a little more formal now, since a model of this type depends on two components: First, we need to select a probability model $\mathscr{F}$, as motivated in Sect. 3.1. Additionally, we have to specify how the input variables x influence the parameters in the probability model. To disentangle these two perspectives, we will change the notation slightly. We define a model $\mathscr{M}$ as a pair, composed of a probability model $\mathscr{F}$ and a structural model $\mathscr{H}$, which defines how the parameters of the distribution depend on the input variables x. To be specific, a model is defined as

$$\mathscr{M} = (\mathscr{F}, \mathscr{H}) \tag{7.1}$$

where the two model components are given through

$$\mathscr{F} = \{F(.;\gamma), \gamma \in \Gamma\} \tag{7.2}$$

$$\mathscr{H} = \{h(.;\theta) \text{ with } h : \mathscr{X} \to \Gamma \text{ and } \theta \in \Theta\}. \tag{7.3}$$

We here call $\mathscr{F}$ the **probability model**, or distribution model, which defines the likelihood with respect to the parametrization of the probability model. The parameter space of the probability models is denoted here with Γ, and for each $\gamma \in \Gamma$, $F(,\gamma)$ is a fully specified distribution function. In particular, if the density exists, then we can equivalently use

$$\mathscr{F} = \{f(.;\gamma), \gamma \in \Gamma\} \tag{7.4}$$

with $f(.,\gamma)$ as density associated with the distribution function $F(.,\gamma)$. This is exactly the setup that we introduced in Sect. 3.1 and employed in the previous chapters. There is a strong connection to machine learning as the choice of a probability model is closely related to the choice of a loss function. We refer to Grünwald (1999) or Wang et al. (2022) for details.

Besides the probability model, we also have a **structural model** denoted by $\mathscr{H}$. In machine learning, this is also called the "hypothesis space." The model defines how the input variables x influence the model's parameters. The function $h(.,\theta)$ thereby depends on parameters θ. The structure of the function $h(.,\theta)$ itself is assumed to be known, but the parameters θ are unknown and need to be estimated. In particular, the value range of $h(.;\theta)$, i.e., the set of all possible values for $h(x;\theta)$, is equal or a subset of Γ for all possible values of θ and all possible values of x, that is,

$$\{h(x,\theta) \text{ for } x \in \mathscr{X} \text{ and } \theta \in \Theta\} \subseteq \Gamma \ . \tag{7.5}$$

The constraint (7.5) guarantees that for any value of x and any value of θ we obtain $F(.,h(x,\theta)) \in \mathscr{F}$, which means we obtain a proper distribution.

The notation above can be motivated by looking at the logistic regression discussed in Sect. 6.3. Here, the probability model $\mathscr{F}$ is a binomial distribution with $\pi \in [0,1]$ as the only parameter. The structural model $\mathscr{H}$ is built from the logistic function $h(.)$ given in (6.14) and the linear structure for the linear predictor η given in (6.13). Defining $x = (1, x_1, x_2, \ldots, x_q)^\top$ and $\theta = (\beta_0, \beta_1, \beta_2, \ldots, \beta_q)^\top$, we obtain

$$h(x;\theta) = \frac{\exp(x^\top\theta)}{1+\exp(x^\top\theta)}.$$

As can be seen, the structure of h is known, but the parameters θ are not.

The broad definition in (7.1)–(7.3) encompasses several different types of tools used to model dependencies, both in statistics and in machine learning. In this chapter, we approach the problem of choosing an appropriate model $\mathscr{M}$ from a general perspective. In principle, this encompasses selecting both a probability model $\mathscr{F}$ and a structural model $\mathscr{H}$. We will

introduce the Akaike information criterion (AIC), which allows the comparison of different models and, hence, the selection of a better one. We also motivate the idea of model selection by dividing the data into one part, which is used for model fitting, and another part used for model selection, commonly known as cross-validation. In most practical applications, the focus is thereby to select the "best" structural model $\mathscr{H}$, while the probability model $\mathscr{F}$ is considered fixed. Strategically, however, it would be a considerable restriction if we consider model selection only as a selection of the structural model.

Now that we have clearly distinguished between the structural model component $\mathscr{H}$ and the probability component $\mathscr{F}$, we want to simplify the notation for better readability. We therefore write the model class $\mathscr{M}$ as

$$\mathscr{M} = \{F(\cdot;\theta); \theta \in \Theta\} \tag{7.6}$$

which formally stands for

$$\mathscr{M} = \{F(\cdot; h(.;\theta)); F \in \mathscr{F} \text{ and } h(\cdot,\cdot) \in \mathscr{H}\}.$$

This notation accommodates that models comprise structural and probability components, without explicitly listing them. While (7.6) simplifies notation, one should bear in mind that any model choice requires selecting $\mathscr{F}$ as well as $\mathscr{H}$.

7.2 Training and Test Data

Let us look more closely at the problem of model selection. As before, we assume the data-generating process $G(\cdot)$ given in (2.2), but we allow now that $G(\cdot)$ depends on input quantities x, that is, we have $G(\cdot \mid x)$ with $x \in \mathscr{X}$. For simplicity, we look at two possible models which we denote as $\mathscr{M}_1 = (\mathscr{F}_1, \mathscr{H}_1)$ and $\mathscr{M}_2 = (\mathscr{F}_2, \mathscr{H}_2)$. Often, these models will be nested, meaning that one model is a subset of the other. A nesting structure is most naturally defined in the space of the structural model while using the same probability model, i.e., $\mathscr{F}_1 = \mathscr{F}_2$ and $\mathscr{H}_2 \subset \mathscr{H}_1$, or vice versa with $\mathscr{H}_1$ being nested in $\mathscr{H}_2$. Assuming $\mathscr{H}_1$ to be the larger model, this can formally be written as

$$\mathscr{H}_1 = \{h(.;\theta) \text{ with } \theta \in \Theta_1\} \text{ and } \mathscr{H}_2 = \{h(\cdot;\theta) \text{ with } \theta \in \Theta_2\} \text{ with } \Theta_2 \subset \Theta_1. \tag{7.7}$$

We call (7.7) a **nested model structure**.

Alternatively, we may consider non-nested models. If $\mathscr{F}_1 = \mathscr{F}_2$, this can be the case if

$$\mathscr{H}_1 = \{h(\cdot;\theta), \theta \in \Theta_1\} \text{ and } \mathscr{H}_2 = \{h(\cdot;\theta), \theta \in \Theta_2\} \text{ with } \Theta_2 \not\subset \Theta_1 \text{ and } \Theta_1 \not\subset \Theta_2. \tag{7.8}$$

In this case, the probability model is the same, but the structural models differ, and they are not subsets of another, i.e., $\mathscr{H}_1 \not\subset \mathscr{H}_2$ and $\mathscr{H}_2 \not\subset \mathscr{H}_1$. But we may even consider models where the probability models differ, that is, $\mathscr{F}_1 \neq \mathscr{F}_2$. In this general case, we write $F_1(\cdot;\theta_1)$ for elements of $\mathscr{F}_1$ and accordingly $F_2(\cdot;\theta_2)$ for elements of $\mathscr{F}_2$. Note that for nested models the distributional models are the same, and in this case, we omit the subscript and write $F(.;\theta_1)$ and $F(\cdot;\theta_2)$, respectively. The possible different setups for model selection, including nested and non-nested scenarios, are sketched in Fig. 7.1.

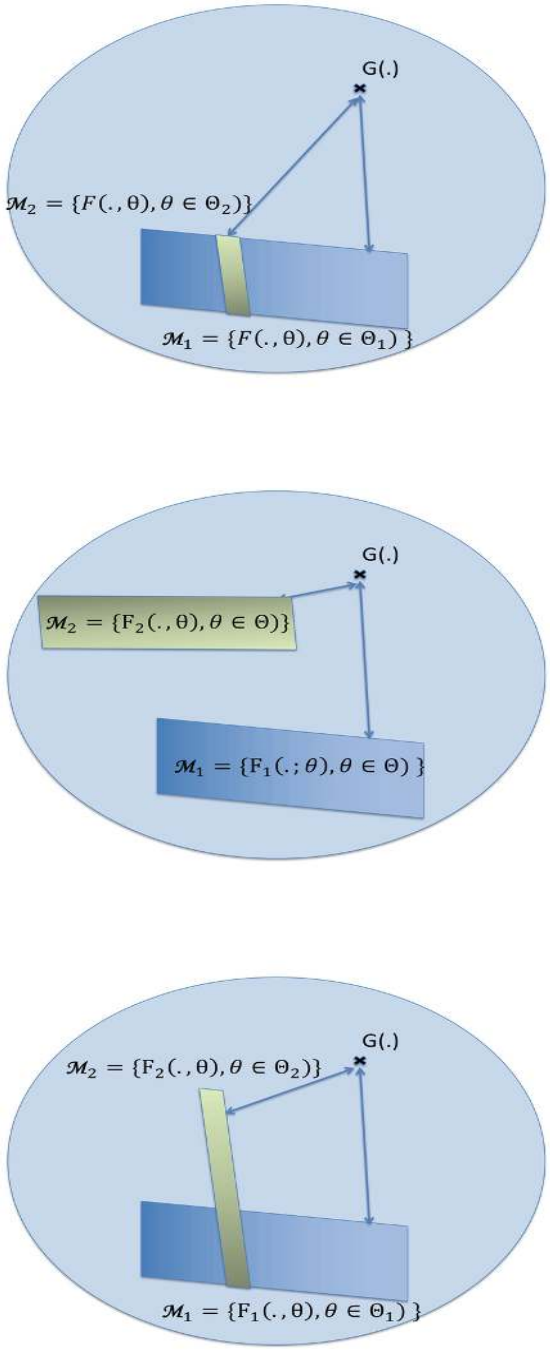


Fig. 7.1 Different settings for model selection. The top row shows nested models, the middle row shows disjoint models, and the bottom row visualizes non-nested and non-disjoint models

We refer to our results from Sect. 2.1, where we introduced the Kullback-Leibler divergence to define the best model in a set of models $\mathscr{F}$. We extend this now by allowing for the comparison of models. If the Kullback-Leibler divergence is taken as a distance measure between F and the true data-generating process G, the model with the smaller distance to G should be preferred. This leads to the comparison of

$$KL(G(\cdot); F_1(\cdot; \hat{\theta}_1)) = \int \log \frac{g(y)}{f_1(y;\hat{\theta}_1)} dG(y)$$

and

$$KL(G(\cdot); F_2(\cdot; \hat{\theta}_2)) = \int \log \frac{g(y)}{f_2(y;\hat{\theta}_2)} dG(y),$$

where $\hat{\theta}_1$ and $\hat{\theta}_2$ are the maximum likelihood estimates in model $\mathscr{M}_1$ and $\mathscr{M}_2$, respectively. We should prefer model 1 if its divergence is smaller, that is, we prefer model $\mathscr{M}_1$ if

$$\begin{aligned} & KL(G(\cdot); F_1(\cdot; \hat{\theta}_1)) - KL(G(\cdot); F_2(\cdot; \hat{\theta}_2)) < 0 \\ \Leftrightarrow \quad & \int \log \frac{f_1(y;\hat{\theta}_1)}{f_2(y;\hat{\theta}_2)} dG(y) > 0 \\ \Leftrightarrow \quad & \text{prefer model } \mathscr{M}_1. \end{aligned} \tag{7.9}$$

Note that we cannot calculate the integrals since we do not know $G(\cdot)$. One might be tempted to replace the integral with its empirical version, since data $y_1, \ldots, y_n$ have been drawn from $G(\cdot)$. This would lead to the quantity

$$\frac{1}{n} \sum_{i=1}^{n} \log \frac{f_1(y_i;\hat{\theta}_1)}{f_2(y_i;\hat{\theta}_2)}. \tag{7.10}$$

However, this quantity does not converge to the integral in (7.9) since we violate the fundamental rule, which states "*Never make use of the same data twice.*" Note that the integral in (7.9) refers to a new variable $Y \sim G$ and in fact we can write (7.9) as the selection rule:

$$\text{If } E\left(\log \frac{f_1(Y;\hat{\theta}_1)}{f_2(Y;\hat{\theta}_2)} \right) > 0, \text{ we prefer } \mathscr{M}_1 \tag{7.11}$$

Now, it becomes obvious that the expectation is with respect to Y, some new variable, but $\hat{\theta}_1$ and $\hat{\theta}_2$ are based on the data $y_1, \ldots, y_n$. In contrast, in (7.10), we make use of the data twice, first in the sum, where we sum over $y_1, \ldots, y_n$, as well as in the estimates $\hat{\theta}_1$ and $\hat{\theta}_2$, which are also calculated from the data $y_1, \ldots, y_n$. This is a conceptual flaw that, in fact, will inevitably lead to what is called **overfitting**. This is easily seen for nested models. If $\mathscr{F}_1 = \mathscr{F}_2$ and $\mathscr{H}_2 \subset \mathscr{H}_1$, we obtain

$$l(\hat{\theta}_2) = \sum_{i=1}^{n} \log f(y_i; \hat{\theta}_2) \leq \sum_{i=1}^{n} \log f(y_i; \hat{\theta}_1) = l(\hat{\theta}_1),$$

since $\hat{\theta}_2 \in \Theta_2 \subset \Theta_1$. Putting it differently, the maximum of the likelihood function in the larger space Θ_1 will always be larger than or equal to the maximum of the likelihood in the smaller space Θ_2. Inevitably, overfitting results since complex models are always preferred. Let us therefore correct the mistake occurring in (7.10). This can be done in two ways. We start by making the summands explicitly independent of the estimates. This is achieved by using $n-1$ observations for fitting and the remaining observation for estimating the integral. Going through all observations, this would lead to

$$\text{if } \frac{1}{n} \sum_{i=1}^{n} \log \frac{f_1(y_i;\hat{\theta}_{1,-i})}{f_2(y_i;\hat{\theta}_{2,-i})} > 0, \text{ we prefer } \mathscr{M}_1, \tag{7.12}$$

where $\hat{\theta}_{k,-i}$ is the maximum likelihood estimate in model *k* calculated from all observations except of the *i*-th one. Hence, we take the data $(y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n)$ to calculate $\hat{\theta}_{k,-i}$, where $k = 1$ or $k = 2$, and use the left-out observation y_i to estimate the expectation in (7.11). This approach is called *leave-one-out cross-validation*, as sketched below.

Definition 7.1 (Leave-One-Out Cross-Validation)

1. *For* $i = 1, \ldots, n$*, refit the two models by excluding the i-th observation leading to estimates* $\hat{\theta}_{1,-i}$ *and* $\hat{\theta}_{2,-i}$.
2. *Calculate the average log-likelihood ratio given in* (7.12)*. If this is larger than* 0*, prefer model* $\mathscr{M}_1$*; otherwise, select model* $\mathscr{M}_2$.

The approach requires refitting each model *n* times. This can be burdensome or even infeasible if the fitting process takes too long or is too computer-intensive. If this is the case, then it is advisable to divide the data generally into training and test data. The training data are used to fit the model, that is, to calculate the maximum likelihood estimate. The test data are used for model comparison. Formally, this leads to the following procedure.

Definition 7.2 (Training and Test Data)

1. *We divide n into* $n_{train} \in \mathbb{N}$ *and* $n_{test} \in \mathbb{N}$ *such that* $n = n_{train} + n_{test}$*. A typical choice is a ratio of approximately 4:1 for training to test data.*
2. *From the database, we draw a sample without replacement of size* n_{test} *yielding the test data. We denote with* $\mathscr{N}_{test} \subset \{1, \ldots, n\}$ *the index set of selected data points.*
3. *We train the models on* $\mathscr{N}_{train}$ *where* $\mathscr{N}_{train} = \{1, \ldots, n\} \setminus \mathscr{N}_{test}$ *and obtain estimates* $\hat{\theta}_1$ *and* $\hat{\theta}_2$.
4. *We calculate*
$$\sum_{i \in \mathscr{N}_{test}} \log \frac{f_1(y_i; \hat{\theta}_1)}{f_2(y_i; \hat{\theta}_2)}.$$
If this is > 0*, we prefer model* $\mathscr{M}_1$*; otherwise, we select model* $\mathscr{M}_2$.

It appears that such a procedure is somewhat data-hungry since we omit all observations in $\mathscr{N}_{\text{test}}$ for fitting the model. This means that if the database is small and the train-test decomposition makes it even smaller, it is advisable to work with a different approach. Figure 7.2 shows the two versions of splitting the data. The upper plot demonstrates the leave-one-out cross-validation. The middle plot visualizes the idea of splitting the data into test and training data. As shown in the bottom plot of Fig. 7.2, it is also possible to select multiple sets of observations for testing and fitting the data on the remaining observations. This is called *k*-fold cross-validation and refers to the concept that the data are divided into *k* groups of similar size. This reduces the numerical burden of refitting the model from *n* times to *k* times. Formally, one proceeds as follows.

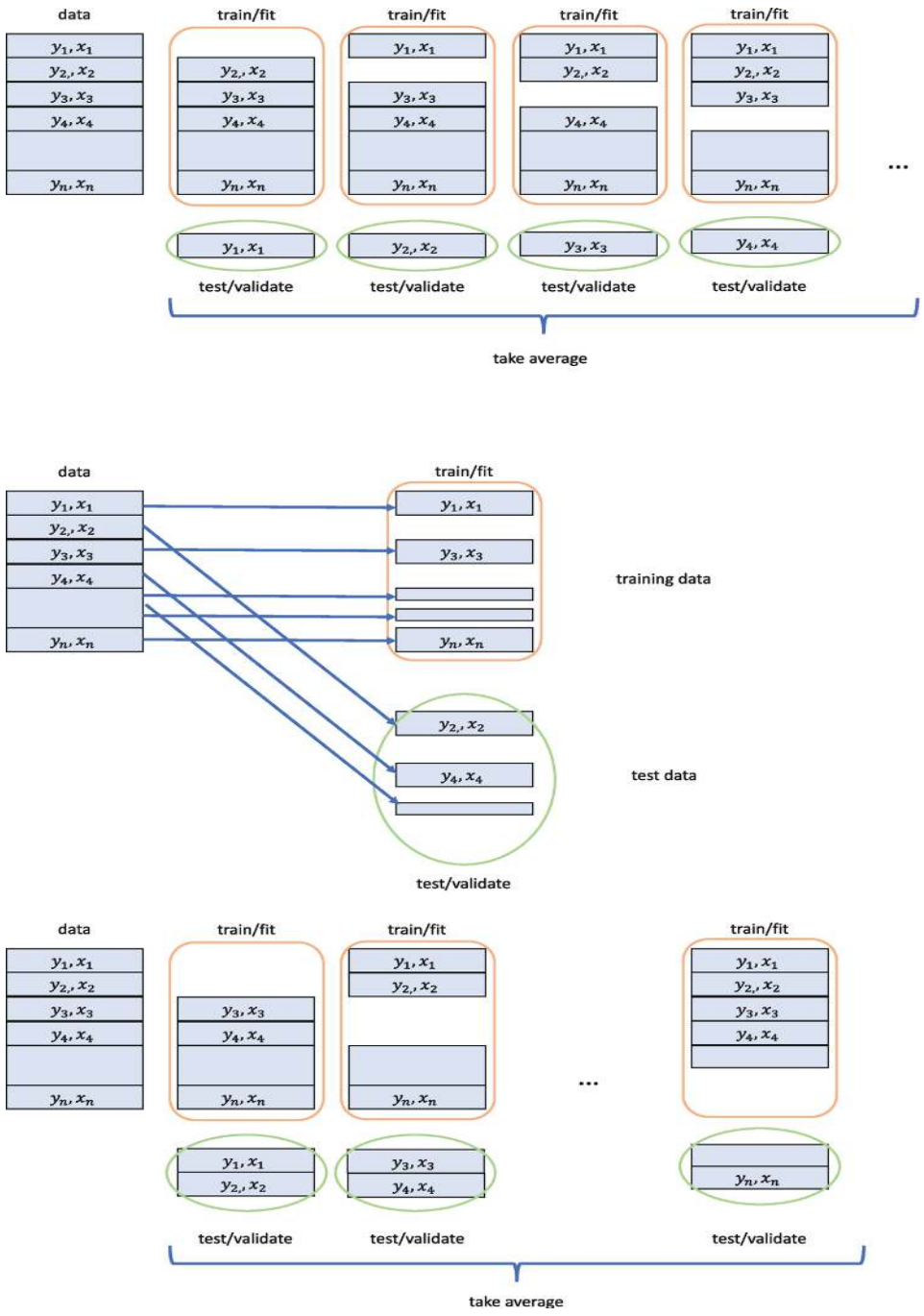


Fig. 7.2 Different versions of splitting data. The top row shows leave-one-out cross-validation. The middle row shows a single split into training and test data. The bottom plot shows *k*-fold cross-validation

Definition 7.3 (*k*-Fold Cross-Validation)

1. *We divide the data into k groups and denote the resulting index sets with* $\mathcal{N}_1, \ldots, \mathcal{N}_k$ *such that*

   (i) $\mathcal{N}_j \cap \mathcal{N}_l = \emptyset$,

   (ii) $|\mathcal{N}_j| \approx |\mathcal{N}_l|$,

(iii) $\bigcup_{l=1}^{k} \mathcal{N}_l = \{1, \ldots, n\}$.

2. *We fit the two models on* $\mathcal{N} \setminus \mathcal{N}_j$ *separately for all* $l = 1, \ldots, k$, *leading to estimates* $\hat{\theta}_{1,-\mathcal{N}_l}$ *and* $\hat{\theta}_{2,-\mathcal{N}_l}$.
3. *We calculate*

$$\sum_{l=1}^{k} \sum_{i \in \mathcal{N}_l} \log \frac{f_1(y_i; \hat{\theta}_{1,-\mathcal{N}_l})}{f_2(y_i; \hat{\theta}_{2,-\mathcal{N}_l})}.$$

If this is > 0, *we prefer model* $\mathcal{M}_1$; *otherwise, we select model* $\mathcal{M}_2$.

It is easy to see that if $|\mathcal{N}_l| = 1$ for all l, we get $k = n$, which leads to leave-one-out cross-validation.

7.3 AIC and BIC

The above procedures require refitting the model in one way or another. This can be avoided by using the Akaike information criterion (AIC) or the related Bayesian information criterion (BIC). We, therefore, return to formula (7.10) and suggest a correction which has been proposed already more than 50 years ago by Akaike (1973). Before going into technical details, we remind the reader to take a second look at Fig. 3.1. We observe that the Kullback-Leibler divergence $KL(G(\cdot), F(\cdot; \hat{\theta}))$ is larger than for the optimal parameter, that is,

$$KL(G(\cdot), F(\cdot; \hat{\theta})) > KL(G(\cdot), F(\cdot; \theta_0)).$$

In other words, the overfitting behavior discussed before can be linked to estimation, in that estimation remains imperfect due to the limited data availability. Hence, we will not uncover θ_0 and hence the minimal Kullback-Leibler divergence exactly, unless we have an infinite amount of data. Consequently, $KL(G(\cdot), F(\cdot; \hat{\theta}))$ is a biased estimator of $KL(G(\cdot), F(\cdot; \theta_0))$, and the bias is strictly positive.

The subsequently introduced Akaike information criterion aims to find a correction for this bias. The exact proof is a little bit clumsy to follow and will therefore be given at the end of this section.

Definition 7.4 *The* **Akaike information criterion** *(AIC) for model class* $\mathcal{M}$ *is defined as*

$$AIC(\mathcal{M}) = -2 \sum_{i=1}^{n} \log f(y_i; \hat{\theta}) + 2p, \tag{7.13}$$

where p is the dimension of the parameter θ *and* $\hat{\theta}$ *is the maximum likelihood estimate in model class* $\mathcal{M}$ *as defined in* (7.6). *The AIC can be seen as an estimate of the Akaike information* (7.14)

$$
\begin{aligned}
& 2\ E\{KL(G(\cdot), F(\cdot;\hat{\theta}))\} - 2\int \log\,(g(y))dG(y) \\
= \quad & -2\ E_{\widetilde{Y}}\{E_Y\{\log\, f(Y, \hat{\theta}(\widetilde{Y}))\}\},
\end{aligned}
$$

where the expectation is taken over two different sets of data $Y = (Y_1, \ldots, Y_n)$, $\widetilde{Y} = (\widetilde{Y}_1, \ldots, \widetilde{Y}_n)$ *stemming both from G, with* $\widetilde{Y}$ *being used for estimation of* $\hat{\theta}$ *and Y for evaluating the log likelihood.*

The factor 2 in (7.13) and (7.14) is arbitrary and, in principle, not necessary but used for conventional reasons. Note that with the result (7.14), we can employ the AIC for model selection. In particular, this means that $\mathrm{AIC}(\mathscr{M}_1) - \mathrm{AIC}(\mathscr{M}_2)$ is an estimate for the following quantity:

$$
\begin{aligned}
& 2E\{KL(G(\cdot), F_1(\cdot;\hat{\theta}_1))\} - 2E\{KL(G(\cdot), F_2(\cdot;\hat{\theta}_2))\} \\
= \quad & -2E\left\{\int \log\left(\frac{f_1(y;\hat{\theta}_1)}{f_2(y;\hat{\theta}_2)}\right) dG(y)\right\}
\end{aligned}
$$

This suggests that if the term is negative, then model $\mathscr{M}_1$ should be preferred, and if it is positive, then $\mathscr{M}_2$ is preferable.

The main advantage of the AIC is that it does not require refitting the model, unlike cross-validation. Hence, if the model training is burdensome, then the AIC has advantages compared to cross-validation. However, the AIC requires quantifying the number of parameters, which itself requires approximations when parameters are constrained. Moreover, maximum likelihood estimation needs to be carried out. This can be difficult in high-dimensional models, where, for numerical stability, some regularization needs to be applied. In fact, regularization has implicit consequences on the dimension of the parameter space and hence on *p*, the quantity involved in (7.13). We will pick up this point again in a later chapter.

Let us write the AIC in the case of normally distributed data. Assume $Y_i \sim \mathrm{N}(\mu, \sigma^2)$, where the mean μ might depend on some input variables *x* and some parameter θ. We denote this as $\mu = h(x;\theta)$. The resulting log likelihood equals

$$
l(\theta;\sigma^2) = \sum\nolimits_{i=1}^{n}\left(\tfrac{1}{2}\ \log\,(\sigma^2) - \tfrac{(y_i - h(x_i;\theta))^2}{2\sigma^2}\right). \tag{7.15}
$$

Differentiating (7.15) with respect to θ and σ^2 yields the maximum likelihood estimates, and we find that the variance is estimated through

$$
\widehat{\sigma}^2 = \tfrac{1}{n}\sum\nolimits_{i-1}^{n}\left(y_i - h(x_i;\hat{\theta})\right)^2. \tag{7.16}
$$

Further calculations show that the log likelihood at its maximum (ignoring terms not dependent on the parameters) takes the form

$$
l\left(\hat{\theta};\widehat{\sigma}^2\right) = -\tfrac{n}{2}\ \log\left\{\sum\nolimits_{i=1}^{n}\left(y_i - h\left(x_i;\hat{\theta}\right)\right)^2\right\}. \tag{7.17}
$$

Consequently, the AIC results in

$$
\mathrm{AIC}(\mathscr{M}) = n\,\log\left\{\sum\nolimits_{i=1}^{n}\left(y_i - h\left(x_i;\hat{\theta}\right)\right)^2\right\} + 2p, \tag{7.18}
$$

where p is the dimension of θ. This is an interesting formula, which allows interpreting the AIC from a different angle. Note that the first component in (7.18) measures the goodness of fit, that is, the empirical prediction error. How well is the fitted model $h(x_i, \hat{\theta})$ able to predict the unknown y_i? As we stated before, the more complex the model, the smaller the prediction error for the data used for fitting, also called "in-sample" prediction error. The goodness of fit term is now penalized by the complexity of the model, measured as the number of parameters. Hence, one aims to balance the goodness of fit with the complexity of the model. The two components and the resulting AIC are visualized in Fig. 7.3, upper plot.

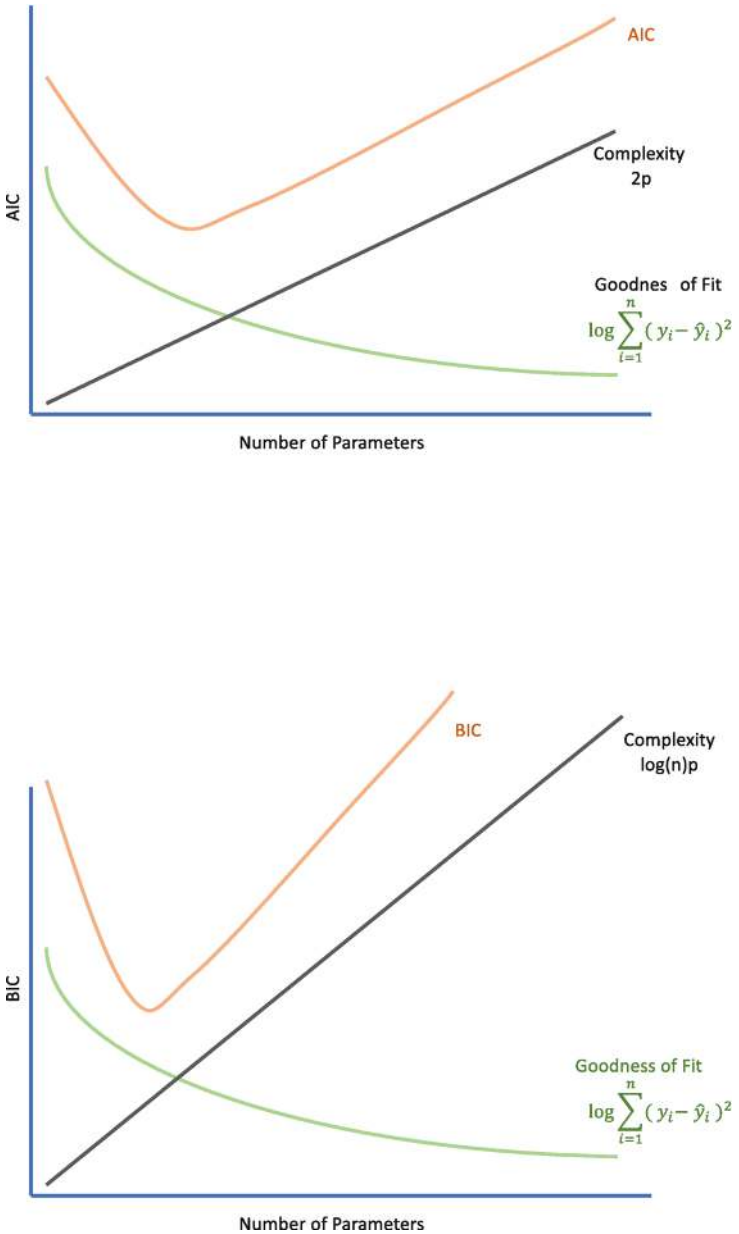


Fig. 7.3 AIC and BIC behavior by balancing goodness of fit with the complexity of a model

A companion of the AIC is the BIC, short for Bayesian information criterion. This is defined as follows.

Definition 7.5 *The* **Bayesian information criterion** *(BIC) for model class $\mathscr{M}$ is defined as*

$$BIC(\mathscr{M}) = -2 \sum_{i=1}^{n} \log f(y_i; \hat{\theta}) + \log (n) p \ , \tag{7.19}$$

where p is the dimension of the parameter θ and $\hat{\theta}$ is the maximum likelihood estimate in model class $\mathscr{M}$ as defined in (7.6).

Hence, we replace the "2" in the latter term of the AIC in (7.13) by the factor $\log (n)$. The BIC received its name from its derivation, which utilizes ideas of Bayesian inference. However, the BIC is an asymptotic measure and not a Bayesian model selection criterion.

Looking at the BIC in the normal distribution case discussed above, we obtain

$$\text{BIC}(\mathscr{M}) = n \log \left\{ \sum_{i=1}^{n} \left(y_i - h\Big(x_L; \hat{\theta}\Big)\right)^2 \right\} + \log (n) p. \tag{7.20}$$

Since for $n \geq 8$ we have $\log (n) \geq 2$, we see that the BIC favors models with lower complexity, that is, with lower-dimensional parameters. This is also visualized in Fig. 7.3.

We will subsequently prove that AIC as well as the BIC can be formally derived as model selection criteria. This is, however, somewhat clumsy and technical and can therefore be omitted upon first reading.

Proof Derivation of the AIC: We look at the expected Kullback-Leibler divergence

$$\begin{aligned} & E_{Y_1,\ldots,Y_n}\Big\{KL\big(G(\cdot), F(\cdot; \hat{\theta}(Y_1, \ldots, Y_n))\big)\Big\} \\ = \quad & \text{const} - E_{Y_1,\ldots,Y_n}\left\{ \int \log f\big(y; \hat{\theta}(Y_1, \ldots, Y_n)\big) dG(y) \right\}, \end{aligned} \tag{7.21}$$

where we made the dependence of our estimate $\hat{\theta}$ on the data $y_1, \ldots, y_n$ explicit. The constant term can be ignored in the following. Note that we have two integrals in the second term in (7.21): the inner integral over y in the Kullback-Leibler divergence and the outer integral resulting from the expectation with respect to $Y_1, \ldots, Y_n$. To be explicit, we can write the expectation in the second component of (7.21) as

$$\int \left[\int \log \Big(f\big(y; \hat{\theta}(y_1, \ldots, y_n)\big) \Big) dG(y) \right] dG(y_1, \ldots, y_n) \tag{7.22}$$

where $G(y_1, \ldots, y_n) = \prod_i G(y_i)$. Our aim is now to estimate the integral. We begin by disentangling the double integral and approximating $f(y; \hat{\theta})$ with a second-order Taylor series expansion around θ_0. To simplify notation, from now on, we write $\hat{\theta}(y_1, \ldots, y_n)$ as $\hat{\theta}$, but it should be kept in mind that $\hat{\theta}$ depends on $y_1, \ldots, y_n$ and not on y. This gives the approximation of (7.22) as

(7.23)

$$\int\int \left\{ \log f(y;\theta_0) + \frac{\partial \log f(y;\theta_0)}{\partial \theta^\top}(\hat{\theta}-\theta_0) - \frac{1}{2}(\hat{\theta}-\theta_0)^\top J(\theta_0)(\hat{\theta}-\theta_0) \right\}$$

$$dG(y) \prod_{i=1}^{n} dG(y_i) \approx \int \log\big(f(y;\theta_0)\big) dG(y)$$

$$-\frac{1}{2}\int (\hat{\theta}-\theta_0)^\top I(\theta_0)(\hat{\theta}-\theta_0) \prod_{i=1}^{n} dG(y_i).$$

where $J_i(\theta_0, y_i) = -\partial^2 \log f(y_i;\theta_0) \,/\, \partial\theta\partial\theta^\top$, and we utilized (3.3). Let us first look at the final component in (7.23). Making use of the asymptotic normality of the maximum likelihood estimate, this can be asymptotically approximated by $tr(I^{-1}(\theta_0)V(\theta_0))$ where $V(\theta_0)$ is the variance of the score function (see (4.28)). Note that neither the integral in the first component in (7.23) nor $V(\theta_0)$ and $I(\theta_0)$ can be calculated explicitly, as they depend on the unknown data-generating process $G(\cdot)$. Our aim is therefore to estimate the above quantity based on the data $y_1, \ldots, y_n$. We do know that the data $y_1, \ldots, y_n$ are drawn from $G(\cdot)$. Therefore, we replace the integral with its arithmetic mean. To do so, we look at the empirical version of (7.22) but leave $\hat{\theta}$ fixed for now, that is, we replace the expectation over $G(y)$ with the arithmetic mean using $y_1, \ldots, y_n$. To be specific, we calculate

$$\frac{1}{n}\sum_{i=1}^{n} \log f(y_i;\hat{\theta}). \tag{7.24}$$

This is clearly just the likelihood function at its maximum divided by the sample size. We again expand (7.24) around θ_0 and obtain an approximation of (7.24) with

$$\frac{1}{n}\sum_{i=1}^{n}\left\{ \log f(y;\theta_0) + s_i(\theta_0;y_i)^\top(\hat{\theta}-\theta_0) - \frac{1}{2}(\hat{\theta}-\theta_0)^\top J_i(\theta_0;y_i)(\hat{\theta}-\theta_0) \right\}. \tag{7.25}$$

Note that the first component in (7.25) can approximate the first component in (7.23). This means

$$\frac{1}{n}\sum_{i=1}^{n} \log f(y_i;\theta_0) \quad \to \quad \int \log f(y;\theta_0) dG(y) \tag{7.26}$$

for increasing *n*, which can in fact be proven to be a consistent estimate. Moreover, because with increasing sample size *n*

$$\frac{1}{n}\sum_{i} J_i(\theta_0;y) \quad \to \quad I(\theta_0), \tag{7.27}$$

we can argue that the integrand of the third component in (7.25) converges to

$$\frac{1}{2}(\hat{\theta}-\theta_0)^\top I^{-1}(\theta_0)(\hat{\theta}-\theta_0).$$

Taking the expectation with respect to $Y_1, \ldots, Y_n$ and using the asymptotic distribution of the maximum likelihood estimate, we get

$$E_{Y_1,\ldots,Y_n}\left(\frac{1}{2}(\hat{\theta}-\theta_0)^\top I^{-1}(\theta_0)(\hat{\theta}-\theta_0)\right) = \mathrm{tr}(I^{-1}(\theta_0)V(\theta_0)).$$

What remains is the second component in (7.25). If we subtract this and take (7.26), (7.27) and the approximation (7.23), we get

$$\frac{1}{n}\sum_{i=1}^{n}\log f(y_i;\hat{\theta}) - \frac{1}{n}\sum_{i=1}^{n} s_i^{\top}(\theta_0;y_i)(\hat{\theta}-\theta_0)$$

$$\rightarrow \quad \int\int \log f(y;\hat{\theta})dG(y)\prod_{i=1}^{n} dG(y_i).$$

This suggests the use of $\frac{1}{n}\sum_{i=1}^{n} s_i^{\top}(\theta_0;y_i)(\hat{\theta}-\theta_0)$ as a bias correction, which we now try to simplify further. Using (4.18) gives

$$\tfrac{1}{n}\sum\nolimits_i s_i^{\top}(\theta_0;y_i)(\hat{\theta}-\theta_0) \approx \tfrac{1}{n}\sum\nolimits_i s_i^{\top}(\theta_0;y_i)I^{-1}(\theta_0)\sum\nolimits_j s_j(\theta_0;y_j).$$

Bearing in mind that $E(s_i(\theta_0;y_i))=0$, if we take the expectation with respect to $Y_1,\ldots,Y_n$, we get

$$E_{Y_1,\ldots,Y_n}\left(\tfrac{1}{n}\sum s_i(\theta_0;y_i)^{\top}(\hat{\theta}-\theta_0)\right) \approx \tfrac{1}{n}tr\big(I^{-1}(\theta_0)V(\theta_0)\big). \tag{7.28}$$

As both $I(\theta_0)$ and $V(\theta_0)$ depend on the unknown distribution $G(\cdot)$, we cannot calculate their values. But if $G(\cdot)\in\mathscr{F}$, we have $V(\theta_0)=I(\theta_0)$ so that

$$\tfrac{1}{n}tr(I^{-1}(\theta_0)V(\theta_0)) = \tfrac{1}{n}p,$$

where *p* is the dimension of the parameter. This approximation is attractively simple and was proposed by Akaike (1973). Combining the above derivations gives the famous Akaike information criterion (AIC). □

Proof Derivation of the BIC: Very similar to the AIC is the Bayesian information criterion (BIC). Assume we have a set of models $\mathscr{M}_1,\ldots,\mathscr{M}_K$, each of which corresponds to a distributional model, such that

$$\mathscr{M}_k \Leftrightarrow Y_i \sim f_k(y;\theta_k) \quad i.i.d.$$

Model selection can now be interpreted as a decision problem, and a plausible strategy is to select the most likely model or, in Bayesian terminology, the model with the highest posterior probability. To this end, we need to calculate the posterior probability of each model. For model $\mathscr{M}_k$, this is given by

$$P(\mathscr{M}_k \mid y) \;=\; \frac{f(y|\mathscr{M}_k)P(\mathscr{M}_k)}{f(y)} \;=\; \frac{\int f_k(y;\vartheta_k)f_{\theta_k}(\vartheta_k)d\vartheta_k}{f(y)}P(\mathscr{M}_k), \tag{7.29}$$

where $P(\mathscr{M}_k)$ is the prior belief in model $\mathscr{M}_k$. The denominator $f(y)$ is calculated by summing over all models

$$f(y) = \sum\nolimits_k \int f_k(y;\vartheta_k)f_{\theta_k}(\vartheta_k)d\vartheta_k.$$

Although (7.29) allows us to quantify the posterior model probability, it is often complicated, or even infeasible, to calculate. We therefore pursue a simplifying approach and apply a Laplace approximation to the above integral. To begin, let

$$l_k(\theta_k) = \sum\nolimits_{i=1}^{n} \log\, f_k(y_i;\theta_k)$$

be the log likelihood for model $\mathscr{M}_k$ and define with $\hat{\theta}_k$ the corresponding maximum likelihood estimate. With Laplace approximation, the integral component in (7.29) is given by

$$\int \exp\left(l_k(\vartheta_k)\right) f_{\theta_k}(\vartheta_k) d\vartheta_k \approx \left(\tfrac{2\pi}{n}\right)^{\frac{p_k}{2}} \exp\left(l_k(\hat{\theta}_k)\right) f_{\theta_k}(\hat{\theta}_k) \left|\tfrac{1}{n} I_k(\hat{\theta}_k)\right|^{-\frac{1}{2}}$$

where p_k is the dimension of θ_k. Taking the log of the right-hand side gives

$$\tfrac{p_k}{2} \log(2\pi) - \tfrac{p_k}{2} \log(n) + l_k(\hat{\theta}_k) - \tfrac{1}{2} \log\left|\tfrac{1}{n} I_k(\hat{\theta}_k)\right| + \log f_{\theta_k}(\hat{\theta}_k).$$

If we collect all the components that grow with order *n*, i.e., the second and third, and multiply them by -2, we obtain the Bayesian information criterion (BIC):

$$BIC_k = -2 l_k(\hat{\theta}_k) + \log(n) p_k.$$

□

7.4 AIC, BIC, and Hypothesis Testing

For nested models, we may consider model selection also as a testing problem. This means we test on hypotheses stating that one or multiple components of the parameter vector θ can be ignored. To connect the two ideas, consider the setting where we assume two nested statistical models $\mathscr{M}_0$ and $\mathscr{M}_1$ with $\mathscr{M}_0 \subset \mathscr{M}_1$. We set

$$\mathscr{M}_k = \left\{f(y;\theta), \theta \in \Theta_k\right\} \text{ for } k = 0, 1$$

with $\Theta_0 \subset \Theta_1$. Assume that $p_0 = \left|\Theta_0\right|$ and $p_1 = \left|\Theta_1\right|$ are the dimensions of the corresponding parameter spaces, and let $p = p_1 - p_0 > 0$. In Sect. 4.4, we derived the distribution of the log-likelihood ratio for parameter testing. This can be extended to the multivariate case as follows. Let $\hat{\theta}_0$ be the maximum likelihood estimate in model $\mathscr{M}_0$, and accordingly $\hat{\theta}_1$ is the corresponding estimate in $\mathscr{M}_1$. It can be shown (see the end of the section) that if model $\mathscr{M}_0$ holds then

$$2\left\{l\left(\hat{\theta}_1\right) - l\left(\hat{\theta}_0\right)\right\} \sim \mathscr{X}^2_p. \tag{7.30}$$

Formulating the hypothesis $H_0 : \mathscr{M} = \mathscr{M}_0$ and the alternative $H_1 : \mathscr{M} = \mathscr{M}_1$, we obtain the decision rule

$$\text{“H}_1\text{”} \Leftrightarrow 2\left\{l\left(\hat{\theta}_1\right) - l\left(\hat{\theta}_0\right)\right\} > \mathscr{X}^2_{p,1-\alpha}, \tag{7.31}$$

where $\mathscr{X}^2_{p,1-\alpha}$ is the $1-\alpha$ quantile of the Chi-squared distribution with *p* degrees of freedom.

Instead of testing, we can also use AIC or BIC for model selection. Note that the decision rule of the AIC, when selecting either H_0 or H_1, can be written as

$$\text{“H}_1\text{”} \Leftrightarrow 2\left\{l\left(\hat{\theta}_1\right) - l\left(\hat{\theta}_0\right)\right\} > 2p, \tag{7.32}$$

while the BIC corresponds to

$$\text{“H}_1\text{”} \Leftrightarrow 2\left\{l\left(\hat{\theta}_1\right) - l\left(\hat{\theta}_0\right)\right\} > \log(n) p. \tag{7.33}$$

The different critical values, i.e., the threshold values to the right in the decision rules (7.31), (7.32), and (7.33), respectively, are shown in the upper plot in Fig. 7.4 for different values of

p. For the BIC, we plot this with $n = 100$. We see that the BIC has the strongest increase, which demonstrates again that the BIC tends to prefer small models. The AIC looks comparable to the test decision rule but also gets larger and hence prefers smaller models if p, the difference in dimensions, gets large. For small values of p, however, we see that the AIC prefers larger models.

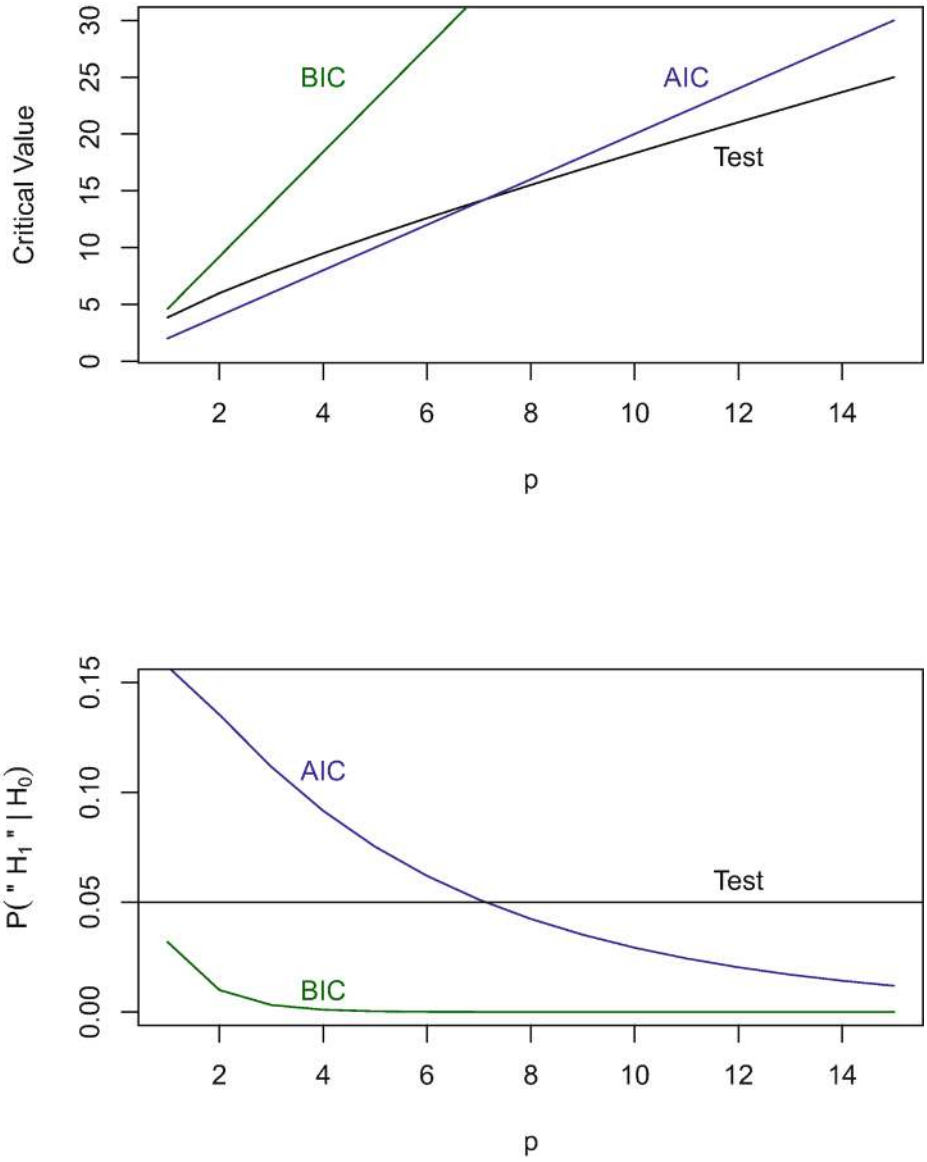


Fig. 7.4 Penalization values for AIC, BIC, and hypothesis testing (upper plot) and resulting significance level (bottom plot)

We could now calculate the probability of a type I error if we base a test on the decision rules from above. In other words, we calculate $P(\text{“H}_1\text{”} \mid H_0)$ using the three decision rules given above. This is shown in the bottom plot of Fig. 7.4. We see that both AIC and BIC lead to smaller models for larger p, meaning that the resulting significance level is small.

The above result may leave us puzzled, and we need to conclude that model selection based on AIC or BIC, respectively, and model selection based on testing are two different

and not generally matching ideas. This is important to note and should not come as a surprise. The two concepts are based on two different methodological grounds and follow different strands of reasoning. While statistical testing is more designed toward decision problems, with one of the decision errors limited to happen with small probability only, model selection based on AIC or BIC aims to avoid overfitting by balancing the goodness of fit with the complexity of a model. These are apparently two different strategies. Generally, if the focus is on model selection and not on inference with respect to particular parameters, it is advisable to use concepts like cross-validation or AIC/BIC, and these should not be combined with statistical tests in post-model selection inference. We refer to Claeskens and Hjort (2008) or Rügamer and Greven (2018) for a more detailed discussion in this direction.

Proof We derive the asymptotic distribution of the likelihood ratio (7.30). Let us for simplicity assume

$$\theta = (\underbrace{\theta_1, \ldots, \theta_{p_0}}_{=:\theta^{(0)}}, \underbrace{\theta_{p_0+1}, \ldots, \theta_{p_0+p}}_{=:\theta^{(1)}}),$$

so that for $\theta \in \Theta_0$ it holds that $\theta_0 = (\theta_0^{(0)}, \theta_0^{(1)})$ with $\theta_0^{(1)} \equiv 0$. Let $H_0 : \theta = \theta_0$ and assume that H_0 holds. We have

$$l(\hat{\theta}_1) - l(\theta_0) \overset{a}{\sim} \mathcal{X}^2_{p_0+p}. \tag{7.34}$$

Following similar arguments as in Sect. 4.4 and using that under H_0, we get $\hat{\theta}_0^{(0)} \approx \hat{\theta}_1^{(0)}$; it follows

$$\begin{aligned} 2\{l\left(\hat{\theta}_1\right) - l\left(\hat{\theta}_0\right)\} &= 2\{l\left(\hat{\theta}_1\right) - l(\theta_0) - \left(l\left(\hat{\theta}_0\right) - l(\theta_0)\right)\} \\ &\approx (\hat{\theta}_1 - \theta_0)^T I^{-1}(\theta_0)(\hat{\theta}_1 - \theta_0) - (\hat{\theta}_0 - \theta_0)^T I^{-1}(\theta_0)(\hat{\theta}_1 - \theta_0) \\ &\approx (\hat{\theta}_1^{(1)} - \theta_0^{(1)})^T \left[I^{-1}(\theta_0)\right]^{(1)(1)} (\hat{\theta}_1^{(1)} - \theta_0^{(1)}) \end{aligned}$$

where $\left[I^{-1}(\theta_0)\right]^{(1)(1)}$ is the submatrix corresponding to the elements in subvector $\theta^{(1)}$. Under H_0, this vector is zero, and with the asymptotic normality for the maximum likelihood estimator, we obtain the result. □

7.5 Forward Selection and Small Data Adjustment

The AIC and the BIC are based on asymptotic arguments. Hence, we need to assume that *n* is large. In fact, asymptotic arguments require that $n \to \infty$. But in real data examples, the amount of data is always limited, and sometimes we are faced with small samples, where *n* is so small that asymptotic arguments do not necessarily hold. We will demonstrate how a small sample correction for the AIC can circumvent problems in the model selection step. We will present this in so-called forward selection, meaning we start with simple models and extend these successively.

We have used AIC and BIC for comparing two models above. Let us extend this approach here to a cascade of *K* nested models $\mathcal{M}_1 \subset \mathcal{M}_2 \subset \ldots \mathcal{M}_K$. For demonstration purposes,

we motivate the idea using a simple linear regression model. Let $x_i = (x_{i1}, \ldots, x_{iK})$ be a set of covariates (for the *i* observation) leading to the overall linear model

$$Y_i = \underbrace{\theta_0 + x_{i1}\theta_1 + x_{i2}\theta_2 + \ldots x_{iK}\theta_K}_{=:\, h(x;\,\theta)} + \epsilon_i. \tag{7.35}$$

We aim to find the "relevant" input variables by excluding "irrelevant" covariates. To do so, we require some notation. Let $\mathscr{M}_K$ be the model (7.35) with all covariates included. Excluding a single covariate can be notated by the *K* index sets $\mathscr{I}_{(K-1)(k)} = \{1, \ldots, K\} \setminus \{k\}$, and we define with $\mathscr{M}_{(K-1)(k)}$ the regression model with covariates $k \in \mathscr{I}_{(K-1)(k)}$ included. This means

$$\mathscr{M}_{(K-1)(k)} \Leftrightarrow Y_i = \theta_0 + \sum\nolimits_{j \in \mathscr{I}_{(K-1)(k)}} x_{ij}\theta_j + \epsilon_i. \tag{7.36}$$

Excluding two covariates leads to $\binom{K}{2}$ different index sets which we index with $\mathscr{I}_{(K-2)(k)}$ where $k = 1, \ldots, \binom{K}{2}$. Accordingly, we can define the index sets $\mathscr{I}_{(k)(l)}$ where $k = 1, \ldots K$ and $l = 1, \ldots, \binom{K}{k}$, and the corresponding model sets are denoted as $\mathscr{M}_{(k)(l)}$. We may now calculate the AIC values for all models and choose the one with the smallest value. Hence, we select the model such that

$$\widehat{\mathscr{M}} = \text{argmin}_{k,l}\{\text{AIC}(\mathscr{M}_{(k)(l)})\}. \tag{7.37}$$

It should be clear that this can lead to a formidable numerical effort. It is therefore advisable to proceed successively, starting with the null model, i.e., with no covariates included, and then proceed step by step by including a single variable that decreases the AIC or BIC value most. This is called *forward selection*, and it can be written formally as follows:

1. Calculate the AIC for the model
$$Y_i = \mu + \epsilon_i,$$
and denote this with AIC_0. Let $\mathscr{I}_0 = \emptyset$ and set $k = 0$.
2. Let $\mathscr{M}_k = \{1, \ldots, K\} \setminus \mathscr{I}_k$, calculate the AIC for the $|\mathscr{M}_k|$ different models
$$Y_i = \mu + \sum\nolimits_{k \in \mathscr{I}_k} x_{ik}\theta_k + x_{ij}\theta_j + \epsilon_i$$
for $j \in \mathscr{M}_k$, and choose the one with the minimal AIC value. Let the corresponding index of the covariate be *l* and define $\mathscr{I}_{k+1} = \mathscr{I}_k \cup \{l\}$. The corresponding AIC value is defined as AIC_{k+1}.
3. If $\text{AIC}_{k+1} < \text{AIC}_k$, set $k = k + 1$ and jump back to step 2. Otherwise, select the model with $\mathscr{M}L_k$ as the best model.

We visualize the procedure in Fig. 7.5 and demonstrate it with a simulation setup. We therefore simulate from

$$Y_i = \sum\nolimits_{k=1}^{K} x_{ik}\theta_k + \epsilon, \tag{7.38}$$

where $\epsilon \sim \mathrm{N}(0,1)$. The covariates are also drawn from a standard normal distribution, and all draws are independent. The parameters θ are such that $K \,/\, 2$ of the values are set to zero, while the remaining $K \,/\, 2$ are again drawn from a $\mathrm{N}(0,1)$ distribution. In Fig. 7.6—left-hand side—we plot the resulting smallest AIC values for a sample size of 1000 and $K = 100$. We find an exposed minimum at model dimension 59. The optimal model contains most of the $K \,/\, 2$ covariates with truly nonzero values (meaning existent covariates) but also additional spurious ones. The true coefficients are shown in the bottom left plot in Fig. 7.6, where we see the selected coefficients as bullet points and the unselected as circles. We experience a well-known weakness of the AIC, which has the tendency to choose too complex models.

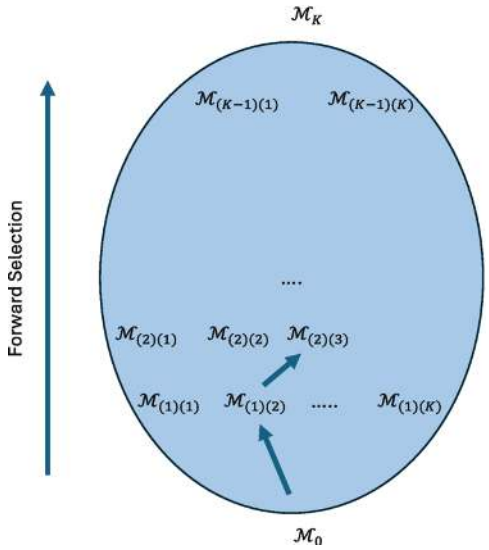


Fig. 7.5 Sketch of forward selection procedure

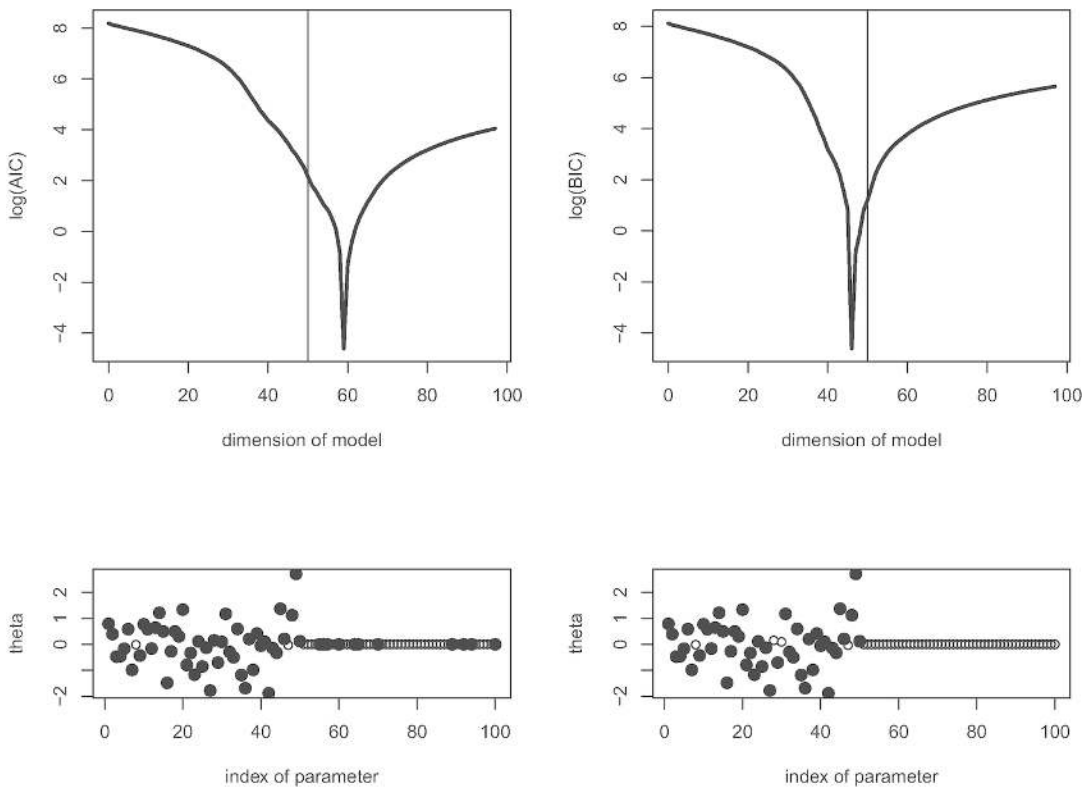


Fig. 7.6 Performance of AIC (left-hand side) and BIC (right-hand side) in the simulation example. The upper row shows the AIC and BIC values, respectively. The bottom shows the true coefficients with the selected one indicated with points

If we replace the factor 2 in AIC by $\log(n)$, we obtain the BIC. Applied to the data, we get the plot shown on the right-hand side of Fig. 7.6. We find that the selected model is smaller, with just 46 parameters out of the 50 existing ones included. The unselected parameters in fact have small values as can be seen in the bottom plot of Fig. 7.6. More importantly, the BIC has selected none of the spurious variables. This is also a known performance of the BIC as it tends to select smaller models compared to the AIC.

The performance of the AIC (as well as the BIC) depends on the sample size. If the sample size n is small, then AIC can perform in a very unsatisfactory way. We demonstrate this with the above example by reducing the sample size to 100. The corresponding plot is shown in Fig. 7.7 on the left-hand side. Now, all variables are included, even all spurious ones, and apparently, the performance is not acceptable. This has led to a small sample size correction for the AIC. For small samples, the AIC needs to be replaced by the corrected AIC, which is defined as follows:

$$\mathrm{AIC}_c = -2l(\hat{\theta}) + 2p\left(\frac{n}{n-p-1}\right) \tag{7.39}$$

The correction was proposed by Hurvich and Tsai (1989). We also refer to DelSole and Tippett (2021) for further discussion with respect to the small sample correction of the AIC. Note that if the dimension p gets large, the denominator gets small, and hence the second term in (7.39) gets large. We demonstrate the effect of the correction using the above simulation. The results are shown in Fig. 7.7 on the right-hand side. The improvement is

clearly visible, so one should generally work with the corrected AIC, in particular, if the sample size n is small.

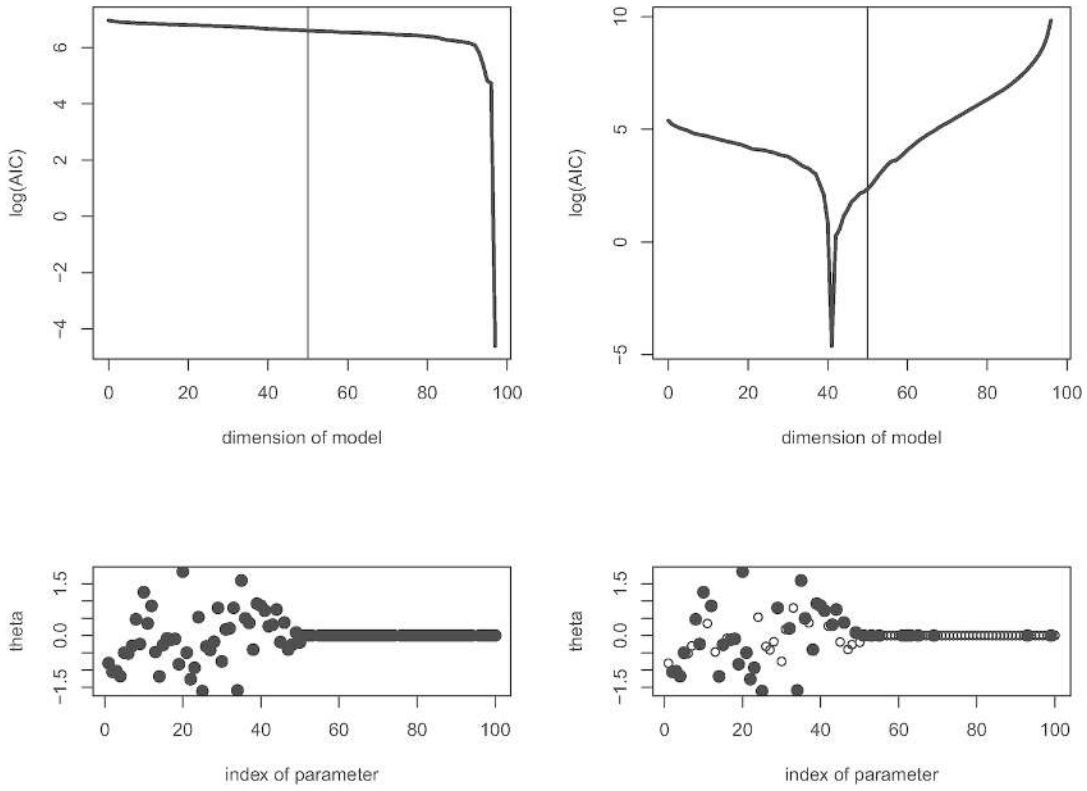


Fig. 7.7 Performance of AIC (left-hand side) and corrected AIC (right-hand side) in the simulation example for small n. Upper row shows the AIC and corrected AIC values, respectively. The bottom shows the true coefficients with the selected one indicated as filled points; unselected are unfilled

7.6 Model Selection and Dimension of the Model

Let us define the dimension of a model $\mathscr{M}$ as the dimension of the related parameter space Θ, which we denote as $|\Theta|$. Classical statistical inference is based on asymptotic arguments, which means that the data size n is increasing, while the model dimension $|\Theta|$ itself is fixed. Apparently, though, with more and more data, one can fit more complex models, and hence, in principle, it seems useful and plausible to allow the model dimension $|\Theta|$ to depend on n as well. We want to initiate a discussion here in the case where the data size n is large (or even very large) and set this in comparison to the model size. In general, we want to look at the ratio

$$\frac{|\Theta|}{n}. \tag{7.40}$$

We can think of three constellations for the above ratio, namely,

$$\text{(a) } \frac{|\Theta|}{n} \ll 1 \quad \text{or} \quad \text{(b) } \frac{|\Theta|}{n} \lesssim 1 \quad \text{or} \quad \text{(c) } \frac{|\Theta|}{n} > 1. \tag{7.41}$$

The three cases need to be considered separately. Model selection allows choosing the dimension of the model data-driven, but that means that the selected model with size $|\Theta|$ depends on n. Assuming that n is large (or even very large), we are in principle in the framework of considering asymptotic statements. But this does require that $|\Theta|$ is limited,

i.e., we are in situation (a) in (7.41). In this case, all asymptotic arguments derived in the book so far hold.

But what happens if the dimension of the parameter space is large as well, maybe even larger than n? This brings us to constellations (b) and (c) in (7.41). A comprehensive discussion of these scenarios goes well beyond the scope of this book. But we want to refer to some ideas that are found in later chapters.

Scenario (b) in (7.41) is often dealt with by using regularization techniques. The latter guarantees that the numerical performance behaves stably, i.e., the optimization problem is (to a large extent) concave. Statistically, this can be linked to Bayes estimation, which we motivate in Sect. 9.7.

The latter case (c) is beyond the classical statistical framework, and classical statistical statements are void. It means that the model has more parameters than observations. Recent applications in machine learning and, in particular, in deep learning have shown that such models can have good, and in fact better, prediction properties compared to models where the ratio (7.40) is smaller than 1 (LeCun et al., 2015). We will return to relations to machine learning in Chap. 13.

7.7 Exercises

Exercise 1

(a) Briefly define the Akaike and the Bayesian Information Criteria. What is their function? What is the main (empirical) difference between the two?

(b) To model the number of observed daily deaths related to COVID-19 in Germany over the year 2021, a Poisson distribution could be used. The resulting BIC when applying this model is 3020.443. What is the value of the likelihood for the model?

(c) As an alternative to the Poisson, a researcher proposes to use the negative binomial distribution, a parent of the Poisson distribution with a second parameter to account for potential over- or under-dispersion. The BIC of that model is 3023.909. Which model should be chosen according to the BIC?

(d) Suppose we decide to select our model based on AIC instead of BIC. What are the AIC values for the two models?

(e) Do the two criteria agree? Which of the two criteria imposes a heavier penalty on the number of parameters?

Exercise 2 The dataset `teengamb` in the **R**-package `faraway` contains information on the gambling behavior of teenagers in Great Britain and contains the following variables:

Table

(a) You can fit a linear model in *R* using the function `lm`. Use the AIC in order to pick

between *Model 1*

$$\texttt{gamble} = \gamma_0 + \gamma_1 \texttt{status} + \gamma_2 \texttt{income} + u$$

and *Model 2*

$$\texttt{gamble} = \psi_0 + \psi_1 \texttt{verbal} + \psi_2 \texttt{sex} + \epsilon.$$

Which model would you prefer?

(b) Now, use leave-one-out cross-validation in order to compare the mean squared errors of prediction (MSEPs) for *Model 1* and *Model 2*. Do the two methods agree?

(c) What are the main differences between AIC and leave-one-out cross-validation? What similarities do the two methods share? What are the advantages/disadvantages of each approach?

Exercise 3 The file `Portland.rds` contains data on the income (in US dollars) of a sample of 1000 randomly selected single women and men residing in the city of Portland, Oregon. We here want to model the data by using an inverse Gaussian distribution, which possesses the positive skew that is typical of income distributions. The inverse Gaussian distribution has the following form:

$$f(x, \mu, \lambda) = \frac{\lambda}{\sqrt{2\pi x^3}} \exp\left\{-\frac{\lambda(x-\mu)^2}{2\mu^2 x}\right\}$$

(a) Assuming λ to be fixed to a constant (positive) value, derive the maximum likelihood estimator (MLE) for the parameter μ, as well as an asymptotic 95% confidence interval for it.

(b) We are now interested in how well our model would fit to new, unseen income data from Portland. To estimate this, fit the inverse Gaussian model using leave-one-out cross-validation, and calculate the mean squared error of prediction (MSEP). You can orient your implementation on the code provided in Exercise 11 or Tutorial 4 for this.

(c) Leave-one-out cross-validation is often infeasible in settings with large amounts of data or with complex models that take a lot of time to compute. Therefore, we frequently employ k-fold cross-validation instead. Implement a simple version of fourfold cross-validation that splits the data into four equally sized blocks (you can assume that the data can be evenly divided by four), iteratively fits four models using three of the four blocks, and determines the MSEP on the leftover fourth block, respectively. Then, calculate the average of the four MSEPs computed in each iteration and return this value.

References

Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle. In *Proceedings of the 2nd*

international symposium on information theory (pp. 267–281).

Claeskens, G., & Hjort, N. L. (2008). *Model selection and model averaging*. Cambridge Books.

DelSole, T., & Tippett, M. K. (2021). Correcting the corrected aic. *Statistics & Probability Letters,173*, 109064. [MathSciNet][https://doi.org/10.1016/j.spl.2021.109064]

Grünwald, P. (1999). Viewing all models as "probabilistic". In *Proceedings of the twelfth annual conference on Computational learning theory* (pp. 171–182).

Hurvich, C. M., & Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika,76*(2), 297–307. [MathSciNet][https://doi.org/10.1093/biomet/76.2.297]

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature,521*(7553), 436–444. [https://doi.org/10.1038/nature14539]

Rügamer, D., & Greven, S. (2018). Selective inference after likelihood-or test-based model selection in linear models. *Statistics & Probability Letters,140*, 7–12. [MathSciNet][https://doi.org/10.1016/j.spl.2018.04.010]

Wang, Q., Ma, Y., Zhao, K., & Tian, Y. (2022). A comprehensive survey of loss functions in machine learning. *Annals of Data Science,9*(2), 187–212. [https://doi.org/10.1007/s40745-020-00253-5]

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
G. Kauermann et al., *Statistics—Coping with Uncertainty*, Textbooks in Statistical Science
https://doi.org/10.1007/978-3-032-33658-3_8

8. Simulating Random Variables and Simulation-Based Inference

Göran Kauermann[1] ✉, Giacomo De Nicola[2], Cornelius Fritz[3] and David Rügamer[4]
(1) Department of Statistics, LMU Munich, Munich, Germany
(2) Harvard T.H. Chan School of Public Health, Harvard University, Boston, MA, USA
(3) School for Computer Science and Statistics, Trinity College Dublin, Dublin, Ireland
(4) Munich Uncertainty Quantification AI Lab, LMU Munich, Munich, Germany

The concepts and ideas of statistical reasoning presented so far are mostly based on known or asymptotic distributions combined with mathematical results, which were developed in times before computers were easily accessible. Now, with computing power generally available, the view changes, and simulation-based strategies come to the foreground. Imagine Mendel, with his inheritance experiments described in Example 5.1, had a computer available at that time. He could have easily simulated how much variation can occur in the considered properties of peas if the inheritance law were valid. Similarly, Fisher and Pearson could have sorted out their argument about the degree of freedom in Chi-squared tests quite easily by running extensive simulations. Indeed, today, scientific progress is heavily based on simulations, and we want to devote ourselves to the topic here as well. Our focus is thereby on simulations of random variables, be they univariate or multivariate. We start from basic ideas and end with rather complex simulation strategies; some of these will be reused again, once we get into Bayesian reasoning.

8.1 Simulating Uniformly Distributed Random Variables

Today, it is natural to generate random numbers, but we tend to forget that, apparently, numbers which are produced by computers and hence algorithms are

not random but deterministic. In this respect, it would be more exact to call random numbers generated by a computer program “pseudo-random numbers.” However, we do not want to be that precise here and use the terminology random numbers throughout this chapter. We refer to Ross (2022) for a general discussion and readable introduction to simulation and random number generators. We here only give the main ideas.

The first task is to generate uniformly distributed random numbers on $[0, 1]$. This is a mathematical and algorithmic problem which has been tackled in the middle of the twentieth century with generators proposed by Lehmer (1951) and Thomson (1958). The approach, known as the multiplicative congruential method, works as follows. For some starting value x_0 and values a and m, one produces

$$x_n := (a x_{n-1}) \bmod m \tag{8.1}$$

where “mod” is the modulo operator, that is, we divide $a x_{n-1}$ by m and take the remaining full integer. The numbers a and m need to be chosen according to the computing environment, and choices $m = 2^{31} - 1$ and $a = 7^5$ have proven to work appropriately. From these numbers, we calculate $U_n = x_n \,/\, m$, which lies between 0 and 1. We show the values of U_n for $x_0 = 1$ in Fig. 8.1 in the top plot. Calculating the resulting empirical distribution function leads to the middle plot shown in Fig. 8.1. We see a reasonable concordance of U_n with a uniform distribution on $[0, 1]$. It is important to note, though, that the numbers given are deterministic and will always result as shown. To demonstrate this, we change the initial value to 5 and denote the two resulting series with $x_{n,1}$ and $x_{n,2}$, where $x_{0,2} = 5$ while $x_{0,1} = 1$ as before. When we plot $x_{n,1} \,/\, m$ against $x_{n,2} \,/\, m$, we see the pattern shown on the bottom plot of Fig. 8.1, which clearly visualizes that the whole process is deterministic. Still, we are able to produce pseudo-random variables whose distribution resembles a uniform distribution on the interval $[0, 1]$. It is also relevant to remark that the random number generator (8.1) produces random numbers which are close to independence, meaning that x_n and x_{n+1} can be treated as (quasi) independent draws, even though apparently they are functionally dependent.

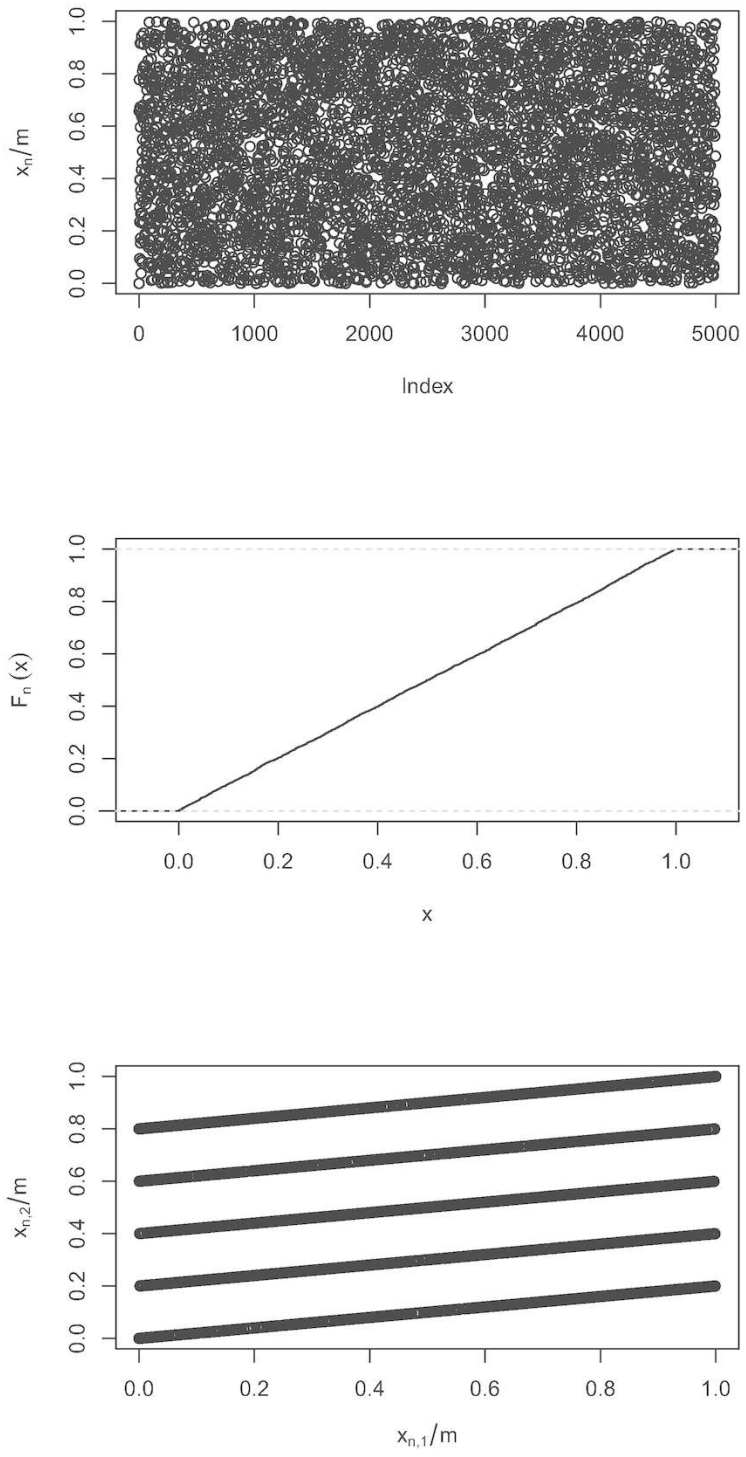


Fig. 8.1 Pseudo-random numbers generated through (8.1) with starting value $x_0 = 1$ (top plot) and the resulting empirical distribution function (middle plot). The bottom plot shows two sequences of random numbers with different starting values

From now on, we drop the term "pseudo" and just speak about generated random variables U_n on $[0, 1]$. We can utilize these random numbers in various

forms. To start with, assume that we are interested in calculating the integral of a function $g(y)$, where $g : \mathbb{R} \to \mathbb{R}^+$. Consider, for instance, $g()$ to be some function that is proportional to some density function $f()$. In other words, we know the shape of the density, but the function $g()$ does not integrate to one. In this case, we are interested in

$$\int_{-\infty}^{\infty} g(y) \mathrm{d}\, y \tag{8.2}$$

assuming that the function is integrable. We want to calculate this integral using simulations. To do so, we first transform y to some value lying in the interval $[0, 1]$. Let therefore $h(.) : \mathbb{R} \to [0, 1]$ be some invertible transformation and set $u = h(y)$ and correspondingly $y = h^{-1}(u)$, where $u \in [0, 1]$. A convenient choice is to work with the logit function (also called the sigmoid function), which takes the form

$$u = h(y) = \frac{\exp(y)}{1+\exp(y)} \quad \Leftrightarrow y = h^{-1}(u) = \log\left(\frac{u}{1-u}\right)$$

Then using standard substitution rules integral (8.2) can be rewritten to

$$\int_{-\infty}^{\infty} g(y) \mathrm{d}\, y = \int_0^1 g(y(u)) \frac{\mathrm{d}\, y(u)}{\mathrm{d}u} \mathrm{d}u = \int_0^1 \underbrace{g(y(u)) \frac{1}{u(1-u)}}_{=:\, \tilde{g}(u)} \mathrm{d}u \tag{8.3}$$

With this trick, we have replaced the calculation of an integral with the calculation of an expected value. To see this, assume $U \sim \text{Uniform}[0, 1]$. Then

$$E\Big(\tilde{g}(U)\Big) = \int_0^1 \tilde{g}(u) \mathrm{d}u. \tag{8.4}$$

Drawing now independent uniformly distributed random variables U_i with $i = 1, \ldots N$, we obtain, based on the law of large numbers (see Sect. 1.2.2),

$$\overline{\tilde{g}} := \frac{1}{N} \sum_{i=1}^{N} \tilde{g}(U_i) \to E\Big(\tilde{g}(U)\Big) \tag{8.5}$$

as N increases. Hence, we can generate a sample U_i, with $i = 1, \ldots, n$, to approximate the interval. If the random numbers are independent, which we assume, we can even assess the quality of the approximation using the tools from the previous chapters. That is, with a 95% confidence, we obtain based on the central limit theorem

$$E\Big(\tilde{g}(U)\Big) \in \left[\tilde{g} - 1.96\, \frac{\widehat{\sigma}}{\sqrt{N}}, \tilde{g} + 1.96\, \frac{\widehat{\sigma}}{\sqrt{N}}\right] \tag{8.6}$$

where

(8.7)

$$\widehat{\sigma}^2 = \tfrac{1}{N-1} \sum_{i=1}^{N} \big(\tilde{g}(u_i) - \tilde{g}\big)^2.$$

Interestingly, statistical concepts can be used for numerical calculations, and we can also utilize the derived results to quantify uncertainty.

Example 8.1 We want to demonstrate the above idea by calculating the integral

$$\int_{-\infty}^{\infty} \exp\left\{-\tfrac{1}{2}(y-\mu)^2\right\} \mathrm{d}\, y \tag{8.8}$$

where $\mu \in \mathbb{R}$. Apparently, we know the value of the integral exactly from the normalization constant of the normal distribution, and it results in $\sqrt{2\pi} = 2.506628$. But let us pretend that we do not know the integral and draw $N = 1000$ uniformly distributed random variables as proposed above. In Fig. 8.2, we show the resulting values $\tilde{g}(U_i)$ for $\mu = 0$ on the top plot, while the bottom plot shows $\tilde{g}(U_i)$ when setting $\mu = 4$. The resulting confidence intervals for the value of the integral are based on the 1000 samples, and they result in $[2.425826, 2.594262]$ for the case $\mu = 0$ and $[2.034159, 3.535753]$ for the case that $\mu = 4$. Note that the value of the integral in (8.8) does not depend on μ, but apparently, the quality of the approximation does depend on the value of μ. In particular, the quality of the approximation depends on the variance, which is much larger for the case $\mu = 4$, as can also be seen in Fig. 8.2. We conclude that while mathematically the value of μ does not matter, in simulations, we need to take the variance into account in order to achieve reliable results. We can also conclude that integration can be carried out by simulations, and classical statistical inference allows us to assess the quality of the approximation. ▷

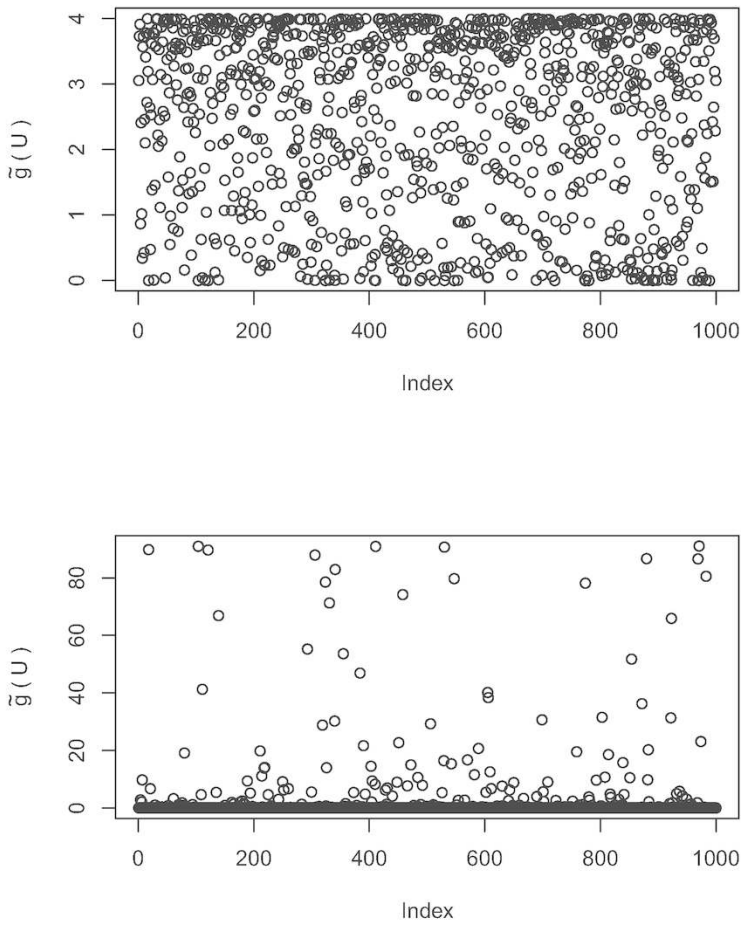


Fig. 8.2 Random numbers $\tilde{g}(U_i)$ for $\mu = 0$ (top) and $\mu = 4$ (bottom)

8.2 Simulating from a Distribution Function and a Density

In the previous section, we discussed how to simulate uniformly distributed random variables. We now turn to the more general problem of simulating random variables from arbitrary continuous distributions. Specifically, we aim to simulate a random variable $Y \sim F(\cdot)$, where $F(\cdot)$ is the distribution function of a continuous random variable. We consider three distinct cases:

1. The distribution function $F(\cdot)$ is known.
2. The distribution function $F(\cdot)$ is unknown, but the related density $f(\cdot)$ is known.
3. Neither the distribution function $F(\cdot)$ nor the density $f(\cdot)$ are known. We just know the density up to an unknown normalization constant, that is, $\widetilde{f}(y)$ is known with $\widetilde{f}(y) \propto f(y)$

known with $f(y) \propto f(y)$.

In this section, we address the first two scenarios. The third case will be discussed in Sect. 8.3.

We start with the setting where the distribution function $F(y)$ of a continuous random variable Y is known. We also assume that this is invertible, which allows us to easily draw a random sample by utilizing the inverse $F^{-1}(.)$ of the distribution function.

Property *Let U be uniformly distributed on* $[0, 1]$*. We define* $Y = F^{-1}(U)$*. Then* Y *is distributed with distribution function* $F(\cdot)$*, that is*

$$Y \sim F(.) \tag{8.9}$$

□

The proof of the statement is rather simple.

Proof Let *U* be uniformly distributed on $[0, 1]$, and define $Y = F^{-1}(U)$. Note that one has $P(U \leq u) = u$ for $u \in [0, 1]$. Hence

$$\begin{aligned} P(Y \leq y) &= P\big(F^{-1}(U) \leq y\big) \\ &= P\big(F(F^{-1}(U)\big) \leq F(y)) \\ &= P\big(U \leq F(y)\big) = F(y). \end{aligned}$$

This concludes the proof. □

The simulation idea is sketched in Fig. 8.3 (top plot), and it is not difficult to see that the result extends to discrete random variables, as sketched in the bottom plot of Fig. 8.3 for a Poisson distribution with $\lambda = 2$. The inverse of the distribution function is, in this case, formally defined as

$$F^{-1}(u) = \inf\{y : F(y) \geq u\}.$$

We can conclude that if the (cumulative) distribution function $F(\cdot)$ is known and available in analytic form, we can easily simulate from it using Eq. (8.9).

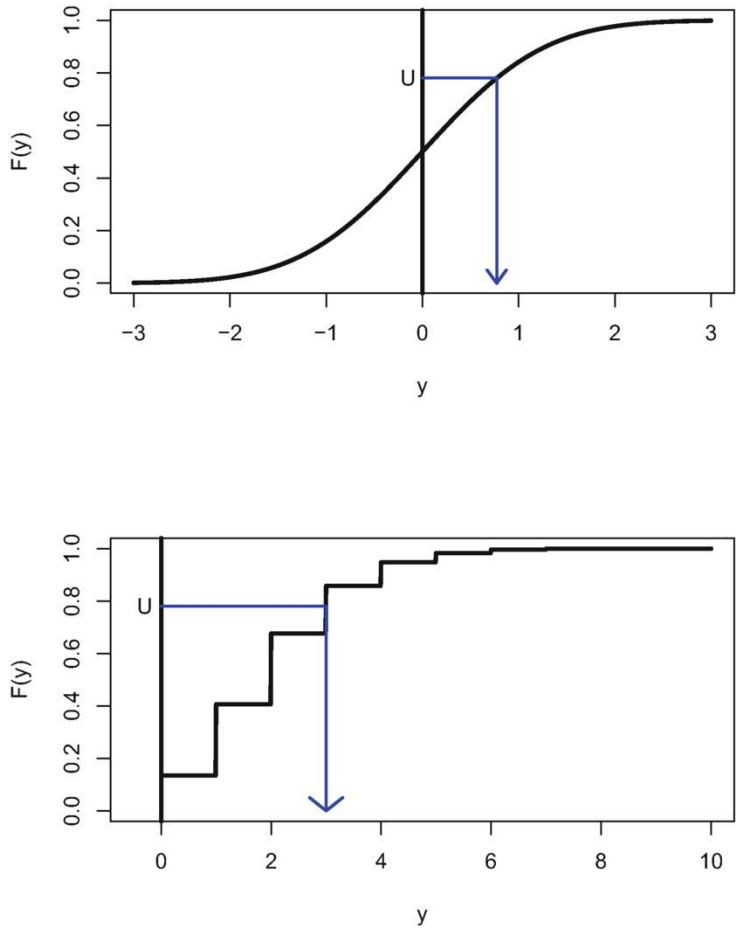


Fig. 8.3 Sketch for drawing random variables with known distribution

Let us now go a step ahead and assume that the distribution function of the random variable Y is not available, but the density $f(y)$ is. Hence, we know $f(y)$, but the integral $F(y) = \int_{-\infty}^{y} f(\tilde{y})d\tilde{y}$ is not available in analytic (and hence invertible) form. Consequently, we cannot apply Eq. (8.9).

In this case, we can use **rejection sampling**, a technique described subsequently. Suppose we are able to efficiently simulate from a so-called **proposal distribution** $F^*(\cdot)$, meaning that we can use Eq. (8.9) to obtain simulations $Y^* \sim F^*(.)$. Assume further that the related density $f^*(\cdot)$ is known. In other words, $F^*(\cdot)$, its inverse $F^{*-1}(.)$ and the corresponding density $f^*(\cdot)$ are available, and we can (numerically) evaluate them. The proposal distribution is thereby chosen such that there exists a scalar $a > 1$ with $f(y) < af^*(y)$ for all $y \in \mathbb{R}$. The latter means that if we multiply the proposal density $f^*(y)$ by some $a > 1$, the resulting value is always larger than $f(y)$, the density we want to simulate from, for all $y \in \mathbb{R}$. Hence $af^*(\cdot)$ lies like an "umbrella" over $f(\cdot)$, which

is why $f^*(\cdot)$ is also called **umbrella density**. We visualize this in the top plot in Fig. 8.4. The density $f(\cdot)$ from which we want to simulate is bimodal. If we use the density of the normal distribution $\mathrm{N}(0, 3^2)$ and multiply it by $a = 2$, we obtain an umbrella density, as shown in the figure. The idea is now to simulate from $f^*(\cdot)$ but reject the simulated values according to the target density $f(\cdot)$.

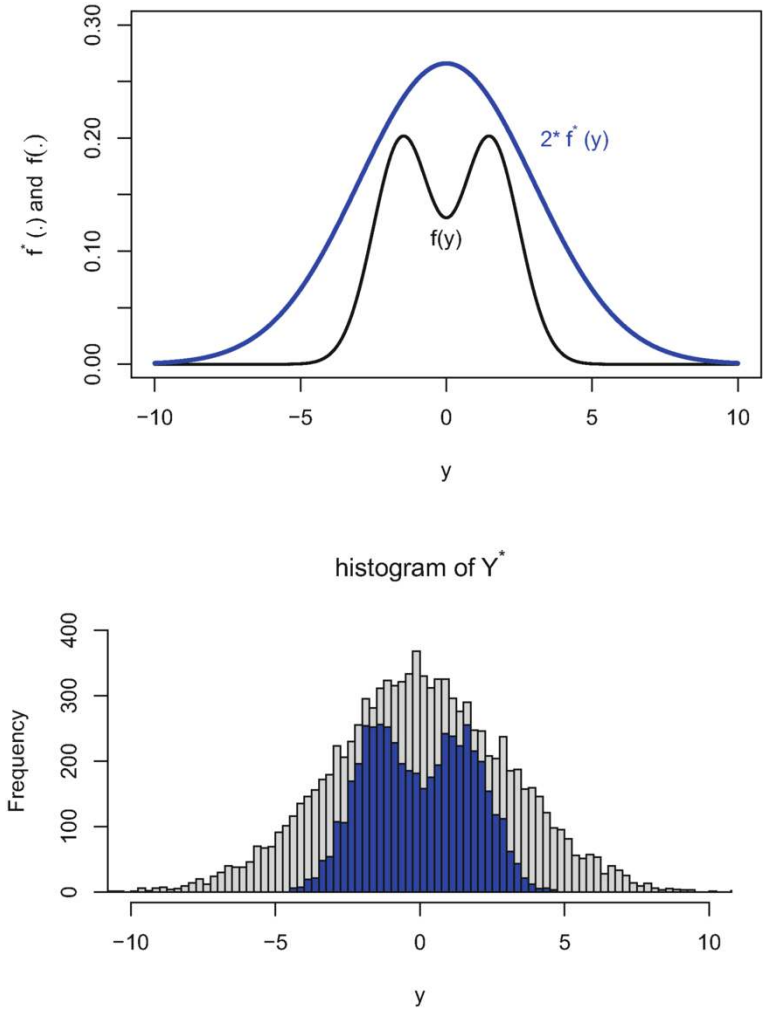


Fig. 8.4 Sketch of rejection sampling. The top plot shows the "umbrella" density $f^*(\cdot)$ and the density $f(\cdot)$ from which we want to sample. Bottom plot shows the simulated values Y^* from $f^*(\cdot)$ as a histogram. The accepted values are shown as blue

The rejection sampling scheme proceeds as follows:

1. Draw Y^* from $F^*(\cdot)$ using Eq. (8.9).
2. Generate a uniformly distributed random variable U on $[0, 1]$.
3. If

$$U \leq \frac{f(Y^*)}{af^*(Y^*)}$$

accept Y^*; otherwise, go back to Step 1.

The procedure is called "rejection" sampling, since the proposed value Y^* needs to be accepted and is otherwise "rejected." Before going through a formal proof, we demonstrate the performance with the umbrella density shown in Fig. 8.4. The bottom plot shows a histogram of 10.000 simulated values Y^* from $f^*(\cdot)$, which is a $\mathrm{N}(0, 3^2)$ distribution. Not all values are accepted. We color the accepted values in the histogram as blue. It appears that the accepted values follow the bimodal density $f(\cdot)$.

We want to show the impression resulting from Fig. 8.4 more rigorously. Hence, we need to prove that the accepted values Y^*, which pass Step 3, are drawn from the distribution $F(\cdot)$ and not from the proposal distribution $F^*(\cdot)$.

Proof Note that values, which pass Step 3, have the distribution function

$$P\left(Y^* \leq y \middle| U \leq \frac{f(Y^*)}{af^*(Y^*)}\right). \tag{8.10}$$

We need to prove that this equals $F(y)$. Following Bayes' rule, we obtain

$$P\left(Y^* \leq y \middle| U \leq \frac{f(Y^*)}{af^*(Y^*)}\right) = \frac{P\left(Y^* \leq y, U \leq \frac{f(Y^*)}{af^*(Y^*)}\right)}{P\left(U \leq \frac{f(Y^*)}{af^*(Y^*)}\right)}. \tag{8.11}$$

Let us first look at the numerator. This can be rewritten as

$$\begin{aligned} P\left(Y^* \leq y, U \leq \frac{f(Y^*)}{af^*(Y^*)}\right) &= \int_{-\infty}^{y} P\left(U \leq \frac{f(y^*)}{af^*(y^*)}\right) f^*(y^*)\mathrm{d}\ y^* \\ &= \int_{-\infty}^{y} \frac{f(y^*)}{af^*(y^*)}) f^*(y^*)\mathrm{d}\ y^* \\ &= \frac{1}{a} F(y) \end{aligned}$$

The denominator in (8.11) is the same as the numerator with y set to infinity. Since $F(\infty) = 1$, the result follows immediately. □

The rejection sampling algorithm sounds simple, but one should not forget that one needs to pass Step 3 to obtain a single simulated value. Note that the denominator in (8.11) gives the overall acceptance probability, which results in $1 / a$. Hence, the larger a, the smaller the acceptance probability. The general intention should, therefore, be to find a distribution $F^*(\cdot)$, which is as close as possible to the unknown distribution $F(\cdot)$. We visualize this in Fig. 8.5, where we show two possible proposal densities $f^*(\cdot)$ to simulate from the bimodal distribution $f(\cdot)$. For version 1, we use a normal distribution with $\mu = 0$ and

$\sigma = 2$. Here, a is set to 1.55. For the second version, we use a normal distribution with $\mu = 0$ and $\sigma = 3$. In this case, $a = 1.8$. Hence, the version with the smaller value of a is preferable. The quantity a can thereby also be interpreted as the expected number of loops through Steps 1 and 2 until Step 3 leads to an acceptance. The number of steps required follows a geometric distribution with parameter $p = 1 \,/\, a$. Hence, the smaller a, the smaller is the proportion of rejected simulated values, and hence the more efficient is the simulation scheme.

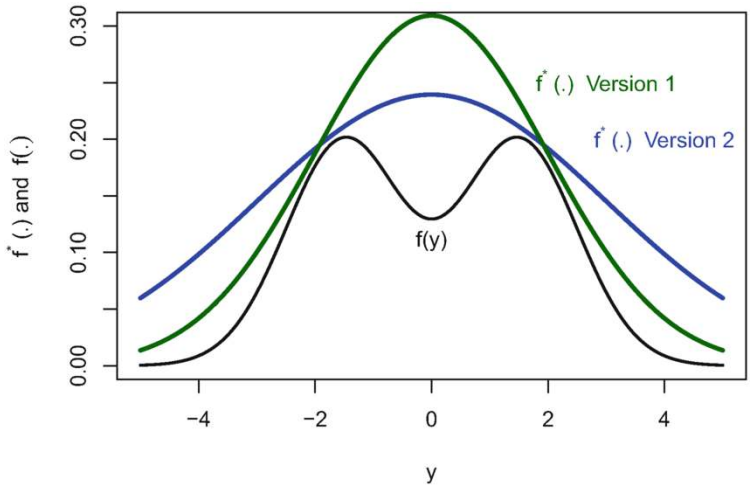


Fig. 8.5 Sketch for rejection sampling with two different proposal distributions

An alternative to rejection sampling is **importance sampling**. In this case, we are interested in calculating an expected value $E\big(h(Y)\big)$ for some known function $h(\cdot)$ and random variable Y, which is distributed according to $F(\cdot)$. Instead of calculating the integral analytically, we aim to use simulations.

If we could simulate from $F(\cdot)$ directly, it is easy to see that for i.i.d. samples $Y_i^*, i = 1, \ldots, N$, we get

$$\tfrac{1}{N} \sum_{i=1}^{N} h(Y_i^*) \;\; \to E_F\Big(h(Y)\Big) = \int_{-\infty}^{\infty} h(y) f(y) \mathrm{d}\, y$$

for N increasing. However, while we can evaluate the density $f(\cdot)$, we cannot simulate from it, since $F(\cdot)$ is not available. Like in rejection sampling, we choose a proposal distribution $F^*(.)$ from which we can easily simulate. And, like above, we assume that the corresponding density $f^*(\cdot)$ is available and can be evaluated numerically. The idea is now to rewrite the expected value as follows:

$$\begin{aligned} E_F\Big(h(Y)\Big) &= \int_{-\infty}^{\infty} h(y) f(y) \mathrm{d}\ y \\ &= \int_{-\infty}^{\infty} h(y) \tfrac{f(y)}{f^*(y)} f^*(y) \mathrm{d}\ y \\ &= E_{F^*}\Big(h(Y) \tfrac{f(Y)}{f^*(Y)}\Big), \end{aligned}$$

where the index at the expectation refers to the distribution the expectation is taken over. Using now $i.i.d.$ samples $Y_i^*, i = 1, \ldots, N$ drawn from $F^*(\cdot)$, we obtain

$$\tfrac{1}{N} \sum\nolimits_{i=1}^{N} h(y_i^*) \tfrac{f(y_i^*)}{f^*(y_i^*)} \to E\Big(h(Y)\Big).$$

The accuracy of the approximation again depends on the variance, and it is easy to see that the closer $F(\cdot)$ and $F^*(\cdot)$, the smaller the variance. Hence, both rejection and importance sampling possess small accuracy if the distribution $F^*(.)$ from which we sample is not a good proxy for the distribution $F(\cdot)$ we want to sample from.

8.3 Simulating Proportional to a Density (MCMC)

The simulation schemes above work if we know the distribution function $F(\cdot)$ or at least the density $f(\cdot)$. We extend this now to simulation settings where the density is only known up to an unknown proportionality factor. Hence, instead of the true density $f(y)$, one only knows $\widetilde{f}(y)$, which is proportional to the density, i.e., $\widetilde{f}(y) \propto f(y)$. Put differently, we have

$$f(y) = \tfrac{1}{a} \widetilde{f}(y) \quad \Leftrightarrow \quad \widetilde{f}(y) = a f(y) \tag{8.12}$$

where the constant *a*, however, is unknown. The procedure proposed below results from physics goes back to Metropolis et al. (1953). The idea was extended in Hastings (1970) and brought into the statistical domain, nowadays known as the Metropolis-Hastings approach. . The procedure is heavily used in Bayesian statistics, as further discussed in Chap. 9, and we will make full use of it there. We present the procedure here in its general form as a simulation technique.

The underlying idea is to construct a Markov chain $Y^{*(1)}, Y^{*(2)}, Y^{*(3)}, \ldots, Y^{*(T)}$ as sketched in Fig. 8.6. The figure suggests to simulate the random variables $Y^{*(t)}$ sequentially, where the simulation of $Y^{*(t)}$ can depend on the previous simulation $Y^{*(t-1)}$, but is independent of previous simulations $Y^{*(t-1)}, Y^{*(t-2)}, \ldots$. Hence, the Markov chain is an iterative simulation process.

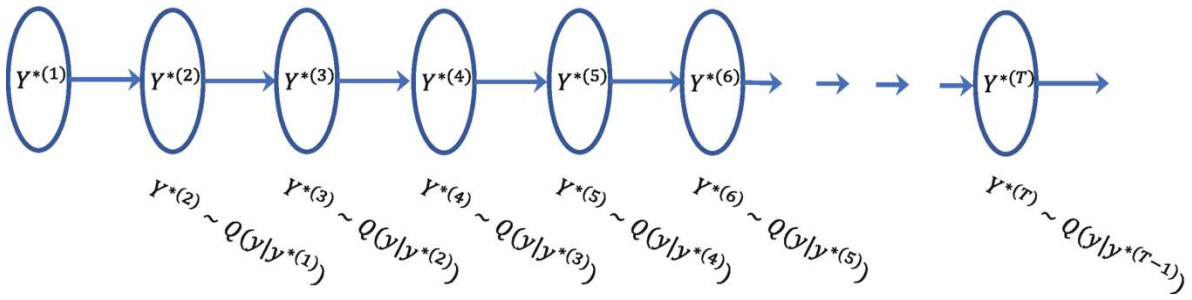


Fig. 8.6 Markov chain with transition probabilities $Q(y^{*(t)} \mid y^{*(t-1)})$

At the *t*-step, a new value $Y^{*(t+1)}$ is simulated from a so-called transition distribution $Q(y \mid Y^{*(t)})$. The transition distribution is allowed to depend on the current value $Y^{*(t)}$, but it is independent of all previous values. Hence, given the current value of the Markov chain, we draw the next value from $Y^{*(t+1)} \sim Q(y \mid Y^{*(t)})$. The transition probability leads (under certain technical assumptions) to a stationary distribution of the Markov chain. The stationary distribution is thereby defined as the limit as the Markov chain proceeds, that is, the distribution of $Y^{*(t)}$ as $t \to \infty$, assuming that this distribution exists. The aim is now to choose the transition probability $Q(y \mid Y^{*(t)})$ such that $Y^{(t)} \sim F(\cdot)$ as $t \to \infty$. To achieve this limit, we need (a) a transition probability from which we can easily simulate, and (b) the stationary distribution should be equal to $F(y)$.

Returning to our simulation problem, while we cannot simulate from $F(y)$ directly, since we only know the density up to a proportionality factor, we aim to construct a transition distribution $Q(y \mid Y^{*(t)})$ such that the Markov chain, if we let it run long enough, converges to the desired distribution $F(y)$. This is achieved with so-called **Markov Chain Monte Carlo**, or in short **MCMC** methods. To formally prove that the procedure works requires several technical details that lie beyond the scope of the material in this book. We, therefore, present the procedure and refer to the classical literature on Markov Chains, like Grimmett and Stirzaker (2001) or a more technical discussion on requirements and properties as provided by Robert and Casella (2004).

The MCMC procedure presented here is called Metropolis-Hastings algorithm. The goal is to derive a transition probability $Q(y \mid Y^{(t)})$ such that the Markov chain does have a stationary distribution that is in fact $F(y)$, where $F(y)$ is the distribution function to density $f(y)$, which, as said above, is only known up to a proportional function $\widetilde{f}(y)$. The procedure is thereby as follows:

1. Select a starting value $Y^{*(0)}$, and set $Y^{*(t)} = Y^{*(0)}$ in the following.
2. Based on $Y^{*(t)}$, propose a new value Y^{*} from a proposal distribution

$$Y^* \sim G(. \mid Y^{*(t)}) \tag{8.13}$$

where $G(\cdot \mid Y^{*(t)})$ is a known distribution from which we can easily simulate. Note that $G(\cdot \mid Y^{*(t)})$ can depend on $Y^{*(t)}$, and let $g(y \mid Y^{*(t)})$ denote the corresponding density.

3. Define the acceptance probability
$$\alpha(Y^*, Y^{*(t)}) = \min\left(1, \frac{\widetilde{f}(Y^*)}{\widetilde{f}(Y^{*(t)})} \frac{g(Y^{*(t)}|Y^*)}{g(Y^*|Y^{*(t)})}\right) \tag{8.14}$$

4. Draw $U^* \sim \text{Uniform}[0, 1]$.

5. If $U^* > \alpha(Y^*, Y^{*(t)})$, reject the proposed value Y^*, and go back to Step 2. Otherwise accept the new value and set $Y^{*(t+1)} = Y^*$ and proceed to Step 6.

6. Go back to Step 2 until t reaches the prespecified maximum number of iterations.

We will not prove that the resulting stationary distribution is in fact equal to $F(\cdot)$ but refer to the literature cited above. We want, however, to demonstrate that the procedure is easily applicable. Note first that we do not know the normalization constant a in (8.12). The normalization constant, however, cancels out when calculating the acceptance probability in (8.14). Hence, even though the density $f(\cdot)$ is only known up to a multiplicative constant, this constant cancels out and is not needed in the algorithm. Secondly, note that if the proposal distribution is symmetric, then the second ratio in (8.14) is equal to 1 and can be ignored. Finally, note that the procedure shows similarities to rejection sampling, and the Markov chain can remain on the same value without changes if the acceptance probability is small. We also want to note that there are many variants of the algorithm, and we refer to Andrieu and Thoms (2008).

Let us now demonstrate the performance of the algorithm with a simple example. In Fig. 8.7, we show $f^*(y)$, which is a bimodal function proportional to a density. We aim to draw random variables from this distribution. To do so, we set the starting value to $Y^{*(0)} = -5$ and work with a normal distribution as a proposal, which is centered at $\mu = Y^{*(t)}$. Hence we simulate $Y^* \sim \text{N}(Y^{*(t)}, \sigma^2)$. This is a symmetric distribution so the ratio $g(Y^{*(t)} \mid Y^*) \,/\, g(Y^* \mid Y^{*(t)}) \equiv 1$ in (8.14). For the variance of the proposal, we choose two options. First, we take a small variance $\sigma^2 = 0.2^2$ (setting 1), and second, we take $\sigma^2 = 3^2$ (setting 2). The two settings are visualized in Fig. 8.8 in the top row, where the left-hand side refers to setting 1. We run the algorithm up to $T = 10.000$ and plot the resulting simulated values in the second row in Fig. 8.8. Apparently, in setting 1, the Markov

chain remains in the left mode and does not jump over to the right side mode. This is in contrast to setting 2, where the larger variance allows for larger steps. We plot the empirical distribution functions of the simulated values $Y^{*(t)}$ in the third row, left-hand side. Apparently, the run with $\sigma = 0.2$ was not successful and ignored large parts of the distribution. This can be avoided by making larger jumps, as in setting 2. Here, both modes are visited by the Markov chain, and in fact, the empirical distribution is indistinguishable from the true distribution. Finally, on the right-hand side in the third row, we show the values of the acceptance probabilities (8.14). Setting 1 has a median of the acceptance probability, which is almost 1, and hence more than 50% of the proposed values have been accepted. Contrary, for setting 2 (right-hand side boxplot), this is in the order of 20%. We conclude that the procedure works best if the acceptance probability is not too high, with a rough rule of thumb stating it should be in the order of 20 to 30%, see also Neal and Roberts (2006).

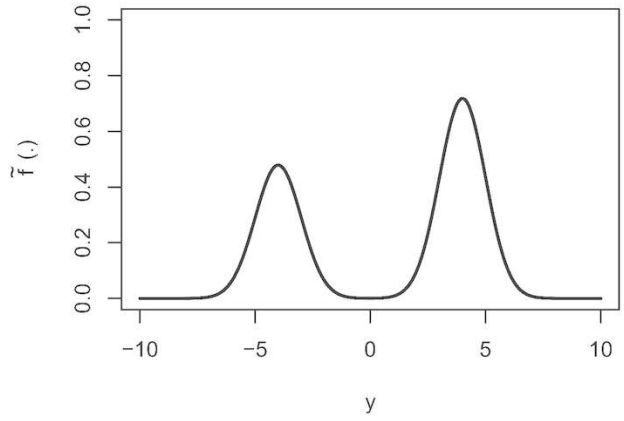


Fig. 8.7 Function $f^*(y)$ which is proportional to a density

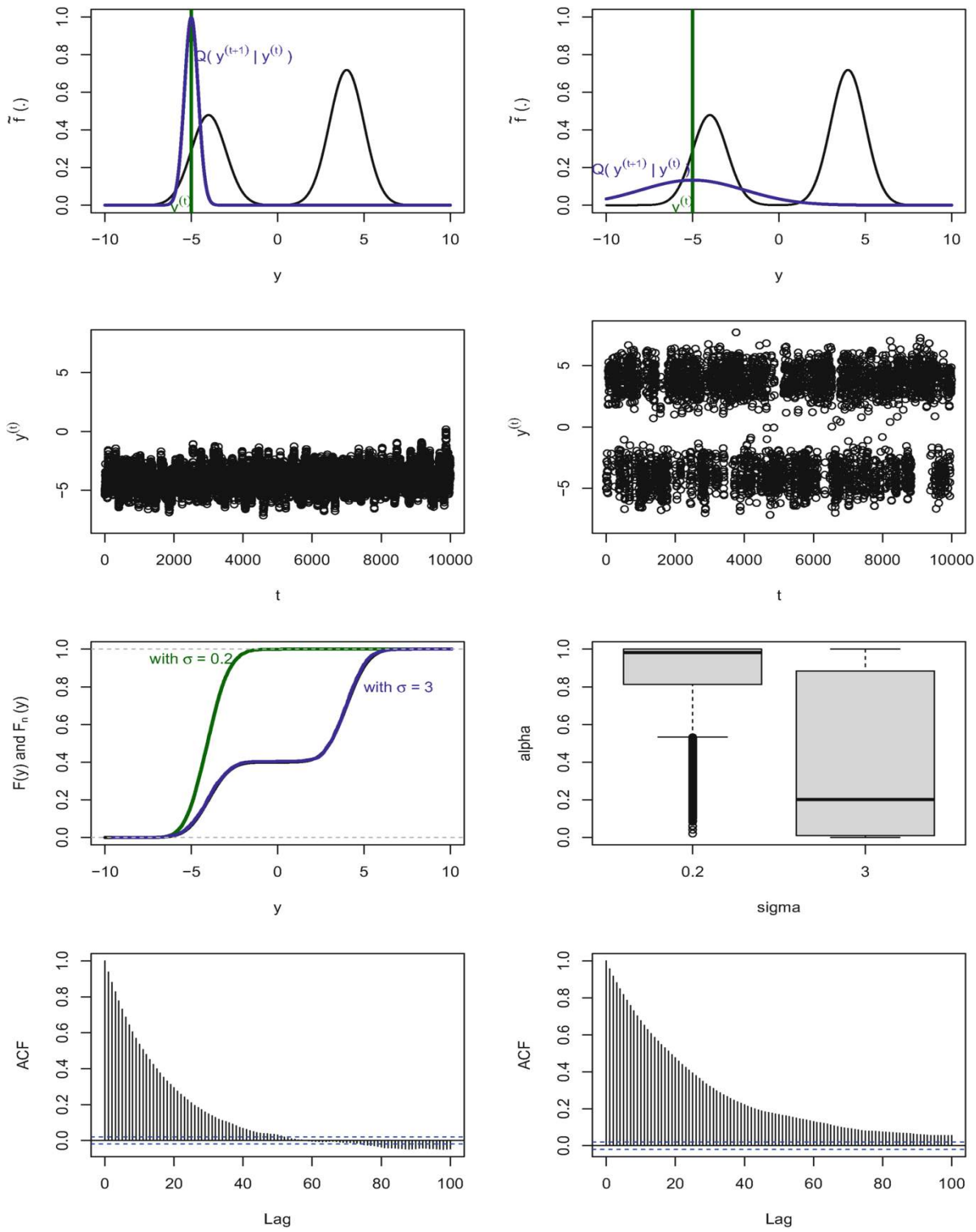


Fig. 8.8 Top row: Function $f^*(y)$ and two normal proposal distributions. Left-hand side with $\sigma = 0.2$, right-hand side with $\sigma = 3$. Second row: Simulated values $y^{(t)}$ for the two settings (left for $\sigma = 0.2$, right for $\sigma = 3$). Third row: Empirical distribution function of simulated values (left) and acceptance probabilities (8.14) for the two settings. Markov chain of first (top) and second (bottom) order. Bottom row: Autocorrelation $\mathrm{Cor}(Y^{(t)}, Y^{(t-d)})$ for different lags d

It should be clear that although we have simulated T values, these are not independent. In fact, due to the construction of the Markov chain, we have a strong (serial) correlation. In Fig. 8.8, we include the resulting autocorrelation functions for the two settings. The autocorrelation is thereby defined as $\text{Cor}(Y^{*(t)}, Y^{*(t-d)})$ where d is denoted as lag. These autocorrelations can easily be estimated (see Sect. 12.1) from the simulated data, and we see a longer-lasting correlation for setting $\sigma = 3$ on the right-hand side. To obtain independent values $Y^{*(t)}$ one can apply thinning, that is, one only takes every d-th simulated value, which leads to the (quasi) $i.i.d.$ sample Y_i^* where $Y_i^* = Y^{*(d_0+i\cdot d)}$, where d can be set to 100, for instance, and d_0 is a larger constant which guarantees that the Markov chain has achieved its steady state, meaning that the resulting values are drawn from the stationary distribution $F(y)$. We refer to Green (2001). This point is quite controversially discussed in the literature as it retains fewer samples; see, e.g., Riabiz et al. (2022).

We will return to MCMC samplers in the next chapter, as this is the working machinery for Bayesian statistics. It is, however, important to note that MCMC sampling, while regularly used in Bayesian statistics, is "just" a simulation procedure in situations where the density is only known up to an unknown constant.

8.4 Simulating from a Multivariate Normal Distribution

The above simulation sections dealt with univariate random variables. But how can we simulate multivariate random variables with a prespecified dependence structure? We want to give some general ideas here, and it should be clear that multivariate dependencies can be complex and so are their simulation processes.

Most central is certainly the multivariate normal distribution, which we want to treat first. The aim is to generate simulations from

$$Y \sim \text{N}(\mu, \Sigma)$$

with $Y = (Y_1, \ldots, Y_p)^\top$ and Σ as prespecified $p \times p$ dimensional variance matrix. This task is easily accomplished by utilizing some matrix algebra. We can simulate p independent $N(0,1)$ random variables and stack them to the vector $Z^* = (Z_1^*, \ldots, Z_p^*)^\top$. Next, we decompose Σ through singular value decomposition to

$$\Sigma = L \Lambda L^\top, \tag{8.15}$$

where $L \in R^{p \times p}$ are (orthonormal) eigenvectors of Σ and $\Lambda = diag(\lambda_1, \ldots, \lambda_p)$ is the diagonal matrix of the eigenvalues. Note that in order for Σ to be positive

definite, as required for variance matrices, we have $\lambda_j > 0$ for all $j = 1, \ldots, p$. With this property and (8.15), we can define

$$\Sigma^{\frac{1}{2}} = L\Lambda^{\frac{1}{2}}L^\top, \tag{8.16}$$

where $\Lambda^{1/2} = diag(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_p})$. If we now define

$$Y^* = \Sigma^{\frac{1}{2}}Z^* + \mu, \tag{8.17}$$

we obtain a simulation from the desired distribution, i.e., $Y^* \sim \mathrm{N}(\mu, \Sigma)$. This is easy to see, since $Z^* \sim \mathrm{N}(0, I_p)$ with I_p as p-dimensional identity matrix, so that

$$\begin{aligned}
\mathrm{Var}(Y^*) &= \Sigma^{\frac{1}{2}} I \Sigma^{\frac{1}{2}} \\
&= L\Lambda^{\frac{1}{2}} \underbrace{L^\top L}_{=I_p} \Lambda^{\frac{1}{2}} L^\top \\
&= L\Lambda^{\frac{1}{2}}\Lambda^{\frac{1}{2}}L^\top \\
&= L\Lambda L^\top \\
&= \Sigma.
\end{aligned}$$

A different approach to simulate multivariate random variables is to factorize the distribution. This utilizes standard results from multivariate normal distributions. Assume that a normally distributed random vector Y is decomposed into two subvectors Y_A and Y_B, so that $Y = (Y_A^\top, Y_B^\top)^\top$. With

$$Y = \begin{pmatrix} Y_A \\ Y_B \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \mu_A \\ \mu_B \end{pmatrix}, \begin{pmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{pmatrix} \right) \tag{8.18}$$

we denote the multivariate distribution, with mean subvectors denoted with μ and covariance submatrices with Σ. Then the following two properties hold

$$Y_A \sim \mathrm{N}(\mu_A, \Sigma_{AA}) \tag{8.19}$$

$$Y_B \mid (Y_A = y_A) \sim \mathrm{N}(\mu_B + \Sigma_{BA}\Sigma_{AA}^{-1}(y_A - \mu_A), \Sigma_{BB} - \Sigma_{BA}\Sigma_{AA}^{-1}\Sigma_{AB}). \tag{8.20}$$

With these results, we can now easily derive the separate components for sequential simulations. For instance, for $p = 2$, we first simulate

$$Y_1^* \sim \mathrm{N}(\mu_1, \Sigma_{11})$$

and with simulated value Y_1^* use (8.20) and draw Y_2^* from

$$Y_2^* \mid Y_1^* \sim \mathrm{N}(\mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(Y_1^* - \mu_1), \Sigma_{22} - \Sigma_{21}^2 \,/\, \Sigma_{11}),$$

which, taken together, provides a random variable drawn from

$$Y^* = \begin{pmatrix} Y_1^* \\ Y_2^* \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right). \tag{8.21}$$

The procedure can be extended beyond the bivariate setting, i.e., to $p > 2$. Note that we can rewrite any multivariate distribution through

$$
\begin{aligned}
f_Y(y_1, \ldots, y_p) = \; & f_{Y_1}(y_1) \\
& \times f_{Y_2|Y_1}(y_2 \mid y_1) \\
& \times f_{Y_3|Y_1,Y_2}(y_3 \mid y_1, y_2) \\
& \cdots \times f_{Y_p|Y_1,\ldots,Y_{p-1}}(y_p \mid y_1, \ldots y_{p-1}).
\end{aligned} \tag{8.22}
$$

Hence, we can decompose the simulation of a p-dimensional multivariate normal distribution into p one-dimensional simulations. Indeed, the idea of decomposing multivariate simulations into a sequence of univariate simulations can be used for different distributions as well, which will be demonstrated in the next section.

8.5 Gibbs Sampling and Markov Fields

The previous section was constrained to multivariate normal distributions. We want to extend the view now toward simulations from general multivariate distributions. Assume the multivariate random vector $Y = (Y_1, \ldots, Y_p)^\top$ which we want to simulate. We assume that

$$
Y^* \sim F(\cdot) = F_{Y_1,\ldots,Y_p}(\cdot). \tag{8.23}
$$

The index notation used (8.23) refers to the set of random variables covered by the distribution function. The notation will be helpful, as seen in the following. The multivariate distribution function in (8.23) might not be known or might just be too complex or complicated. But we assume the univariate conditional distributions to be known for each component of Y, given the remaining variables. That is, we assume that we know

$$
F_{Y_j|Y_{-j}}(\cdot \mid Y_{-j}) \tag{8.24}
$$

where subscript "$-j$" refers to all components of the vector except the j-th one, that is, $Y_{-j} = (Y_1, \ldots, Y_{j-1}, Y_{j+1}, \ldots, Y_p)$. We want to iteratively simulate Y_j^* from (8.24), for $j = 1, \ldots, p$, such that the simulated entire vector $Y^* = (Y_1^*, \ldots, Y_p^*)$ comes from the distribution target multivariate distribution $F(\cdot) = F_{Y_1,\ldots,Y_p}(.)$

The idea is to construct a Markov chain of simulations $Y^{*(t)}$, comparable to the MCMC sampling scheme proposed above. In every single step in the chain, we simulate only a single component of Y while conditioning on the others. This means we simulate from the conditional distribution

$$
Y_j^* \mid Y_{-j}^* \sim F_{Y_j|Y_{-j}}(\cdot \mid Y_{-j}^*)
$$

and iterate over the indices $j = 1, \ldots p$. The simulation procedure is called **Gibbs sampling**, and the idea was proposed in Geman and Geman (1984); see also Gelfand (2000) or Suess and Trumbo (2010).

The algorithm proceeds as follows:

1. Take any starting value $Y^{*(0)}$, and set $Y^{*(t)} = Y^{*(0)}$ subsequently.

2. Set $Y^* = Y^{*(t)}$.

3. For $j = 1, \ldots, p$, simulate (and hence overwrite) the j-th component of Y^* through
$$Y_j^* \sim F_{Y_j|Y_{-j}}(\cdot \mid Y_{-j}^*). \tag{8.25}$$

4. Set $Y^{*(t+1)} = Y^*$ and proceed with step 2 until a sufficient number of iterations have been carried out.

The algorithm is sketched in Fig. 8.9 for two variables. A formal proof that the Markov Chain converges to the desired target distribution $F(\cdot)$ is not given here. We refer to the literature cited above. We demonstrate the performance with a simple example.

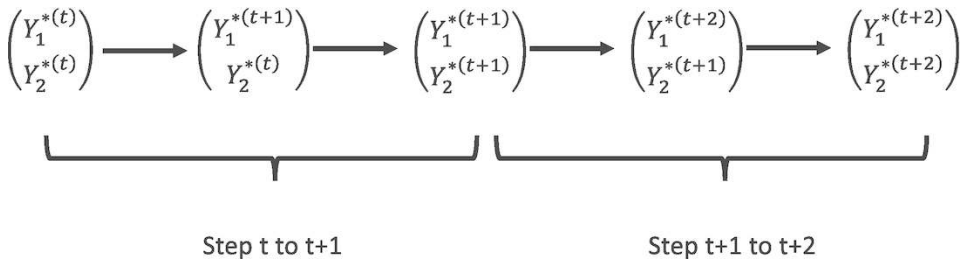


Fig. 8.9 Sketch of Gibbs sampling

Example 8.2 We consider a bivariate normally distributed random variable with conditional distributions
$$Y_1 \mid Y_2 \sim \mathrm{N}(0.75Y_2, 1) \text{ and } Y_2 \mid Y_1 \sim \mathrm{N}(0.75Y_1, 1). \tag{8.26}$$
We use these two conditional distributions to start a Markov chain as described in the Gibbs sampling algorithm above. We take two different starting values, namely, $Y^{*(0)} = (0, 0)$ and $Y^{*(0)} = (-50, -50)$. The results of the first 1000 steps are shown in Fig. 8.10. We see speedy convergence of the algorithm, even if the starting point is far away from the center of the distribution.

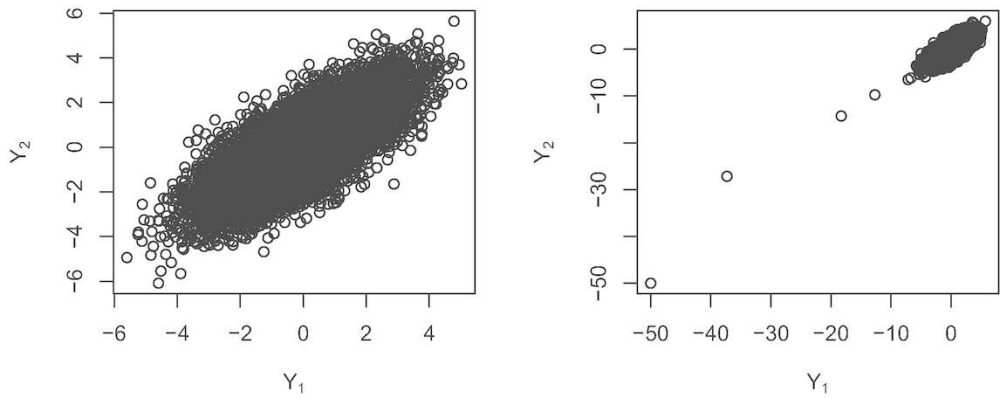


Fig. 8.10 Example of Gibbs sampling with two different starting points

The simulated data in this example come from a joint normal distribution, and in this case, we can even calculate the parameters exactly. We write

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix} \right). \tag{8.27}$$

Using (8.19) and (8.20) and the simulation setup (8.26) leads to

$$\sigma_{12} \,/\, \sigma_{22} = 0.75 \text{ and } \sigma_{11} - \sigma_{12}^2 \,/\, \sigma_{22} = 1.$$

Due to the symmetry of the simulation setup (8.26), we have $\sigma_{11} = \sigma_{22} =: \sigma^2$, and apparently, we also generally have $\sigma_{12} = \sigma_{21}$. This allows us to solve the two equations above, and we obtain $\sigma^2 = 1 \,/\, (1 - 0.75^2) \approx 2.3$ and $\sigma_{12} = 0.75 * \sigma^2 \approx 1.7$. These numbers correspond to the empirical numbers calculated from the simulations.▷

The Gibbs sampling approach also allows for the simulation of more complex distributions, but one should keep in mind that it is not generally guaranteed that each set of conditional distributions leads to a stable joint distribution. We here consider the case where a joint distribution exists, but the joint distribution is not known; only the univariate marginal distributions are available.

Example 8.3 We modify the previous example by replacing the simulation steps with

$$Y_1 \mid Y_2 \sim \mathrm{N}(0.75 Y_2, 1) \text{ and } Y_2 \mid Y_1 \sim \mathrm{N}(0.75 Y_1, \exp\,(Y_1) \,/\, (1 + \exp\,(Y_1)). \tag{8.28}$$

Hence, we induce variance heteroscedasticity for the second component. The resulting simulated data (with starting point $(0, 0)$) are shown in Fig. 8.11. In this case, the joint distribution has no simple structure. ▷

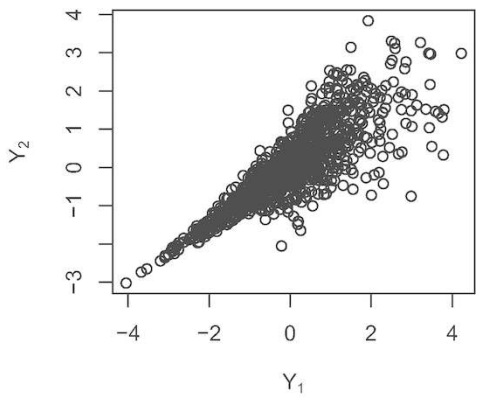


Fig. 8.11 Example of Gibbs sampling more complex conditional distributions

The previous example, along with the simulation steps in Eq. (8.25), highlights a key feature of the Gibbs sampling algorithm: We do not need the full joint distribution to perform simulations. Instead, it is sufficient to know the conditional distributions for each individual component of the random vector. In other words, if we have access to the conditional distribution of each variable Y_j given all other variables Y_{-j}, we can still use Gibbs sampling to generate samples from the full (unconditional) distribution.

In fact, it can be shown that by iteratively sampling from these conditionals, as described in Eq. (8.24), the resulting sequence converges to samples from the joint distribution in Eq. (8.23), provided the underlying Markov chain has a stationary distribution. A sketch to prove this result is provided at the end of this section.

The Gibbs sampling idea can easily be combined with the different sampling strategies that we discussed in Sect. 8.1. Note that we assumed in (8.25) that we know the conditional distribution functions for Y_j given Y_{-j}. If we instead only know the condition density $f_{Y_j|Y_{-j}}(\cdot \mid y_{-j})$, we can replace the simulation step (8.25) in the algorithm above by rejection sampling. If we know the conditional densities $f_{Y_j|Y_{-j}}(\cdot \mid y_{-j})$ only up to an unknown multiplicative constant, we can use Metropolis-Hastings simulations in the steps of the Gibbs sampler. The combination of Gibbs sampling and Metropolis-Hastings is frequently used in Bayesian computation, and we will return to this in Chap. 9. We motivate the idea with an example below.

Example 8.4 We consider the case where the density $f(y)$ is of peculiar shape, as shown in Fig. 8.12 in the top left plot. We apply a Gibbs-sampling Metropolis-Hastings-based MCMC approach, where in each step of the Gibbs-sampling, we use a Metropolis-Hastings step for simulation. We start with $Y_1^{*(0)} = -1.1$ and $Y_2^{*(0)} = 2$, which is indicated as dashed lines in the top left plot. The conditional

densities for Y_1^* given $Y_2^* = 2$ and Y_2^* given $Y_1^* = -1.1$ are shown in the next two plots. We also include our proposal densities as dashed blue lines, which are normal distributions, centered at the current value and with variances equal to $\sigma^2 = 0.1$. We use these to start the Markov chain, where in each step a Metropolis-Hastings algorithm is used. The resulting MCMC samples are shown in the bottom right plot as a heatmap. We see that the simulated values Y^* uncover the "banana" shape of the distribution quite well.▷

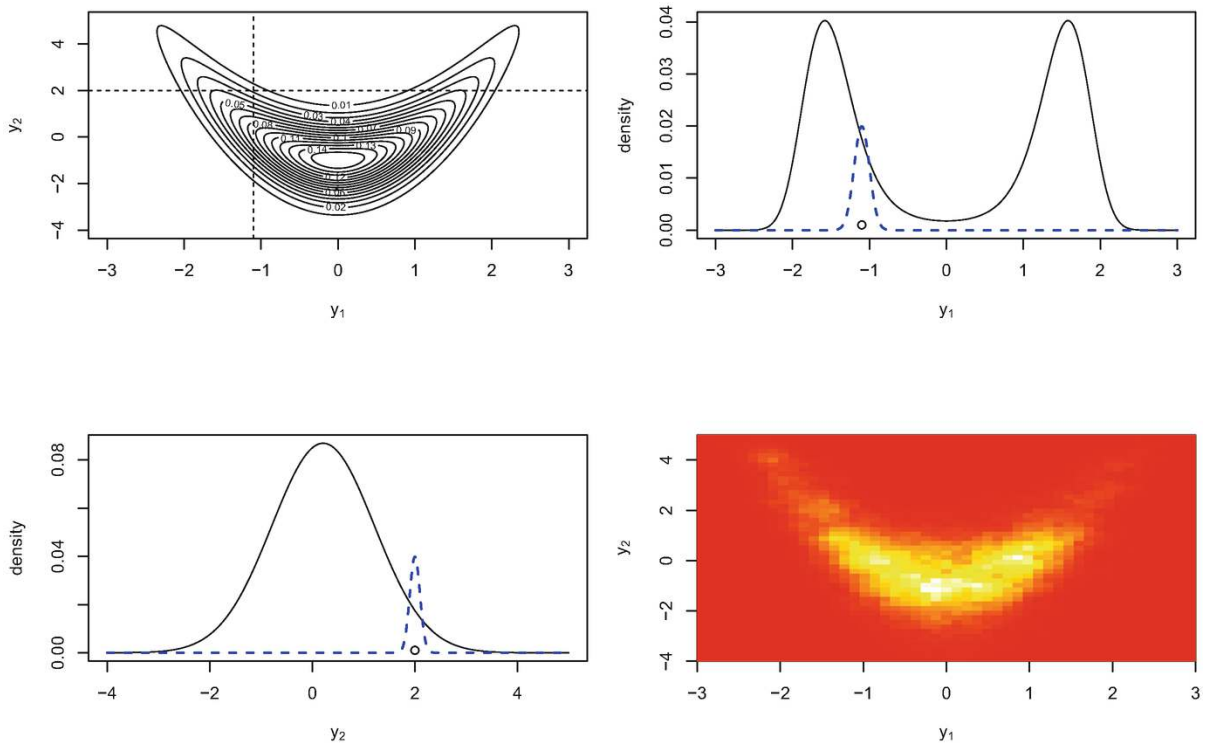


Fig. 8.12 Example of Gibbs sampling using MCMC Metropolis-Hastings. The density is shown in the top left plot. The starting point $(Y_1^*, Y_2^*) = (-1.1, 2)$ is indicated with vertical and horizontal dashed lines. The top right plot shows the conditional density of Y_1^* given $Y_2^* = 2$, and the bottom right shows the same for Y_2^* given $Y_1^* = -1.1$. This is the density cut along the dashed lines in the top left plot. The blue dashed lines in the two plots represent the proposal densities. The results of a Gibbs-sampling MCMC run of length 10.000 are shown in the bottom right plot.

If p is large, it can be quite burdensome to construct p conditional distributions. In this situation, it is advisable to rely on more locally structured simulation techniques, which require conditional independence assumptions. For instance, for $Y = (Y_1, \ldots, Y_p)$, we could assume some neighborhood structure in that Y_j is conditionally independent of Y_k for all indices $|k - j| > d$ given the values Y_l for $0 < |l - k| \leq d$. The formula is less easy to understand than a visualization in the form of a simple graph, which expresses the (conditional) independence structure. This is shown in Fig. 8.13. In the top plot, we show a Markov structure of first

order. This means that Y_j is dependent on the neighbors Y_{j-1} and Y_{j+1}, but given these values, it is independent of all other values. The setup is typically used for classical time series analysis (see also Sect. 12.1), and it allows for the factorization

$$f_{Y_1,\ldots,Y_p}(y_1,\ldots,y_p) = f_{Y_1}(y_1)\prod_{j=2}^{p} f_{Y_j|Y_{j-1}}(y_j \mid y_{j-1}).$$

The bottom plot in Fig. 8.13 extends the setting toward a Markov chain of second order. In this case, one can use the factorization

$$f_{Y_1,\ldots,Y_p}(y_1,\ldots,y_p) = f_{Y_1,Y_2}(y_1,y_2)\prod_{j=3}^{p} f_{Y_j|Y_{j-1},Y_{j-2}}(y_j \mid y_{j-1},y_{j-2}).$$

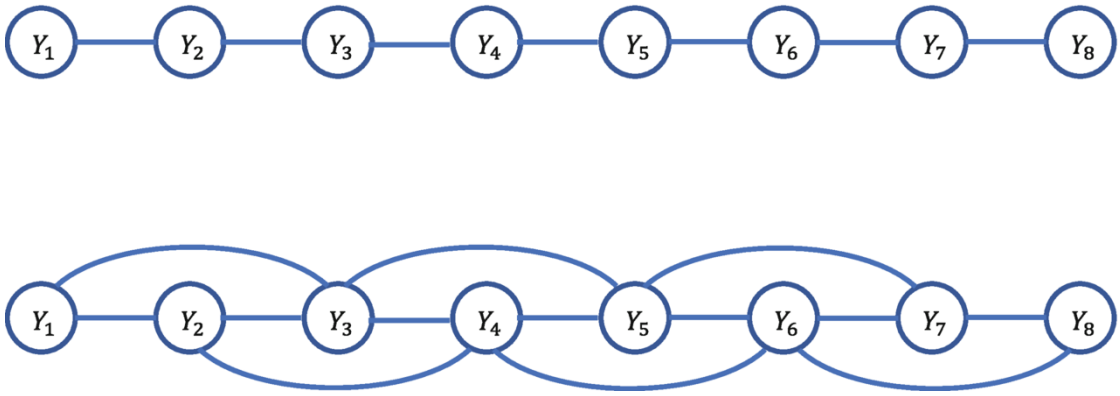


Fig. 8.13 Markov chain of first (top) and second (bottom) order

The Markov structure can be extended further, which formally refers to the Hammersley-Clifford theorem ; see Grimmett (1973). Let therefore Y be indexed by some index set $\mathcal{J}$, and for each $j \in \mathcal{J}$, let $\mathcal{N}_j$ be the "neighbours" of *j*. We assume that Y_j is directly dependent on Y_l with $l \in \mathcal{N}_j$ but conditionally independent of the remaining components given the neighbors. We give an example in Fig. 8.13. Such a dependence structure will be defined more rigorously in Sect. 11.2, where we denote this generally as Graphical Models. For now, we treat the structure only for our simulation setup. To do so, we simplify the simulation step (8.25) by just conditioning on the "neighbors," that is,

$$Y_j^* \mid Y_{-j} \sim F_{Y_j|Y_{\mathcal{N}_j}}(\cdot \mid y_{\mathcal{N}_j}).$$

The framework is often used in so-called Markov fields, which is sketched in Fig. 8.14, top plot. The setup is very suitable if the data carry some "spatial" structure. We can now write Y in matrix form
$Y = (Y_{ij}, i = 1,\ldots,N_1, j = 1,\ldots,N_2)$ and induce a neighbor structure on Y. For instance, we may assume a first-order Markov structure, which means that Y_{ij} is independent of all Y_{lk} with $|i-l| > 1$ or $|k-l| > 1$ given the direct neighbors,

as visualized in Fig. 8.14 in the bottom plot. Sampling can now be carried out with Gibbs sampling by replacing the simulation step in the algorithm above through

$$Y_j^* \mid Y_{-j}^* \sim F_{Y_j|Y_{\mathcal{N}_j}}(\cdot \mid Y_{\mathcal{N}_j}^*) \tag{8.29}$$

Iterating over all indexed variables concludes one step in the Gibbs sampler.

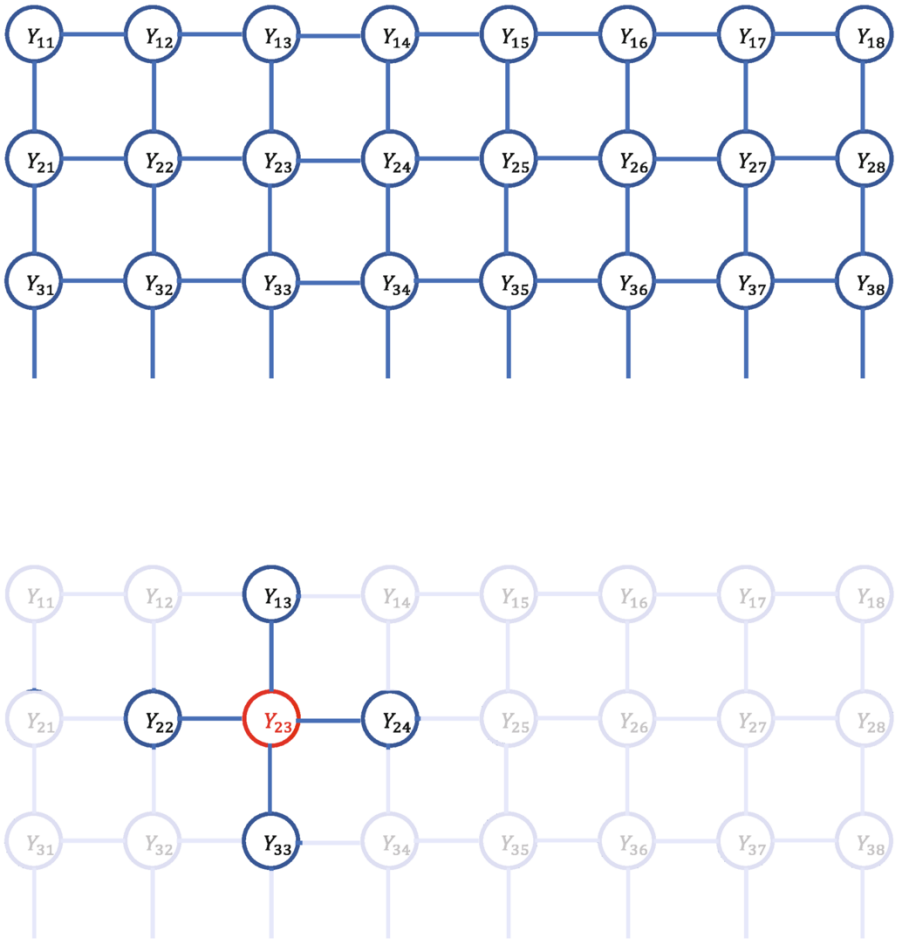


Fig. 8.14 Markov field (top) and dependence structure when conditioning on the neighbors

We want to demonstrate the idea with a small example. Assume that we are interested in finding the maximum of a bivariate function. Let $h(y_1, y_2)$ be the function, which we can evaluate at given values (y_1, y_2). The evaluation of the function is considered to be time and/or computer-intensive, so we aim to evaluate it at a small number of points only. As an example, we consider the function shown in Fig. 8.15 in the top left plot. We evaluate the function at 25 points, sketched together with the isolines of the function in Fig. 8.15 in the top right-hand side plot. We take these 25 points and construct a Markov field around them with 100×100 cells, which we call pixels subsequently. The pixels are

connected as shown in Fig. 8.14; that is, a pixel depends on its neighbors. We assume that Y_{ij}, given the neighbors follows a normal distribution

$$Y^*_{ij} \mid Y^*_{\mathcal{N}_{ij}} \sim \mathrm{N}(Y_{\mathcal{N}_{ij}}, \sigma^2_t), \tag{8.30}$$

where $Y_{\mathcal{N}_{ij}}$ is the arithmetic mean of the neighbor values. This setting is called **Gaussian Markov Random Field (GMRF)**. For the 25 pixels evaluated from the function, we do not simulate the values but keep the evaluated value from the function. For all other pixels, we simulate from (8.30). To start, we just simulate from a normal distribution $Y^{*(0)}_{ij} \sim \mathrm{N}(0.5, 0.5^2)$, where again, the 25 pixels from the function remain unchanged. This gives an image as shown in the middle left plot of Fig. 8.15. Though this looks like random noise, bear in mind that 25 pixels are, in fact, fixed. We then start the Gibbs sampling using (8.30), except for the 25 evaluated points. We run the algorithm for 600 loops, where we set the variance in (8.30) to shrink in t with order $\sigma_t = 1/\log(t)$. We plot the simulated values $Y^{*(600)}_{ij}$ in the right-hand side plot in the middle row. Apparently, the functional shape looks comparable to the original function. One should, however, bear in mind that the function has been evaluated just at 25 pixels; i.e., only for these 25 pixels, we know the exact value of the function.

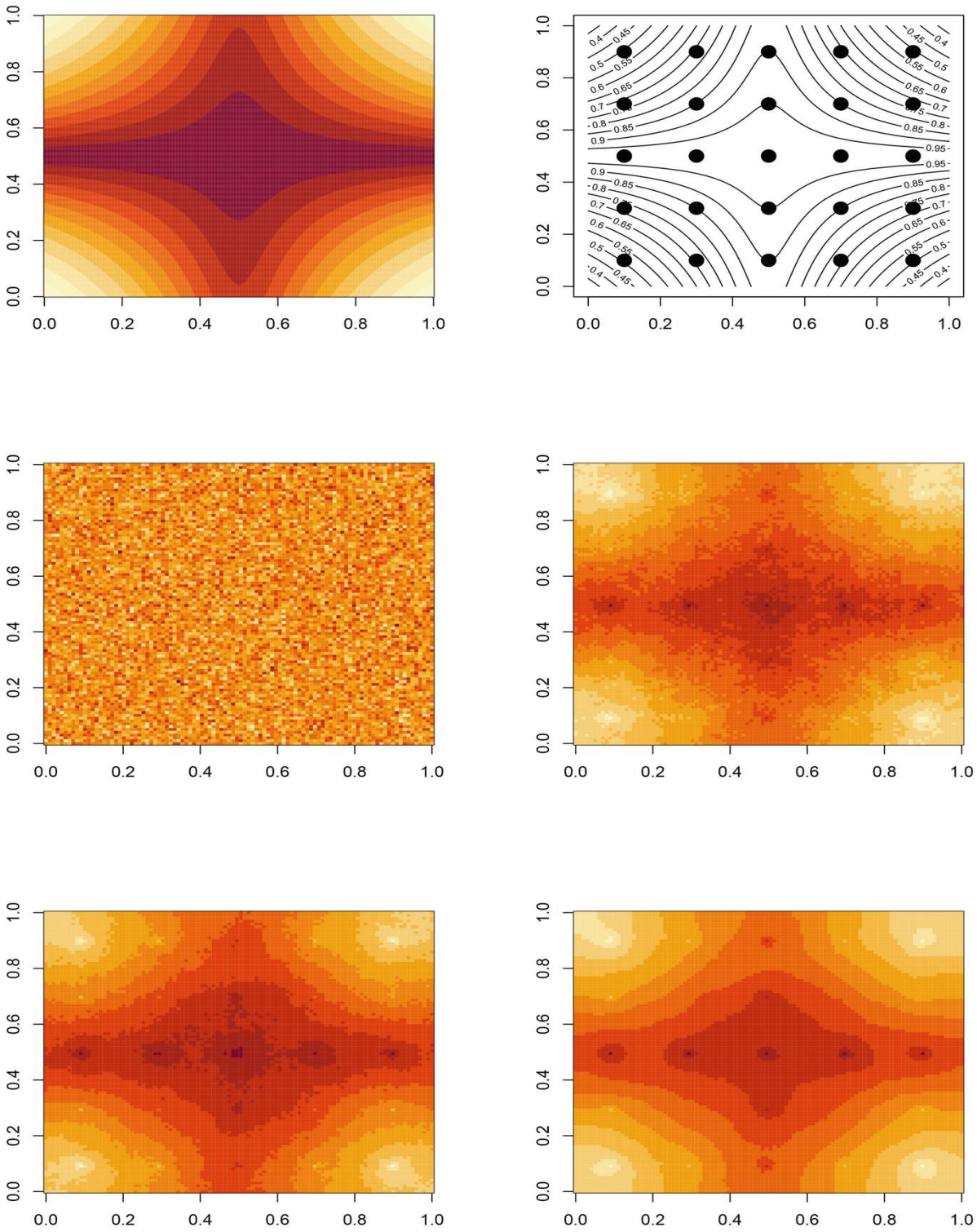


Fig. 8.15 Top row: Function $h(y_1, y_2)$ and the 25 data points at which the function is evaluated. Middle row: Simulated Gaussian Markov Random Field. Starting value to the left, after 1000 iterations to the right. Bottom row: maximum (left) and mean (right) of the last 100 simulated values

We can now also analyze the simulated values $Y_{ij}^{*(t)}$ of the Markov field. We could, for instance, calculate the maximum or the mean of the simulations. This is shown in the bottom row plots in Fig. 8.15. We thereby only take the last 100 simulations into account to guarantee that the Markov chain has achieved its stationary distribution. All in all, we see that even with a small number of points for which the function is evaluated, we already obtain reasonable information about the shape of the function through proper simulations. The topic is also known as response surface methodology, and we refer to Dean et al. (2017) for more details. Generally, we can conclude that simulations provide useful insight and can be used in numerous application fields.

Proof We want to sketch that a single Gibbs step has the unconditional distribution as the stationary distribution, if it exists, which we assume here. For simplicity, we do this for a two-dimensional vector (Y_1, Y_2). Let $F_1(\cdot)$ denote the marginal (unconditional) distribution of Y_1 with density $f_1(\cdot)$. Analogously, we notate with $f_2(\cdot)$ the marginal density for Y_2 and $f_{1|2}(\cdot \mid y_2)$ stands for the conditional density of variable 1 given variable 2.

Assume now that Y_1 is drawn from $F_1(\cdot)$, and we conduct a single step in the Gibbs algorithm. This means we simulate Y_2^* given Y_1, and afterward, we simulate Y_1^* given the value of Y_2^*. We want to show that $P(Y_1^* \leq y_1^*)$ is equal to $F_1(y_1^*)$, which means the marginal distribution is the stationary distribution in the Gibbs algorithm.

Note that

$$\begin{aligned}
P(Y_1^* \leq y_1^*) &= \int P(Y_1^* \leq y_1^* \mid y_1) f_1(y_1) \mathrm{d}\, y_1 \\
&= \int \int F_{1|2}(y_1^* \mid y_2^*) f_{2|1}(y_2^* \mid y_1) f_1(y_1) \mathrm{d}\, y_1 \mathrm{d}\, y_2^* \\
&= \int F_{1|2}(y_1^* \mid y_2^*) \Big(\int f_{2|1}(y_2^* \mid y_1) f_1(y_1) \mathrm{d}\, y_1 \Big) \mathrm{d}\, y_2^* \\
&= \int F_{1|2}(y_1^* \mid y_2^*) f_2(y_2^*) \mathrm{d}\, y_2^* \\
&= F_1(y_1^*)
\end{aligned}$$

As can be seen, the reasoning is rather straightforward. □

8.6 Uncertain Input Parameters

Usually, each simulation requires the specification of input (or tuning) parameters, which we can denote as θ. These parameters are often not completely known, and

there is uncertainty about properly specifying the parameters for the simulation. We may quantify this uncertainty concerning some probability distribution $f_\theta(\theta)$ for $\theta \in \Theta$. Given θ, we draw Y^* from

$$Y \mid \theta \sim F(.;\theta) \tag{8.31}$$

The uncertainty about the parameter leads to the distribution

$$Y \sim F()$$

where

$$F(y) = P(Y \leq y) = \int_{\theta \in \Theta} \int_{\infty}^{y} f(y;\theta) f_\theta(\theta) d\theta \mathrm{d}\, y.$$

Instead of calculating the distribution $F(\cdot)$ exactly, we can simulate from it by first simulating the parameter and then simulating from (8.31). To be specific:

1. Draw θ^* from $f_\theta(.)$;
2. Draw Y^* from $F(;\theta^*)$.

Example 8.5 Assume we aim to draw a bivariate normal distribution, but we are uncertain about the correlation. We assume the correlation ρ^* to be distributed uniformly between -1 and 1. Given ρ^*, we draw (Y^*, Z^*) as bivariate normal with variance 1 and correlation ρ^*. We show the resulting simulated random variables in Fig. 8.16. The left-hand side shows 10000 simulated values, and the right-hand side shows the resulting contour plot. We see that although every single simulated pair (Y^*, Z^*) is drawn from a normal distribution, the uncertainty concerning ρ induces a non-normal distribution. ▷

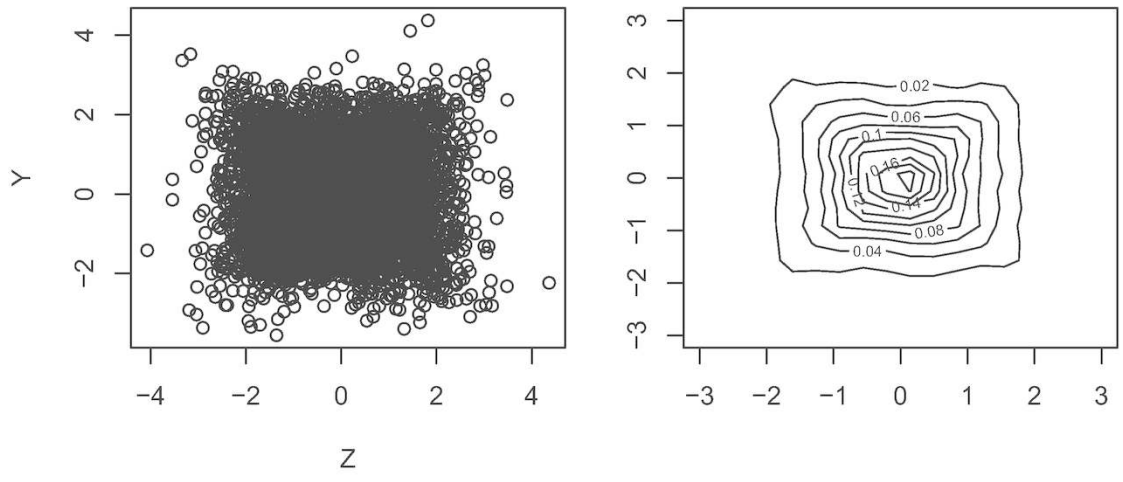


Fig. 8.16 Bivariate normal random variables with varying correlation

8.7 Bootstrapping

8.7.1 The Plug-In Principle

We now turn to a fundamentally different simulation technique known as **bootstrapping**. At its core, bootstrapping involves sampling from the observed data itself. While this may initially seem counterintuitive, the method is supported by strong theoretical justification, as we will explore in the following sections.

The idea of bootstrapping was originally proposed by Efron (1979) and has seen many extensions since then. We refer to Tibshirani and Efron (1993), Davison and Hinkley (1997) or Dikta and Scheer (2021). With the material derived in this chapter, we can quite easily motivate the bootstrap by referring to the Glivenko-Cantelli theorem (see Sect. 1.2.2). We assume a sample drawn from the general data-generating process (2.3):

$$Y_i \sim G() \text{ i.i.d. for } i = 1, \dots n. \tag{8.32}$$

For notational reasons, it is helpful here to relabel the unknown distribution $G(\cdot)$ subsequently with $F(\cdot)$, where we emphasize that $F(\cdot)$ is unknown and in most parts of what follows not necessarily a member of any statistical model $\mathscr{F}$. Hence, we misuse our previous notation slightly and rewrite the data-generating process (8.32) to

$$Y_i \sim F() \text{ i.i.d. for } i = 1, \dots n, \tag{8.33}$$

where $F(\cdot)$ is unknown. Let $F_n(y)$ denote the empirical distribution function defined as

$$F_n(y) = \frac{1}{n} \sum_{i=1}^{n} 1_{\{y_i \leq y\}}.$$

The Glivenko-Cantelli theorem states that for n increasing, we obtain that $F_n(y)$ converges to $F(y)$ for all $y \in R$. We want to utilize this now for simulations.

Assume that we have the intention to draw random numbers from (8.33). If we were able to do so, we could obtain full information about $F(\cdot)$ just by drawing more and more random variables. Apparently, this is seldom possible due to time or computing constraints. But even more importantly, we don't know $F(\cdot)$, and hence we cannot draw random numbers from it. The only information we have about $F(\cdot)$ is the concrete sample $y_1, \dots y_n$, or to be more specific, we have the empirical distribution function $F_n(\cdot)$. The idea is now to replace (8.33) by its empirical version. That is, we propose to draw (new) samples from

$$Y_i^* \sim F_n() \text{ i.i.d. for } i = 1, \dots n. \tag{8.34}$$

Note that this is possible, unlike drawing from $F(\cdot)$, since we know $F_n()$ and can directly apply the ideas from the earlier parts of this chapter; see also Fig. 8.3. It is not difficult to see that (8.34) is equal to

$$\text{draw } Y_i^* \text{ from the data } (y_1, \ldots y_n) \textbf{ with } \text{replacement, for } i = 1, \ldots, n. \quad (8.35)$$

The procedure is known as **bootstrapping**, or more specifically as nonparametric bootstrapping. With Glivenko-Cantelli, we see that the approach is justifiable, since we just replace $F(\cdot)$ by its estimate $F_n()$. We visualize the idea in Fig. 8.17, where we show the empirical distribution function for exponentially distributed random variables for two different sample sizes.

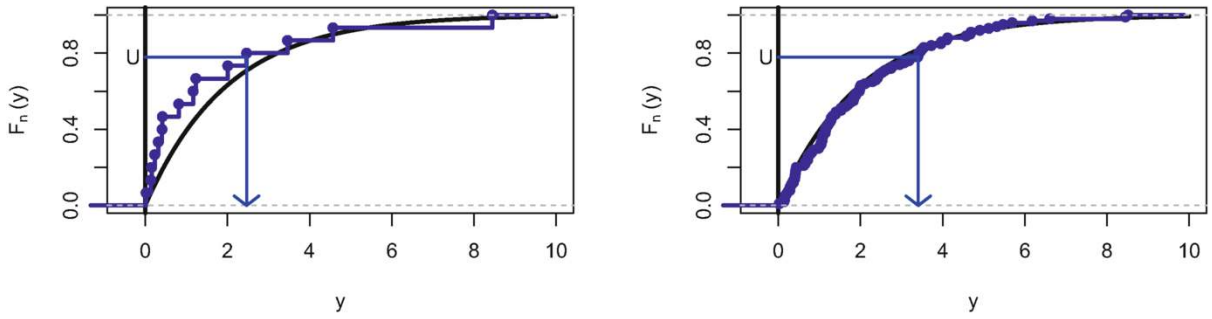


Fig. 8.17 Sampling from the empirical distribution function with sample size $n = 15$ (left) and $n = 100$ (right)

Example 8.6 We want to clarify the idea of bootstrapping with a simple example. In Fig. 8.18, we show in the top left plot 100 standard normally distributed random variables. These lead to the empirical distribution function shown as blue step function in the top right plot. We include the true normal distribution function as a reference in black.

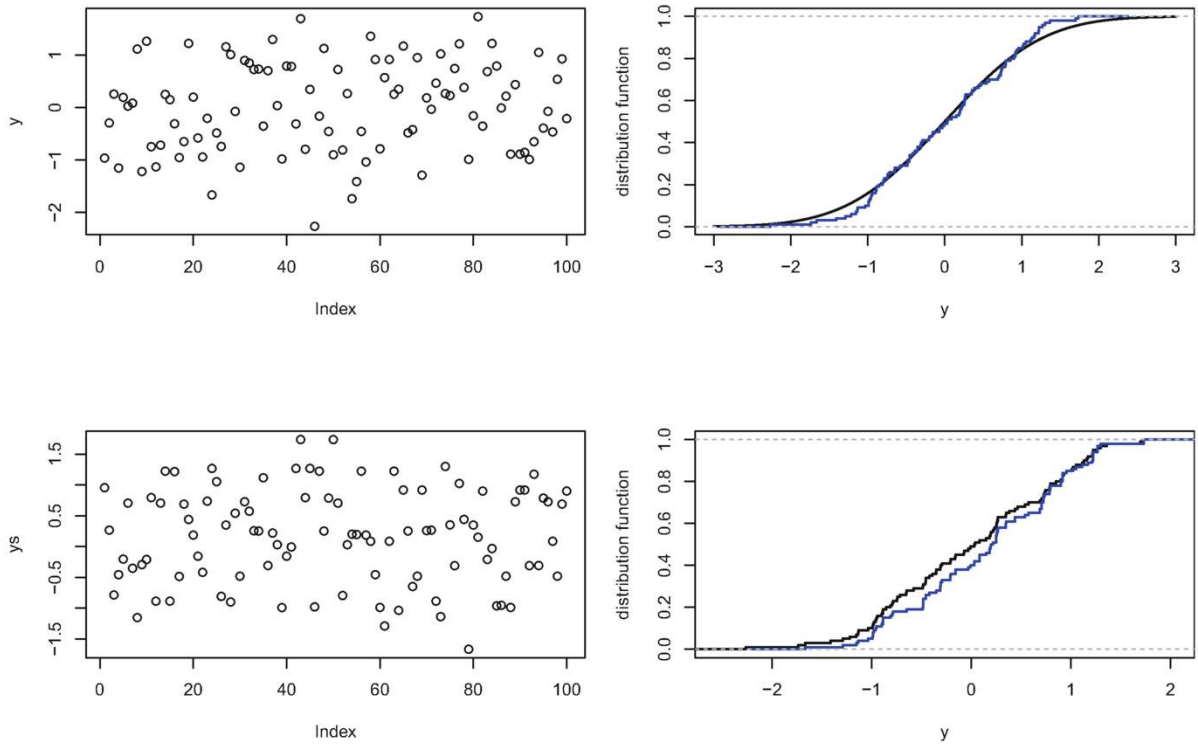


Fig. 8.18 Demonstration of bootstrapping. The top left plot shows 100 $\mathrm{N}(0,1)$ random variables. The resulting empirical distribution function is shown in the top right plot in blue with the true distribution as reference in black. The bottom row shows bootstrapped values. The left plot shows simulated values with replacement, and the right-hand side plot shows the empirical distribution function in blue with the original (empirical) distribution function shown in black as a reference

From the original data, we now draw again 100 samples with replacement, which are shown in the bottom left plot. Note that some original values appear multiple times, since we drew with replacement and hence can draw the same numbers more than once. The resulting empirical distribution function is shown in the bottom right plot in blue. Now, the reference is the original empirical distribution function, included as a black curve. ▷

We want to motivate the use of the bootstrap in more detail using the so-called plug-in principle. Let $y = (y_1, \ldots y_n)$ denote the data drawn from (8.33) and assume that we have calculated a statistics $t(y)$ from the data for which we would like to derive some properties. For instance, $t(y)$ could be an estimate of the mean value, and we would like to derive the variance of the mean estimate. First, we rewrite the statistics $t(y)$ as a functional in the form

$$t(y) = T(F_n), \tag{8.36}$$

where T is a functional, that is, a function applied to a function. For instance, the arithmetic mean can be written as

$$T(F_n) = y = \sum\nolimits_{i=1}^{n} \tfrac{y_i}{n} = \int_{-\infty}^{\infty} y \mathrm{d}F_n(y),$$

utilizing the notation introduced in Sect. 1.2.2. If we replace the empirical distribution by its true distribution, we obtain

$$T(F) = \int_{-\infty}^{\infty} y \mathrm{d}F(y) = \mu.$$

With $F_n(\cdot)$ converging to $F(\cdot)$, we have $T(F_n)$ converging to $T(F)$. This is what we call the **plug-in principle**. If we are interested in some quantity $T(F)$, we can estimate this quantity through $T(F_n)$. But the plug-in principle even extends to other calculations.

To demonstrate the idea, we go through the formula for one example quite explicitly (which unfortunately gets quite cumbersome, even though not complicated). Assume, for instance, that we want to estimate the variance of $Y = \dfrac{1}{n} \sum\nolimits_{i=1}^{n} Y_i$. Knowing $F(\cdot)$, this can be calculated through

(8.37)

$$
\begin{aligned}
T(F) = \operatorname{Var}(\overline{Y}) &= \int \dots \int (\overline{y} - \mu)^2 \mathrm{d}F(y_1) \dots F(y_n) \\
&= \int \dots \int \left(\frac{1}{n} \sum_{i=1}^{n} (y_i - \mu) \right)^2 \mathrm{d}F(y_1) \dots F(y_n) \\
&= \frac{1}{n^2} \int \dots \int \sum_{i=1}^{n} (y_i - \mu)^2 \mathrm{d}F(y_1) \dots F(y_n) \\
&\quad + \frac{1}{n^2} \underbrace{\int \dots \int \sum_{i=1}^{n} \sum_{j \neq i}^{n} (y_i - \mu)(y_j - \mu) \mathrm{d}F(y_1) \dots F(y_n)}_{=0} \\
&= \frac{1}{n^2} \sum_{i=1}^{n} \int (y_i - \mu)^2 \mathrm{d}F(y_i) = \frac{1}{n} \sigma^2
\end{aligned}
$$

where $\sigma^2 = \operatorname{Var}(Y_i), i = 1, \dots, n$. We can replace the above formula using the plug-in principle by replacing $F(\cdot)$ with $F_n(\cdot)$. First, we replace the unknown mean μ through the arithmetic mean y, as derived above. Second, we replace the unknown distribution function with the empirical one; meaning we apply the plug-in principle, which gives

$$
T(\widehat{F}_n) = \int \dots \int (y^* - y)^2 \mathrm{d}F_n(y_1^*) \dots F_n(y_n^*),
$$

where $y^* = \sum_{i=1}^{n} y_i^* \Big/ n$, with y_i^* independently drawn from $y_i^* \sim F_n(\cdot)$. We can now carry out the same calculation steps as derived above to obtain (8.37), which leads us to

$$
T(\widehat{F}_n) = \tfrac{1}{n} \widehat{\sigma}^2
$$

where we are again using the plug-in principle

$$
\widehat{\sigma}^2 = \int (y^* - y)^2 dF_n(y) = \tfrac{1}{n} \sum_{i=1}^{n} (y_i - y)^2.
$$

While the plug-in principle is helpful, it is not always applicable. Most often, we cannot calculate the quantities of interest explicitly. In this case, though, we can apply simulations. That is, instead of deriving explicit calculations like above, we draw a number of samples. In other words, we draw *n* variables with replacement from $y_1, \dots, y_n$ which is equivalent to

$$
Y_i^* \sim F_n(), \ i.i.d. \text{ for } i = 1, \dots, n. \tag{8.38}
$$

This step can be repeated several times.

Example 8.7 The above idea can be motivated with a simple example. We assume a sample $(y_1, \ldots, y_n)$ of size n from which we aim to estimate the variance. Figure 8.19 shows our data of size $n = 100$ in the top left plot. Calculating the empirical variance of these data leads $\widehat{\sigma}^2 = 0.92$. We aim to quantify uncertainty for this variance estimate through bootstrapping. Do so we draw B bootstrap samples denoted by $(Y_1^{*(b)}, \ldots, Y_n^{*(b)})$. Each sample is drawn from $y_1, \ldots, y_n$ with replacement which is equivalent to

$$Y_i^{*(b)} \sim F_n(), \ i.i.d. \text{ for } i = 1, \ldots, n \text{ and } b = 1, \ldots, B.$$

Three such bootstrap samples are shown in Fig. 8.19, top right plot. For each of the B bootstrap samples, we calculate the resulting variance estimate, that is,

$$\widehat{\sigma}^{2(b)} = \sum_{i=1}^{n} \left(Y_i^{*(b)} - Y^{*(b)}\right)^2,$$

where $Y^{*(b)}$ is the arithmetic mean in the b-th bootstrap sample. Based on $B = 400$ bootstraps, we show the resulting variance estimates in the bottom left plot of Fig. 8.19. The estimate based on the original sample is shown as a vertical line. From these estimates, we can now construct a confidence interval. This is carried out by taking the distribution of the bootstrapped values $\widehat{\sigma}^{2(b)}$, which is shown in the bottom right plot of Fig. 8.19. Using the $\alpha / 2$ and the $1 - \alpha / 2$ quantiles as boundaries provides a bootstrapped confidence interval for the estimate. In Fig. 8.19, we show this for $\alpha = 0.1$ leading to the interval $[0.7, 1.15]$. ▷

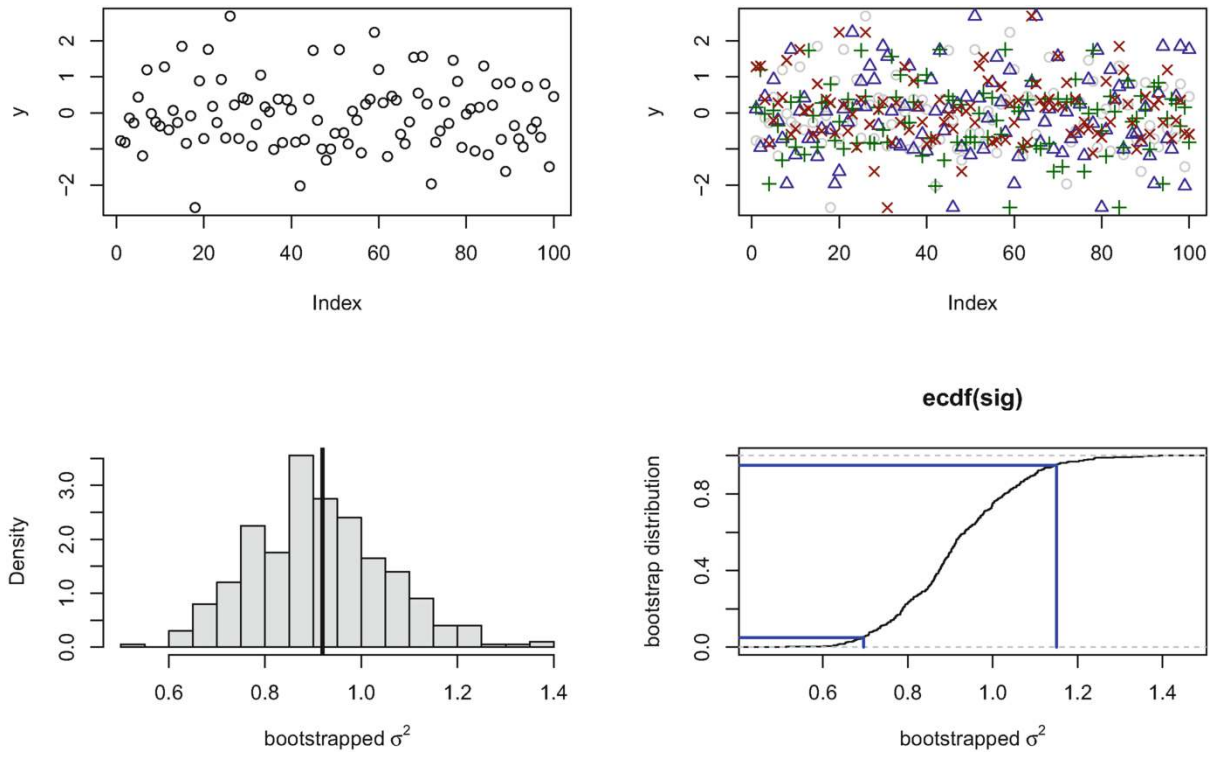


Fig. 8.19 Top row: Sample data of size $n = 100$ (left hand side) and original data (grey) with three bootstrap samples, coded in different colors and shapes of points (right-hand side) Bottom row: Variance of $B = 400$ bootstrap samples and empirical variance of the original data shown as vertical line (left-hand side) and resulting distribution function, which allows constructing confidence intervals (right-hand side)

It should be noted that the bootstrap procedure is very general and easily applicable to a wide variety of different problems. We will also demonstrate its generality by looking at prediction models.

8.7.2 Bootstrapping in Prediction Models

Assume a supervised machine learning framework where input data *x* is used to predict an output variable *y*. We also refer to Sect. 13.2, where we discuss relations to machine learning in more depth. For now, assume that we have (i.i.d.) data (x_i, y_i) with $i = 1, \ldots, n$ which have been used to train a prediction model $\hat{y}(x) = h(x, \hat{\theta})$ with $\hat{\theta}$ as trained (fitted) parameters. We intend to quantify the uncertainty of the model, given that it is trained on finite data. If we consider both x_i and y_i as random, we can group them as pairs and bootstrap directly from these pairs. In other words, we draw *B* bootstrap samples

$$(x_i^{*(b)}, y_i^{*(b)}) \sim \{(x_i, y_i),\ i = 1, \ldots, n\} \text{ with replacement} \tag{8.39}$$

for $b = 1, \ldots B$. From each bootstrap sample $(x_i^{*(b)}, y_i^{*(b)}),\ i = 1, \ldots n$, we retrain the model leading to $h(x, \hat{\theta}^{(b)})$ bootstrapped models. The *B* trained models now allow us to quantify the uncertainty of the models based on the finite sample. We motivate the idea with a concrete example.

Example 8.8 Assume we have two machine learning classifiers and want to compare them. Consider, for instance, the case where two machine learners were trained for classifying images of "cats" and "dogs," to provide a simple example. The results of the two procedures applied to some test data of size $n = 1000$ are shown as confusion matrices in Table 8.1. Looking at the numbers, it looks as if classifier 2 is slightly better. It classifies more cats as cats and more dogs as dogs. Calculating the accuracy for these tables, one gets that classifier 1 predicts 86.1% of the cases right while classifier 2 increases this number to 88%. The question arises whether these numbers support the hypothesis that classifier 2 is performing better. This results in a classical test scenario, where the hypothesis can be formulated as

$$H_0: \text{ Both classifiers perform equally good.}$$

Putting it differently, we want to check whether the difference in the accuracy measures of order $|0.88 - 0.861| = 0.019$ is significantly different from 0. This question can be tackled using bootstrapping.

Table 8.1 Error matrices for two classifiers. Classifier 1 on the left and classifier 2 on the right, applied to the same test data

The main idea to make the bootstrap work in this scenario is that if the hypothesis holds. In that case, we can classify with either classifier 1 or classifier 2, since they perform exchangeable well. We utilize this idea when constructing the bootstrap. Note that the data have the structure

$$(z_i, c_{1,i}, c_{2,i}) \text{ for } i = 1, \dots n \tag{8.40}$$

where z_i is the true class (either cat or dog) and $c_{k,i}$ is the predicted class based on classifier k for $k = 1, 2$. We now draw a bootstrap sample as follows. First, we draw the triples

$$(z_i^*, c_{1,i}^*, c_{2,i}^*) \tag{8.41}$$

from the data with replacement, with $i = 1, \dots n$. Then, for each bootstrapped observation, we randomly swap the classifiers, which we can do if the hypothesis holds. This defines the final bootstrap data

$$(z_i^*, c_{(1+R_i^*),i}^*, c_{(2-R_i^*),i}^*) \tag{8.42}$$

where $R_i^* \sim \mathrm{B}(1, 0.5)$ are binomial and i.i.d., which guarantee the random swaps between the classifiers. Based on the bootstrap, we can calculate the bootstrap accuracy of both classifiers, denoted as A_k^*, for $k = 1, 2$, from which we get the absolute distance $D^* = |A_1^* - A_2^*|$. Repeating this step B times yields a bootstrap distribution of D^*, which is generated under H_0 and allows evaluating the observed value from above $d = 0.019$. We show the resulting empirical distribution function of the bootstrapped values D^* in Fig. 8.20 and indicate the observed value as a vertical line. We conclude that the differences found in Table 8.1 can solely be explained by random variation, and in fact, the resulting *p*-value is on the order of 0.15. ▷

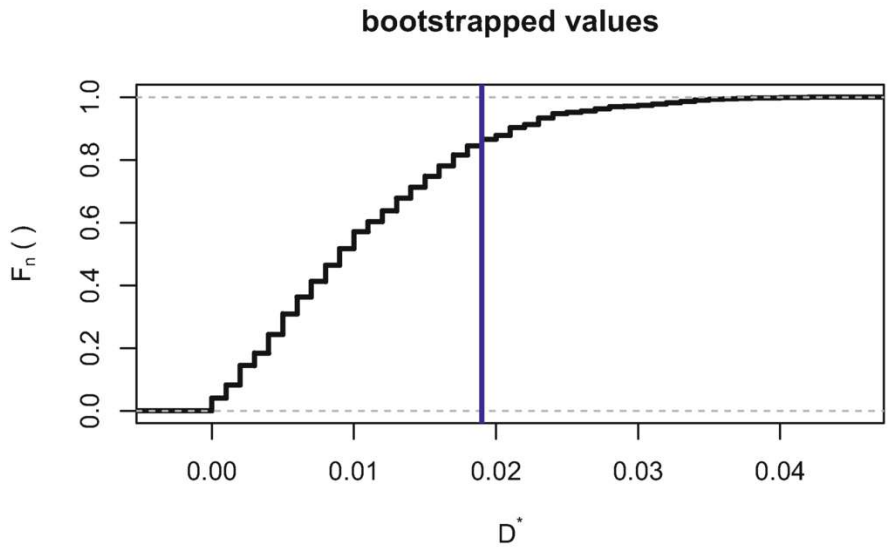


Fig. 8.20 Bootstrapped absolute differences in accuracy

8.8 Permutation

To complete the chapter, we want to mention permutation-based inference as well. Permutation is primarily used in statistical testing in order to generate the distribution under the hypothesis through simulations. We refer to Berry et al. (2011) or Berry et al. (2019) for a more elaborate discussion of the topic. We here only motivate the idea with a simple example. Assume we have data from two settings or two populations, that is,

$$Y_i \sim F_Y(\cdot) \text{ i.i.d.}, i = 1, \dots n \text{ and } Z_j \sim F_Z(\cdot) \text{ i.i.d.}, j = 1, \dots m. \tag{8.43}$$

We want to test the hypothesis $H_0 : F_Y(\cdot) = F_Z(\cdot)$. First, we need to construct some test statistics, where we could use the results of Chap. 5 and use the Wilcoxon test statistics, the Kolmogorov-Smirnoff statistics, or any other suitable measure. To exemplify the idea, we use the Kolmogorov-Smirnoff statistics

$$D = \sup_y |F_{n,Y}(y) - F_{m,Z}(y)|, \tag{8.44}$$

where $F_{n,Y}(\cdot)$ and $F_{m,Z}(\cdot)$ are the empirical distribution functions based on data $y_1, \dots y_n$ and $z_1, \dots z_m$. We do not derive the distribution of D under H_0 but aim to estimate it based on the appropriate permutation of the data. To do so, we stack the data to obtain vector $x = (y_1, \dots, y_n, z_1, \dots z_m)^\top$. If the hypothesis holds, then all elements of x are drawn from the same distribution, which we may denote as

$$x_i \sim F(), \text{ i.i.d. for } i = 1, \dots (m + n). \tag{8.45}$$

Test statistics D in (8.44) results by taking the empirical distribution function of the first n elements of x and comparing this with the distribution function of the

last *m* elements of *x*. Again, if the hypothesis holds, then the order of the elements in *x* does not matter, and we can permute the values on *x*. In other words, we draw a random sample without replacement leading to x^*. Note that the elements in x^* are the same as in *x* but just in a different order. From these numbers, we define the first *n* elements as y^*, that is, $y^* = (x_1^*, \ldots x_n^*)^\top$ and the latter *m* elements as z^*, i.e., $z^* = (x_{n+1}^*, \ldots x_{n+m}^*)^\top$. Using these permuted samples leads to

$$D^* = \sup_y |F_{n,Y^*}(y) - F_{m,Z^*}(y)|,$$

where $F_{n,Y^*}(\cdot)$ and $F_{m,Z^*}(\cdot)$ are the empirical distribution functions based on data $y_1^*, \ldots y_n^*$ and $z_1^*, \ldots z_m^*$. Repeating this step *B* times provides *B* values $D^{*(b)}$, with $b = 1, \ldots, B$, generated under the hypothesis. We can then calculate the *p*-value from these simulated (through permutation) values. We demonstrate this with a real data example.

Example 8.9 In Fig. 8.21, we show rents for apartments in Munich for surveys taken in 2 different years (left-hand side). Apparently, we see an increase in the rents, which is primarily driven by inflation. We therefore increase the rents in the earlier year by multiplying them by the inflation factor. This gives the plot on the right-hand side. We question now whether the distribution of rents, once inflation-adjusted, is the same in the 2 years. We can now calculate the Kolmogorov-Smirnoff distance of the distributions shown on the right-hand side in Fig. 8.21, resulting in $D = 0.0124$. To check whether this is significant, we apply a permutation test with the simulated values D^* shown in the bottom plot in Fig. 8.21. The observed value is included as a vertical line, and the simulated *p*-value results to 0.96. There is no evidence that the structure of the rents has changed, once corrected for the increase based on regular inflation. ▷

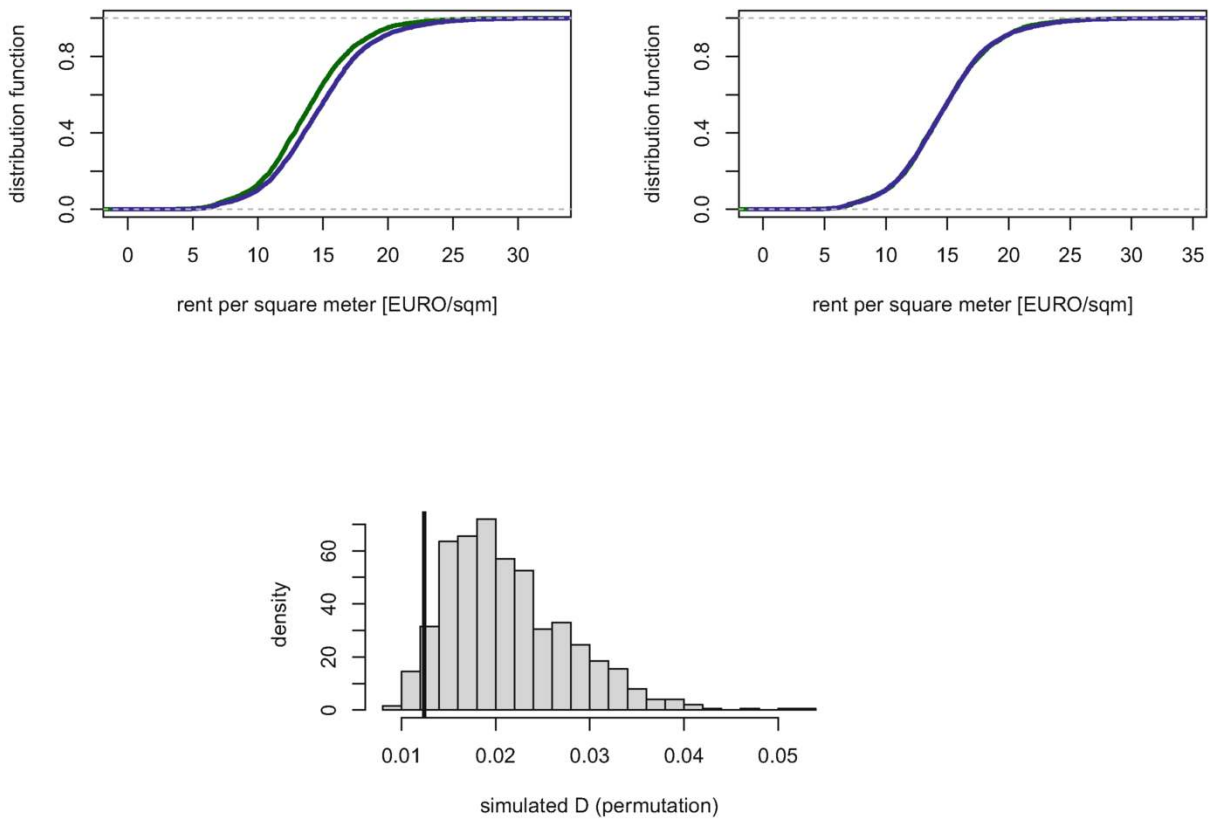


Fig. 8.21 Top row: Rent data (left) and corrected by inflation factor (right). Bottom row: Simulated distances D^* through permutation and observed value D shown as vertical line

8.9 Exercises

Exercise 1 A simple way of modelling the weather is to assume that it can take states with different probabilities of staying the same or changing to a different condition. Let us assume three possible weather states, "Sunny" (S), "Cloudy" (C), and "Rainy" (R). The probabilities of the weather on the next day given today's weather are denoted in the table below. For example, if it is sunny today, for tomorrow there is a 25% probability of it being Cloudy, 25% of being Rainy, and 50% of being Sunny again.

Table

(a) Calculate the probability of the weather being sunny in 2 days, given that it is sunny today.

(b) Write `R`-code to simulate a Markov Chain of the weather on n days. Choose "Sunny" as the starting value and $n = 1000$. What is the (marginal) distribution of the three weather states, i.e., what is the probability of the

weather being Sunny, Cloudy, or Rainy on a given day without any information on the prior days?

(c) Visualize the Markov Chain you simulated by plotting the transitions. Can you identify a "burn-in" period from the plot, and why (not)?

Exercise 2 Let X be a random variable with some unknown distribution function F. You are given the following sample:

$$x = (2.49, 1.35, 2.48, 1.54, 3.84)^\top$$

We want to estimate the median θ. A possible estimate is the sample median $\hat{\theta} = 2.48$.

(a) We would like to quantify the uncertainty of $\hat{\theta}$. Implement a nonparametric bootstrap method to estimate $se(\hat{\theta})$.

(b) Assume now that the observations follow a normal distribution. Repeat the task from point (a), this time using a parametric bootstrap. What is an advantage of this type of bootstrap?

(c) Estimate the bias of $\hat{\theta}$ using nonparametric bootstrapping. Compute a bias-corrected estimate as well.

(d) Is the estimation of the bias via bootstrap useful for an *unbiased* estimator, like the sample mean for the mean of the underlying distribution? What would be the result?

Exercise 3

(a) Briefly describe how and why the bootstrap method can be used for hypothesis testing. Try to relate your explanation to permutation testing.

(b) Test whether the number of fans watching games of the TSV 1860 Munich (recent spectators: 1450, 1700, 5850) is different from that of other clubs in the same category. 20 spectator numbers from other clubs' games are available.

Ordered fan numbers:

1780, 2420, 2580, 2650, 2770, 3030, 3070, 3240, 3380, 3450, 3490, 3600, 3610, 3610, 3750, 3780, 3790, 4080, 4230, 4450

Answer the question using the bootstrap, setting $\alpha = 0.1$. Specify both hypotheses, and interpret the results.

(c) A different way of answering the question from point (b) may be a two sample t-test. Compute the t-test for the given data, and compare the results with the previous ones. What test would you prefer for this kind of data? In your answer, take into account the assumptions for both tests and the exact hypotheses they test.

References

Andrieu, C., & Thoms, J. (2008). A tutorial on adaptive MCMC. *Statistics and Computing,18*, 343–373. [MathSciNet][https://doi.org/10.1007/s11222-008-9110-y]

Berry, K. J., Johnston, J. E., & Mielke, P. W. (2019). *A primer of permutation statistical methods*. Springer. [https://doi.org/10.1007/978-3-030-20933-9]

Berry, K. J., Johnston, J. E., & Mielke, P. W. Jr. (2011). Permutation methods. *Wiley Interdisciplinary Reviews: Computational Statistics,3*(6), 527–542. [https://doi.org/10.1002/wics.177]

Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their application* (Cambridge series in statistical and probabilistic mathematics). Cambridge University Press. [https://doi.org/10.1017/CBO9780511802843]

Dean, A., Voss, D., & Draguljić, D. (2017). Response surface methodology. In *Design and analysis of experiments* (pp. 565–614). Springer.

Dikta, G., & Scheer, M. (2021). *Bootstrap methods*. Springer. [https://doi.org/10.1007/978-3-030-73480-0]

Efron, B. (1979). Bootstrap methods: Another look at the Jackknife. *The Annals of Statistics,7*(1), 1–26. [MathSciNet][https://doi.org/10.1214/aos/1176344552]

Gelfand, A. E. (2000). Gibbs sampling. *Journal of the American Statistical Association,95*(452), 1300–1304. [MathSciNet][https://doi.org/10.1080/01621459.2000.10474335]

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence,PAMI-6*(6), 721–741. [https://doi.org/10.1109/TPAMI.1984.4767596]

Green, P. J. (2001). A primer on Markov chain Monte Carlo. *Monographs on Statistics and Applied Probability,87*, 1–62. [MathSciNet]

Grimmett, G., & Stirzaker, D. (2001). *Probability and random processes* (3rd ed.). Oxford University Press. [https://doi.org/10.1093/oso/9780198572237.001.0001]

Grimmett, G. R. (1973). A theorem about random fields. *Bulletin of the London Mathematical Society,5*(1), 81–84. [MathSciNet][https://doi.org/10.1112/blms/5.1.81]

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika,57*(1), 97–109. [MathSciNet][https://doi.org/10.1093/biomet/57.1.97]

Lehmer, D. H. (1951). Mathematical methods in large-scale computing units. *Annals of the Computation*

Laboratory of Harvard University,26, 141–146.
[MathSciNet]

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & E. Teller (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics,21*(6), 1087–1092.
[https://doi.org/10.1063/1.1699114]

Neal, P., & Roberts, G. (2006). Optimal scaling for partially updating MCMC algorithms. *The Annals of Applied Probability,16*(2), 475–515.
[MathSciNet][https://doi.org/10.1214/105051605000000791]

Riabiz, M., Chen, W. Y., Cockayne, J., Swietach, P., Niederer, S. A., Mackey, L., & Oates, C. J. (2022). Optimal thinning of MCMC output. *Journal of the Royal Statistical Society Series B: Statistical Methodology, 84*(4), 1059–1081.
[MathSciNet][https://doi.org/10.1111/rssb.12503]

Robert, C., & Casella, G. (2004). *Monte Carlo statistical methods*. Springer.
[https://doi.org/10.1007/978-1-4757-4145-2]

Ross, S. M. (2022). *Simulation*. Academic.

Suess, E. A., & Trumbo, B. E. (2010). *Introduction to probability simulation and Gibbs sampling with R* (Vol. 307). Springer.

Thomson, W. E. (1958). A modified congruence method of generating pseudo-random numbers. *The Computer Journal,1*(2), 83–83.
[https://doi.org/10.1093/comjnl/1.2.83]

Tibshirani, R. J., & Efron, B. (1993). *An introduction to the bootstrap* (Monographs on statistics and applied probability, Vol. 57, pp. 1–436). Chapman & Hall/CRC.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
G. Kauermann et al., *Statistics—Coping with Uncertainty*, Textbooks in Statistical Science
https://doi.org/10.1007/978-3-032-33658-3_9

9. Bayesian Inference

Göran Kauermann[1] ✉, Giacomo De Nicola[2], Cornelius Fritz[3] and David Rügamer[4]
(1) Department of Statistics, LMU Munich, Munich, Germany
(2) Harvard T.H. Chan School of Public Health, Harvard University, Boston, MA, USA
(3) School for Computer Science and Statistics, Trinity College Dublin, Dublin, Ireland
(4) Munich Uncertainty Quantification AI Lab, LMU Munich, Munich, Germany

The terminology Bayesian statistics is used if the uncertainty about parameters is formulated with distributions. This is a general setup that can be employed for nearly all models discussed so far. While the conceptual step from parametric models to Bayesian models is simple in so far, that we just aim to quantify our knowledge of parameter values through distributions, the numerical step carries multiple obstacles, since posterior distributions can be burdensome and not easy to work with. This chapter introduces the main ideas of Bayesian statistics but also extends this to Approximate Bayes Computing (ABC) as well as variational approaches in Bayesian statistics.

9.1 The Bayesian Principle

What is known as Bayesian statistics today has a long tradition, which is in fact even older than classical statistics. It refers less to drawing inference from data but more to probability statements and how to predict future events based on knowledge achieved from the past. The patron of the name is reverend Thomas Bayes, who worked in the eighteenth century on the principles of inverse probability, today called the Bayes rule. The result was published posthumously, see Bayes (1763), primarily driven forward by Richard Price, a philosopher and collaborator of Bayes. Following Stigler (2018), the work published under Bayes's name was largely written by Price, in particular, the appendix of the work. The results did not find wide attention at that time, and more than 50 years later, Laplace independently derived the same formula under more general conditions; see de Laplace (1820). Whether Bayes or Laplace count as "inventors" of inverse reasoning remains an open question, and Stigler (1975) and Dale (1982) provide an interesting discussion and historical findings on

this topic. For a general discussion on Bayes statistics and its recent development, we refer to van de Schoot et al. (2021).

Let us now clarify the fundamental principles of Bayesian statistics. Following the notation of Chap. 7, we consider the model $\mathscr{M} = \{f(.;\theta), \theta \in \Theta\}$, but instead of treating the parameter as unknown, which needs to be estimated, we reformulate our knowledge about θ using probability theory. In other words, we consider the parameter to be a (vector-valued) random variable. It should be stressed that this is a conceptual step, but we learned in Chap. 2 that uncertainty is well quantifiable with probability arguments.

Let us demonstrate the idea with a simple example. Assume a binomial model with

$$Y \sim \mathrm{B}(n, \pi)$$

such that π is the parameter of interest. This leads to the likelihood

$$f(y;\pi) = \binom{n}{y}\pi^y(1-\pi)^{n-y}. \tag{9.1}$$

We postulate our prior knowledge (or prior uncertainty) on π with a beta distribution, i.e.,

$$\pi \sim \mathrm{Beta}(\alpha, \beta).$$

The beta distribution is a continuous distribution on the interval $[0, 1]$ with density depending on the parameters $\alpha > 0$ and $\beta > 0$:

$$p(\pi;\alpha,\beta) = \frac{\pi^{\alpha-1}(1-\pi)^{\beta-1}}{B(\alpha,\beta)} \tag{9.2}$$

where $B(\alpha, \beta)$ is the so-called Beta-function. The Beta-function serves as a normalization constant. It formally results in

$$B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} = \int_0^1 u^{\alpha-1}(1-u)^{\beta-1}\mathrm{d}u,$$

with $\Gamma(.)$ as the Gamma function, i.e., the continuous version of the factorial function. The exact shape of the Beta-function is of little interest right now. More relevant is the fact that the normalization constant is analytically available for distributions taking the shape (9.2).

We are interested in the posterior distribution of π after having observed y. Following Bayes' rule, this results in

$$p(\pi \mid y) = \frac{f(y;\pi)p(\pi)}{\int_0^1 f(y;\tilde{\pi})p(\tilde{\pi})d\tilde{\pi}} \propto \pi^{\alpha+y-1}(1-\pi)^{\beta+n-y-1} \tag{9.3}$$

with $\propto$ referring to proportionality, that is, omitting all multiplicative constants that do not depend on π. With (9.2) and (9.3), we readily find that

$$\pi \mid y \sim \mathrm{Beta}(\alpha + y, \beta + n - y). \tag{9.4}$$

We call (9.2) the *prior distribution*, or sometimes in short just prior, and (9.3) the *posterior distribution*. The prior distribution expresses the uncertainty before having

seen any data. The posterior describes the information achieved about the parameter after having observed the data.

We demonstrate the idea in Fig. 9.1. We use the beta prior with $\alpha = \beta = 1$, which is the uniform distribution on $[0, 1]$ as shown in the top left plot. Assume now that we obtain a single observation with $y = 1$, i.e., $n = 1$. The resulting posterior distribution is shown in the top right plot in Fig. 9.1. We also visualize 95 probability mass, which gives an interval previously defined as a credibility interval in Definition 3.13. If $n = 10$ and $y = 10$, that is, in all 10 trials, we observe success, and we obtain the density and credibility interval as shown in the bottom left plot. We emphasize that in none of the cases, maximum likelihood reasoning would have worked, since $\overline{y} = 1$ and consequently the central limit theorem fails to hold. This highlights a quite central advantage of Bayesian reasoning, since we readily obtain quantifiable inference, even if we just observe a single observation. Finally, for $n = 100$ and $y = 90$, we show the resulting posterior density on the bottom right plot. We see that with increasing data, the posterior distribution gets more condensed, which in turn means that the uncertainty reduces.

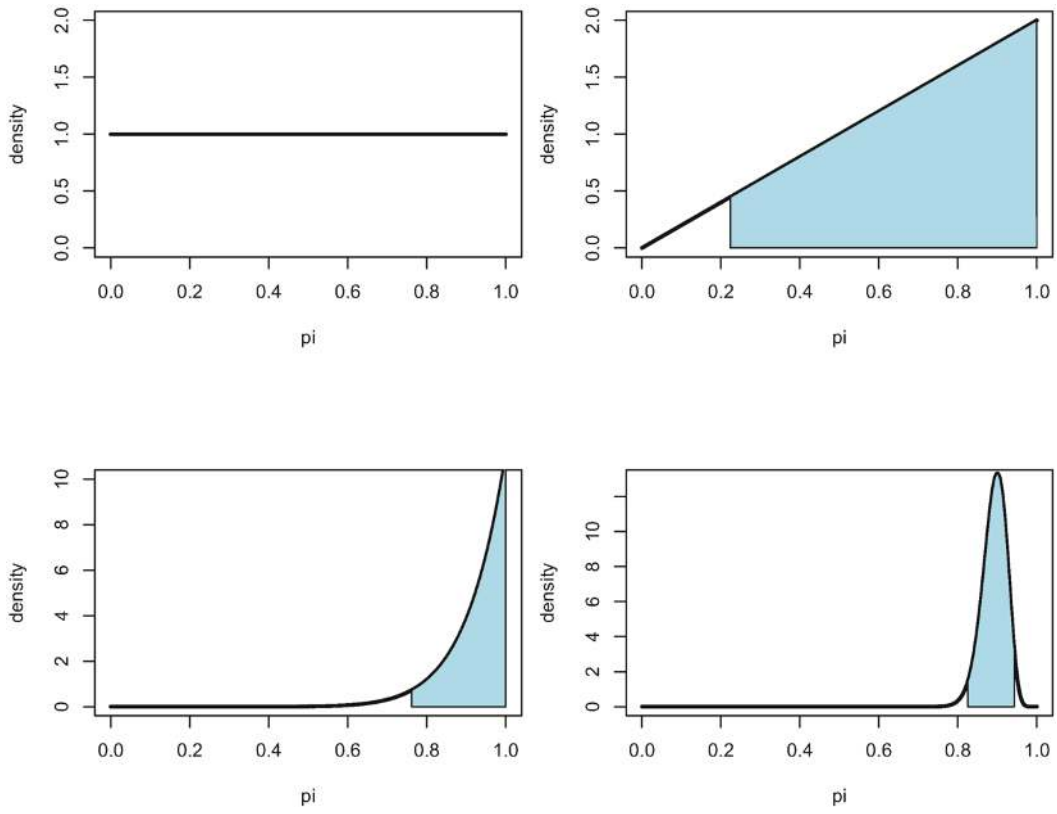


Fig. 9.1 Uniform prior beta distribution (top left). The posterior distribution for $n = 1$ and $y = 1$ (top right). Posterior for $n = 10$ and $y = 10$ (bottom left). Posterior for $n = 100$ and $y = 90$ (bottom right). Blue-shaded areas give 95% probability mass

It seems helpful to summarize the main ingredients of Bayesian statistics.

1. We assume a distribution model $\mathscr{M} = \{f(y;\theta), \theta \in \Theta\}$ for the data. This means that Bayesian statistics does require the specification of a probability model.
2. We impose a **prior distribution** on the parameters, i.e., $\theta \sim p(\cdot, \omega)$, where the prior distribution itself might depend on additional parameters $\omega \in \Omega$, which are called hyperparameters.
3. We are interested in the **posterior distribution** given the data, that is formally $p(\theta \mid y; \omega)$
4. The posterior distribution allows drawing inference, e.g., by calculating intervals with 95% probability mass.

9.2 Prior and Posterior

Let us get more formal now. With $\mathscr{M} = \{f(\cdot;\theta); \theta \in \Theta\}$, we define our probability model for the data $y_1, \ldots, y_n$, which are assumed to be $i.i.d.$ For θ, we formulate our knowledge (or better missing knowledge) as prior distribution

$$\theta \sim p(\cdot;\omega), \tag{9.5}$$

where parameter $\omega \in \Omega$ is called hyperparameter with Ω as hyperparameter space. We define the prior structure as the model class

$$\mathscr{P} = \{p(\cdot;\omega), \omega \in \Omega\}.$$

The choice of the hyperparameters ω will be of interest later in the chapter. For now, we assume that ω is fixed and does express the prior knowledge that one might have on the parameter θ. Based on (9.5) and model $\mathscr{M}$, we obtain the posterior distribution

$$p_{\theta|y}(\theta \mid y_1, \ldots, y_n; \omega) = \frac{\prod_{i=1}^n f(y_i;\theta)\ p(\theta;\omega)}{f(y;\omega)} \tag{9.6}$$

where

$$f(y;\omega) = \int_{\theta\in\Theta} \prod_{i=1}^n f(y_i;\theta) p(\theta;\omega) \mathrm{d}\theta. \tag{9.7}$$

We will see that the normalization $f(y;\omega)$ requires integration over the parameter θ, which can be numerically demanding and keeps us from utilizing (9.6) immediately for drawing inference. A numerically simple strategy results if $\mathscr{F}$ and $\mathscr{P}$ are chosen such that the posterior is again in $\mathscr{P}$. We call this the conjugate prior.

Definition 9.1 *For a given probability model $\mathscr{F}$, we call $\mathscr{P}$ the set of conjugate prior distributions, if it exists and if the posterior $p_{\theta|y}(\cdot) \in \mathscr{P}$. That is, the posterior is from the same family of distributions as the prior distribution.*

We have already seen an example for conjugate priors in Sect. 3.5, namely, the beta distribution for the binomial distribution for y, but there are more examples as discussed below. We list familiar conjugate prior distribution pairs $\mathscr{F}$ and $\mathscr{P}$ in Table 9.1.

Table 9.1 Examples for conjugate distributions

Example 9.1 Let $Y \sim \text{Poisson}(\lambda)$, and we assume that λ follows a Gamma distribution, that is,

$$\begin{aligned} \lambda &\sim \text{Gamma}(\alpha, \beta) \\ &\sim \frac{\lambda^{\alpha-1}\exp(-\lambda\beta)\beta^{\alpha}}{\Gamma(\alpha)}, \end{aligned}$$

with $\Gamma(\alpha)$ a factorial function. The hyperparameters are $\omega = (\alpha, \beta)$. Note that if

$$p(\lambda; \alpha, \beta) \propto \lambda^{\alpha-1} \exp(-\lambda\beta),$$

then $\lambda \sim \text{Gamma}(\alpha, \beta)$. This, in turn, shows for the posterior

$$p_{\lambda|y;\text{alpha},\beta}(\lambda \mid y) \propto \left\{\prod_{i=1}^{n}\left(\lambda^{y_i} \exp(-\lambda)\right)\right\}\lambda^{\alpha-1} \exp(-\lambda\beta),$$

with

$$\left\{\prod_{i=1}^{n}\left(\lambda^{y_i} \exp(-\lambda)\right)\right\} \propto \lambda^{\sum_{i=1}^{n} y_i+\alpha-1} \exp(-\lambda(n+\beta)),$$

which readily shows

$$\lambda \mid y \sim \text{Gamma}\left(\alpha + \sum_{i=1}^{n} y_i, \beta + n\right).$$

Hence, prior and posterior distributions come from the same model class. ▷

Example 9.2 Assume $Y_i \sim \text{N}(\mu, \sigma^2)$, and let σ^2 be known for simplicity. We assume the prior $\mu \sim \text{N}(\delta, \tau^2)$. Note that for normality with mean δ and variance τ^2, we have

$$\begin{aligned} \mu &\sim \frac{1}{\sqrt{2\pi}\sqrt{\tau^2}} \exp\left\{-\frac{1}{2}\frac{(\mu-\delta)^2}{\tau^2}\right\} \\ &\propto \exp\left\{-\frac{1}{2}\frac{(\mu-\delta)^2}{\tau^2}\right\} \\ &\propto \exp\left\{-\frac{1}{2\tau^2}\mu^2 + \frac{\delta}{\tau^2}\mu\right\}. \end{aligned}$$

For the posterior, we get

$$\begin{aligned}\mu \mid y \;&\propto \left\{\prod_{i=1}^{n} \exp\left(-\tfrac{1}{2}\tfrac{(y_i-\mu)^2}{\sigma^2}\right)\right\} \exp\left\{-\tfrac{1}{2\tau^2}\mu^2 + \tfrac{\delta}{\tau^2}\mu\right\} \\ &\propto \exp\left\{-\tfrac{1}{2}\sum_{i=1}^{n}\tfrac{(y_i-\mu)^2}{\sigma^2} - \tfrac{1}{2\tau^2}\mu^2 + \tfrac{\delta}{\tau^2}\mu\right\} \\ &\propto \exp\left\{-\tfrac{1}{2}\left[\tfrac{n}{\sigma^2} + \tfrac{1}{\tau^2}\right]\mu^2 + \left[\sum_{i=1}^{n}\tfrac{y_i}{\sigma^2} + \tfrac{\delta}{\tau^2}\right]\mu\right\}.\end{aligned}$$

Consequently, by collecting the terms and using the above proportionality statements, we get

$$\mu \mid y \sim \mathrm{N}\left(\left(\tfrac{n}{\sigma^2} + \tfrac{1}{\tau^2}\right)^{-1}\left(\sum_{i=1}^{n}\tfrac{y_i}{\sigma^2} + \tfrac{\delta}{\tau^2}\right), \left(\tfrac{n}{\sigma^2} + \tfrac{1}{\tau^2}\right)^{-1}\right).$$

Note that if we set $\delta = 0$, that is we assume that the mean value μ is "a priori" centered around 0, and setting $\tau^2 \to \infty$, that is, we assume a prior such that possible values of μ range from very small to very large values, then we obtain approximately

$$\mu \mid y \sim \mathrm{N}\left(\overline{y}, \tfrac{\sigma^2}{n}\right).$$

This result looks familiar when we exchange μ and $\overline{y}$ and do not take a Bayesian view, but let μ be a parameter. Note also that for $\tau^2 \to \infty$, the prior distribution gets degenerated, meaning that the integral $\int p(\mu; \delta = 0, \tau^2 \to \infty) d\mu \to \infty$. We call this an *improper prior*. Still, even with an improper prior, the posterior distribution is well-defined in this case.▷

Conjugate prior distributions are very special cases. Let us therefore now look at (9.6) and derive alternatives for drawing inference in a Bayesian setting. First, we take an asymptotic view in the sense that $n \to \infty$ and assume that the parameter space Θ is finite-dimensional.

Theorem 9.1 (Bernstein—von Mises Theorem) *For increasing sample size n and an appropriately chosen prior, we find*

$$\theta - \hat{\theta} \overset{a}{\sim} N(0, I^{-1}(\hat{\theta})).$$

We sketch the proof at the end of this section. The Bernstein-von Mises theorem is a useful result. It states that with massive data, Bayesian inference (that is, treating the parameter as random) corresponds to maximum likelihood inference. The result, however, does not say anything about finite samples, which occur in practical applications all the time. We will therefore look at non-asymptotic inference tools. In particular, we will later suggest the use of simulation-based techniques that we introduced in Sect. 8.3.

Proof of Theorem 9.1 Looking at (9.6), we find

$$
\begin{aligned}
p_{\theta|y}(\theta \mid y) \quad &\propto \exp\left\{\sum_{i=1}^{n} \log f(y_i;\theta) + \log p(\theta;\gamma)\right\} \\
&\propto \exp\left\{l_p(\theta)\right\}
\end{aligned} \tag{9.8}
$$

where

$$
\begin{aligned}
l_p(\theta) \quad &= \sum_{i=1}^{n} \log f(y_i;\theta) + \log p(\theta;\gamma) && (9.9)\\
&= l(\theta) + \log p(\theta;\gamma). && (9.10)
\end{aligned}
$$

We occasionally call (9.10) a penalized likelihood, since we add a component, the so-called penalty, to the likelihood. Defining $\hat{\theta}_p$ as the so-called penalized maximum likelihood estimate through

$$
0 = \frac{\partial l_p(\hat{\theta}_p)}{\partial \theta},
$$

we can approximate (9.8) through

$$
\exp\left\{l_p(\theta)\right\} \approx \exp\left\{l_p(\hat{\theta}_p) + \frac{1}{2}(\theta - \hat{\theta}_p)^\top \frac{\partial^2 l_p(\hat{\theta})}{\partial\theta\partial\theta^\top}(\theta - \hat{\theta}_p)\right\}.
$$

This takes the form of a normal distribution. Replacing the second-order derivative by the Fisher Information and reflecting that the $l_p(\hat{\theta})$ is fixed and hence can be treated as a multiplicative constant shows the asymptotic distribution

$$
\theta \mid y \overset{a}{\sim} \mathrm{N}\left(\hat{\theta}_p, I_p^{-1}(\hat{\theta}_p)\right).
$$

For *n* increasing, the influence of the prior diminishes, so that the penalty component in (9.9) does not play any role. Consequently, $\hat{\theta}_p$ becomes the maximum likelihood estimate $\hat{\theta}$ and the result follows. □

9.3 Hyperparameters and Empirical Bayes

We have introduced prior distributions that depend on additional parameters ω, labeled as hyperparameters. Their concrete choice is up to the user of the Bayesian approach and hence needs to be treated with some care. The general advice is certainly to set ω such that $p(\theta;\omega)$ expresses the prior knowledge about θ, where prior knowledge may result from previous analyses, expert knowledge, or any kind of prior belief. However, this is often not properly quantifiable. A plausible strategy is to choose a so-called "flat prior." This means that

$$
p(\theta;\omega) \propto \text{ const } . \tag{9.11}
$$

In other words, every value of θ seems equally likely before having seen any data. We stress that this often results in an *improper prior* in that the density $p(\theta;\omega)$ is not integrable. This immediately happens if the parameter space Θ is not bounded. Nonetheless, even though the prior is not proper, the posterior usually is, at least for a

sufficiently large sample size. The posterior in this case is just proportional to the likelihood since with $p(\theta;\omega) = \text{const}$ it follows

$$p_{\theta|y}(\theta \mid y) \propto \prod_{i=1}^{n} f(y_i;\theta) \cdot \text{ const } . \tag{9.12}$$

A further strategy to select the hyperparameters is called **empirical Bayes**. If we consider parameter θ as random and treat the hyperparameter ω as fixed, we can integrate out θ and derive the distribution for y based on parameters ω. This results through

$$\widetilde{f}(y;\omega) := \int_{\theta\in\Theta} f(y;\theta)p(\theta;\omega)\mathrm{d}\theta. \tag{9.13}$$

In other words, instead of working with the distribution $f(y;\theta)$, we work with the probability model $\widetilde{\mathscr{F}} = \{\widetilde{f}(;\omega); \omega \in \Omega\}$. Choosing a value for hyperparameter ω can be done by utilizing Maximum Likelihood estimation from Chap. 4 and selecting ω to maximize the resulting likelihood

$$\tilde{l}(\omega) = \prod_{i=1}^{n} \widetilde{f}(y_i;\omega) = \int_{\theta\in\Theta} \prod_{i=1}^{n} f(y_i;\theta)p(\theta;\omega)\mathrm{d}\theta$$

The approach is called *empirical Bayes*, even though strictly speaking, it is not Bayesian at all, since no parameters are treated as random. In fact, empirical Bayes completely contradicts the Bayesian principle. The prior expresses the knowledge about θ*before* we have seen any data. Selecting the hyperparameters of the prior distribution based on the data is a clear violation of the Bayesian paradigm. Still, in practice, the approach has been shown to be quite useful, and it, for instance, is dominantly used in smooth regression; see, e.g., Wood (2017).

9.4 Inference Based on MCMC

We now get to the heart of the Bayesian machinery. Looking at formula (9.6), we find

$$p_{\theta|y}(\theta \mid y_1, \ldots, y_n) \propto \underbrace{\prod_{i=1}^{n} f(y_i;\theta)}_{L(\theta)}\, p(\theta;\omega). \tag{9.14}$$

Both components on the right-hand side in (9.14) are numerically available: the first is the likelihood and the second is the prior. The numerical obstacle is the proportionality constant, which we omitted in (9.14). Hence, we have the density of the posterior up to an unknown constant. This is exactly the setup that we discussed in Sect. 8.3, where we introduced Monte Carlo Markov Chain (MCMC) simulations. It allows us to simulate from the posterior, even though we know it only up to a constant. The idea is now to simulate a Markov chain $\theta^{*(t)}$ based on MCMC simulations. The simulated values $\theta^{*(t)}$ (after some burning in) can then be used to

draw inference about θ given the data. We might look at the distribution of the simulated values or just calculate, e.g., the posterior mean

$$\hat{\theta} = \tfrac{1}{T-d} \sum_{t=d+1}^{T} \theta^{*(t)},$$

where d determines the burning in phase.

The algorithm is then as follows:

1. Let $\theta = (\theta_1, \ldots, \theta_p)$ be the p-dimensional parameter of interest, and choose $\theta^{*(0)} \in \Theta$ as some starting value. Set $\theta^{*(t)} = \theta^{*(0)}$ subsequently.
2. Simulate θ^* from some proposal distribution
$$\theta^* \sim g(\cdot \mid \theta^{*(t)}).$$
3. Calculate the acceptance probability through
$$\alpha(\theta^*, \theta^{*(t)}) = \min\left\{1, \tfrac{L(\theta^*)p(\theta^*;\omega)}{L(\theta^{*(t)})p(\theta^{*(t)};\omega)} \tfrac{g(\theta^{*(t)}|\theta^*)}{g(\theta^*|\theta^{*(t)})}\right\}.$$
4. Simulate $U^* \sim \text{Uniform}[0, 1]$.
5. If $U \leq \alpha(\theta^*, \theta^{*(t)})$ set $\theta^{*(t+1)} = \theta^*$ and proceed to Step 6. Otherwise, go back to Step 2.
6. Go back to Step 2 until t takes the required number of steps has been reached.

Unless the number of components in θ is small, the acceptance probability calculated in Step 3 will typically be rather low. This is because multivariate densities can take small values, making the algorithm extremely slow due to a small acceptance rate. If $p > 2$, it is therefore advisable to combine the MCMC with Gibbs sampling as described in Sect. 8.5. This means that the simulation Steps 2–5 are replaced by componentwise simulation steps, where we simulate one component of θ after the other. We thereby keep all other components fixed. The reason why we can do this is that the posterior for component θ_j, keeping all other components θ_{-j} fixed, follows the proportionality

$$\begin{aligned} p_{\theta_j|y,\theta_{-j}}(\theta_j \mid y_1, \ldots, y_n, \theta_{-j}) &\propto p_{\theta|y}(\theta \mid y_1, \ldots, y_n) \\ &\propto \underbrace{\prod_{i=1}^{n} f(y_i; \theta)}_{L(\theta)} p(\theta; \omega), \end{aligned} \tag{9.15}$$

where θ_{-j} is the $p-1$ dimensional subvector of θ excluding the j component. The MCMC chain with Metropolis-Hastings-Gibbs sampling is then constructed as follows:

1. Let $\theta = (\theta_1, \ldots, \theta_p)$ be the p-dimensional parameter of interest, and choose $\theta^{*(0)} \in \Theta$ as some starting value. Set $\theta^{*(t)} = \theta^{*(0)}$ subsequently
2. Define $\theta^* = \theta^{*(t)}$
3. For $j = 1, \ldots p$ run through the following steps:
   (a) Define $\theta^{**} = \theta^*$
   (b) Simulate (and hence overwrite) the j the component θ_j^{**} from some proposal distribution, that is,
   $$\theta_j^{**} \sim g(\cdot \mid \theta_j^*)$$
   (c) Calculate the acceptance probability through
   $$\alpha(\theta^{**}, \theta^*) = \min\left\{1, \frac{L(\theta^{**})p(\theta^{**};\omega)}{L(\theta^*)p(\theta^*;\omega)} \frac{g(\theta^*|\theta^{**})}{g(\theta^{**}|\theta^*)}\right\}$$
   (d) Simulate $U^* \sim \text{Uniform}[0, 1]$
   (e) If $U \leq \alpha(\theta^{**}, \theta^*)$ replace the j-th component in θ^* through $\theta_j^* = \theta_j^{**}$ and proceed. Otherwise, go back to step a.
4. Set $\theta^{*(t+1)} = \theta^*$.
5. Go back to Step 2 until t takes the required number of steps has been reached.

The algorithm has an outer loop, which provides a simulation for all elements in θ and an inner loop, which provides univariate simulations for the individual components in θ.

Example 9.3 We consider zero-inflated count data. Figure 9.2 gives a typical example. Zero inflation occurs if we observe more zeros than assumed by any standard model, e.g., the Poisson distribution. Consider, for instance, the number of claims of customers of an insurance company per given time interval. Most customers will not submit any claim; that is, the number of claims is zero. For some customers, however, one will observe claims and possibly more than one. This is visualized with simulated data in Fig. 9.2 (top left). We can model the data by using the Poisson

distribution but by inflating the zero probability. With $Y \in \{0, 1, \dots\}$ being the count, this leads to the model

$$P(Y = k; \pi, \lambda) = \begin{cases} \pi + (1 - \pi) \frac{\lambda^0}{0!} \exp(-\lambda), & \text{for } k = 0 \\ (1 - \pi) \frac{\lambda^k}{k!} \exp(-\lambda), & \text{for } k > 0 \end{cases}, \tag{9.16}$$

with $\lambda \geq 0$ and $\pi \in [0, 1]$. Note that for $\pi = 0$, we obtain a Poisson distribution, while for $\pi > 0$, the zero outcome is inflated. This leads to a two-parameter distribution, which we aim to model in a Bayesian way. We thereby assume a uniform prior for π and an exponential distribution for λ, that is,

$$\pi \sim \text{Uniform}[0, 1] = \text{Beta}(1, 1) \tag{9.17}$$

$$\lambda \sim p(\lambda; \omega) = \omega \exp(-\lambda\omega) \tag{9.18}$$

with ω as hyperparameter. For ω fixed and data $y_1, \dots, y_n$, the posterior is then proportional to

$$\lambda, \pi \mid y_1, \dots, y_n \propto \prod_{i=1}^{n} P(Y = y_i; \lambda, \pi) \omega \exp(-\lambda\omega), \tag{9.19}$$

with $P(\cdot; \lambda, \pi)$ as given in (9.16). For the proposal, we assume independence of λ and π and therefore model the proposals separately. In both cases, we need to observe the constraints on the parameters, namely, $\pi \in [0, 1]$ and $\lambda > 0$. This can be numerically clumsy for the proposal distribution, and it is therefore advisable to work with unconstrained, transformed parameters. To do so, we parameterize the model with $\eta = \text{logit}(\pi) = \log(\pi / (1 - \pi))$ instead of π. Note that the logit transformation yields the prior for η through

$$\eta \sim p_\eta(\eta) = \frac{\partial \pi}{\partial \eta} \underbrace{p_\pi(\pi(\eta))}_{=\text{const}} = \pi(\eta)(1 - \pi(\eta)) \tag{9.20}$$

with

$$\pi(\eta) = \text{logit}^{-1}(\eta) = \frac{\exp(\eta)}{1 + \exp(\eta)}.$$

As proposal density for η, we use a normal distribution centered at the current value $\eta^{*(t)}$, that is,

$$\eta^* \mid \eta^{*(t)} \sim \text{N}(\eta^{*(t)}, \sigma_\eta^2),$$

where we set $\sigma_\eta = 0.5$.

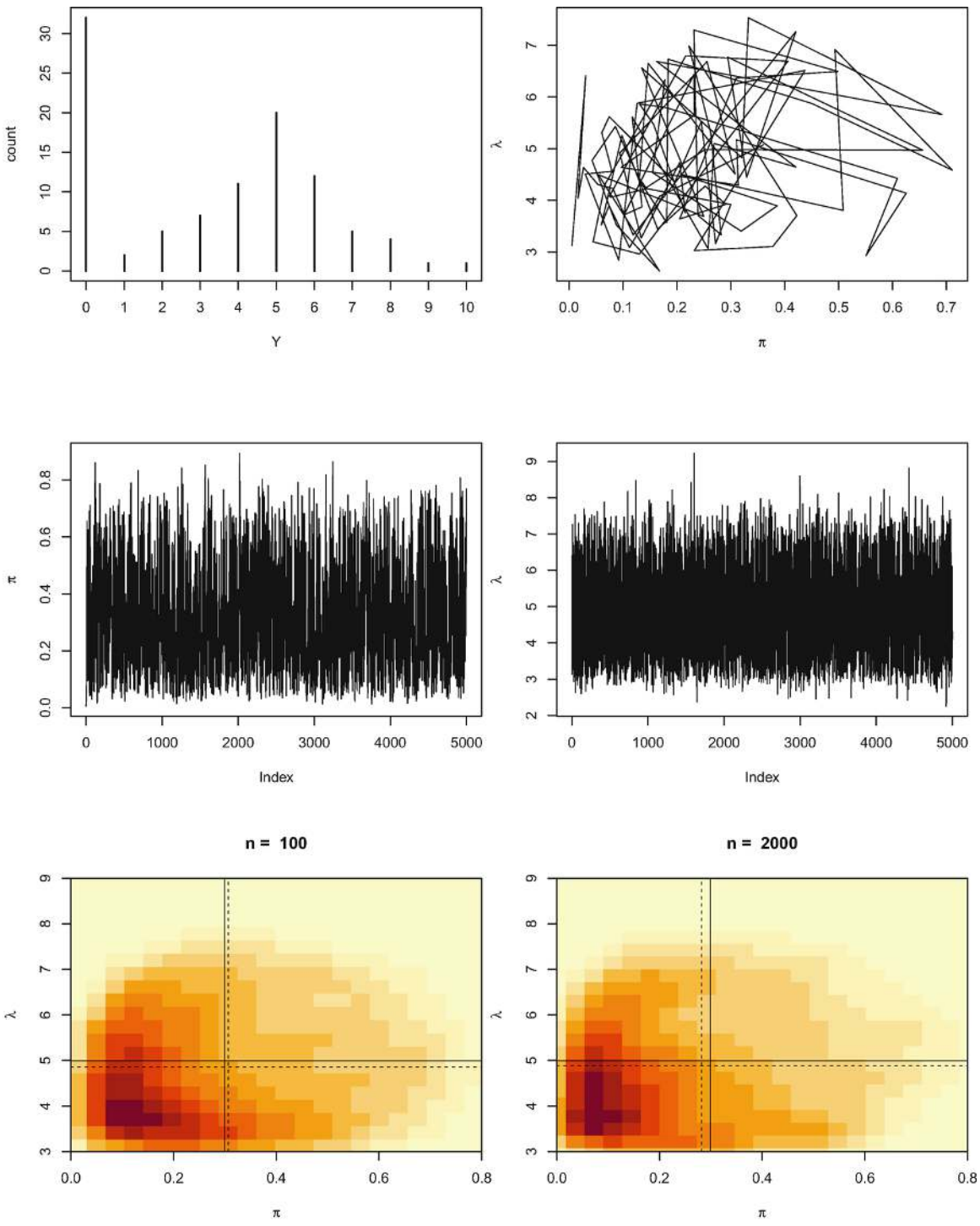


Fig. 9.2 Posterior distribution through MCMC for zero-inflated count data. Data (top left) and the first 300 steps of the MCMC chain (top right). MCMC steps for π (middle left) and λ (middle right). Joint posterior distribution for (π, λ) for $n = 100$ and for $n = 2000$

Similarly, we proceed for λ. Instead of modeling λ, we take the log and model $\delta = \log(\lambda)$. This leads to the prior

$$\delta \sim p_\delta(\delta; \omega) = \frac{\partial \lambda}{\partial \delta} p_\lambda(\lambda(\delta); \omega) = \lambda\omega \exp(-\lambda\omega) \tag{9.21}$$

with $\lambda = \exp(\delta)$. For the proposal, we choose $\delta^* \mid \delta^{*(t)} \sim \mathrm{N}(\delta^{*(t)}, \sigma_\delta^2)$ with $\sigma_\delta = 0.5$. We can now start the MCMC for η and δ. Since we use a symmetric proposal for η and δ, we obtain the acceptance probability through

$$\alpha((\delta^*, \eta^*), (\delta^{*(t)}, \eta^{*(t)})) = \min\left\{1, \frac{\prod_{i=1}^n P(y_i; \lambda^*, \pi^*)}{\prod_{i=1}^n P(y_i; \lambda^{*(t)}, \pi^{*(t)})} \; \frac{p_\delta(\delta^*, \omega)}{p_\delta(\delta^{*(t)}, \omega)} \; \frac{p_\eta(\eta^*)}{p_\eta(\eta^{*(t)})}\right\}$$

with the ratio of the likelihood components rewritten as

$$\frac{\prod_{i=1}^n P(y_i; \lambda^*, \pi^*)}{\prod_{i=1}^n P(y_i; \lambda^{*(t)}, \pi^{*(t)})} = \exp\left\{\sum_{i=1}^n \left[\log\left(P(y_i; \lambda^*, \pi^*)\right) - \log\left(P(y_i; \lambda^{*(t)}, \pi^{*(t)})\right)\right]\right\},$$

where $\lambda^* = \exp(\delta^*)$ and $\pi^* = \exp(\eta^*) / (1 + \exp(\eta^*))$. For large *n*, it is advisable to calculate the ratio of the probabilities as suggested.

We look at the performance of the MCMC steps in Fig. 9.2. The top right plot shows the first 300 steps of the MCMC samples. The quantity $\pi = \exp(\eta) / (1 + \exp(\eta))$ is on the horizontal axis and $\lambda = \exp(\delta)$ is on the vertical axis. We see good mixing, that is, the Markov chain jumps around. This is also seen in the middle plots, where we show the MCMC steps for π on the left and for λ on the right. The bottom plots show a density plot of the simulated π and λ for the original data ($n = 100$) and for an increased data set ($n = 2000$). The vertical lines indicate the true values used to simulate the data, while the dotted lines give the mean value of the MCMC samples. We recognize decent concordance. We also find that the posterior density is slightly skewed, so that the mode of the posterior density is found for small values of π and λ, while the values used for simulations are $\pi = 0.3$ and $\lambda = 5$. All in all, the example shows that the MCMC approach is easy to implement and allows for estimating complex distributional models. ▷

We demonstrated how MCMC sampling can be used to simulate from the posterior, even though this is known up to a constant only. The resulting simulations $\theta^{*(t)}$ can be used to draw inference, as sketched in the example. We may estimate the posterior mean through the mean of the MCMC samples, i.e.,

$$\hat{E}(\theta \mid y) = \frac{1}{T-d} \sum_{t=d+1}^{T} \theta^{*(t)}$$

where *d* defines the length of the burn-in phase. We may also calculate the posterior variance to assess uncertainty, i.e.,

$$\widehat{\mathrm{Var}}(\theta \mid y) = \frac{1}{T-k} \sum_{t=k+1}^{T} \left\{\theta^{*(t)} - \hat{E}(\theta \mid y)\right\}^2.$$

One should keep in mind, though, that the MCMC samples are not independent and, in fact, they are strongly correlated through the structure of the Markov chain. This is, however, not that relevant, but we need to bear in mind that the *T* simulations do not carry the same amount of information as *T* independent replicates.

Instead of calculating moments of the posterior sample, we can also directly look at the posterior density as done in the example above. The density, be it univariate or bivariate, gives good insight into the amount of uncertainty, but for dimensions beyond 2, it is not really presentable.

We will now extend the above Bayesian approaches to two further situations. First, we question whether we can neglect the likelihood and find simulation parameters that lead to data comparable to the observed ones. Second, we will look at useful approximations if the posterior distribution is too complex or numerically too demanding.

9.5 Approximate Bayes Computation (ABC)

Approximate Bayes computing, in short ABC, can be applied if the likelihood is intractable or only known up to some constant. We consider the setup where data y are available and $L(\theta) = L(\theta; y)$ is the corresponding likelihood. For notational purposes, it is advisable to explicitly take the dependence of the likelihood on *y* into account. Note that the posterior results through

$$p(\theta \mid y) \propto L(\theta; y) p_\theta(\theta).$$

In some constellations, the likelihood itself can be clumsy and difficult to calculate. We consider a simple example later in this chapter. The idea of approximate Bayes computing is quite simple, and we refer to Marin et al. (2012) for a general overview. The approach traces back to Rubin (1984) and Tavaré et al. (1997), and the main idea is as follows:

1. Generate θ^* from some prior distribution $p_\theta(\theta)$.
2. Generate Y^* from the assumed distribution, i.e., the likelihood function $f_Y(\cdot; \theta^*)$.
3. If $Y^* = y$ accept θ^*, otherwise go back to Step 1.
4. Repeat Steps 1–3 until the required number of simulated values $\theta^{*(1)}, \ldots, \theta^{*(N)}$ are reached.

The algorithm itself is not really feasible, since it is very unclear whether the acceptance criterion in Step 3 is ever met. But the idea of the algorithm easily shows that an accepted value θ^* is in fact drawn from the posterior distribution $p_\theta(\theta \mid y)$. To see this, note that for simulations that passed step 3, we denote the sampling probability as $\widetilde{p}_\theta$. Then

$$\begin{aligned} \theta^* \sim \widetilde{p}_\theta(\theta^*) &\propto p_\theta(\theta^*) f(y; \theta^*) \\ &\propto p_\theta(\theta^* \mid y). \end{aligned}$$

Instead of postulating equality in Step 3, it is preferable to request proximity for the proposed value θ^* to be accepted. In other words, one may replace Step 3 through

3. If $d(Y^*, y) \leq \epsilon$, then accept θ^*; otherwise, go back to Step 1.

The quantity $d(\cdot, \cdot)$ is thereby some distance measure, and ϵ needs to be chosen appropriately. We demonstrate the idea by continuing with Example 9.3 on zero-inflated count data.

Example 9.4 Considering Fig. 9.2, we have plotted zero-inflated count data in the top left plot. Using model (9.16) together with the priors (9.19) and (9.20), we can simulate π^*, λ^* and subsequently Y^*. We define the distance between the data y, seen in Fig. 9.2, and the simulated values through

$$d(Y^*, y) = \sum_{k=0}^{10} \left(\overline{x}_k^* - x_k\right)^2$$

where

$$\overline{x}_k = \frac{1}{n} \sum_{i=1}^{n} 1\{y_i = k\} \text{ and } \overset{*}{\overline{x}}_k = \frac{1}{n} \sum_{i=1}^{n} 1\{\overset{*}{Y}_i = k\}.$$

Hence, we compare the relative frequencies of the observed and simulated data up to an outcome 10. In Fig. 9.3, we plot the simulated values π^* and λ^* based on the distance $d(Y^*, y)$. The thicker the points, the smaller the distance. Simulations with $d(Y^*, y) > 0.05$ are omitted. The dashed vertical and horizontal lines indicate the weighted arithmetic mean of the simulated π^* and λ^*, where the weights are defined as the reciprocal of the distance $d(Y^*, y)$. The solid lines are the values used for simulation. We see again a reasonable concordance, which demonstrates that the ABC approach is doing the right thing.▷

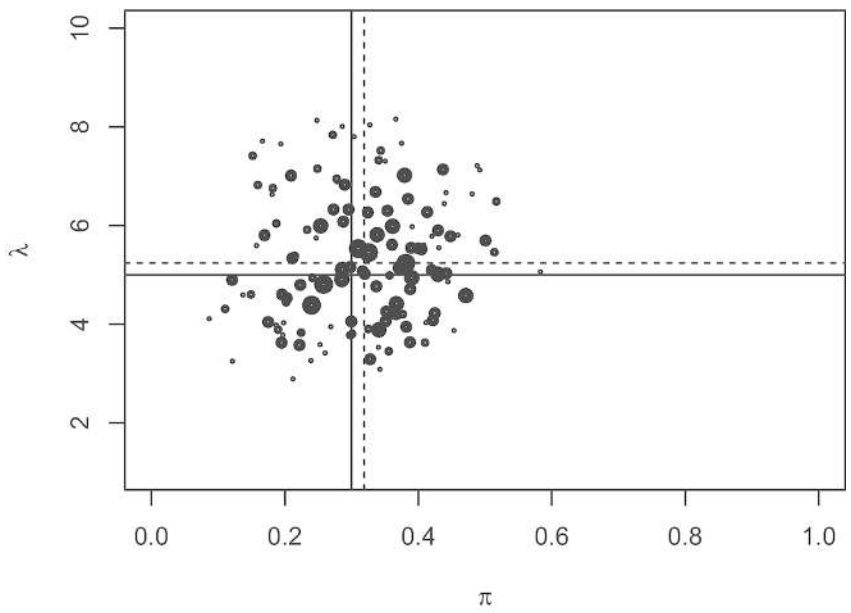


Fig. 9.3 Approximate Bayesian computation. Simulated parameters π and λ, weighted by the reciprocal distance to the observed data

9.6 Variational Bayes Reasoning

We now look at an approach to approximate the posterior distribution. The simplest setup for this is the Bernstein-von Mises Theorem 9.1 or its penalized version, which we state again below. For large samples, i.e., for n sufficiently large, we have

$$\theta \mid y \overset{a}{\sim} \mathrm{N}(\hat{\theta}_p, I_p^{-1}(\hat{\theta}_p)), \tag{9.22}$$

where $\hat{\theta}_p$ is the maximum of the penalized log likelihood

$$l_p(\theta) = \sum_{i=1}^{n} \log f(y_i; \theta) + \log p(\theta).$$

Note that the prior distribution $p(\theta)$ might depend on further hyperparameters. The penalized Fisher information takes the form

$$I_p(\theta) = I(\theta) + \frac{\partial^2 p(\theta)}{\partial\theta\partial\theta^\top}.$$

Besides approximation through a normal distribution, we want to look at the idea of approximating the posterior in more technical detail. The idea is to replace the exact posterior distribution $p(\theta \mid y)$ with some approximation $q(\theta)$, where $q(\cdot)$ has a simpler structure than the original $p(\theta \mid y)$. For example, we might restrict $q(\theta)$ to be a normal distribution and hence aim to find the "closest" normal distribution to $p(\theta \mid y)$. Apparently, this resembles (9.22). We may also simplify the distribution further and assume independence. That is for $\theta = (\theta_1, \ldots, \theta_p)$, we assume $q(\theta) = q_1(\theta_1)q_2(\theta_2) \cdots q_p(\theta_p)$. In general, let

$$q(\theta; \gamma) \in \mathcal{Q} = \{q(\cdot; \gamma); \gamma \in \Gamma\}$$

be some set of easy-to-handle distributions for which we aim to find the element $q(\cdot;\widehat{\gamma})$ which has the smallest distance to the true posterior distribution $p(\theta \mid y)$. As distance, we use the Kullback-Leibler divergence as in Sect. 3, but now we replace the order of the densities. Hence, we define

$$q(\cdot;\widehat{\gamma}) = \arg\min_{q\in\mathcal{Q}} \Big\{ \mathrm{KL}\big(q(\cdot;\gamma), p(\cdot \mid y)\big) \Big\}$$

with

$$\mathrm{KL}\big(q(\cdot;\gamma), p(\cdot \mid y)\big) = \int \log \tfrac{q(\theta;\gamma)}{p(\theta|y)} q(\theta;\gamma)\mathrm{d}\theta.$$

Note that the integral is with respect to the density $q(\theta;\gamma)$, which will be utilized later. With $p(\theta \mid y) = f(y;\theta)p(\theta) \,/\, f(y)$, we can rewrite the above to

$$\mathrm{KL}\big(q(\cdot;\gamma),\, p(\cdot \mid y)\big) = -\int \log \frac{f(y;\theta)p(\theta)}{q(\theta;\gamma)} q(\theta;\gamma)\mathrm{d}\theta + \underbrace{\int \log f(y) q(\theta;\gamma)\mathrm{d}\theta}_{=\log f(y)}. \tag{9.23}$$

The second component in (9.23) is typically unknown, since the marginal distribution of $f(y)$ requires unfeasible integration. The first component, however, contains only analytically available components. Both the likelihood $f(y;\theta)$ as well as the prior $p(\theta)$ are known. Moreover, $q(\cdot;\gamma)$ is chosen to be "simple," meaning that we can easily calculate $q(\theta;\gamma)$. The main idea behind variational Bayes can be seen from (9.23). Minimizing the Kullback-Leibler divergence is equivalent to maximizing the so-called lower bound

$$\mathrm{LB}\big(q(\cdot;\gamma)\big) = \int \log \tfrac{f(y;\theta)p(\theta)}{q(\theta;\gamma)} q(\theta;\gamma)\mathrm{d}\theta \tag{9.24}$$

with respect to γ.

As mentioned before, the class of distributions $\mathcal{Q}$ is chosen such that calculations are feasible, so that the lower bound $\mathrm{LB}(\cdot)$ can either be calculated analytically or obtained through simulations. This also applies to the derivative, which we aim to set to zero. We get

$$\begin{aligned}\tfrac{\partial \mathrm{LB}(q(\cdot;\gamma))}{\partial\gamma} &= \int \log f(y;\theta)p(\theta) \tfrac{\partial q(\theta;\gamma)}{\partial\gamma}\mathrm{d}\theta - \int \tfrac{\partial \log q(\theta;\gamma)}{\partial\gamma} q(\theta;\gamma)\mathrm{d}\theta \\ &\quad - \int \log q(\theta;\gamma) \tfrac{\partial q(\theta;\gamma)}{\partial\gamma}\mathrm{d}\theta.\end{aligned}$$

The second component above is zero for Fisher regular distributions; see (4.5). The remaining two components can be rewritten to

$$\tfrac{\partial \mathrm{LB}(q(\cdot;\gamma))}{\partial\gamma} = E_{q(\cdot;\gamma)}\left\{ \tfrac{\partial \log q(\theta;\gamma)}{\partial\gamma} \Big[\log\big(f(y;\theta)p(\theta)\big) - \log\big(q(\theta;\gamma)\big) \Big] \right\}$$

where the expectation is taken with respect to the density $q(\cdot;\gamma)$. The expectation can be replaced by simulated values, i.e., for given hyperparameters $\gamma^{(t)}$; we draw

$$\theta^{*(b)} \sim q(\cdot;\gamma^{(t)}) \text{ for } b = 1,\ldots,B$$

and calculate the estimate of the derivative through

$$\begin{aligned}\frac{\partial \widehat{\text{LB}(q(\cdot;\gamma^{(t)}))}}{\partial\gamma} &= \frac{1}{B}\sum_{b=1}^{B}\left\{\frac{\partial \log q(\theta^{*(b)};\gamma^{(t)})}{\partial\gamma}\left[\log\left(f(y;\theta^{*(b)})p(\theta^{*(b)})\right)\right.\right.\\ &\left.\left. - \log\left(q(\theta^{*(b)};\gamma^{(t)})\right)\right]\right\}\end{aligned} \tag{9.25}$$

Following standard numerical optimization techniques, one updates $\gamma^{(t)}$ through

$$\gamma^{(t+1)} = \gamma^{(t)} + r^{(t)}\frac{\partial \widehat{\text{LB}(q(\cdot;\gamma^{(t)}))}}{\partial\gamma} \tag{9.26}$$

with $r^{(t)}$ some (possibly vector-valued) learning rate. We refer exemplarily to Kingma and Welling (2014) and Kucukelbir et al. (2017). We emphasize that iteration (9.26) is in practice more complicated than it looks, but one can use standard results from optimization, such as stochastic gradient ascent leading to stochastic variational inference; see, e.g., Hoffman et al. (2013) or Blei et al. (2017).

Example 9.5 We apply the idea to the example from above on the zero-inflated count data. For given π and λ, the distribution (likelihood) of Y is given in (9.16). For π, we assume a flat beta distribution, i.e.,

$$p(\pi) = \begin{cases}1, & \text{for } \pi \in [0,1]\\ 0, & \text{otherwise}\end{cases}$$

and for λ, we assume a "flat" exponential family, i.e.,

$$\lambda \sim p(\lambda;\omega) = \omega\exp\left(-\lambda\omega\right)$$

with $\omega = 0.01$. As before, we look at transformed parameters, i.e.,
$\eta = \text{logit}(\pi) = \log\left(\pi\,/\,(1-\pi)\right)$ and $\delta = \log\left(\lambda\right)$ with resulting prior distributions given in (9.20) and (9.21). We aim to approximate the posterior through the product of two normal distributions, i.e.,

$$\begin{aligned}q(\eta,\delta;\gamma) = &\underbrace{\text{N}(\eta;\mu_\eta,\sigma_\eta^2)}_{=:\,q_1(\eta;\mu_\eta,\sigma_\eta^2)} * \underbrace{\text{N}(\delta;\mu_\delta,\sigma_\delta^2)}_{=:\,q_2(\delta;\mu_\delta,\sigma_\delta^2)}.\end{aligned}$$

The parameters are collected in $\gamma = (\mu_\eta, \log\left(\sigma_\eta^2\right), \mu_\delta, \log\left(\sigma_\delta^2\right))$. We aim to choose γ such that the lower bound (9.24) is minimized. To do so, we set $\gamma^{(0)} = (0,0,0,0)$, which leads to the product of two standard normal distributions and set $\gamma^{(t)} = \gamma^{(0)}$ in the following. The derivative $\partial\,\log\,q(\cdot;\gamma)\,/\,\partial\gamma$ results through

$$\begin{aligned}\frac{\partial \log q(\cdot;\gamma)}{\partial\gamma_1} &= \frac{\partial \log q_1(\eta;\mu_\eta,\sigma_\eta^2)}{\partial\mu_\eta} = -\frac{1}{2}\frac{\eta-\mu_\eta}{\sigma_\eta^2}\\ \frac{\partial \log q(\cdot;\gamma)}{\partial\gamma_2} &= \frac{\partial\sigma_\eta^2}{\partial\gamma_2}\frac{\partial \log q_1(\eta;\mu_\eta,\sigma_\eta^2)}{\partial\sigma_\eta^2} = \sigma_\eta^2\left\{-\frac{1}{2\sigma_\eta^2} + \frac{1}{2}\frac{(\eta-\mu_\eta)^2}{\sigma_\eta^4}\right\}\end{aligned}$$

and accordingly for the derivatives with respect to γ_3 and γ_4. We now simulate

$$\begin{aligned}\eta^{*(b)} &\sim \mathrm{N}(\mu_\eta^{(t)}, \sigma_\eta^{2(t)}) \\ \delta^{*(b)} &\sim \mathrm{N}(\mu_\delta^{(t)}, \sigma_\delta^{2(t)})\end{aligned}$$

for $b = 1, \ldots, B$ with B set to 1000. With these simulated values, we can calculate (9.25) from which we then obtain an update $\gamma^{(t+1)}$ through (9.26). The step size is chosen to be small, i.e., $r^{(t)} = 0.001$, and we only update the $\gamma^{(t)}$ in its steepest direction. Apparently, there is ample space for numerical improvement, but we aim to keep it simple to demonstrate the idea. The final estimate is denoted by $\widehat{\gamma}$, and we can plot the resulting density $q(\cdot; \widehat{\gamma})$ which we visualize in Fig. 9.4 for λ and π. ▷

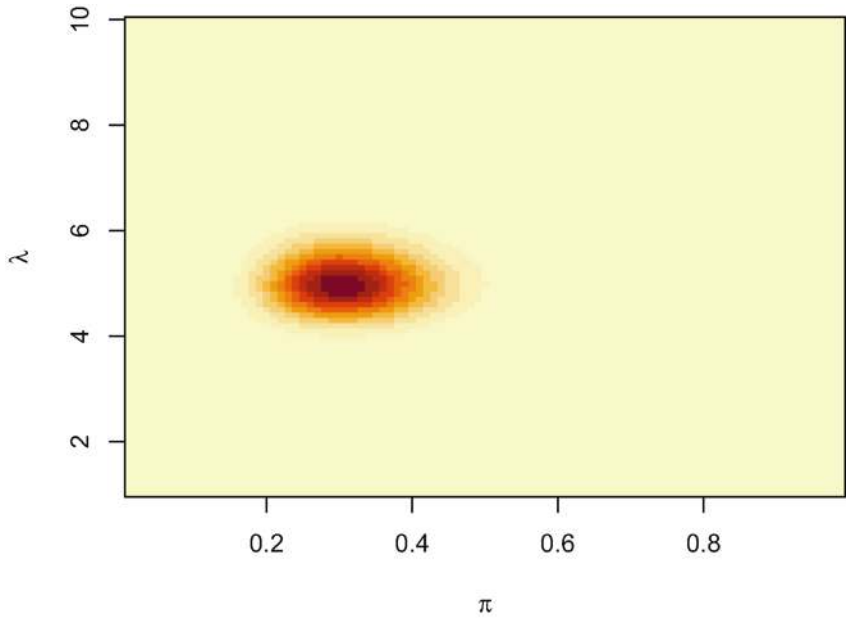


Fig. 9.4 Variational approximation of the posterior, based on the product of two normal distributions for $\mathrm{logit}(\pi)$ and $\log(\lambda)$

Variational Bayes methods have become more attractive over the last years, also because they scale better to high-dimensional parameters. Apparently, one issue remains, which becomes obvious if we look at the Kullback-Leibler divergence (9.23) again. We aim to minimize the first component, which corresponds to maximizing the lower bound (9.24). The second component, though, the distribution of the data $f(y)$ remains unknown. In other words, in general, we have no measure that quantifies how close the approximation $q(\cdot; \gamma)$ is to the true posterior distribution, even though we can derive some bounds. We refer to Fox and Roberts (2012) for more details.

9.7 Bayes and Regularization

Bayesian formulation is today often used for complex models with multiple parameters, where estimation can become unstable due to the high dimension of the parameter θ. In such settings, we are faced with the fact that the number of

parameters that we need to estimate is (too) large compared to the available data. Let us therefore consider the *p* dimensional parameter $\theta = (\theta_1, \theta_2, \ldots, \theta_p)$, where *p* is large relative to sample size *n*. This means we have $p \approx n$ or even $p > n$. In the latter case, classical likelihood approaches fail. But Bayesian principles appear as numerically stable alternatives. To demonstrate the idea, we look at a constructive example with a small sample, which is in the example below, or order $n = 1$ or $n = 2$. Proper data-driven statistical inference is hardly possible with such small samples. The example here serves as a demonstration to visualize how the likelihood behaves if $n < p$ or $n \approx p$.

Example 9.6 We look at the situation $Y_i \sim \mathrm{N}(\mu, \sigma^2)$ with $i.i.d.$ samples, and we set $n = 1$. Hence, there are two parameters that need to be estimated based on a simple observation. The log likelihood results in

$$l(\mu, \sigma^2) = \tfrac{1}{2} \log(\sigma^2) - \tfrac{1}{2} \tfrac{(y_1 - \mu)^2}{\sigma^2}.$$

We plot the likelihood (not the log likelihood) for a single observation in Fig. 9.5, top left plot. We can see, and it is also easy to show that for $n = 1$, the maximum of the likelihood is achieved for $\widehat{\mu} = y_1$ and $\widehat{\sigma}^2 = 0$. This is not a useful result, since we obtain a degenerated distribution with variance 0. In other words, likelihood estimation fails if *p*, the dimension of the parameter, is larger than *n*, the data size.

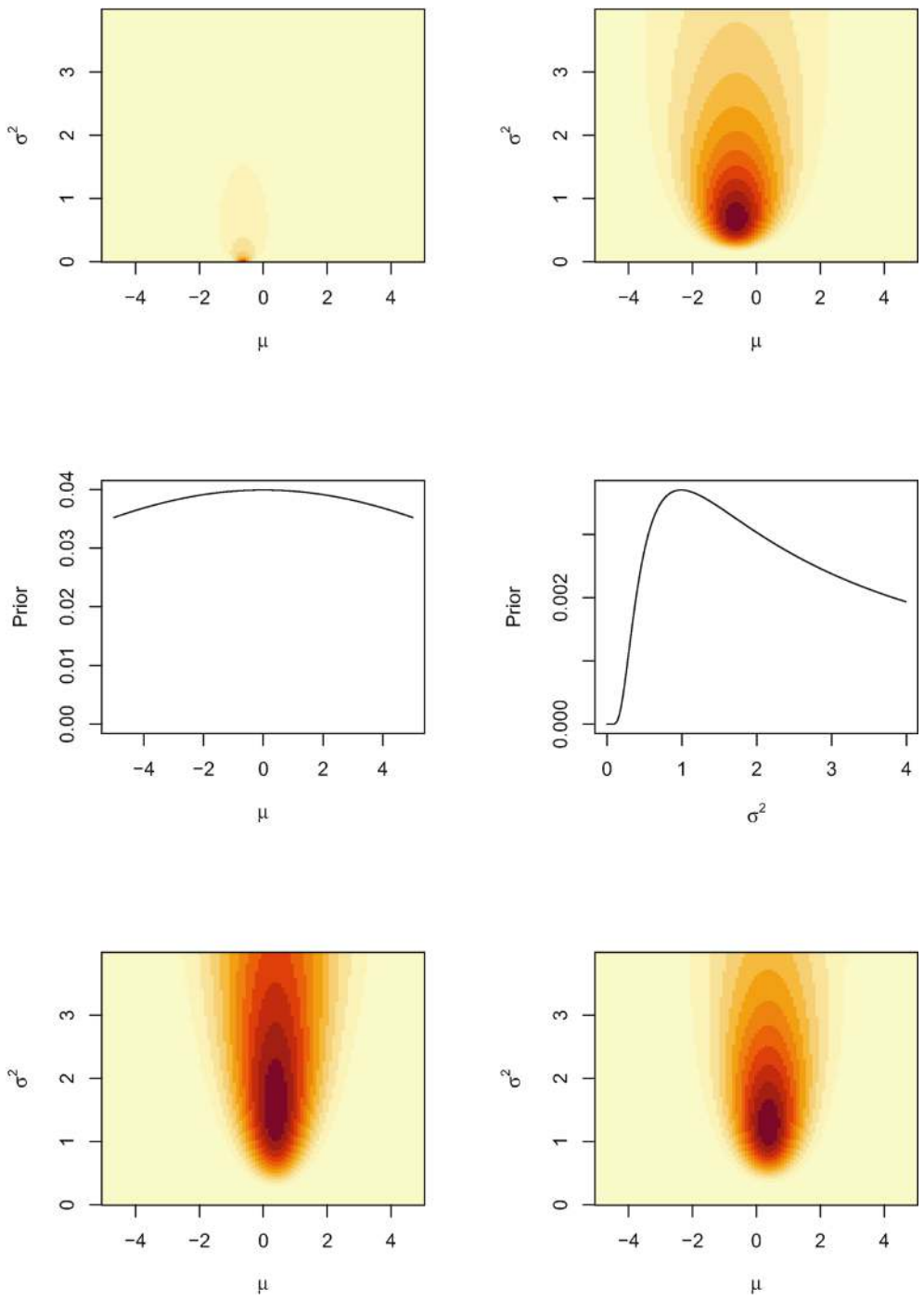


Fig. 9.5 Likelihood function for $n = 1$ (left) and $n = 2$ (right) for normally distributed data (top row). Prior structure for μ and σ^2 (middle row) and resulting posterior distribution for $n = 1$ (left) and $n = 2$ (right) shown in bottom row

We therefore assume weak informative prior distributions for both the mean value μ and the variance σ^2, as shown in the middle plot of Fig. 9.5. For the mean value μ, we assume a normal prior with large variance ($\sigma_{mu} = 10$), and for σ^2, we use an inverse gamma distribution with a large range of values. Note that we plot the

prior in the same range as the likelihood in the upper left plot. This means that we show a small fraction of the prior for μ only, and in fact the prior 95% probability range for μ ranges from -20 to 20. In other words, the priors are taken to be rather uninformative.

Using these priors, we obtain the posterior distribution shown in the bottom left plot. The function stabilizes and the maximum, which would correspond to the maximum posterior mode, is clearly exposed, even for a situation where $p > n$. We also show the resulting function for $n = p = 2$, visualized in the top right and bottom right plots. While there is a substantial difference in the pure likelihood (top row), the plots show less discrepancy for the posterior distribution given in the bottom row. ▷

The idea of stabilizing the likelihood can be written down formally and is also known as penalized likelihood. We refer to Scheipl et al. (2013) or Kauermann et al. (2009) for further reading. Formally, we define the penalized log likelihood as

$$l_p(\theta, \omega) = l(\theta) + \log\left(p(\theta; \omega)\right), \tag{9.27}$$

where ω are some hyperparameters. The maximum penalized likelihood, corresponding to the posterior mode, is achieved by setting

$$0 = \frac{\partial l_p(\theta,\omega)}{\partial \theta} = \frac{\partial l(\theta)}{\partial \theta} + \frac{\partial \log(p(\theta;\omega))}{\partial \theta}. \tag{9.28}$$

The remaining question is the specification of the hyperparameters ω. In principle, these should be set such that the prior is uninformative. Sometimes, it is also possible to estimate the parameters, but that requires some more simplified distribution model. We refrain from going into more details here and refer to the above-cited literature.

9.8 Exercises

Exercise 1 A large home appliances company plans to conduct a market study to find out whether people are interested in the new Toast-O-Matic 9000—a revolutionary toaster. In the preplanning phase, the internal departments have already considered this issue. The estimates from four departments for the proportion of people interested in the toaster are available:

$$\begin{aligned} \text{Finance } (D_1): &\quad P_{D1} &&= 0.31 \\ \text{Product Development } (D_2): &\quad P_{D2} &&= 0.50 \\ \text{Research } (D_3): &\quad P_{D3} &&= 0.35 \\ \text{Marketing } (D_4): &\quad P_{D4} &&= 0.70 \end{aligned}$$

The departments have varying track records, with some being right more often than others. In particular, the departments are correct with the following prior probabilities:

$$P(D_1) = 0.4$$
$$P(D_2) = 0.1$$
$$P(D_3) = 0.3$$
$$P(D_4) = 0.2$$

(a) An initial survey E1 showed that out of 20 people, 7 were interested in the product. Which department does the data speak in favor of after the first survey?

(b) In the final E2 survey, 71 out of 200 respondents were interested. Does the best department change?

Exercise 2 Let $X_1, \ldots, X_n$ be an i.i.d. sample from a geometric distribution with parameter π, i.e., $X_i \sim \text{Geom}(\pi),\ i = 1, \ldots, n$.

(a) We assume a beta distribution as prior for the unknown parameter π; i.e., it holds $X_i \mid \pi \sim \text{Geom}(\pi)$ and $\pi \sim \text{Beta}(\alpha, \beta)$. Show that the beta distribution is a conjugate prior for the geometric likelihood by computing the posterior distribution.

Hint: The density of a beta distribution $\text{Beta}(\alpha, \beta)$ is given by $f(x; \alpha, \beta) = \frac{1}{c} x^{\alpha-1}(1-x)^{\beta-1}$, where *c* is a constant chosen in a way such that the density integrates up to one. The expectation of a $\text{Beta}(\alpha, \beta)$ distributed random variable *X* is $E(X) = \frac{\alpha}{\alpha+\beta}$.

(b) Compute the posterior mean of π.

Exercise 3 Consider a discrete random variable *Y* with three different states (H), (R), and (W) with given probabilities:

$$\pi_H = P(H) = 0.3, \quad \pi_R = P(R) = 0.1, \quad \pi_W = P(W) = 0.6.$$

We will use the Metropolis (Hastings) algorithm to construct a Markov chain $\{Y_t^*,\ t \in \mathbb{R}_0^+\}$ with stationary distribution $\pi = (\pi_H, \pi_R, \pi_W)$. We assume a symmetric proposal density.

The formula for the transition probability in the discrete case, i.e., the probability of going from state Y_t^* to Y_{t+1}^*, is given by

$$\begin{aligned} p_{ij} &= P(Y_{t+1}^* = j | Y_t^* = i) = P(Y_1^* = j | Y_0^* = i) \\ &= \begin{cases} q_{ij}\alpha_{ij} & i \neq j, \\ 1 - \sum_{k \in S \setminus \{i\}} q_{ik}\alpha_{ik} & i = j, \end{cases} \quad i, j \in S, \end{aligned}$$

where α_{ij} is the probability that we accept for $Y_t^* = i$ a simulated value $Y_{t+1}^* = j$ and $q_{ij} = q(i,j)$ is the proposal probability for *j* given state *i*.

The following matrix of proposal densities *Q* is given:

$$Q = \begin{pmatrix} 1-2q & q & q \\ q & 1-2q & q \\ q & q & 1-2q \end{pmatrix}, \quad q \in (0, 0.5].$$

(a) Compute the acceptance matrix *A*, where each of the nine entries α_{ij} is probability of the chain moving from state Y_t^* to Y_{t+1}^* given that the latter state Y_{t+1}^* was already proposed.

(b) Using the matrix *A* just obtained, compute the transition matrix *P*, i.e., the matrix containing the nine probabilities of the chain moving from state Y_t^* to Y_{t+1}^* given that the latter state Y_{t+1}^* was not yet proposed.

(c) Show that π is an invariant distribution for *P*.

(d) Write down in pseudo-code a Metropolis (Hastings) algorithm that constructs a Markov chain with stationary distribution π.

References

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, FRS communicated by Mr. Price, in a letter to John Canton, AMFR S. *Philosophical Transactions of the Royal Society of London,* (53), 370–418.

Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association,112*(518), 859–877.
[MathSciNet][https://doi.org/10.1080/01621459.2017.1285773]

Dale, A. I. (1982). Bayes or laplace? An examination of the origin and early applications of Bayes' theorem. *Archive for History of Exact Sciences, 27*, 23–47.
[MathSciNet][https://doi.org/10.1007/BF00348352]

de Laplace, P. S. (1820). *Théorie analytique des probabilités* (Vol. 7). Courcier.

Fox, C. W., & Roberts, S. J. (2012). A tutorial on variational Bayesian inference. *Artificial Intelligence Review,38*, 85–95.
[https://doi.org/10.1007/s10462-011-9236-8]

Hoffman, M. D., Blei, D. M., Wang, C., & Paisley, J. (2013). Stochastic variational inference. *Journal of Machine Learning Research, 14*, 1303–1347.
[MathSciNet]

Kauermann, G., Krivobokova, T., & Fahrmeir, L. (2009). Some asymptotic results on generalized penalized spline smoothing. *Journal of the Royal Statistical Society Series B: Statistical Methodology,71*(2), 487–503.
[MathSciNet][https://doi.org/10.1111/j.1467-9868.2008.00691.x]

Kingma, D. P., & Welling, M. (2014). Auto-encoding variational Bayes. In Bengio & Lecun (Eds.), *2nd international*

conference on learning representations.

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., & Blei, D. M. (2017). Automatic differentiation variational inference. *Journal of Machine Learning Research*, *18*, 1–45.
[MathSciNet]

Marin, J.-M., Pudlo, P., Robert, C. P., & Ryder, R. J. (2012). Approximate bayesian computational methods. *Statistics and Computing,22*(6), 1167–1180.
[MathSciNet][https://doi.org/10.1007/s11222-011-9288-2]

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics, 12*, 1151–1172.
[MathSciNet][https://doi.org/10.1214/aos/1176346785]

Scheipl, F., Kneib, T., & Fahrmeir, L. (2013). Penalized likelihood and bayesian function selection in regression models. *AStA Advances in Statistical Analysis,97*(4), 349–385.
[MathSciNet][https://doi.org/10.1007/s10182-013-0211-3]

Stigler, S. M. (1975). Studies in the history of probability and statistics. xxxiv: Napoleonic statistics: The work of laplace. *Biometrika,62*(2), 503–517.
[MathSciNet]

Stigler, S. M. (2018). Richard price, the first Bayesian. *Statistical Science,33*(1), 117–125.
[MathSciNet][https://doi.org/10.1214/17-STS635]

Tavaré, S., Balding, D. J., Griffiths, R. C., & Donnelly, P. (1997). Inferring coalescence times from dna sequence data. *Genetics,145*(2), 505–518.
[https://doi.org/10.1093/genetics/145.2.505]

van de Schoot, R., Depaoli, S., King, R., Kramer, B., Märtens, K., Tadesse, M. G., Vannucci, M., Gelman, A., Veen, D., Willemsen, J., et al. (2021). Bayesian statistics and modelling. *Nature Reviews Methods Primers,1*(1), 1.
[https://doi.org/10.1038/s43586-020-00001-2]

Wood, S. N. (2017). *Generalized additive models: An introduction with R*. CRC Press.
[https://doi.org/10.1201/9781315370279]

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
G. Kauermann et al., *Statistics—Coping with Uncertainty*, Textbooks in Statistical Science
https://doi.org/10.1007/978-3-032-33658-3_10

10. Inference in Extreme Data

Göran Kauermann[1] ✉, Giacomo De Nicola[2], Cornelius Fritz[3] and David Rügamer[4]
(1) Department of Statistics, LMU Munich, Munich, Germany
(2) Harvard T.H. Chan School of Public Health, Harvard University, Boston, MA, USA
(3) School for Computer Science and Statistics, Trinity College Dublin, Dublin, Ireland
(4) Munich Uncertainty Quantification AI Lab, LMU Munich, Munich, Germany

Our world is profoundly shaped by extreme events. Financial markets occasionally experience crashes that reshape global economies. Rare technological breakthroughs, such as the advent of the internet or artificial intelligence, fundamentally transform societies. Global pandemics disrupt health systems, labor markets, and political institutions. Although such events occur infrequently, their consequences are often disproportionate and long-lasting. In addition to these economic, technological, and societal shocks, we are witnessing an increasing frequency and severity of extreme weather events, including prolonged droughts, intense rainfall, and exceptional river water levels. These developments further highlight the importance of understanding rare but high-impact occurrences. This chapter presents the main ideas for dealing with data that arise in the tails of a distribution, where rare events reside. Because such events occur only seldom, statistical inference and the quantification of uncertainty cannot rely on classical asymptotic tools such as the central

limit theorem. Instead, specialized methods are required. The tools and principles underlying these methods are the focus of this chapter.

10.1 Rare Events

The influential book “The Black Swan” (Taleb, 2007) discusses and highlights events that are individually highly improbable and difficult to predict, yet capable of exerting large and lasting effects on the systems under study. While the probability of any particular such event may be extremely small, the probability that some rare event (among the many possible ones) occurs over a sufficiently long time horizon, or across a sufficiently large number of units, can be substantial. In this sense, rarity at the microlevel does not preclude near inevitability at the macro level. Examples include extreme environmental episodes such as exceptional river levels or severe heat waves, as well as very infrequent technological or industrial accidents such as nuclear incidents or plane crashes, but also positive events, such as revolutionary scientific discoveries. In Sect. 10.2, we turn to statistical methods designed for modeling extreme values, that is, unusually large or small realizations in the tails of a distribution. In contrast, the present chapter addresses situations in which the event of interest occurs only infrequently, so that empirical samples contain very few, or even no, observed occurrences, and classical asymptotic approximations become unreliable.

We motivate the idea with a simple example. Assume we are interested in the proportion π of students in a certain high school who are addicted to drugs. We draw a sample of n students and ask whether each student is addicted to drugs ($Y = 1$) or not ($Y = 0$). For simplicity, we assume that students answer truthfully and correctly.[1] It is reasonable to assume that this probability is nonzero, but quite low. In a small sample of size $n = 10$ students, it might well happen that not a single student has answered that they are addicted to drugs, that is, $y_i = 0$ for all $i = 1, \ldots, n$. In this context, inference methods derived in

Sect. 4 break down. The log-likelihood function, assuming a binomial distribution, takes the form

$$l(\pi) = 10 \log(1 - \pi), \tag{10.1}$$

which is visualized in the left-hand plot of Fig. 10.1. The maximum is achieved at $\widehat{\pi} = 0$. Note that the first-order derivative at the maximum of likelihood is unequal to zero, which contradicts the derivations from Sect. 4.2. Were we to blindly follow the asymptotic results presented in Sect. 4.2, our inference would yield useless results. With $\widehat{\pi} = 0$, the resulting variance of the estimate is zero as well, which implies that $\pi = 0$. In most realistic scenarios, this is not the case. Just because we have not observed drug-addicted students does not mean that there are no people that are addicted to drugs in the entire population of students.

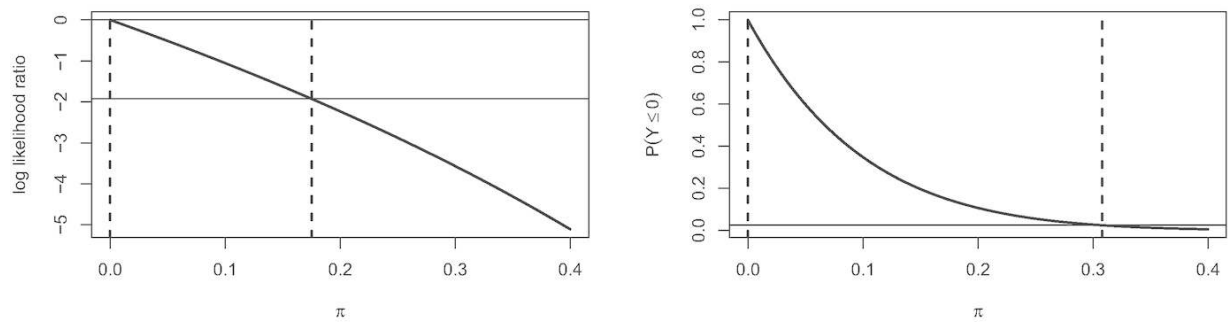


Fig. 10.1 Confidence intervals based on (void) asymptotic Chi-squared approximation (left) and exact calculation (right-hand side) for binomial data with $n = 10$ and $y = 0$

The proportion π is likely small, but our current tools cannot address how small. As a first idea, let's study the log likelihood plotted on the left-hand side in Fig. 10.1. If we use the results derived for the likelihood ratio, we would state by Eq. (4.21) that

$$2[l(\widehat{\pi}) - l(\pi)] \sim \mathscr{X}_1^2.$$

Hence, any test on $H_0 : \pi = \pi_0$ would be rejected if

$$l(\widehat{\pi}) - l(\pi) > \frac{\mathscr{X}^2_{1,1-\alpha}}{2},$$

with $\mathscr{X}^2_{1,1-\alpha}$ being the $1 - \alpha$ quantile of a Chi-squared distribution for the significance level $\alpha \in (0, 1)$. Using the equivalence between testing

and confidence intervals (see Sect. 5.6) and considering the shape of the likelihood above (10.1) implies the following confidence interval

$$\left[0, \exp\left(\frac{\chi^2_{1,1-\alpha}}{2n}\right)\right],$$

visualized as a vertical line on the left-hand side in Fig. 10.1. We should highlight that the interval lacks any mathematical justification and, in fact, is too small. In other words, the asymptotic arguments derived in Sect. 4.2 do not hold since $\partial l(\hat{\theta}) \Big/ \partial\theta \neq 0$. This means we have to derive our inferential tools on different grounds.

First derived by Clopper and Pearson (1934), the shape of the Binomial distribution allows us to calculate exact confidence intervals and hence to draw valid statistical inference even for rare events. Let us assume that, for *n* trials, we observe $y_i = 0$ for all $i = 1, \ldots, n$. We can calculate the probability of observing $Y = 0$ under a binomial distribution with given *n* and varying π. We plot this curve for $n = 10$ on the right-hand side of Fig. 10.1, where the threshold value $\alpha \,/\, 2$ is also indicated. For the confidence interval, we therefore want to find π_r such that

$$P(Y \leq 0; \pi_r) = \tfrac{\alpha}{2}$$

holds. We can interpret this as the right-hand side of a confidence interval (upper value). The lower bound in this case is 0, since π cannot be smaller than zero.

The above approach can be developed into a general construction principle for different values of *y*. For $Y = y$, the left-hand side of the confidence interval is

$$c_l = \begin{cases} 0 & \text{for } y = 0 \\ \pi_l \text{ with } P(Y \geq y; \pi_l) = \frac{\alpha}{2} & \text{else} \end{cases}$$

and the upper end (right-hand side) of the interval is

$$c_r = \begin{cases} 1 & \text{for } y = n \\ \pi_r \text{ with } P(Y \leq y; \pi_r) = \frac{\alpha}{2} & \text{else} \end{cases}$$

The resulting interval $[c_l, c_r]$ is called an *exact confidence interval*, which is shown for different values of n and different outcomes of y / n in Fig. 10.2. We also refer to Somerville and Brown (2013) and Thulin (2014) for further details and a discussion of alternative approaches.

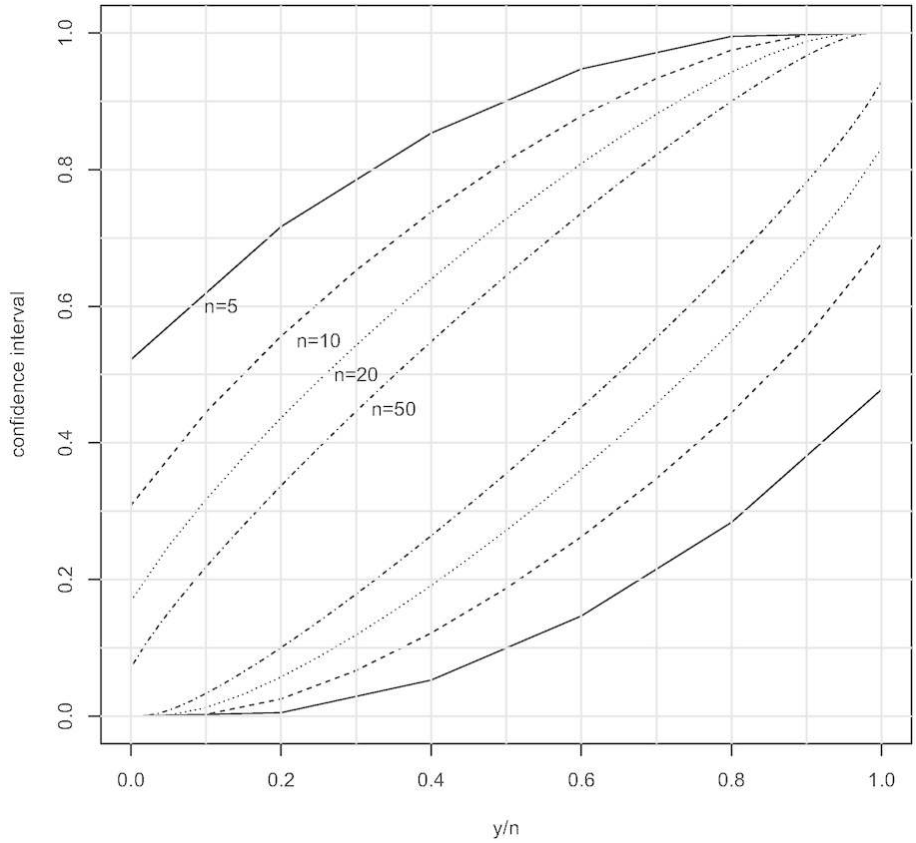


Fig. 10.2 Exact confidence intervals based on Clopper-Pearson approach

If n increases and y is bounded away from 0 and n, we find the exact intervals to be comparable to the approximate versions derived through asymptotic arguments from Sect. 4.

This approach can also be applied to count outcomes as follows: Consider the case with n independent Poisson-distributed random variables

$$Y_i \sim \text{Poisson}(\lambda) \; i.i.d..$$

For rare events, the intensity parameter λ is typically small. In the most extreme case, no events are observed in the sample, that is, $y_i = 0$ for $i = 1, \ldots, n$. Though this gives a clear indication that λ is small, it does

not imply $\lambda = 0$, which would be the value of the maximum likelihood estimate. Hence, we want to derive a confidence interval for λ without relying on the asymptotic arguments given in Sect. 4. Before doing so, let us first look at the likelihood function. We obtain

$$L(\lambda) \;\; = \prod_{i=1}^{n} \frac{\lambda^{y_i}}{y_i!} \, \exp\left(-\lambda\right) \;\propto \quad \lambda^{\sum_{i=1}^{n} y_i} \, \exp\left(-n\lambda\right).$$

This shows that $Y = \sum_{i=1}^{n} Y_i$ is Poisson distributed with parameter $n\lambda$, so that we can derive the confidence interval for the single observation $Y \sim \text{Poisson}(\lambda_n)$ with $\lambda_n = n\lambda$. If the focus is on obtaining a confidence interval for λ instead of λ_n, we just need to divide the interval limits by *n*.

Exact confidence intervals for the Poisson distribution date back to Garwood (1936). In line with the confidence interval derived for the binomial setting, the left-hand side of the interval for λ_n results through

$$c_l = \begin{cases} 0 & \text{for } Y = 0 \\ \lambda_l \text{ with } P(Y \geq y; \lambda_n = \lambda_l) = \frac{\alpha}{2} & \text{else} \end{cases}$$

and the upper right-hand side end equals

$$c_r = \lambda_r \text{ with } P(Y \leq y; \lambda_n = \lambda_r) = \frac{\alpha}{2}.$$

We plot the resulting intervals in Fig. 10.3. Note that one needs to divide the values by *n* in the case of $i.i.d.$ samples with intensity λ.

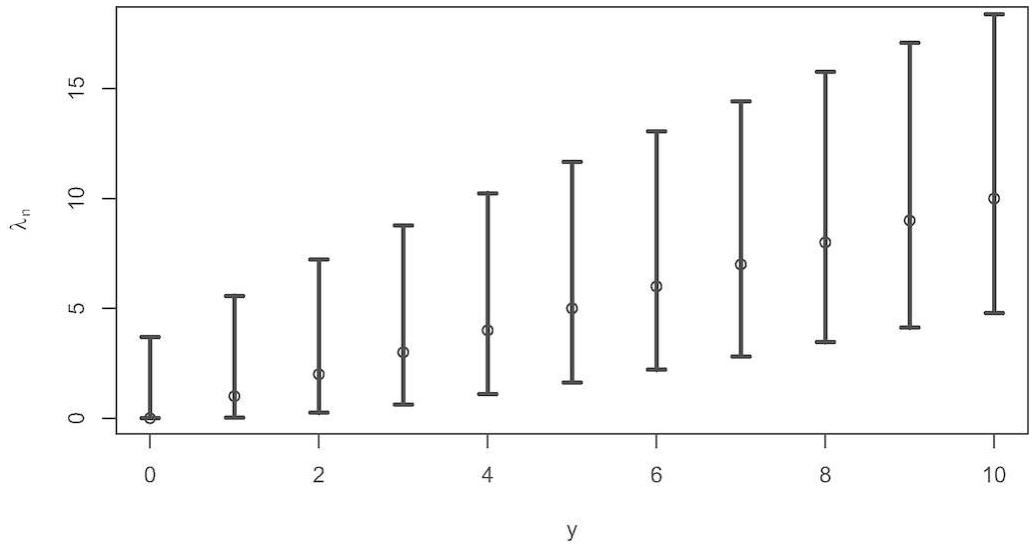


Fig. 10.3 Exact confidence intervals for Poisson distribution

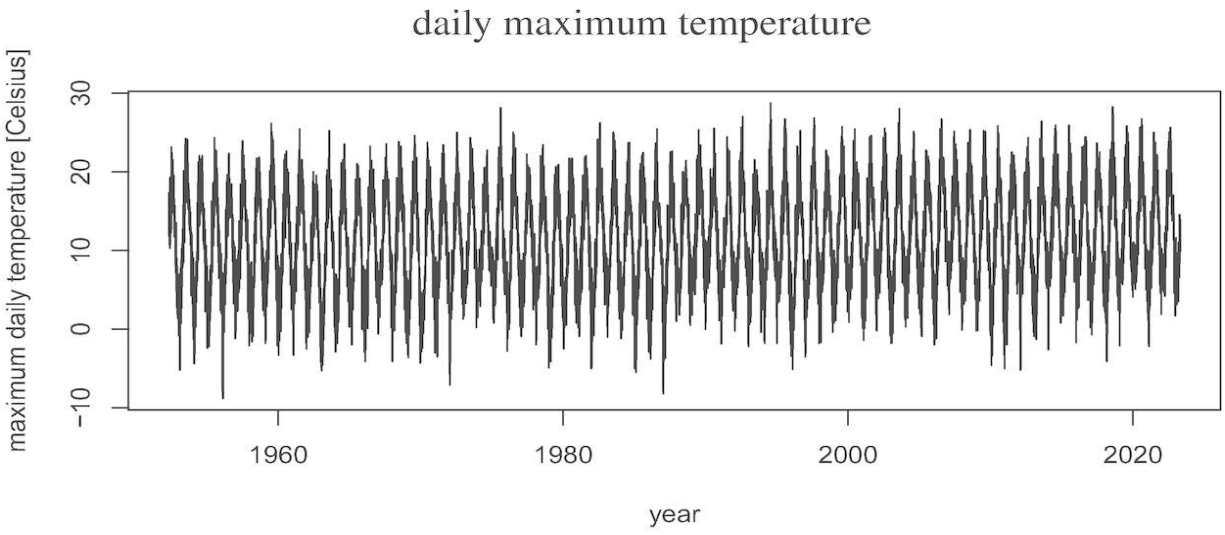
We conclude that, for both the Poisson and the binomial distributions, valid confidence intervals can be constructed even when no events are observed. The construction of exact confidence intervals is somewhat more involved and less intuitive than in the asymptotic setting. However, modern statistical software readily implements these calculations, so this poses no practical difficulty.

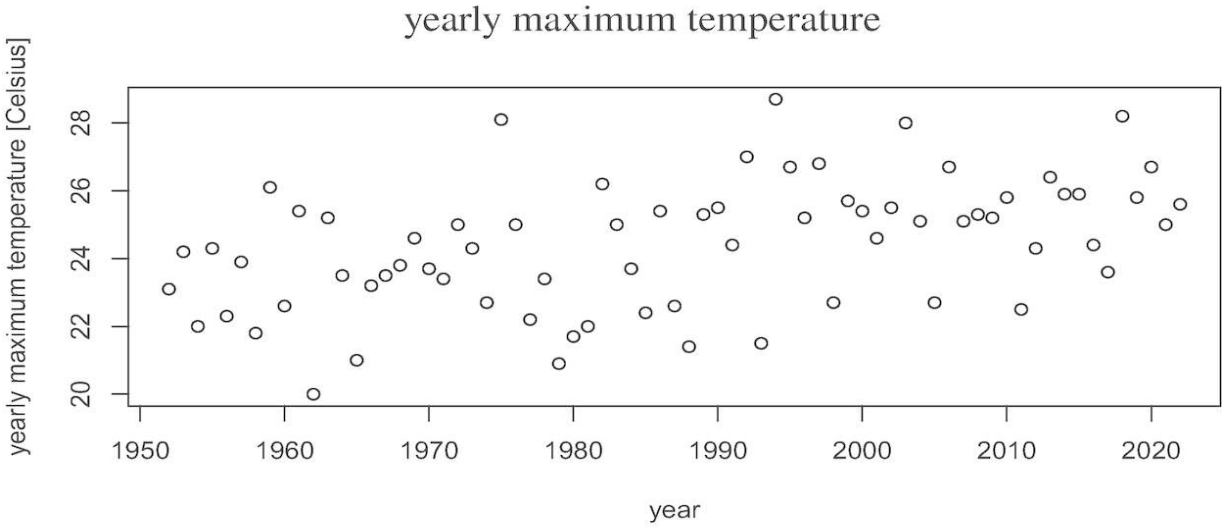
10.2 Extreme Data

In the previous section, we considered rare events. We now turn to extreme events, that is, unusually large or small realizations in the tails of a distribution. Examples include exceptionally high river water levels or unusually high temperatures.

Example 10.1 The upper panel of Fig. 10.4 shows the daily maximum temperature recorded on Helgoland, a German island in the North Sea, in the last years. The annual variation is clearly visible. First, we calculate the yearly maximum of the daily maximum temperatures, shown in the middle panel. While an overall upward trend is visible, we do not aim to discuss this at this point. Second, we compute, for each year, the minimum of the daily maximum temperatures, displayed in the bottom panel. The horizontal zero line marks the freezing point;

values above this line correspond to years without a single day of continuous frost. In this section, our objective is to identify an appropriate distributional model for these data. ▷


daily maximum temperature
maximum daily temperature [Celsius]
−10
0
10
20
30
1960
1980
2000
2020
year


yearly maximum temperature
yearly maximum temperature [Celsius]
20
22
24
26
28
1950
1960
1970
1980
1990
2000
2010
2020
year

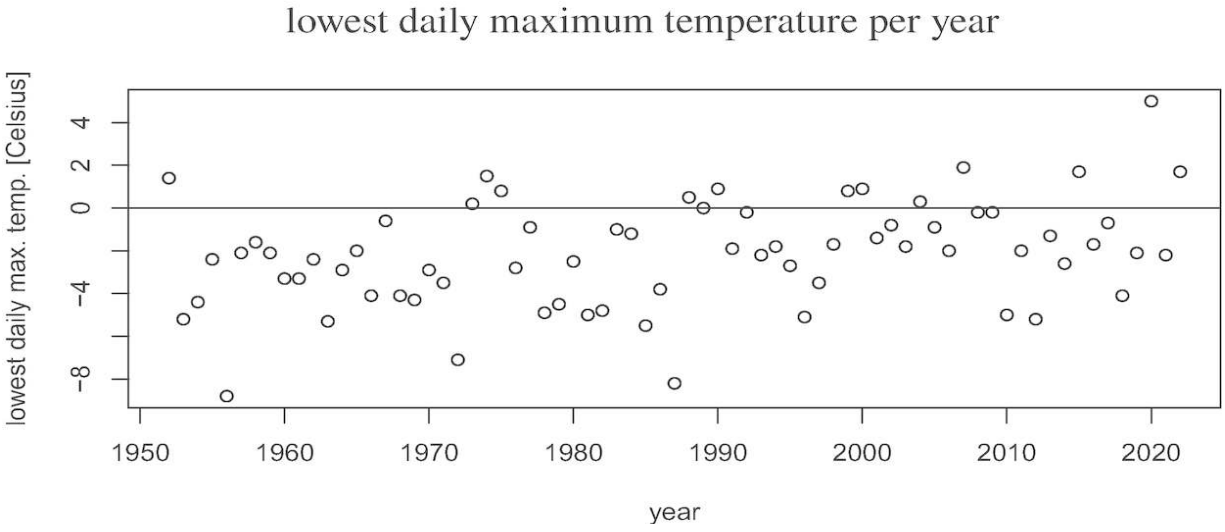

lowest daily maximum temperature per year
lowest daily max. temp. [Celsius]
−8
−4
0
2
4
1950
1960
1970
1980
1990
2000
2010
2020
year

Fig. 10.4 Temperature on the island Helgoland in the North Sea

Formally, we are interested in the maximum (or minimum) of data in the previous example. We, therefore, assume the classical setup

$$Y_i \sim F(\cdot)\ i.i.d.,\ i = 1, \ldots, n,$$

but we want to study the distribution of

$$Y_{\max,n} = \max\{Y_1, \ldots, Y_n\}.$$

It is not difficult to derive the distribution function of $Y_{\max}$, since based on the $i.i.d.$ assumption, we get

$$\begin{aligned} F_{Y_{\max,\mathrm{n}}}(y) &= P(Y_{\max,n} \leq y) \\ &= P(Y_1 \leq y, \ldots, Y_n \leq y) \\ &= P(Y_1 \leq y) \cdots P(Y_n \leq y) \\ &= F^n(y). \end{aligned}$$

Considering the limit for $n \to \infty$, we do not obtain a useful limit distribution: Assuming the support of $F(\cdot)$ is bounded, that is, there exists an overall maximum value *c*, such that

$$c = \sup\{y : F(y) < 1\},$$

it follows that

$$\lim_{n\to\infty} F_{Y_{\max,n}}(y) = \begin{cases} 0 & \text{for } y < c \\ 1 & \text{for } y \geq c \end{cases}.$$

In other words, the limit distribution converges to a one-point distribution, and with infinite data, the maximum value $Y_{\max}$ will equal *c*. If, in contrast, the support is unbounded, then we will always obtain larger maximum values as *n* increases, such that $F_{Y_{\max,n}}(y)$ becomes degenerate. We visualize this in Fig. 10.5, where we plot the maximum of $n = 10^k$ simulated $\mathrm{N}(0, 1)$ variables. With *k* increasing, the maximum increases and goes to ∞ with increasing *n*. Due to these observations, we need to normalize the data in some reasonable way. Let a_n and b_n be sequences that can depend on the sample size, such that

$$\frac{Y_{\max,n}-b_n}{a_n}$$

converges to some non-degenerated distribution with $n \to \infty$. Hence,

$$\begin{aligned}\lim_{n\to\infty} P\left(\frac{Y_{\max,n}-b_n}{a_n} \leq y\right) &= \lim_{n\to\infty} P(Y_{\max,n} \leq a_n y + b_n) \\ &= \lim_{n\to\infty} F^n(a_n y + b_n) = H(y)\end{aligned}$$

where $y \in \mathbb{R}$ and H is some non-degenerated distribution. The latter is called **extreme value distribution**, and is further specified in the subsequent statement.

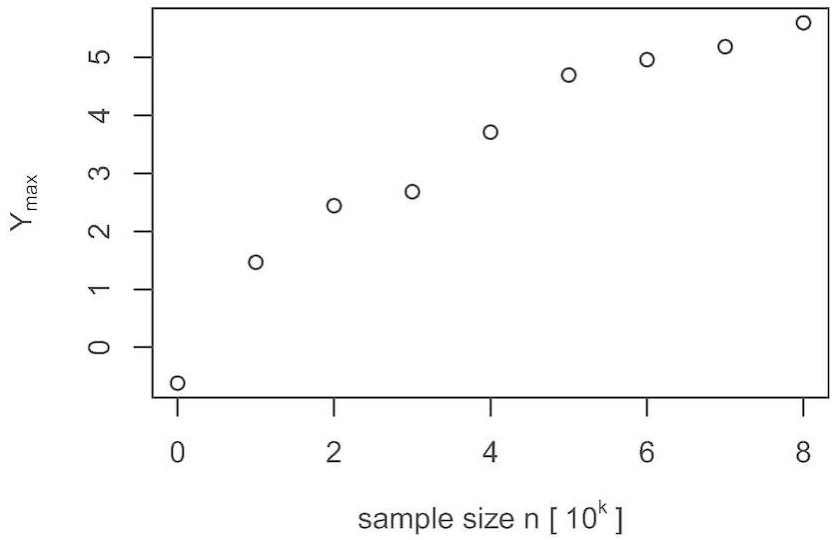


Fig. 10.5 Maximum of n standard normally distributed random variables with $n = 10^k$

Property (Extreme Value Distributions) *Let* $\{Y_i\}_{i=1}^{\infty}$ *be* $i.i.d.$ *random variables and* $Y_{\max,n} := \max(Y_1, \ldots, Y_n)$. *If there are sequences of real numbers* $\{a_n\}_{n=1}^{\infty}, \{b_n\}_{n=1}^{\infty}$ *with*

$$\frac{Y_{\max,n}-b_n}{a_n} \xrightarrow{d} Z \text{ as } n \to \infty, \tag{10.2}$$

then Z *follows an extreme value distribution, i.e.,* $Y \sim H(c)$. *The extreme value distribution* $H(\cdot)$ *takes one of the following three forms:*

1. *Gumbel distribution:*

$$H_1(z) = \exp(-\exp(-z))$$

2. *(Inverted) Weibull distribution:*

$$H_2(z) = \begin{cases} \exp\left(-(-z)^{1/k}\right) & for\ z \leq 0 \\ 1 & for\ z > 0 \end{cases}$$

3. *Fréchet Pareto distribution:*

$$H_3(z) = \begin{cases} 0 & for\ z < 0 \\ \exp\left(-z^{-1/k}\right) & for\ z \geq 0 \end{cases}$$

We can consider Eq. (10.2) as a limit theorem for the tails in the distribution. In other words, if the focus is not on the center of the distribution but on the tails, be it maxima or minima, there is a limit theorem and a natural distributional model to work with. A sketch of a proof for Eq. (10.2) is given at the end of this section.

The methodology goes back to Gumbel (1958), a German statistician who was politically persecuted and escaped Nazi Germany during the Second World War by fleeing to the USA. Once there, he studied extreme values and is now regarded as the founding father of extreme value theory. For further details about Gumbel's life and contribution, see Rendtel et al. (2021).

The three distributions $H_1(\cdot)$, $H_2(\cdot)$ and $H_3(\cdot)$ can be jointly written in one format, which is called the generalized extreme value (GEV) distribution.

Definition 10.1 *The* **Generalised Extreme Value** *distribution (GEV) has three parameters, the location parameter* μ*, the scale parameter* σ*, and the shape parameter* γ*. The distribution function is given by*

$$H(z) = \begin{cases} \exp\left(-(1+\gamma z)^{-1/\gamma}\right) & for\ \gamma \neq 0 \\ \exp\left(-\exp\left(-z\right)\right) & for\ \gamma = 0 \end{cases} \tag{10.3}$$

with $z = (x - \mu) \,/\, \sigma$.

Looking at (10.3), we can interpret the parameters similar to other distributions. Parameter μ is the location parameter, and σ is the scale

parameter. The shape parameter γ characterizes the type of the distribution. We have

1. $\gamma = 0$ gives the Gumbel distribution $H_1()$;
2. $\gamma > 0$ gives the (inverted) Weibull distribution $H_2()$;
3. $\gamma < 0$ gives the Fréchet-Pareto-distribution $H_3()$.

The parameterization of the GEV opens the possibility of estimating the parameters of the distribution using maximum likelihood, which in turn estimates the resulting type of extreme value distribution. This demonstrates the generality of the extreme value distribution approach and its suitability for the analysis of extreme events. The field is well elaborated, and we only motivated the main ideas in this chapter. We refer to Coles et al. (2001) or Haan and Ferreira (2006) for further details.

We conclude by demonstrating the ideas by continuing the example from above.

Example 10.2 In Example 10.1, we analyzed the maximum yearly temperature. We now take the maximum yearly values shown in the middle plot in Fig. 10.4 and estimate the parameters of the GEV distribution using standard maximum likelihood. This leads to parameter estimates

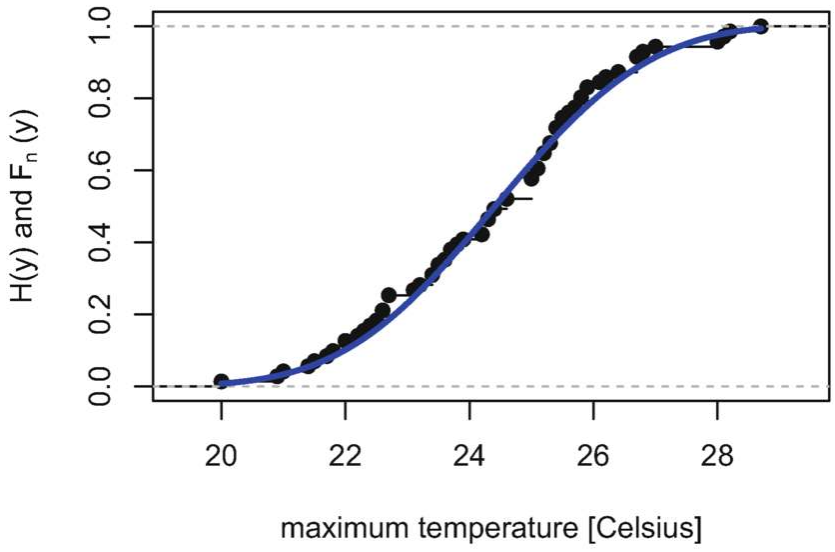
$$\begin{aligned}\widehat{\mu} &= 23.76\ (0.25)\\ \widehat{\sigma} &= 1.88\ (0.17)\\ \widehat{\gamma} &= -0.30\ (0.08)\end{aligned}$$

with standard errors given in brackets. Since the estimate of γ is negative, a Fréchet-Pareto distribution is found in this example. We plot the resulting distribution in Fig. 10.6 (upper plot) and overlay this with the empirical distribution of the yearly maximum temperatures. We see a reasonable concordance between the data and the fitted limit distributions. We can do the same for the lowest maximum

temperature, as shown in the bottom plot of Fig. 10.6. The corresponding estimates (and standard errors) are

$$\begin{aligned}\widehat{\mu} &= -0.25\ (0.05)\\ \widehat{\sigma} &= 2.50\ (0.21)\\ \widehat{\gamma} &= -3.06\ (0.31).\end{aligned}$$

▷



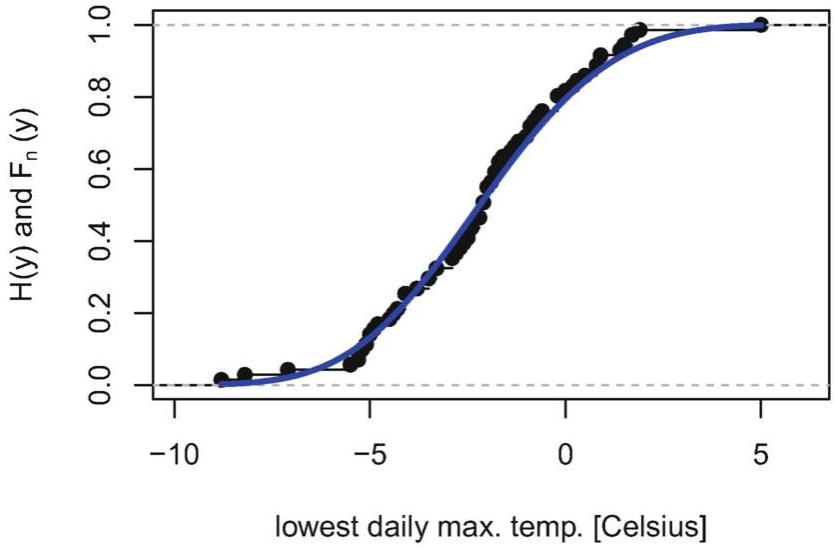


Fig. 10.6 Distribution of yearly maximum temperature and fitted extreme value distribution (upper plot). Distribution of lowest daily maximum temperature and fitted extreme value distribution (bottom plot)

We can now return to the question of whether the data are identically distributed, that is, whether the maximum yearly temperatures in the early years have the same distribution as in the most recent years. We therefore split the data into the years before 1985 and the years after 1985. We show the resulting empirical distribution functions and the fitted distributions in Fig. 10.6. A clear shift toward the right is striking. We list the parameter estimates and their standard errors in Table 10.1. The shift in the location μ is significant and in the order of $2\ ^{\circ}\mathrm{C}$ (Fig. 10.7).

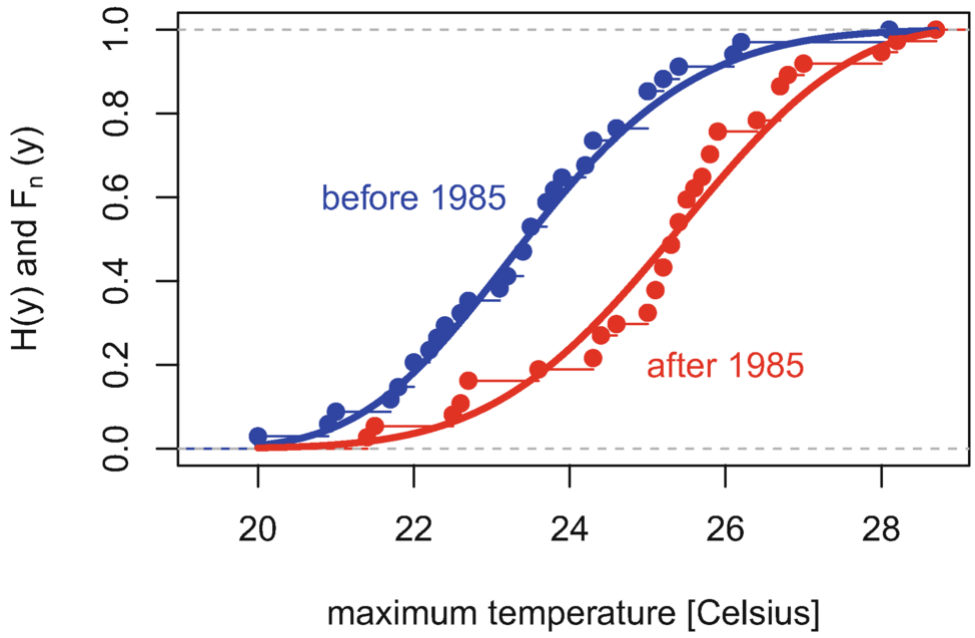


Fig. 10.7 Distribution of yearly maximum temperature and fitted extreme value distribution for years before and after 1985

Table 10.1 Parameter estimates for the two time periods

Proof For a sketch of the proof of Eq. (10.2), it will be helpful to consider $Y_{\max,n}$ as the maximum of data in matrix form. Let therefore Y be written with a double index in matrix form

$$\begin{matrix} Y_{1,1} & Y_{1,2} & \dots & Y_{1,K} \\ Y_{2,1} & Y_{2,2} & \dots & Y_{2,K} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{T,1} & Y_{T,2} & \dots & Y_{T,K} \end{matrix} \tag{10.4}$$

where $n = T \cdot K$. We can then define

$$Y_{t,\max,K} = \max\{Y_{t,1}, \dots, Y_{t,K}\}$$

as the row maxima. The overall maximum is then

$$Y_{\max,n} = \max\{Y_{t,\max,K}\}.$$

With this notation, we have

$$Y_{\max,n} = \max_{1 \le t \le T} \max_{1 \le k \le K} (Y_{tk}) = \max_{1 \le t \le T} \{Y_{t,\max,K}\}. \tag{10.5}$$

If there is a standardizing sequence (a_n, b_n) such that $(Y_{\max,n} - b_n) \,/\, a_n$ converges to a distribution Z, we can write the sequence in double notation (a_{LK}, b_{LK}) such that $(Y_{\max,TK} - b_{LK}) \,/\, a_{LK}$ also converges to Z, for $T \to \infty$ and $K \to \infty$. The same needs to hold for the row maxima, such that each standardized row maximum converges to the same limit. Hence, there is a sequence (a_K, b_K) such that $(Y_{t,\max,K} - b_K) \,/\, a_K$ converges to Z for $K \to \infty$ and all $1 \le t \le T$.

Assuming independence, we can assume $Y_{t,\max,K} \sim \mathrm{F}_{Y_{\max,K}}$; i.e., all row maxima have the same distribution. Using (10.5), we get

$$F_{Y_{\max,TK}}(y) = F^T_{Y_{\max,K}}(y)$$

where $F_{Y_{\max,TK}}(.)$ is the distribution function of $Y_{\max,TK}$. Note that both sequences converge to the same distribution function for a random variable *Z* after appropriate standardization. Applying the standardization again, we can derive the following property for the distribution function of variable Z

$$F_Z(b_T + a_T z) = [F_Z(z)]^T. \tag{10.6}$$

It can now be shown that this property holds for the above extreme value distributions. We show this exemplary for the Gumbel distribution. Setting $a_T = 1$ and $b_T = \log\,(T)$, we get

$$\begin{aligned} H_1(y - \log(T)) &= \exp(-\exp(-(y - \log(T)))) \\ &= \exp(-\exp(-y)\, T) \\ &= \exp(-\exp(-y))^T = H_1(y)^T \end{aligned}$$

Similarly, it can be shown for the other extreme value distributions, which concludes the proof. □

10.3 Exercises

Exercise 1 The following table summarizes the number of major cybersecurity breaches reported by a random sample of 20 technology firms over the past year, with each firm indicated by a Greek letter. A "major breach" is defined as an incident leading to either loss of customer data or operational downtime longer than 24 hours.

Table

(a) Assume the number of breaches per firm follows a Poisson distribution with intensity λ. Derive the log-likelihood function for λ based on the above data.

(b) Find the maximum likelihood estimate $\widehat{\lambda}$. Interpret its meaning in the cybersecurity context.

(c) Construct a 95% exact confidence interval for λ. Comment on why exact intervals are preferred over asymptotic ones in this rare-event setting.

(d) Estimate the probability that a randomly chosen firm experiences at least one major breach in a year.

(e) Suppose a regulator wants to know: what is the probability that at least 3 firms out of 50 experience a breach in a given year? Compute this probability using your estimated model.

Exercise 2 The following table, taken from the classic "Danish Fire Insurance" dataset, contains data regarding the yearly maximal inter-arrival times (measured in days) of insurance claims, between the years 1980 and 1990 (de Carvalho & Marques, 2012). The inter-arrival time is defined as the number of days passing between two subsequent claims.

Table

Assume that in each year, the inter-arrival times are exponentially distributed with intensity $\lambda > 0$ and that the yearly maximal inter-arrival times are *i.i.d.*

(a) Show that the maximum of an *i.i.d.* sequence of exponentially distributed random variables $\{X_1, \ldots, X_n\}$, with intensity $\lambda > 0$, can be approximated with a Gumbel distribution. How do the location and scale parameters of the resulting Gumbel distribution relate to λ and n?

Hint: use the convergence properties for $\max\{X_1, \ldots, X_n\} - \frac{\log n}{\lambda}$.

(b) Use MLE (derive formulas as far as possible) to estimate the extreme value distribution associated with the yearly maximum inter-arrival times.

(c) Based on your estimation of the parameters, evaluate the probability for the yearly maximum inter-arrival time to exceed 25 days.

References

Clopper, C. J., & Pearson, E. S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika,26*(4), 404–413. [https://doi.org/10.1093/biomet/26.4.404]

Coles, S., Bawa, J., Trenner, L., & Dorazio, P. (2001). *An introduction to statistical modeling of extreme values* (Vol. 208). Springer.

de Carvalho, M., & Marques, F. (2012). Jackknife Euclidean likelihood-based inference for spearman's RHO. *North American Actuarial Journal,16*(4), 487–492. [MathSciNet][https://doi.org/10.1080/10920277.2012.10597644]

Garwood, F. (1936). Fiducial limits for the poisson distribution. *Biometrika,28*(3/4), 437–442. [https://doi.org/10.2307/2333958]

Gumbel, E. J. (1958). *Statistics of extremes*. Columbia University Press. [https://doi.org/10.7312/gumb92958]

Haan, L., & Ferreira, A. (2006). *Extreme value theory: An introduction* (Vol. 3). Springer.

Rendtel, U., Wasmuht, U. C., & Wilrich, P.-T. (2021). Emil Julius Gumbel: An innovative statistician and a dedicated pacifist. *AStA Wirtschafts-und Sozialstatistisches Archiv,15*, 273–291. [https://doi.org/10.1007/s11943-021-00293-0]

Somerville, M. C., & Brown, R. S. (2013). Exact likelihood ratio and score confidence intervals for the binomial proportion. *Pharmaceutical Statistics,12*(3), 120–128. [https://doi.org/10.1002/pst.1560]

Taleb, N. N. (2007). *The Black Swan : The impact of the highly improbable*. Random House.

Thulin, M. (2014). The cost of using exact confidence intervals for a binomial proportion. *Electronic Journal of Statistics,8*(1), 817–840. [MathSciNet][https://doi.org/10.1214/14-EJS909]

van den Hout, A., & van der Heijden P. G. (2002). Randomized response, statistical disclosure control and misclassificatio: A review. *International Statistical Review,70*(2), 269–288.

Footnotes

1 Otherwise, we would need to rely on randomized response experiments, which we want to ignore here (see van den Hout & van der Heijden, 2002).

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
G. Kauermann et al., *Statistics—Coping with Uncertainty*, Textbooks in Statistical Science
https://doi.org/10.1007/978-3-032-33658-3_11

11. Multivariate Data

Göran Kauermann[1] ✉, Giacomo De Nicola[2], Cornelius Fritz[3] and David Rügamer[4]
(1) Department of Statistics, LMU Munich, Munich, Germany
(2) Harvard T.H. Chan School of Public Health, Harvard University, Boston, MA, USA
(3) School for Computer Science and Statistics, Trinity College Dublin, Dublin, Ireland
(4) Munich Uncertainty Quantification AI Lab, LMU Munich, Munich, Germany

In this chapter, we study multivariate data, where we assume that y has multiple components that may interact with one another. We do not impose a causal order, such as the first component influencing the second, the first two influencing the third, and so on. Instead, we treat the variables as being on an equal footing, meaning that their relationships are reciprocal rather than directional. Our goal is to introduce methods for quantifying associations in multivariate settings, in particular beyond simple bivariate correlations. We will revisit the multivariate normal distribution, but we begin with discrete-valued multivariate data, which are often summarized in contingency tables.

11.1 Models for Contingency Tables

We now consider multivariate data and let $Y = (Y_1, \ldots, Y_p)$ be a p-dimensional random vector. We begin with discrete-valued variables and assume that $Y_j \in \{1, \ldots, K_j\}$ for $j = 1, \ldots, p$. Hence, Y takes values in a p-dimensional cube with side lengths $K_1, K_2, \ldots, K_p$. We visualize a $4 \times 2 \times 3$-dimensional cube in Fig. 11.1. The outcome $Y = (k_1, k_2, k_3)$ indicates that Y lies in the cell (k_1, k_2, k_3).

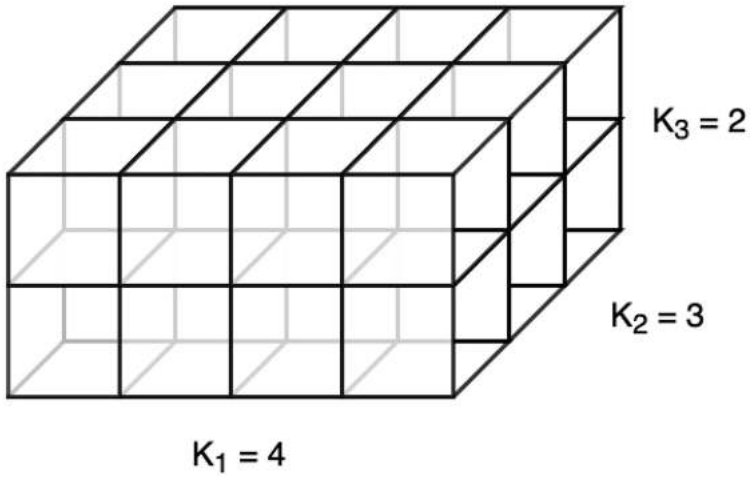


Fig. 11.1 Three-dimensional contingency table represented as 3D object

With multiple multivariate observations, notational ambiguities arise. We need an index for the components of Y as well as an index for the observation number. This would naturally lead to double subscripts, which make expressions cumbersome. In addition, higher-order interactions among the variables require further indexing. To improve readability, we therefore adopt a modified notation in this chapter. Specifically, we denote the observation index by a superscript in parentheses. Thus,

$Y^{(i)} = (Y_1^{(i)}, \ldots, Y_p^{(i)})$ represents the *p*-dimensional random vector of the *i*-th observation. The data are recorded as independent multivariate observations $Y^{(1)}, \ldots, Y^{(n)}$. With this notational convention, we can express the cell counts as

$$n_{k_1 k_2 \ldots k_p} = \sum_{i=1}^{n} 1\{y^{(i)} = (k_1, k_2, \ldots, k_p)\}.$$

We also need counts for the margins of the table, which we denote by replacing the index with a dot, that is, for instance:

$$n_{k_1 k_2 \ldots k_{p-1},\cdot} = \sum_{i=1}^{n} 1\{y_1^{(i)} = k_1\} 1\{y_2^{(i)} = k_2\} \cdots 1\{y_{p-1}^{(i)} = k_{p-1}\}$$

or

$$n_{k_1 \cdots \cdot} = \sum_{i=1}^{n} 1\{\underset{1}{y}^{(i)} = k_1\}.$$

Accordingly,

$$n = n_{\cdots \cdot} = \sum_{i=1}^{n} 1.$$

This corresponds to a multivariate *p*-dimensional contingency table, an example of which is given in the following.

Example 11.1 Consider data from a survey of $n = 3185$ rented households in the city of Munich, regarding rent and equipment of the apartment. To be specific, we consider the categorical variables

- Y_1: floor space (< 50 squared meters, between 50 and 100 squared meters, > 100 squared meters)
- Y_2: center (yes or no)
- Y_3: new contract (yes or no)
- Y_4: district (out of 25 districts of Munich)

This leads to a $3 \times 2 \times 2 \times 25$ dimensional contingency table with resulting 300 cells. It does not make much sense to look at the massive contingency table, which also contains structural zeros, since some districts are central, while others are noncentral. It is more feasible to look at the margins of the table; that is, ignore some of the variables. For instance, looking at Y_1 and Y_2 only gives Table 11.1. Alternatively, we may look at the contract and the different districts, which are shown in Table 11.2, for space reasons, just for 6 out of the 25 districts. It becomes apparent that tabulating multivariate discrete valued variables is clumsy and does not provide any insight if the dimension exceeds the order of 2 or 3. ▷

Table 11.1 Munich rental guide survey, tabulated with respect to floor space and center

Table 11.2 Munich rental guide survey, tabulated with respect to new contract and district

We want to define a general probability model for such kind of data, which in turn will also extend to continuous settings as well as more complex data structures like network data discussed later in the book.

Definition 11.1 *Let $Y = (Y_1, \ldots, Y_p)$ be a discrete-valued random variable. The distribution of Y follows a* **log-linear model** *if*

$$P(Y = y; \theta) \propto \exp\left(\theta_0 + \sum_j s_j(y)\theta_j\right) \tag{11.1}$$

where $s_j(\cdot)$ are some linearly independent statistics calculated from $y = (y_1, \ldots, y_p)$ and $\theta = (\theta_0, \theta_1, \ldots)$ is the parameter vector.

This is a very general setup, which we now specify further for clarity. We begin with the case $p = 2$, which corresponds to a square contingency table. In this setting, it is convenient to introduce a double index for the

statistics in order to distinguish between main effects and interaction terms. Specifically, we define the statistics as

$$\begin{aligned} s_{i_1\cdot}(y) &= 1\{y_1 = i_1\} \text{ for } 1 \le i_1 < K_1 \\ s_{\cdot i_2}(y) &= 1\{y_2 = i_2\} \text{ for } 1 \le i_2 < K_2 \\ s_{i_1 i_2}(y) &= 1\{y_1 = i_1, y_2 = i_2\} \text{ for } 1 \le i_1 < K_1 \text{ and } 1 \le i_2 < K_2. \end{aligned}$$

In other words, $s_{i_1\cdot}(\cdot)$ is an indicator for the first variable that equals 1 if $y_1 = i_1$. Similarly, $s_{i_1 i_2}(\cdot)$ indicates whether $y_1 = i_1$ and $y_2 = i_2$. Thus, the indices specify which variables are involved. Note that the indices above do not range over all possible outcomes, since we restrict to $i_1 < K_1$ and $i_2 < K_2$. These constraints ensure that the statistics are linearly independent and that the model is identifiable.

Now, assume we have independent observations $Y^{(1)}, \dots Y^{(n)}$. The general form of the log-linear model (11.1) for observations $y^{(1)}, \dots, y^{(n)}$ then becomes

$$\begin{aligned} P(Y^{(1)} &= y^{(1)}, \dots, Y^{(n)} = y^{(n)}) \\ &\propto \prod_{i=1}^{n} \exp\Big(\theta_0 + \sum_{i_1=1}^{K_1-1} s_{i_1\cdot}(y^{(i)})\theta_{i_1\cdot} + \sum_{i_2=1}^{K_2-1} s_{\cdot i_2}(y^{(i)})\theta_{\cdot i_2} \\ &\quad + \sum_{i_1=1}^{K_1-1}\sum_{i_2=1}^{K_2-1} s_{i_1 i_2}(y^{(i)})\theta_{i_1 i_2}\Big) \\ &= \exp\Big(\theta_0 + \sum_{i_1=1}^{K_1-1} n_{i_1\cdot}\theta_{i_1\cdot} + \sum_{i_2=1}^{K_2-1} n_{\cdot i_2}\theta_{\cdot i_2} + \sum_{i_1=1}^{K_1-1}\sum_{i_2=1}^{K_2-1} n_{i_1 i_2}\theta_{i_1 i_2}\Big), \end{aligned}$$

where

$$n_{i_1\cdot} = \sum_{i=1}^{n} 1\{y^{(i)} = i_1\}$$

is the number of observations with outcome i_1 for variable Y_1, and accordingly $n_{i_1 i_2}$ the cell count in the two dimensional table. Note that

$$n_{i_1\cdot} = \sum_{i_2=1}^{K_2} n_{i_1 i_2},$$

hence, the points in the subscript refer to the variables we sum over.

To simplify the formula, we set $K_1 = 2$ and $K_2 = 2$ leading to a two-by-two table. In this case, we obtain the following likelihood for the data:

$$P(Y^{(1)} = y^{(1)}, \dots, Y^{(n)} = y^{(n)}) \propto \exp\left(\theta_0 + n_{1\cdot}\theta_{1\cdot} + n_{\cdot 1}\theta_{\cdot 1} + n_{11}\theta_{11}\right)$$

If we now set all parameters involving double indices to zero, which, in the case of a 2×2 table, corresponds to a single parameter, the joint distribution simplifies. It factorizes into one part that depends solely on the outcome of Y_1 and another that depends solely on Y_2. As a result, we get that Y_1 and Y_2 are independent.

The log-linear model does not define a proper distribution, since the normalization constant is missing. This can be addressed by introducing a Poisson model for the cell counts. If we assume an $i.i.d.$ sample $Y^{(1)}, \dots, Y^{(n)}$, we can calculate the (random) counts $N_{i_1 i_2} = \sum_{i=1}^{n} 1\{Y_1^{(i)} = i_1, Y_2^{(i)} = i_2\}$. We then assume

$$N_{i_1 i_2} \sim \text{Poisson}(\lambda_{i_1 i_2})$$

where

$$\lambda_{i_1 i_2} = \exp\left(\theta_0 + \theta_{i_1\cdot} + \theta_{\cdot i_2} + \theta_{i_1 i_2}\right)$$

with the side constraint that $\theta_{K_1\cdot} = 0,\ \theta_{\cdot K_2} = 0$ and $\theta_{i_1 K_2} = \theta_{K_1 i_2} = 0$ for all i_1 and i_2.

This leads to the following likelihood for the observed counts $n_{i_1 i_2}$:

$$
\begin{aligned}
&P(N_{i_1 i_2} = n_{i_1 i_2}, 1 \le i_1 \le K_1, 1 \le i_2 \le K_2) \\
&= \prod_{i_1=1}^{K_1} \prod_{i_2=1}^{K_2} \frac{\lambda_{i_1 i_2}^{n_{i_1 i_2}}}{n_{i_1 i_2}!} \exp\left(-\lambda_{i_1 i_2}\right) \\
&\propto \prod_{i_1=1}^{K_1} \prod_{i_2=1}^{K_2} \exp\left(n_{i_1 i_2} \log \lambda_{i_1 i_2} - \lambda_{i_1 i_2}\right) \\
&= \exp\Bigg(n_{\cdot\cdot}\theta_0 + \sum_{i_1=1}^{K_1-1} n_{i_1\cdot}\theta_{i_1\cdot} + \sum_{i_2=1}^{K_2-1} n_{\cdot i_2}\theta_{\cdot i_2} \\
&+ \sum_{i_1=1}^{K_1-1} \sum_{i_2=1}^{K_2-1} n_{i_1 i_2}\theta_{i_1 i_2} - \sum_{i_1=1}^{K_1-1} \sum_{i_2=1}^{K_2-1} \lambda_{i_1 i_2} \Bigg),
\end{aligned}
$$

where $n_{\cdot\cdot}$ is the number of observations. This is a log-linear model. Note that for the Poisson model, the sample size $n_{\cdot\cdot}$ itself is a random variable, which is not necessarily a plausible assumption. Let us therefore condition on $n_{\cdot\cdot} = n$, that is, the sum of the Poisson distributed cell counts $Y_{i_1 i_2}$ is conditioned to be equal to n. It can be shown (see Lang, 1996) that this results in a multinomial model of the form

$$
\begin{aligned}
&P(N_{i_1 i_2} = n_{i_1 i_2}, i_1 = 1, \ldots, K_1, i_2 = 1, \ldots, K_2 \mid n_{\cdot\cdot} = n) \\
&= \binom{n_{\cdot\cdot}}{n_{11} \quad \ldots \quad n_{K_1 K_2}} \prod_{i_1=1}^{K_1} \prod_{i_2=1}^{K_2} \pi_{i_1 i_2}^{n_{i_1 i_2}}
\end{aligned}
$$

with $\pi_{i_1 i_2} = \lambda_{i_1 i_2} \Big/ \sum_{i_1=1}^{K_1} \sum_{i_2=1}^{K_2} \lambda_{i_1 i_2}$. Hence, knowledge of the overall sample size is not informative with respect to the log-linear structure.

Log-linear models become interesting for higher-dimensional contingency tables, that is, for $p > 2$. In this case, the set of possible parameters increases in both the dimension p and in the number of categories. Let us formally write down a log-linear model for a K_1 by K_2 by K_3 by $\ldots$ by K_p contingency table. Referring again to Fig. 11.1, this corresponds to a p-dimensional cube with side lengths $K_1, K_2, K_3, \ldots, K_p$. We denote the random cell counts by $N_{i_1,\ldots,i_p}$ and the observed counts by $n_{i_1,\ldots,i_p}$. We use bullet notation to indicate summation over an index, that is, a dot in a subscript means that the corresponding index is summed over. This leads to the following rather cumbersome expression

$$
\begin{aligned}
&P(N_{i_1,\ldots,i_p} = n_{i_1,\ldots,i_p}, i_1 = 1, \ldots, K_1; i_2 = 1, \ldots, K_2, \ldots) \\
&\propto \exp\Bigg(\theta_0 + \sum_{i_1=1}^{K_1-1} n_{i_1 \cdot\ldots\cdot} \, \theta_{i_1 \cdot\ldots\cdot} + \sum_{i_2=1}^{K_2-1} n_{\cdot i_2 \cdot\ldots\cdot} \, \theta_{\cdot i_2 \cdot\ldots\cdot} + \cdots + \sum_{i_p=1}^{K_p-1} n_{\cdot\ldots\cdot i_p} \, \theta_{\cdot\ldots\cdot i_p} \\
&\quad + \sum_{i_1=1}^{K_1-1} \sum_{i_2=1}^{K_2-1} n_{i_1 i_2 \cdot\ldots\cdot} \, \theta_{i_1 i_2 \cdot\ldots\cdot} + \cdots + \sum_{i_{p-1}=1}^{K_{p-1}-1} \sum_{i_p=1}^{K_p-1} n_{\cdot\ldots\cdot i_{p-1} i_p} \, \theta_{\cdot\ldots\cdot i_{p-1} i_p} \\
&\quad + \cdots \\
&\quad + \sum_{i_1=1}^{K_1-1} \sum_{i_2=1}^{K_2-1} \cdots \sum_{i_p=1}^{K_p-1} n_{i_1 i_2 \ldots i_p} \, \theta_{i_1 i_2 \ldots i_p} \Bigg).
\end{aligned}
$$

Although the formula appears complicated, its structure is clear. The terms involve main effects (depending on a single variable), all pairwise interactions, all three-way interactions, and so on, up to interactions among all p variables. For example, $\theta_{i_1 i_2 \cdot\ldots\cdot}$ represents the pairwise interaction between Y_1 and Y_2 at levels i_1 and i_2, and $\theta_{i_1 i_2 i_3 \cdot\ldots\cdot}$ represents a three-way interaction. Such interaction structures quickly become cumbersome for multivariate discrete data. Our goal is therefore to develop more parsimonious model specifications. Before doing so, we formalize the model in a simpler setting. For convenience, we write 1 for variable Y_1, 2 for Y_2, and so on. We restrict attention to the case $p = 3$ and consider the binary case $K_1 = K_2 = K_3 = 2$. This results in a $2 \times 2 \times 2$ table with eight cells. The full model, also called the saturated model, contains eight parameters: $\theta_0, \theta_{1\cdot\cdot}, \theta_{\cdot 1\cdot}, \theta_{\cdot\cdot 1}, \theta_{11\cdot}, \theta_{1\cdot 1}, \theta_{\cdot 11}$ and θ_{111}. If we condition on the total sample size, then θ_0 is determined by this constraint. We denote the full model, including all interaction terms, by

$$
1 * 2 * 3.
$$

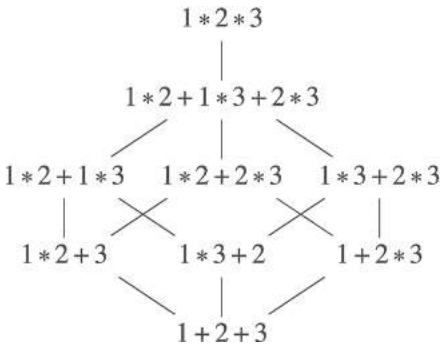
This indicates the model with the three-way interaction between variables Y_1, Y_2 and Y_3 included. If we impose that $\theta_{111} = 0$, that is, we exclude the three-way interaction, then the highest-order interactions are pairwise interactions between Y_1 and Y_2, between Y_1 and Y_3,, as well as between Y_2 and Y_3. Formally, we write this as

$$1 * 2 + 1 * 3 + 2 * 3.$$

If we further restrict $\theta_{11\cdot} = 0$, that is the pairwise interaction between Y_1 and Y_2 should be set to zero, the highest-order interactions are between Y_1 and Y_3 and between Y_2 and Y_3. Consequently, the model can be written as

$$1 * 3 + 2 * 3.$$

We can now further set parameters to zero and just take the highest-order interaction and write this in the given model specification. The cascade of different model leads to a model hierarchy, which is visualized as follows:



The statistical challenge is now to identify an appropriate model for the data in the contingency table. Formally, this can be achieved either by testing whether certain parameters are equal to zero or by applying model selection criteria, such as the Akaike Information Criterion, discussed in Chap. 7. In other words, given observed data, we can use the methods developed so far to determine a suitable—possibly simplified —model.

The approach extends naturally to higher-order contingency tables with $p > 3$. It is also evident that the model hierarchy illustrated above for $p = 3$ becomes even more complex as p increases, which motivates a further simplification of the structure. In fact, some of the models considered above admit an interpretation in terms of conditional independence, which we explore next.

We say that Y_1 and Y_2 are conditionally independent given Y_3 if and only if

$$P(Y_1 = i_1, Y_2 = i_2 \mid Y_3 = i_3) = P(Y_1 = i_1 \mid Y_3 = i_3) P(Y_2 = i_2 \mid Y_3 = i_3)$$

for all $1 \leq i_1 \leq K_1$, $1 \leq i_2 \leq K_2$, and $1 \leq i_3 \leq K_3$. We denote this formally as

$$Y_1 \perp\!\!\!\perp Y_2 \mid Y_3 \text{ or } 1 \perp\!\!\!\perp 2 \mid 3. \tag{11.2}$$

Applying Bayes rule shows

$$P(Y_{i_1} = i_1, Y_{i_2} = i_2 \mid Y_{i_3} = i_3) \propto P(Y_{i_1} = i_1, Y_{i_2} = i_2, Y_{i_3} = i_3)$$

Conditional independence (11.2) therefore implies that the latter component decomposes to the product of two terms dependent on (Y_{i_1}, Y_{i_3}) and (Y_{i_2}, Y_{i_3}). That is, if $Y_1 \perp\!\!\!\perp Y_2 \mid Y_3$, we have

$$\begin{aligned} P(Y_{i_1} = i_1, Y_{i_2} = i_2 \mid Y_{i_3} = i_3) &\propto P(Y_{i_1} = i_1, Y_{i_2} = i_2, Y_{i_3} = i_3) \\ &\propto h_{13}(i_1, i_3) h_{23}(i_2, i_3) \end{aligned}$$

for some functions $h_{13}(\cdot)$ and $h_{23}(\cdot)$ for all possible outcomes i_1, i_2 and i_3.

Considering the structure of the log-linear model, we see that this holds if all parameters that involve components i_1 and i_2 are equal to zero. Putting it differently, we have

$$Y_1 \perp\!\!\!\perp Y_2 \mid Y_3 \Leftrightarrow \theta_{i_1 i_2 \cdot} = 0 \text{ and } \theta_{i_1 i_2 i_3} = 0 \text{ for all } i_1, i_2 \text{ and } i_3.$$

Using the model notation from above, this is the model

$$1 * 3 + 2 * 3.$$

This model implies the conditional independence of Y_1 and Y_2 given Y_3.

The conditional independence also allows for further interpretations. If $Y_1 \perp\!\!\!\perp Y_2 \mid Y_3$ holds, it follows

$$P(Y_1 = i_1 \mid Y_2 = i_2, Y_3 = i_3) = P(Y_1 = i_1 \mid Y_3 = i_3).$$

In other words, given the value of Y_3, additional information about Y_2 does not affect the distribution of Y_1. Hence, Y_2 provides no further information about Y_1 once Y_3 is known. We will examine these information-theoretic aspects in more detail later in this chapter.

We can revisit the model hierarchy above and restrict attention to conditional independence models only. In this case, the hierarchy simplifies, as will be discussed in greater detail in the next section.

Generally, estimation of the parameters can be carried out with maximum likelihood estimation, which is not described in detail here. We refer to Tutz (2011) for technical details and further discussion (see also Agresti, 2012).

Example 11.1 (Continued) We return to the rental guide example from above and run a model selection on the data. Based on the Bayesian Information Criterion, the selected model is

$$\text{floor space} * \text{center} + \text{center} * \text{district} + \text{floor space} * \text{new contract}$$

The model simplifies quite drastically, allowing more straightforward interpretation of the results. ▷

11.2 Graphical Models

We introduced the idea of conditional independence in the previous section; we now formalize it. Let $Y = (Y_1, \ldots, Y_p)$ be a p-dimensional random variable with density or probability mass function $f_{123\cdots p}(y_1, \ldots, y_p)$. We define with A, B and C the following subsets:

$$\left.\begin{array}{l} A \cap B = A \cap C = B \cap C = \emptyset \\ A \cup B \cup C \subset \{1, \ldots, p\} \\ A \neq \emptyset \text{ and } \mathrm{B} \neq \emptyset \end{array}\right\} \tag{11.3}$$

We say that A is conditionally independent of B given C, denoted as

$$A \perp\!\!\!\perp B \mid C$$

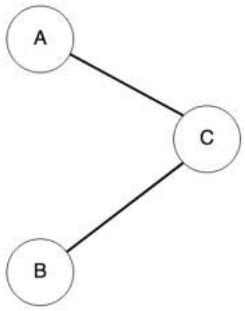
if and only if

$$f_{A \cup B \mid C}(y_A, y_B \mid y_C) = f_{A \mid C}(y_A \mid y_C) f_{B \mid C}(y_B \mid y_C). \tag{11.4}$$

This holds if and only if the density decomposes to

$$f_{A \cup B \cup C}(y_A, y_B, y_C) = h_{A \cup C}(y_A, y_C) h_{B \cup C}(y_B, y_C).$$

Conditional independence can be visualized using a graph, which motivates the term graphical model. To be concrete, we represent the statement that A and B are conditionally independent given C as follows:



In an undirected graphical model, the absence of an edge between A and B indicates conditional independence given the remaining variables (here, C).

The graph above serves as an illustration. We now formalize the concept by considering individual variables. To this end, define a graph G as the pair $G = (V, E)$, where $V = \{1, \ldots, p\}$ denotes the set of nodes and $E \subset V \times V$ the set of edges. We identify V with the variables $Y_1, Y_2, \ldots, Y_p$, but for simplicity, refer to them by their indices. In this context, the edges themselves are less important than the missing edges, as the absence of an edge represents conditional independence. We link these two concepts through

$$i \perp\!\!\!\perp j \mid V \setminus \{i, j\} \Leftrightarrow (i, j) \notin E \text{ and } (j, i) \notin E.$$

In other words, a missing edge in the graph corresponds to conditional independence between the two variables that are not linked. As an example, for $p = 3$, the graph shown in Fig. 11.2 can be interpreted as $1 \perp\!\!\!\perp 2 \mid 3$:

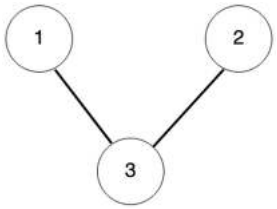


Fig. 11.2 Graphical model expressing the conditional independence $1 \perp\!\!\!\perp 2 \mid 3$

We now turn to higher dimensions. Figure 11.3 displays a graphical model for six variables. There are eight missing edges in the graph, and these correspond to the conditional independence relations listed below:

$$1 \perp\!\!\!\perp 4 \mid \{2, 3, 5, 6\}$$
$$1 \perp\!\!\!\perp 5 \mid \{2, 3, 4, 6\}$$
$$1 \perp\!\!\!\perp 6 \mid \{2, 3, 4, 5\}$$
$$2 \perp\!\!\!\perp 4 \mid \{1, 3, 5, 6\}$$
$$2 \perp\!\!\!\perp 5 \mid \{1, 3, 4, 6\}$$
$$2 \perp\!\!\!\perp 6 \mid \{1, 3, 4, 5\}$$
$$3 \perp\!\!\!\perp 6 \mid \{1, 2, 4, 5\}$$
$$5 \perp\!\!\!\perp 6 \mid \{1, 2, 3, 4\}$$

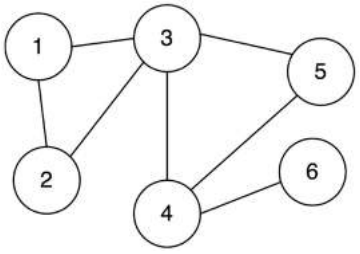


Fig. 11.3 Graphical model for six variables

These statements are also called *pairwise conditional independence*.

Looking at the graph in Fig. 11.3, we can also make other statements, which can be shown to hold equivalently in a conditional independence framework. Note that all paths from the set of nodes $A = \{4, 5, 6\}$ to the set of nodes $B = \{1, 2\}$ go via $C = \{3\}$. This translates to

$$\{1, 2\} \perp\!\!\!\perp \{4, 5, 6\} \mid 3.$$

To prove this, we need to show that the distribution factorizes as

$$f_{123456}(y_1, y_2, y_3, y_4, y_5, y_6) = h_{123}(y_1, y_2, y_3) h_{3456}(y_3, y_4, y_5, y_6).$$

With the eight conditional independence statements above, we have similar decomposition properties, which indeed confirm the above decomposition. This is called *global conditional independence*. The global conditional independence can be formally defined as follows. Consider the three sets of variables A, B and C which fulfill (11.3). If every path from $i \in A$ to $j \in B$ goes over elements in set C, then

$$A \perp\!\!\!\perp B \mid C.$$

For example, in Fig. 11.3, we also have $\{3, 5\} \perp\!\!\!\perp 6 \mid 4$, and there are further global conditional independence statements that can be read from the graph.

Finally, we can consider the properties of the graph locally, by looking at single nodes and their direct neighbors. For instance, variable 2 has neighbors 1 and 3 and hence is not connected to the remaining nodes. This can be formulated to

$$2 \perp\!\!\!\perp \{4, 5, 6\} \mid \{1, 3\}.$$

In general, let $\mathcal{N}(i)$ be the set of neighbors of *i*, that is,

$$\mathcal{N}(i) = \{j : (i, j) \in E \text{ or } (j, i) \in E\}.$$

Then

$$i \perp\!\!\!\perp (V \setminus \{i \cup \mathcal{N}(i)\}) \mid \mathcal{N}(i).$$

This is called *local conditional independence*.

Although the definitions of pairwise, local, and global conditional independence appear conceptually different, and are in fact different in other settings, they are actually equivalent for graphical models.

Property *Given a graphical model, the three conditional independence interpretations pairwise conditional independence, local conditional independence and global conditional independence are equivalent.* □

The proof can, for instance, be found in Whittaker (2009). We also refer to Lauritzen (1996) or Maathuis et al. (2018) for more details.

For given $i.i.d.$ data $Y^{(i)} = (Y_1^{(i)}, \ldots, Y_p^{(i)})$, we can now aim to select a model which has as many conditional independencies as possible based on some model selection criterion. This again induces some model hierarchy, which we show exemplarily for three variables in Fig. 11.4. Apparently, for more than three variables, the set of possible models gets large.

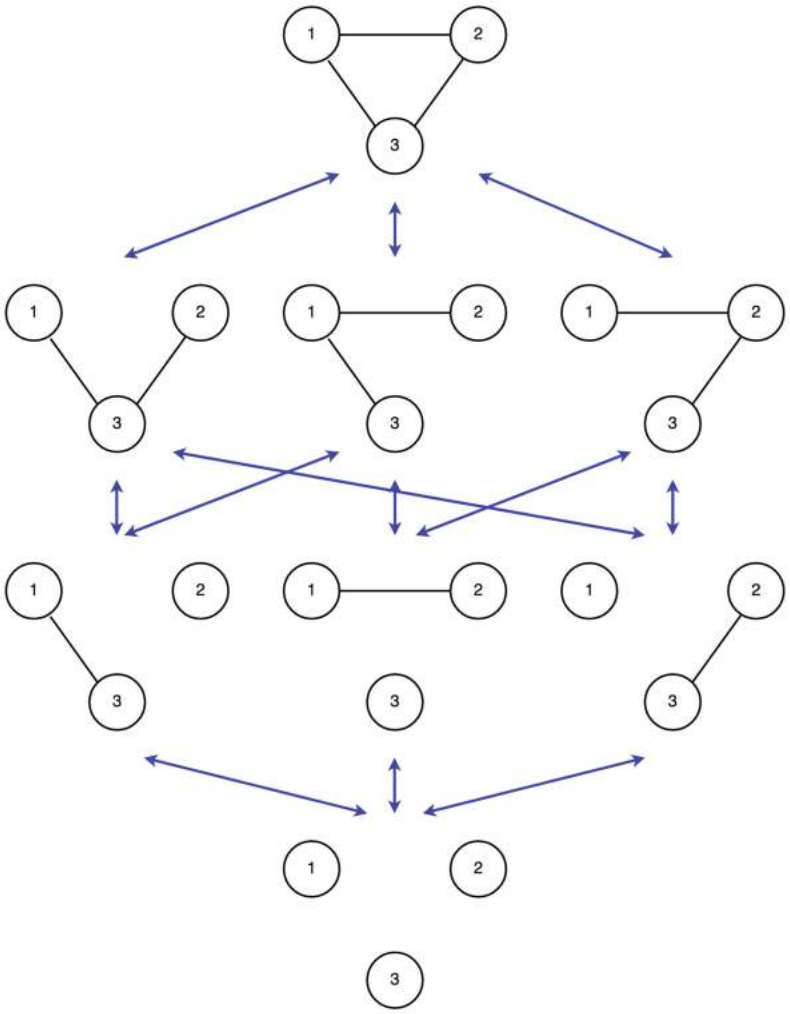


Fig. 11.4 Model hierarchy of graphical models for three variables

Example 11.1 (Continued) We continue with the example of the rental data from above. The model itself becomes quite simple as we only have three interaction terms. This in turn leads to the graphical model shown in Fig. 11.5. We could interpret it as follows. Given that an apartment is in the city center, the explicit location (district) gives no information on rent and contract. Or, given the floor space, the location of the apartment is independent of the contract length. ▷

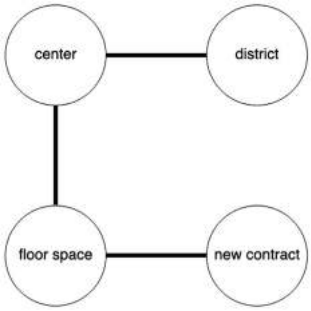
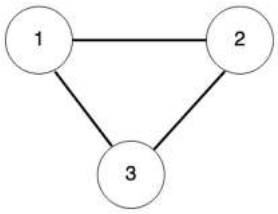


Fig. 11.5 Graphical model rental survey example

Graphical models are a flexible tool, and we utilized the idea already in Chap. 8, when we introduced the simulation of multivariate random variables. Graphs quite easily allow us to visualize the dependence

structure. One should, however, bear in mind that not all sets of conditional independence statements can be visualized with a graph. Assume, for instance, that $1 \perp\!\!\!\perp 2$ and $p = 3$. The corresponding graphical model is



This conclusion follows from the fact that, conditional on Y_3, the variables Y_1 and Y_2 are no longer independent. At first glance, this may seem counterintuitive. Although Y_1 and Y_2 are marginally independent, meaning that knowing the value of one provides no information about the other, this relationship changes once we condition on Y_3. To illustrate this, consider the following example.

Example 11.2 Imagine a technical system consisting of two independent components. Let Y_1 denote the operational status of component 1, where $Y_1 = 1$ if it is functioning and $Y_1 = 0$ otherwise. Similarly, let Y_2 describe the status of component 2 in the same way. We assume that Y_1 and Y_2 are independent Bernoulli random variables, meaning that the functioning of one component has no effect on the other.

However, the entire system operates successfully only if both components are functioning. If either component fails, the system fails. We can model the status of the full system with a third binary variable Y_3 where $Y_3 = 1$ if both components are running (i.e., $Y_1 = Y_2 = 1$) and $Y_3 = 0$ otherwise.

In this setup, even though Y_1 and Y_2 are independent in general, conditioning on the system being operational $Y_3 = 1$ introduces a dependency: Knowing that one component is working increases the likelihood that the other is also working.

This can be modeled formally as follows:

$$\begin{aligned} Y_1 &= \begin{cases} 1 & = \text{ technical failure in component 1} \\ 0 & = \text{normal operation} \end{cases} \\ Y_2 &= \begin{cases} 1 & = \text{ technical failure in component 2} \\ 0 & = \text{normal operation} \end{cases} \\ Y_3 &= \max(Y_1, Y_2) = \begin{cases} 1 & \text{system break down} \\ 0 & \text{system working} \end{cases} \end{aligned}$$

Assuming binomial models leads to

$$\begin{aligned} Y_1 &\sim \mathrm{B}(1, \pi_1) \\ Y_2 &\sim \mathrm{B}(1, \pi_2) \\ Y_3 \mid Y_1, Y_2 &\sim \mathrm{B}(1, \pi_3(y_1, y_2)) \end{aligned}$$

with $\pi_3(y_1, y_2) = \max(Y_1, Y_2)$. Hence π_1 and π_2 are the failure probabilities for components 1 and 2, respectively. Now suppose the technical system has failed, that is, $Y_3 = 1$, and we want to determine which component is responsible for the failure. If we inspect component 1 and find that it is functioning properly, that is $Y_1 = 0$, this leads to an increased probability that the system failure is due to a failure of component 2.

Ignoring variable Y_3 and applying the Bayes rule, we have

$$P(Y_2 = 1 \mid Y_1 = 1) \;\; = \frac{P(Y_1=1, Y_2=1)}{P(Y_1=1)} = \frac{\pi_1 \pi_2}{\pi_1} = \pi_2.$$

Hence we have independence of Y_1 and Y_2. But if we condition on Y_3, we get

$$\begin{aligned} P(Y_1 = 1, Y_2 = 1 \mid Y_3 = 1) &= \frac{P(Y_1=1,Y_2=1,Y_3=1)}{P(Y_3=1)} \\ &= \frac{P(Y_1=1,Y_2=1)}{P(\max(Y_1,Y_2)=1)} \\ &= \frac{\pi_1\pi_2}{1-P(Y_1=0,Y_2=0)} \\ &= \frac{\pi_1\pi_2}{1-(1-\pi_1)(1-\pi_2)} \end{aligned}$$

which does not decompose to a product of probabilities. Or, seeing it differently, if we find component 1 to work but the system failed, then

$$\begin{aligned} P(Y_2 = 1 \mid Y_3 = 1, Y_1 = 0) &= \frac{P(Y_2=1,Y_1=0,Y_3=1)}{P(Y_1=0,Y_3=1)} \\ &= \frac{P(Y_1=0,Y_2=1)}{P(Y_1=0,Y_2=1)} \\ &= 1. \end{aligned}$$

▷

Consequently, we may have marginal (pairwise) independence but conditional dependence. To incorporate marginal as well as conditional independencies, one needs to extend the graphical models toward **directed acyclic graphs** or in short **DAGs**. We sketch the resulting DAG for the above discussed example in Fig. 11.6. Nodes 1 and 2 are not connected, but both lead to node 3.

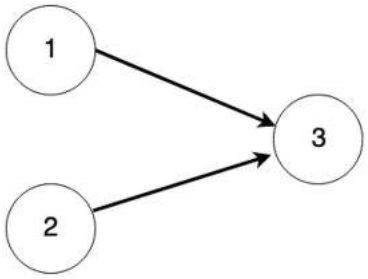


Fig. 11.6 Directed acyclic graph (DAG) for three variables

A DAG is formally defined as a directed graph without cycles or loops. Like before, let $G = (V, E)$ be a (directed) graph with $V = \{1, 2, \ldots, p\}$ being the set of nodes (variables) and $E \subset V \times V$ the set of edges. Elements in *E* fulfill the property that edges are not symmetric, i.e., for $(i, j) \in E$, we have $(j, i) \notin E$, and there are no loops, that is, there is no sequence $v_1, v_2, \ldots v_k \in V$ such that $(v_j, v_{j+1}) \in E$ for $j = 1, 2, \ldots k{-}1$ with $v_1 = v_k$.

The general independence and conditional independence statements in a DAG can be written after some extra definitions. For a node *i* in the graph, we denote with $\mathrm{Pa}(i)$ the parents. For instance, in Fig. 11.6, we have $\mathrm{Pa}(1) = \emptyset$, $\mathrm{Pa}(2) = \emptyset$, and $\mathrm{Pa}(3) = \{1, 2\}$. Moreover, the nodes in the DAG are ordered, so that $\mathrm{Pa}(1) = \emptyset$ holds in general and node *n* has no following node, $n \notin \mathrm{Pa}(j)$ for $j = 1, \ldots, p{-}1$. Moreover, in general, there are no loops or cycles, that is,

$$\mathrm{Pa}(i) \subset \{1, \ldots, i{-}1\}.$$

The conditional independence statements are then

$$i \perp\!\!\!\perp \{1, \ldots, i-1\} \backslash \mathrm{Pa}(i) \quad | \quad \mathrm{Pa}(i)$$

for all $i = 1, \ldots, p$. Hence, if we condition on the parents of a variable Y_i in the DAG, then all variables beyond the direct parents, meaning the ones which lie "before" the parent variables, do not provide any information about Y_i. Moreover, we can also state that two variables are generally (i.e., unconditionally) independent if they have no common ancestors and are not connected by an path given the variables in the graph. In other words, two nodes in the graph are independent if the only path(s) between them go through a shared descendant(s). Note that we looked at simple DAGs already in Chap. 8 and will return to this in the time series examples in Chap. 12.

To fully understand the ideas of a DAG, let us look at a more complex example. Consider the DAG in Fig. 11.7. We can read the following independence statements:

$$
\begin{aligned}
&1 \perp\!\!\!\perp 2\\
&3 \perp\!\!\!\perp 2 \mid 1\\
&4 \perp\!\!\!\perp 3 \mid \{1,2\}\\
&5 \perp\!\!\!\perp \{1,3,4\} \mid 2\\
&6 \perp\!\!\!\perp \{1,2,4\} \mid \{3,5\}\\
&7 \perp\!\!\!\perp \{1,2,4\} \mid \{3,5,6\}\\
&8 \perp\!\!\!\perp \{1,2,3,4\} \mid \{5,6,7\}.
\end{aligned}
$$

As in graphical models, there are equivalent formulations of the resulting independence statements. We refer exemplarily to Cowell et al. (2007) but do not provide a deeper discussion here.

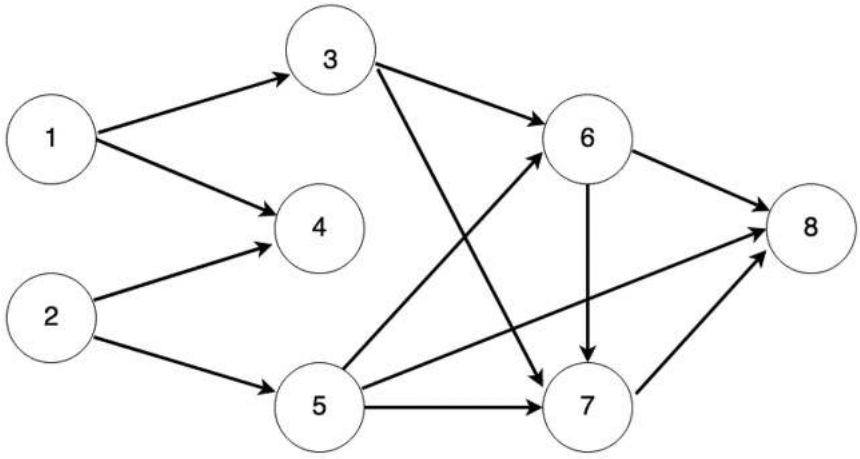


Fig. 11.7 DAG for eight variables

In general, a multivariate distribution on variables $Y_1, \ldots, Y_p$ can be factorized according to a DAG as follows:

$$f_{Y_1 \cdots Y_p}(y_1, \ldots, y_p) = \prod_{j=1}^{p} f_{Y_j|\mathrm{Pa}(j)}(y_j \mid y_{\mathrm{Pa}(j)}).$$

Hence, for the DAG in Fig. 11.7, we get (in short notation)

$$f_{12\cdots 8} = f_1 f_2 f_{3|1} f_{4|12} f_{5|2} f_{6|35} f_{7|356} f_{8|567}.$$

An interesting statement can be derived from the conditional independence structures resulting from a DAG. Note that looking at the DAG in Fig. 11.6, we saw with Example 11.2.

$$Y_1 \not\perp\!\!\!\perp Y_2 \mid Y_3,$$

that is, Y_1 is not independent of Y_2 if we condition on Y_3. We get such statements by "moralizing" the DAG. Moralizing means that we induce an undirected graphical model from the DAG by

1. placing an undirected edge between all nodes in $\mathrm{Pa}(i)$ for $i = 1, \ldots, p$ (i.e., we are "marrying" the parents).
2. replace all directed edges by undirected ones.

More formally, if $G = (V, E)$ is a DAG, then the related undirected moral graph $G^m = (V, E^m)$ results through

1. Set $E^m = E$.

2. For all $i = 1, \ldots p$, include (j, k) in E^m if $\{j, k\} \in \mathrm{Pa}(i)$

3. For all $(i, j) \in E$, include $(j, i) \in E^m$.

We demonstrate this in Fig. 11.8, where we include the "moralized" edges as dashed lines. This graph now describes all conditional independence statements as a graphical model. Apparently, we lose the information of two independencies which disappeared due to looking at all *p* variables. DAGs are a central tool in causality theory, and we refer to Peters et al. (2017) and Pearl (2009) for a more thorough discussion.

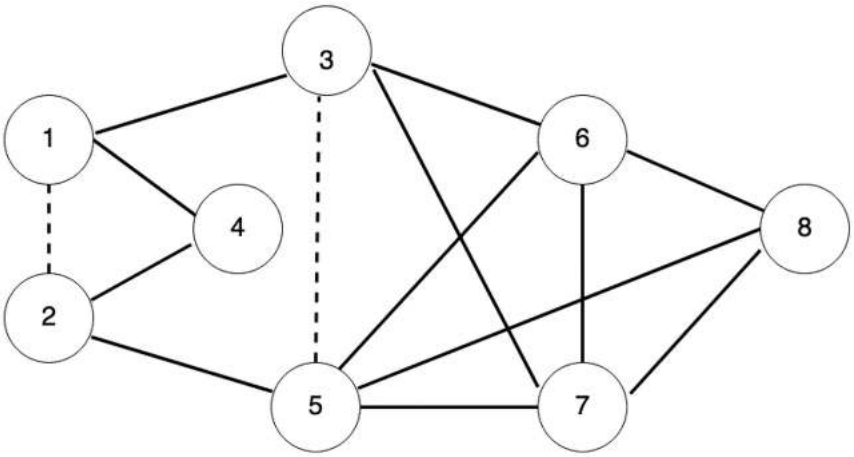


Fig. 11.8 Moralized graph from Fig. 11.7. Induced edges are dashed

11.3 Multivariate Normal Distribution

The most commonly used multivariate distribution is the multivariate normal distribution. We have already worked with this distribution in the simulation section in Sect. 8.4. We now show that the multivariate normal distribution admits a log-linear representation and that, using the results developed above, it can be naturally related to graphical models.

Definition 11.2 *Let* $Y = (Y_1, \ldots Y_p)$ *be a continuous random vector. Y follows a multivariate normal distribution with mean* $\mu = (\mu_1, \ldots \mu_p)^\top$ *and* $p \times p$ *dimensional variance matrix* Σ *if*

$$f(y; \mu, \Sigma) = \frac{1}{(2\pi)^{\frac{p}{2}}} |\Sigma|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(y-\mu)^\top \Sigma^{-1}(y-\mu)\right\}. \tag{11.5}$$

It is thereby assumed that Σ *is positive definite, that is,* $a^\top \Sigma a > 0$ *for all vectors* $a \in R^p$ *with* $a \neq 0$.

Given data $y^{(i)}$ drawn $i.i.d.$ from (11.5), we obtain the maximum likelihood estimates

$$\widehat{\mu} = \frac{1}{n} \sum_{i=1}^n y^{(i)} = \left(\frac{1}{n} \sum_{i=1}^n y_1^{(i)}, \ldots, \frac{1}{n} \sum_{i=1}^n y_q^{(i)} \right)^\top$$

$$\widehat{\Sigma} = \frac{1}{n} \sum_{i=1}^n (y^{(i)} - \widehat{\mu})(y^{(i)} - \widehat{\mu})^\top.$$

Let $\Omega = \Sigma^{-1}$ be the inverse variance matrix, commonly called the **concentration matrix**. Denote with Ω_{jl} the j, l-th element of Ω. This allows to rewrite (11.5) to

(11.6)

$$
\begin{aligned}
f(y;\mu,\Sigma) \quad &\propto \exp\left\{-\tfrac{1}{2}\sum_{j=1}^{p}\sum_{l=1}^{p}\Omega_{jl}(y_j-\mu_j)(y_l-\mu_l)\right\} \\
&\propto \exp\left\{-\tfrac{1}{2}\sum_{j=1}^{p}\sum_{l=1}^{p}\Omega_{jl}\,y_j y_l + \sum_{j=1}^{p}\sum_{l=1}^{p}\Omega_{jl}\mu_l\,y_j\right\} \\
&\propto \exp\left\{-\sum_{j=1}^{p}\sum_{l>j}^{p}\Omega_{jl}\,y_j y_l + \sum_{j=1}^{p}\left(-\tfrac{1}{2}\Omega_{jj}y_j^2 + \Big(\sum_{l=1}^{p}\Omega_{jl}\mu_l\Big)\,y_j\right)\right\}
\end{aligned}
$$

Formula (11.6) uncovers the log-linear structure of a multivariate normal distribution, which allows us to utilize the results derived in the corresponding section. To make this more apparent, we define the statistics

$$
\begin{aligned}
s_j(y) &= (y_j^2, y_j)^\top \\
s_{jl}(y) &= y_j y_l.
\end{aligned}
$$

Then, (11.6) can be rewritten to

$$
f(y;\mu,\Sigma) \quad \propto \exp\left\{\sum_{j=1}^{p}\sum_{l>j}^{p}\theta_{lj}s_{lj}(y) + \sum_{j=1}^{p}\theta_j^\top s_j(y)\right\} \tag{11.7}
$$

with $\theta_j = (-\frac{1}{2}\Omega_{jj}, \sum_{l=1}^{p}\Omega_{jl}\mu_l)^\top$ and $\theta_{jl} = -\Omega_{jl}$. We can derive some properties from (11.7). First, unlike in the analysis of multivariate discrete valued variables, like in Sect. 11.1, the normal distribution does not carry higher-order interaction terms. We only have pairwise interactions in the normal multivariate distribution. Second, we can directly read conditional independencies from the concentration matrix. That is,

$$
\Omega_{jl} = 0 \Leftrightarrow Y_j \perp\!\!\!\perp Y_l \mid Y_{1,\dots,p} \setminus \{j,l\}. \tag{11.8}
$$

Hence, the concentration matrix is informative as it provides information about the conditional independencies.

Example 11.3 We illustrate graphical models for multivariate normally distributed data with a simple example. We use data from the 2025 German federal election. The electoral districts are treated as (independent) observations, and we consider the vote shares of seven parties: SPD (Labour Party), CDU/CSU (Conservatives), Bündnis 90/Die Grünen (Greens), Die Linke (Left), FDP (Liberals), Freie Wähler (Free Voters), and AfD (Right-wing). We assume a normal distribution for the vote shares and select a graphical model using the AIC criterion. The resulting graph is shown in Fig. 11.9. We observe strong dependencies among the vote shares. However, the vote share of the Green Party is conditionally independent of the vote shares of the Free Voters and the Left Party, given the shares of the remaining parties. ▷

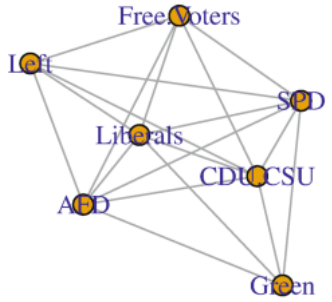


Fig. 11.9 Graphical model for voters share for the 2025 election in Germany

11.4 Exercises

Exercise 1 A survey administered to statistics students gathered data on how they like to spend their free time in a typical week. In particular, for each student, we have information on the following variables:

$$
M = \text{Number of movies watched}
$$

$$B = \text{Number of books read}$$

$$V = \text{Hours spent playing videogames}$$

with $M \leq 7$, $B \leq 2$, and $V \leq 20$. We are interested in studying the behavior of the students with respect to these three aspects.

(a) Can you sketch a contingency table for this data? If so, do it, if not, explain why.

(b) Suppose we want to jointly model the probability of a student watching a certain number of movies, reading a certain number of books, and playing a certain amount of hours of video games in a given week. Specify a log-linear model for this joint probability, assuming independence between the three variables.

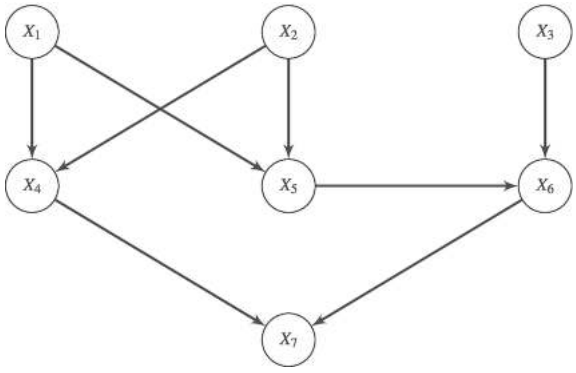
(c) A sociologist comes into the picture and tells you that it is likely for some of these quantities to be dependent on one another. In particular, the number of books read is likely to have some interdependence with both the hours of video games played and the number of movies watched. How can we incorporate these dependencies into the model defined in point (b)?

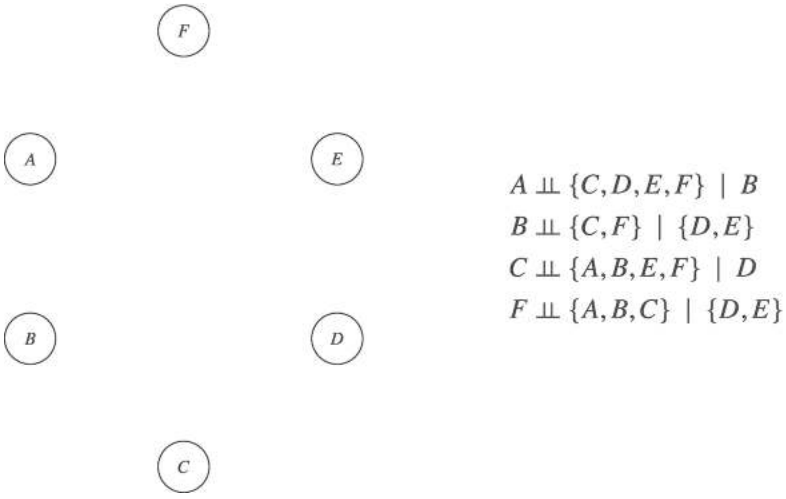
(d) Given the assumptions made in point (c), are the variables *M* and *V* independent? Are they conditionally independent? Sketch a dependency graph of the three variables.

Exercise 2

(a) Below, you are given a directed acyclic graph (DAG) depicting the causal dependence between seven random variables. For each variable $X_i,\ i = 1, \ldots, n$, state the variables that it is independent of. Additionally, factorize the joint distribution $f_{X_1,X_2,X_3,X_4,X_5,X_6,X_7}(x_1, x_2, x_3, x_4, x_5, x_6, x_7)$ of this model; i.e., define it as the product of multiple conditional distributions.

(b) Given the conditional independencies next to the empty graph below, fill in the (undirected) edges, assuming all variables to be dependent unless stated otherwise. *Hint:* Start with a full graph, and then remove edges one by one.





Exercise 3 Let $\boldsymbol{x} \in \mathbb{R}^n$ be a random vector with $E(\boldsymbol{x}) = \mathbf{0}$ and $\boldsymbol{V} = \mathrm{Cov}(\boldsymbol{x})$. Assuming that $\boldsymbol{x}$ follows a multivariate normal distribution and that $\mathrm{Cov}(\boldsymbol{x})$ is a diagonal matrix, show that $x_1, \ldots, x_n$ are independent.

References

Agresti, A. (2012). *Categorical data analysis* (Vol. 792). Wiley.

Cowell, R. G., Dawid, P., Lauritzen, S. L., & Spiegelhalter, D. J. (2007). *Probabilistic networks and expert systems: Exact computational methods for Bayesian networks*. Springer.

Lang, J. B. (1996). On the comparison of multinomial and poisson log-linear models. *Journal of the Royal Statistical Society: Series B (Methodological),58*(1), 253–266.
[MathSciNet][https://doi.org/10.1111/j.2517-6161.1996.tb02079.x]

Lauritzen, S. L. (1996). *Graphical models* (Vol. 17). Clarendon Press.

Maathuis, M., Drton, M., Lauritzen, S., & Wainwright, M. (2018). *Handbook of graphical models*. CRC Press.
[https://doi.org/10.1201/9780429463976]

Pearl, J. (2009). *Causality*. Cambridge University Press.
[https://doi.org/10.1017/CBO9780511803161]

Peters, J., Janzing, D., & Schölkopf, B. (2017). *Elements of causal inference: Foundations and learning algorithms*. MIT Press.

Tutz, G. (2011). *Regression for categorical data* (Vol. 34). Cambridge University Press.

Whittaker, J. (2009). *Graphical models in applied multivariate statistics*. Wiley Publishing.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
G. Kauermann et al., *Statistics—Coping with Uncertainty*, Textbooks in Statistical Science
https://doi.org/10.1007/978-3-032-33658-3_12

12. Non-i.i.d. Data

Göran Kauermann[1] ✉, Giacomo De Nicola[2], Cornelius Fritz[3] and David Rügamer[4]
(1) Department of Statistics, LMU Munich, Munich, Germany
(2) Harvard T.H. Chan School of Public Health, Harvard University, Boston, MA, USA
(3) School for Computer Science and Statistics, Trinity College Dublin, Dublin, Ireland
(4) Munich Uncertainty Quantification AI Lab, LMU Munich, Munich, Germany

The framework underlying the data analyses considered in this book so far was based on the assumption that the observations are *i.i.d.*, that is, independent of one another and generated from the same probabilistic distribution. From a practical perspective, this assumption naturally leads to a likelihood that factorizes into separate, low-dimensional components, as discussed in Chap. 3. At the same time, this assumption forms the foundation of asymptotic theory (Chap. 4). Once this assumption no longer holds, many of the corresponding mathematical properties break down.

We already introduced non-*i.i.d.* data in the regression chapter (Chap. 6), where we allowed the distribution of Y_i to depend on covariates x_i, thereby giving up the assumption of identical distributions. Still, the resulting residuals ϵ_i were assumed to be *i.i.d.*. Moreover, in Chap. 11, we introduced data constellations in which simple independence assumptions are no longer adequate. As a workaround, we studied graphical models as an abstract framework. This chapter will focus on

specific non-*i.i.d.* data types and how to deal with them. More specifically, we cover data which possess a dependence structure in the form of temporal order (Sect. 12.1), spatial dependence (Sect. 12.2), network structure (Sect. 12.3), as well as data which are functions themselves (Sect. 12.4). These four examples represent some of the most commonly encountered dependence structures in real-world data. However, dependence patterns can be highly diverse and complex. Therefore, our treatment remains at a conceptual level, providing a high-level overview without going into all possible details.

12.1 Time Series Models

12.1.1 Notation and Main Ideas

In the following, we assume that the data are collected over time; we refer to such data as **time series data**. We write

$$Y_1, Y_2, Y_3, \ldots, Y_T, \ldots$$

and assume that the time points $t = 1, 2, \ldots, T$ are discrete and equally spaced. We further assume that there is some correlation over time, in that $\text{Cov}(Y_t, Y_{t\prime}) \neq 0$ for $1 \leq t,\ t\prime \leq T$. We define with

$$\mu_Y(t) = E(Y_t)$$

the mean function for $t = 1, 2, \ldots$ and with

$$\gamma_Y(t, t\prime) = \text{Cov}(Y_t, Y_{t\prime})$$

the covariance function. The latter will be of particular interest, since for data collected over time, it's often the case that the current value Y_t for some time point *t* depends on previous observations $Y_{t-1}, Y_{t-2}, \ldots$. The two main questions in time series data are the following:

1. How are the data serially correlated; that is, what is the structure of $\gamma_Y(t, t\prime)$?
2. Given the data up to time point *t*, how can we use these to predict future values $Y_{t+1}, Y_{t+2}, \ldots$?

To address questions about correlation, we clearly need knowledge of $\mu_Y(t)$, and it is important to note that $\gamma_Y(t, t\prime)$ describes covariance, not correlation; meaning it also captures changes in variance over time. Thus, while the two questions posed above are central, they represent only part of the broader modeling tasks involved in time series analysis.

To exemplify the idea of time series, we look at three examples, which will be revisited later.

Example 12.1 Figure 12.1 displays the annual global average temperature from 1850 to 2019, presented as deviations from the overall average across the observed period. A clear upward trend in $\mu_y(t)$ is visible starting around 1975. To highlight this trend, we include an estimate $\widehat{\mu}_y(t)$ in the plot. This enables us to examine the residual fluctuations $y_t - \widehat{\mu}_y(t)$, which are shown in the lower part of Fig. 12.1. Details on the estimation of $\mu_y(t)$ are discussed later in the chapter. ▷

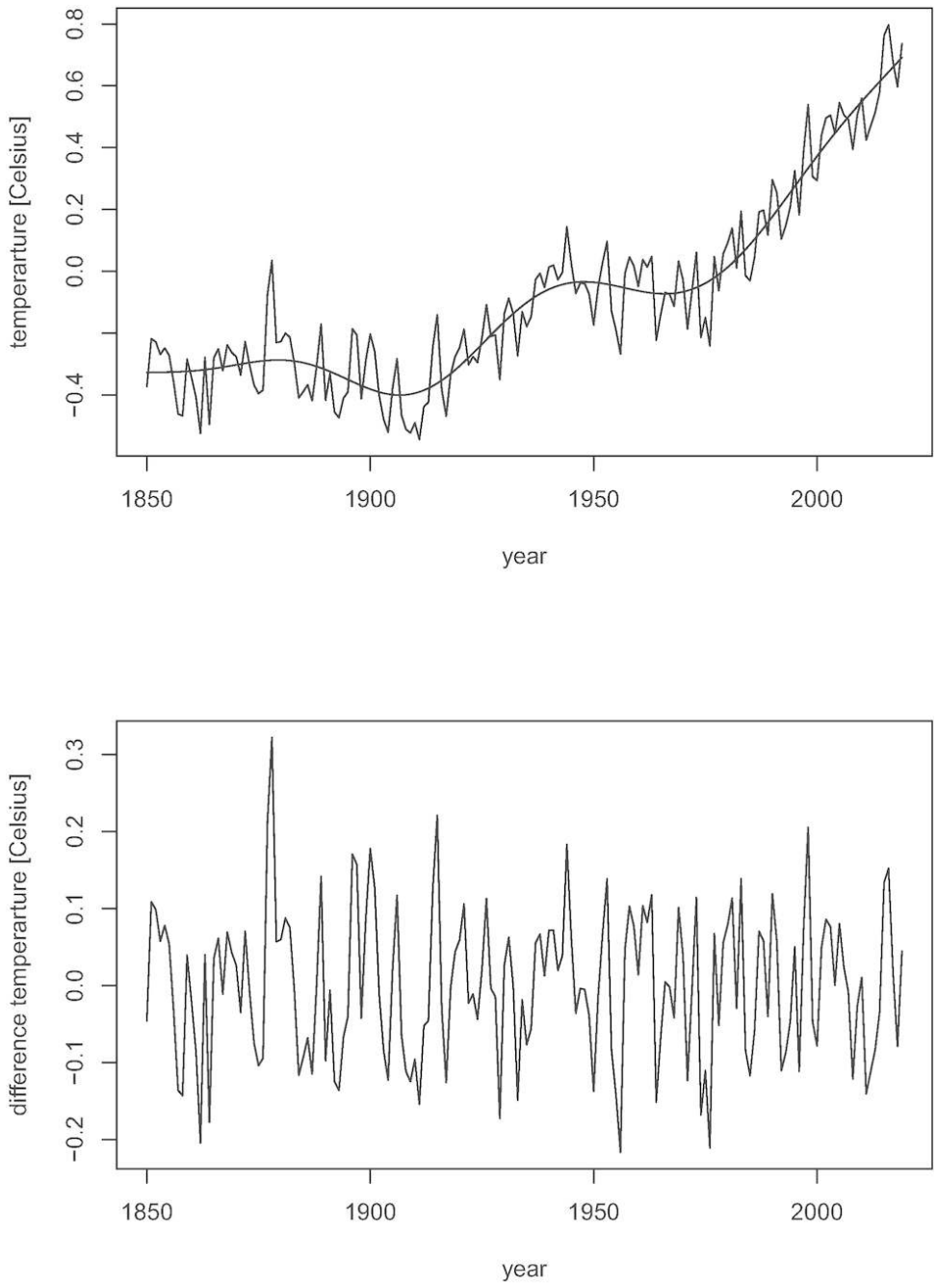


Fig. 12.1 Variation of global annual temperature around the mean value (top plot) and the detrended variation (bottom plot)

Example 12.2 As a second example, we consider the German stock market index DAX over the period from 2010 to 2022, as shown in Fig. 12.2. The data exhibit a modest upward trend, with a notable sharp decline in early 2020, coinciding with the onset of the COVID-19 pandemic, which significantly impacted global financial markets. We can

correct for the trend by looking at so-called log-returns; that is, we define the time series

$$\widetilde{Y}_t = \log\left(\frac{Y_t}{Y_{t-1}}\right).$$

We plot this in the bottom of Fig. 12.2. In this plot, we see additional features of the data, namely, the variation of the DAX. There are two periods that stand out, namely, 2012 and again 2020, where we observe a larger fluctuation of $\widetilde{Y}_t$. ▷

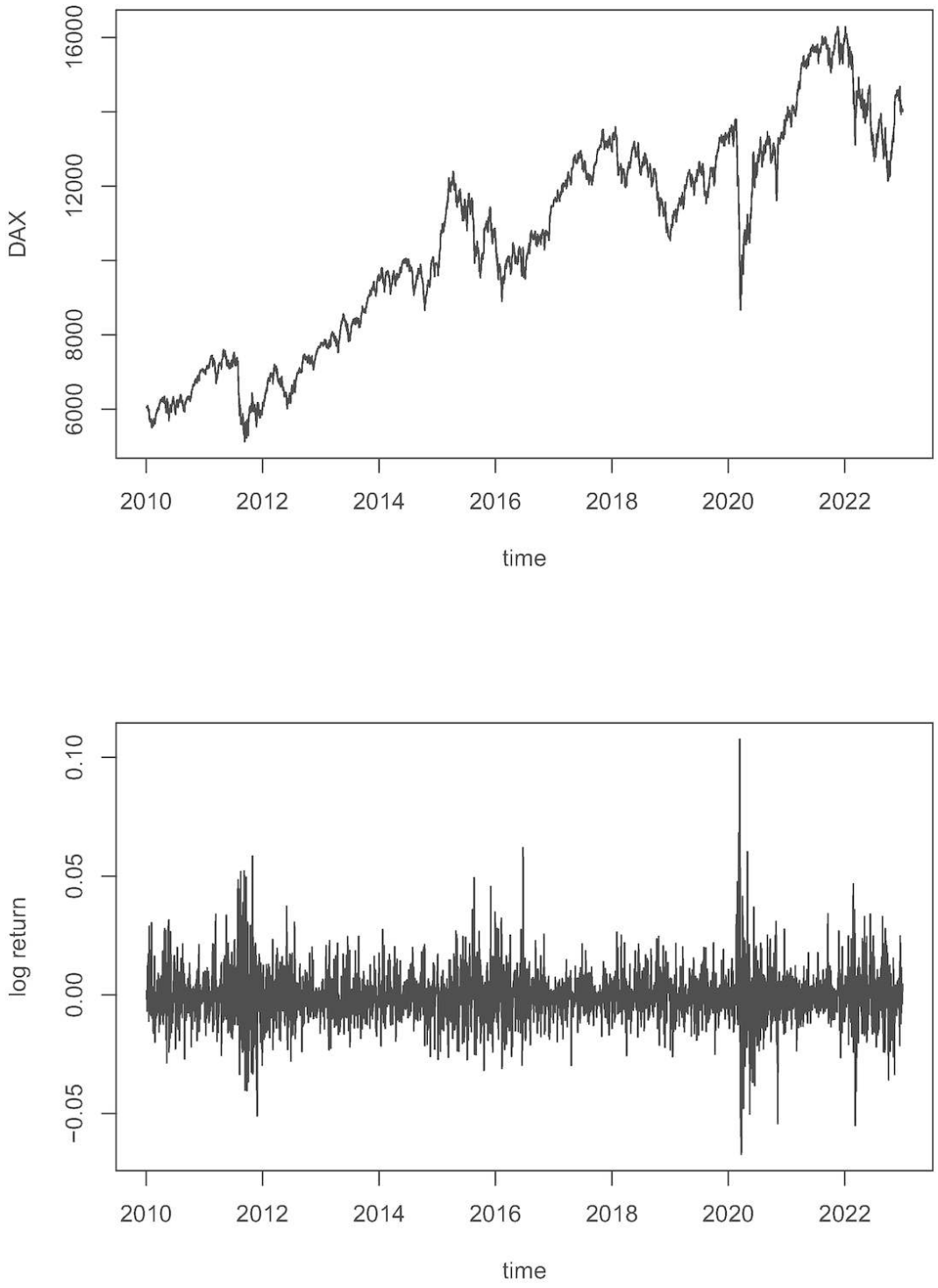
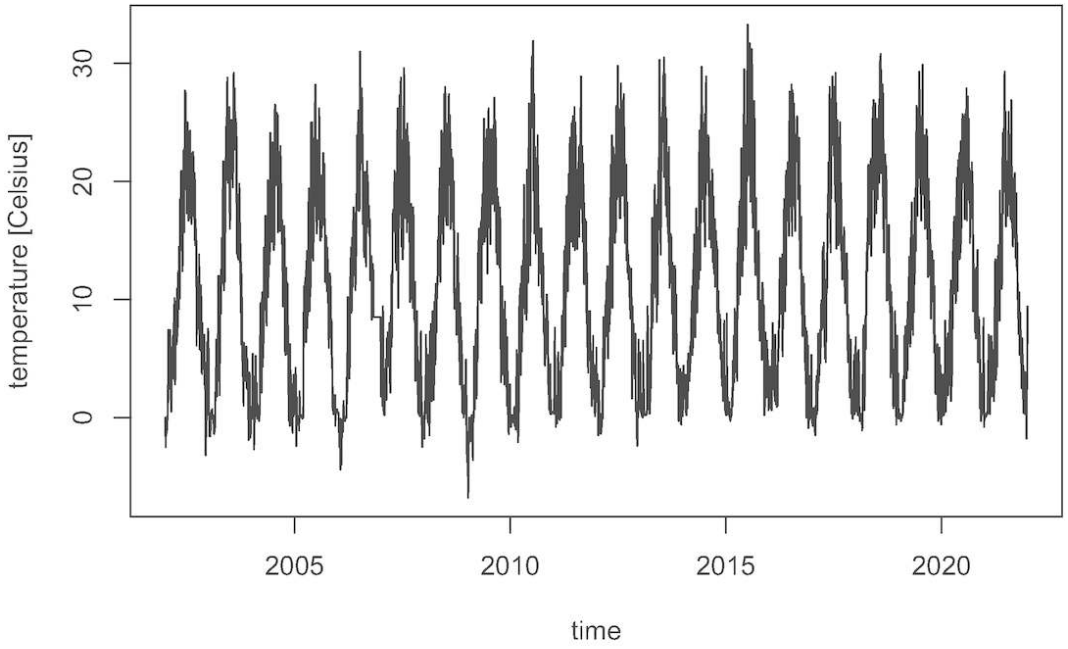


Fig. 12.2 German stock market index (DAX) (top plot) and log returns (bottom plot)

Example 12.3 As a final example, we look at the daily temperature in the city of Augsburg for the years 2003–2022. These are shown in the upper plot of Fig. 12.3. We see an annual variation with winters and

summers clearly standing out. We can extract this annual pattern and look at the residual fluctuation. In other words, we can correct for the seasonal variation and look at the resulting seasonally corrected data. This is shown in the bottom panel of Fig. 12.3. ▷



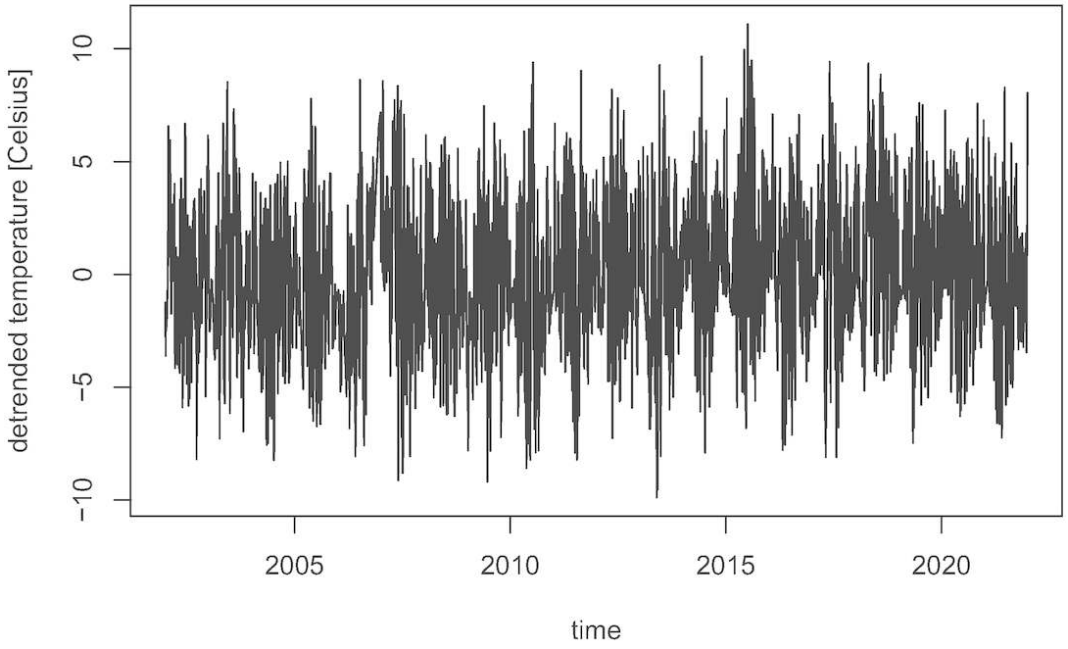


Fig. 12.3 Annual daily temperature in the city of Augsburg (top plot) and deseasoned data (bottom plot)

12.1.2 Trend and Seasonality

Before focusing on the dependence structure in the data, we need to shortly motivate the idea of trend and seasonal components. We can decompose the data into

$$Y_t = \mu(t) + \widetilde{Y}_t, \tag{12.1}$$

where $\widetilde{Y}_t$ has expectation 0. The mean structure $\mu(t)$ is typically composed of two components, which we denote as the trend and the seasonal pattern, meaning we can write

$$\mu(t) = m(t) + s(t). \tag{12.2}$$

The decomposition can be carried out with regression tools, and we refer to Brockwell and Davis (2016) for details.

The disentanglement of trend and seasonality is often not unique, and the corresponding preprocessing steps are referred to as filtering. While this is crucial for time series analyses, we do not go into detail here, since the focus of this chapter is on the modeling of dependence structures. We will therefore assume that our time series data vary around 0; i.e., we assume the mean structure $\mu(t) = 0$, which is possibly achieved after "de-trending" and "de-seasoning" the series.

12.1.3 Autocorrelation

To begin our analysis of serial dependence, as said above, we assume the mean structure to be zero; that is, we postulate $\mu_Y(t) \equiv 0$. We further assume some form of (weak) **stationarity** in the covariance function, meaning that the covariance function only depends on the time lag $|t - t\prime|$ but not on the specific value of t. Hence, $\gamma_Y(t, t\prime)$ can be written as

$$\gamma_Y(t, t\prime) = \gamma_Y(|t - t\prime|).$$

In this case, we call

$$\gamma_Y(h) = \mathrm{Cov}(Y_t, Y_{t+h}) = \mathrm{Cov}(Y_{t+h}, Y_t)$$

the **autocovariance function** with $h = 0, 1, \ldots, T$. This leads to the variance matrix

$$\Gamma = \begin{bmatrix} \gamma(0) & \gamma(1) & \gamma(2) & \dots & \gamma(T) \\ \gamma(1) & \gamma(0) & \gamma(1) & \dots & \gamma(T-1) \\ \gamma(2) & \gamma(1) & \gamma(0) & \dots & \gamma(T-2) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \gamma(T) & \gamma(T-1) & \gamma(T-2) & \dots & \gamma(0) \end{bmatrix}.$$

Since $\gamma_Y(0) = \mathrm{Var}(Y_t)$, we obtain with

$$\rho(h) = \tfrac{\gamma(h)}{\gamma(0)} \tag{12.3}$$

the so-called **autocorrelation function**. Accordingly, this leads to the correlation matrix

$$R = \begin{bmatrix} 1 & \rho(1) & \rho(2) & \dots & \rho(T) \\ \rho(1) & 1 & \rho(1) & \dots & \rho(T-1) \\ \rho(2) & \rho(1) & 1 & \dots & \rho(T-2) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho(T) & \rho(T-1) & \rho(T-2) & \dots & 1 \end{bmatrix}.$$

We typically assume that all these functions are finite. Note that for $i.i.d.$ data, we have

$$\gamma_Y(h) = \begin{cases} \sigma^2 & \text{for } h = 0 \\ 0 & \text{for } h \neq 0 \end{cases}. \tag{12.4}$$

A time series with autocorrelation function of the form (12.4) is typically referred to as **white noise**.

With these definitions in place, the focus turns to estimating the autocovariance or autocorrelation function from the data $y_1, \dots, y_T$. We define $\overline{y} = \sum_{t=1}^{T} y_t \Big/ T$, which leads to the estimate

$$\widehat{\gamma}_Y(h) = \tfrac{1}{T} \sum\nolimits_{t=1}^{T-h} (y_t - \overline{y})(y_{t+h} - \overline{y})$$

for $0 \leq h < T$. Note that the sum above is over $T - h$ elements, but we divide it by T and not by $T - h$. Although this may look counterintuitive, it is intentional, since the estimate yields a positive definite matrix estimate

$$\widehat{\Gamma} = \begin{bmatrix} \widehat{\gamma}(0) & \widehat{\gamma}(1) & \widehat{\gamma}(2) & \dots & \widehat{\gamma}(T) \\ \widehat{\gamma}(1) & \widehat{\gamma}(0) & \widehat{\gamma}(1) & \dots & \widehat{\gamma}(T-1) \\ \widehat{\gamma}(2) & \widehat{\gamma}(1) & \widehat{\gamma}(0) & \dots & \widehat{\gamma}(T-2) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \widehat{\gamma}(T) & \widehat{\gamma}(T-1) & \widehat{\gamma}(T-2) & \dots & \widehat{\gamma}(0) \end{bmatrix}.$$

To see this, let $x_t = y_t - \overline{y}$, then $\widehat{\Gamma}$ can be written as

$$\widehat{\Gamma} = \frac{1}{T} DD^\top,$$

where D is the Tx$(2T-1)$ dimensional matrix with entries

$$D = \begin{bmatrix} 0 & \dots & 0 & x_1 & \dots & x_{T-1} & x_T \\ \vdots & \iddots & x_1 & x_2 & \dots & x_T & 0 \\ 0 & \iddots & \iddots & \vdots & \iddots & \iddots & \vdots \\ x_1 & x_2 & \dots & x_T & 0 & \dots & 0 \end{bmatrix}.$$

This readily shows positive (semi-)definiteness, since

$$a^\top \widehat{\Gamma} a = \frac{1}{T} [aD][aD]^\top \geq 0$$

for all $a \in \mathbb{R}^T$. From $\widehat{\Gamma}$, we immediately obtain an estimate of R through

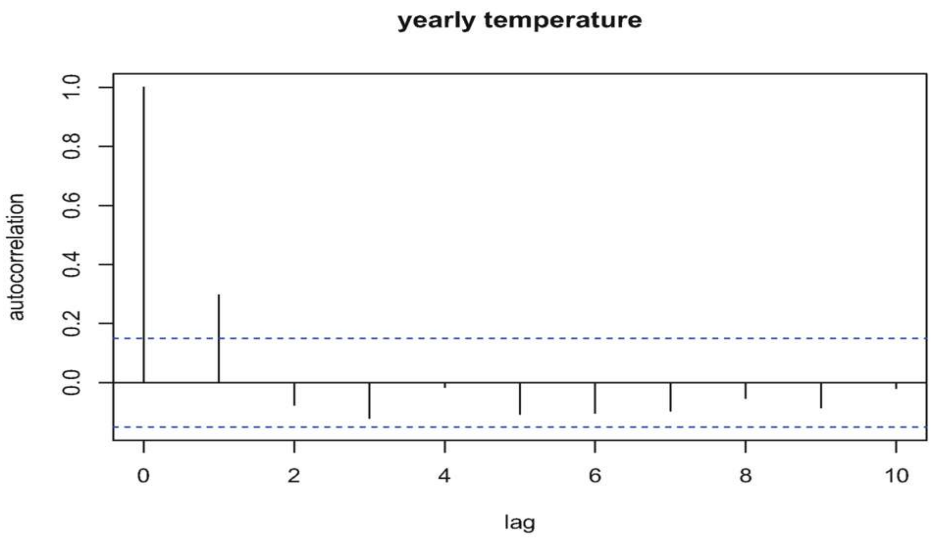
$$\widehat{R} = \frac{\widehat{\Gamma}}{\widehat{\gamma}(0)},$$

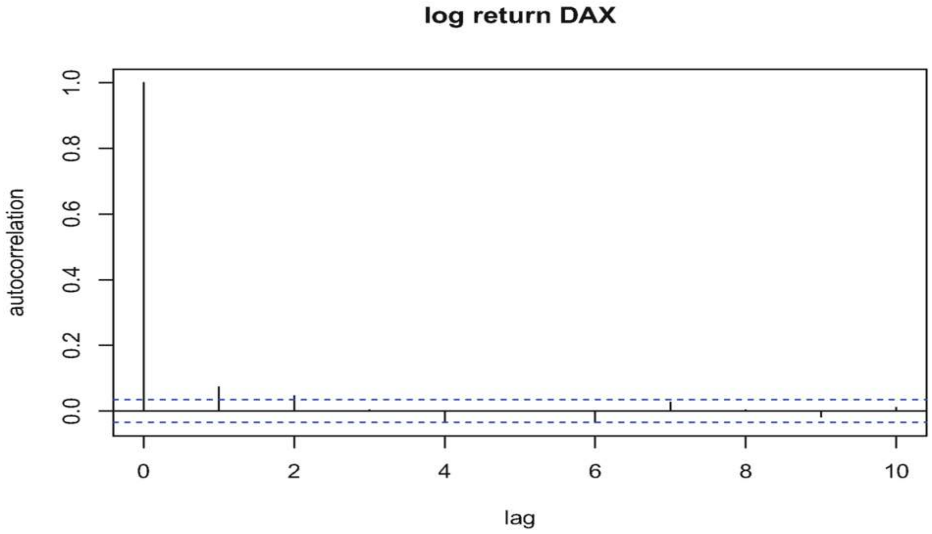
with $\widehat{\gamma}(0)$ as diagonal element of $\widehat{\Gamma}$. We will utilize $\widehat{R}$ later. In fact, the columns of $\widehat{R}$, or, to be more specific, the first column, are also known as the sample autocorrelation function defined as

$$\widehat{\rho}(h) = \frac{\widehat{\gamma}(h)}{\widehat{\gamma}(0)}.$$

The estimate provides information about the serial correlation, as illustrated in the following examples. In Fig. 12.4, we show the resulting estimates $\widehat{\gamma}(h)$ for the three examples from above. The upper panel gives the estimated autocorrelation for the trend-corrected annual temperature, the middle panel is for the log-returns of the DAX, and the bottom panel shows the autocorrelation for the seasonal-corrected temperature in Augsburg. While for the first two examples we have a

quickly decreasing autocorrelation, which is essentially zero for time lags of order two and above, this is not the case for the daily lags of the temperature in Augsburg.


yearly temperature
autocorrelation
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
8
10
lag


log return DAX
autocorrelation
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
8
10
lag

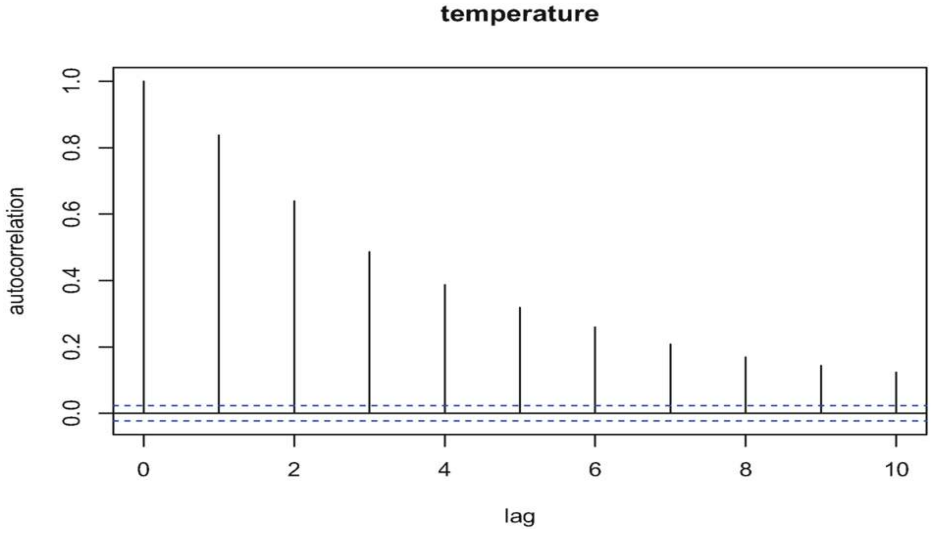

temperature
autocorrelation
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
8
10
lag

Fig. 12.4 Estimated autocorrelation function for the three examples in Figs. 12.1, 12.2 and 12.3, from top to bottom

12.1.4 Autoregressive Process (AR)

A central model class in time series analysis is given by AutoRegressive Moving Average (ARMA) models, which can be extended to AutoRegressive Integrated Moving Average (ARIMA) processes by taking differences of the original data, that is, by considering $Y_t - Y_{t-1}$ instead of the original observations. The model class was originally proposed by Whittle (1951) and made widely accessible by Box and Jenkins (1970) (see also Box et al., 2015). A short historical overview of time series analysis can be found in Tsay (2000).

We start with the AutoRegressive (AR) model. We assume that Y_t is determined by the previous *p* values plus additional white noise.

Definition 12.1 *An autoregressive AR(p) process is defined by*

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + Z_t,$$

where $Z_t \sim N(0, \sigma^2)$ *i.i.d.*.

It is convenient to write this using the so-called **backshift operator** B, defined as

$$\begin{aligned} BY_t &= Y_{t-1} \\ B^2 Y_t &= BY_{t-1} = Y_{t-2}, \end{aligned}$$

or, more generally,

$$B^p Y_t = Y_{t-p}.$$

The backshift operator simplifies notation and theoretical derivations, as will be demonstrated below. Note that the operation also extends to negative powers, e.g.,

$$B^{-1} Y_t = Y_{t+1}.$$

With the help from the backshift operator, we can rewrite the AR(p) process more compactly as

$$\left(1 - \sum\nolimits_{j=1}^{p} \phi_j B^j\right) Y_t = Z_t.$$

The simplest example is the AR(1) process, which often is the first model used to capture serial correlation in real data. In this case, we have

$$Y_t = \phi Y_{t-1} + Z_t,$$

where we dropped the subscript 1 at ϕ for notational simplicity. Note that ϕ plays the role of the autocovariance, since

$$\mathrm{Cov}(Y_t, Y_{t-1}) = \phi \mathrm{Cov}(Y_{t-1}, Y_{t-1}) = \phi \sigma^2.$$

Assuming stationarity, we obtain $\rho(1) = \phi$ for the autocorrelation defined in (12.3). It is not difficult to obtain in general $\rho(h) = \phi^h$. Apparently, restricting to $|\phi| < 1$, the correlation matrix Γ results as

$$\Gamma = \begin{bmatrix} 1 & \phi^1 & \phi^2 & \dots & \phi^\top \\ \phi^1 & 1 & \phi^1 & \dots & \phi^{T-1} \\ \phi^2 & \phi^1 & 1 & \dots & \phi^{T-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \phi^\top & \phi^{T-1} & \phi^{T-2} & \dots & 1 \end{bmatrix}.$$

We now turn to a more detailed analysis of the parameter ϕ. In particular, we aim to explore the range of values that ϕ can plausibly take. We distinguish three cases:

1. $|\phi| < 1$. This will lead to stationary behavior.
2. $|\phi| = 1$. This is called a random walk.
3. $|\phi| > 1$. This will lead to nonstationary behavior.

Let us start with $|\phi| < 1$. We consider the time series Y_t in a more general setting, allowing time t to range from $-\infty$ to ∞, that is,

$$t \in \{\dots, -T, -T+1, -T+2, \dots, -1, 0, 1, \dots, T-2, T-1, T, \dots\}$$

This allows to convert the AR(1) process into an infinite series, which we will later define as moving average process. Note that

$$\begin{aligned} Y_t &= \phi Y_{t-1} + Z_t \\ &= \phi^2 Y_{t-2} + \phi Z_{t-1} + Z_t \\ &= \phi^3 Y_{t-3} + \phi^2 Z_{t-2} + \phi Z_{t-1} + Z_t \end{aligned}$$

and by continuing these equations to infinite steps backward, we obtain

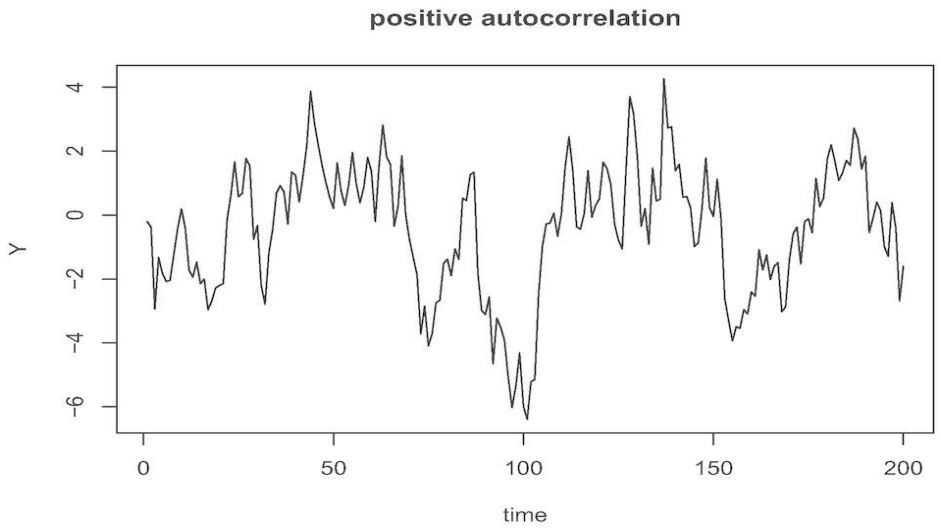
$$Y_t = \sum\nolimits_{j=0}^{\infty} \phi^j Z_{t-j}.$$

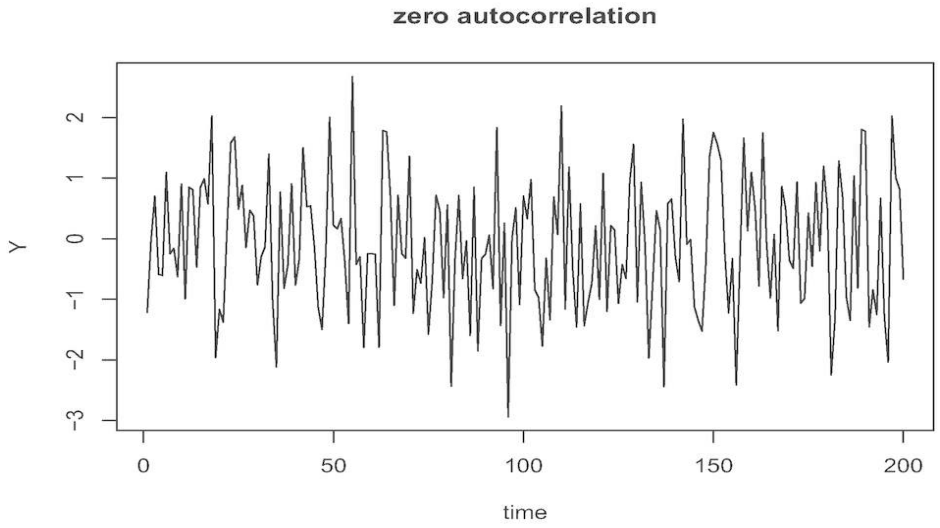
This in turn allows to derive the variance of Y_t through

$$\operatorname{Var}(Y_t) = \sum\nolimits_{j=0}^{\infty} \left(\phi^2\right)^j \sigma^2 = \tfrac{\sigma^2}{1-\phi^2},$$

where the limit follows through the geometric series. Thus, for $|\phi| < 1$, the process has finite variance at each time point *t*.

Figure 12.5 shows three examples of AR(1) processes. The top panel shows a series with $\phi = 0.9$, the middle panel depicts $\phi = 0$, which corresponds to independence, and the bottom panel is for $\phi = -0.9$. For positive autocorrelation, we see phases where the data are positive and phases with negative values. This is quite common for real-world time series; for example, daily temperature data have such features after correcting for seasonality. For zero correlation, these phase patterns disappear, while for negative correlation the opposite, namely, a strongly alternating behavior, occurs.


positive autocorrelation
Y
time
-6
-4
-2
0
2
4
0
50
100
150
200


zero autocorrelation
Y
time
-3
-2
-1
0
1
2
0
50
100
150
200

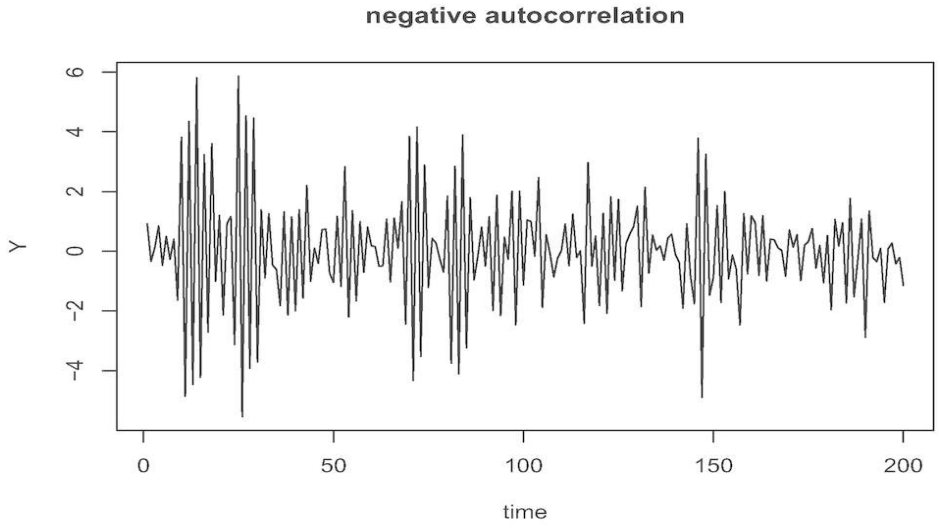

negative autocorrelation
Y
time
-4
-2
0
2
4
6
0
50
100
150
200

Fig. 12.5 Simulated AR(1) process with positive (top), zero (middle), and negative (bottom) autocorrelation

Let us now consider AR(1) processes with $|\phi| > 1$. In this case, the process is said to be nonstationary, as its variance explodes over time. This behavior is illustrated in Fig. 12.6, which shows the trajectories of two AR(1) processes starting from $Y_0 = 0$ with $\phi = 1.02$. The divergence of the process makes it unrealistic for most data analysis applications. We will revisit this process in a more formal mathematical treatment later in the chapter.

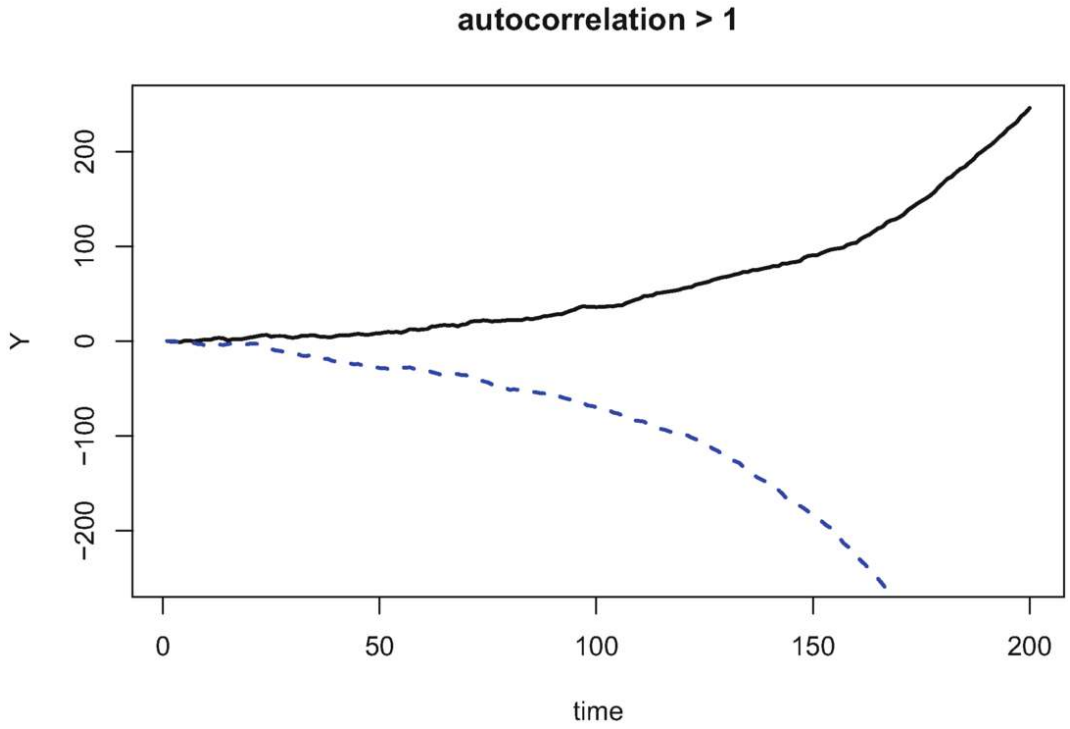


Fig. 12.6 Two simulated AR(1) processes with autocorrelation $\phi > 1$

Finally, for completeness, we also consider the case $|\phi| = 1$. For $\phi = 1$, the process is called a **random walk**, defined as

$$Y_t = Y_{t-1} + Z_t$$

For a process starting at $t = 0$, we obtain

$$\mathrm{Var}(Y_t) = \mathrm{Var}(\sum_{l=1}^{t} Z_t) = \sum_{l=1}^{t} \mathrm{Var}(Z_t) = t\sigma^2.$$

Hence, the variance increases with *t*. The random walk is not of central interest here, as it defines the sum of *i.i.d.* variables. Nevertheless, for completeness, we show the behavior of two sample paths starting at 0 in Fig. 12.7.

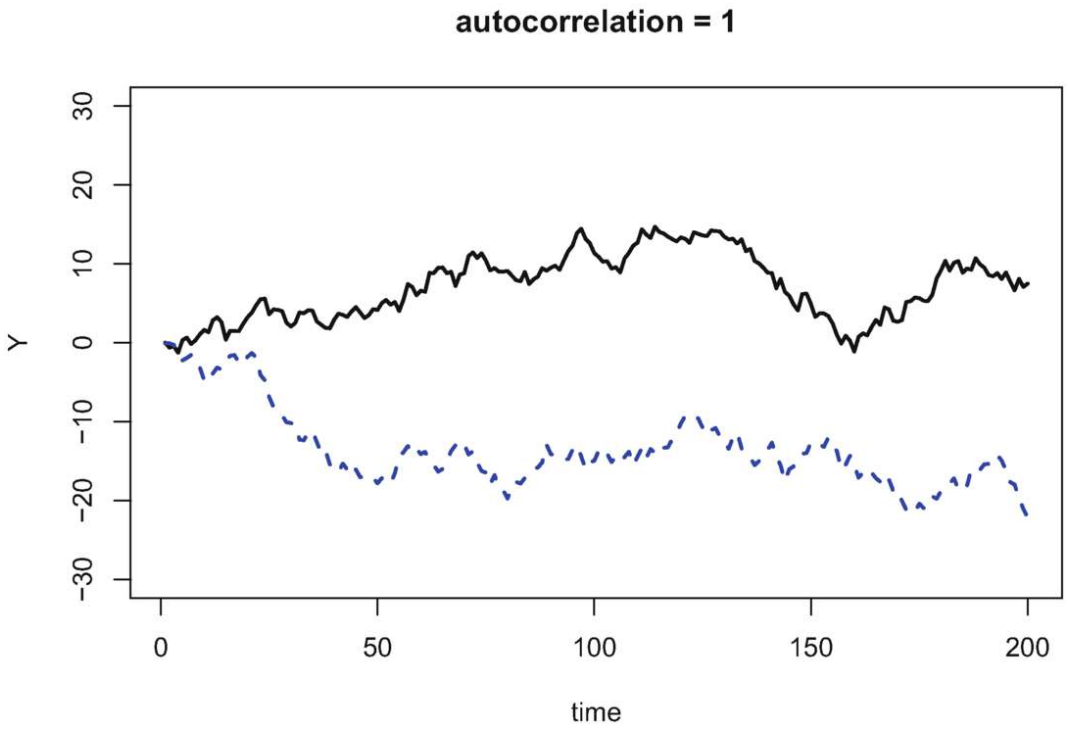


Fig. 12.7 Two random walks, simulated with $\phi = 1$

12.1.5 Moving Average Process (MA)

We now turn to the moving average process.

Definition 12.2 *A moving average process of order q, denoted as MA(q), is defined as*

$$Y_t = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q},$$

where $Z_t \sim N(0, \sigma^2)$ *i.i.d.*.

As above, we can make use of the backshift operator and write

$$Y_t = \sum_{i=0}^{q} \theta_i B^i Z_t$$

with $\theta_0 = 1$.

Let us consider the MA(1) process in more detail. We have

$$Y_t = Z_t + \theta Z_{t-1},$$

where we again dropped the subscript for notational simplicity. For the autocovariance function, we obtain

$$\begin{aligned}
\gamma(0) &= \mathrm{Cov}(Y_t, Y_t) \\
&= \mathrm{Cov}(Z_t + \theta Z_{t-1}, Z_t + \theta Z_{t-1}) \\
&= \sigma^2\left(1 + \theta^2\right) \\
\gamma(1) &= \mathrm{Cov}(Y_t, Y_{t-1}) \\
&= \mathrm{Cov}(Z_t + \theta Z_{t-1}, Z_{t-1} + \theta Z_{t-2}) \\
&= \theta\sigma^2 \\
\gamma(h) &= 0 \text{ for } h > 1
\end{aligned}$$

and accordingly

$$\rho(h) = \begin{cases} 1, & \text{for } h = 0 \\ \frac{\theta}{1+\theta^2}, & \text{for } h = 1. \\ 0, & \text{for } h > 1 \end{cases}$$

Hence, the MA process models directly the autocovariance, which is zero if *h* exceeds the order of the MA process.

In Fig. 12.8, we show two simulated MA(1) processes. The top plot is for $\theta = 0.9$, while the bottom plot is for $\theta = -0.9$. Comparing this with Fig. 12.5 shows that the processes have completely different behavior.

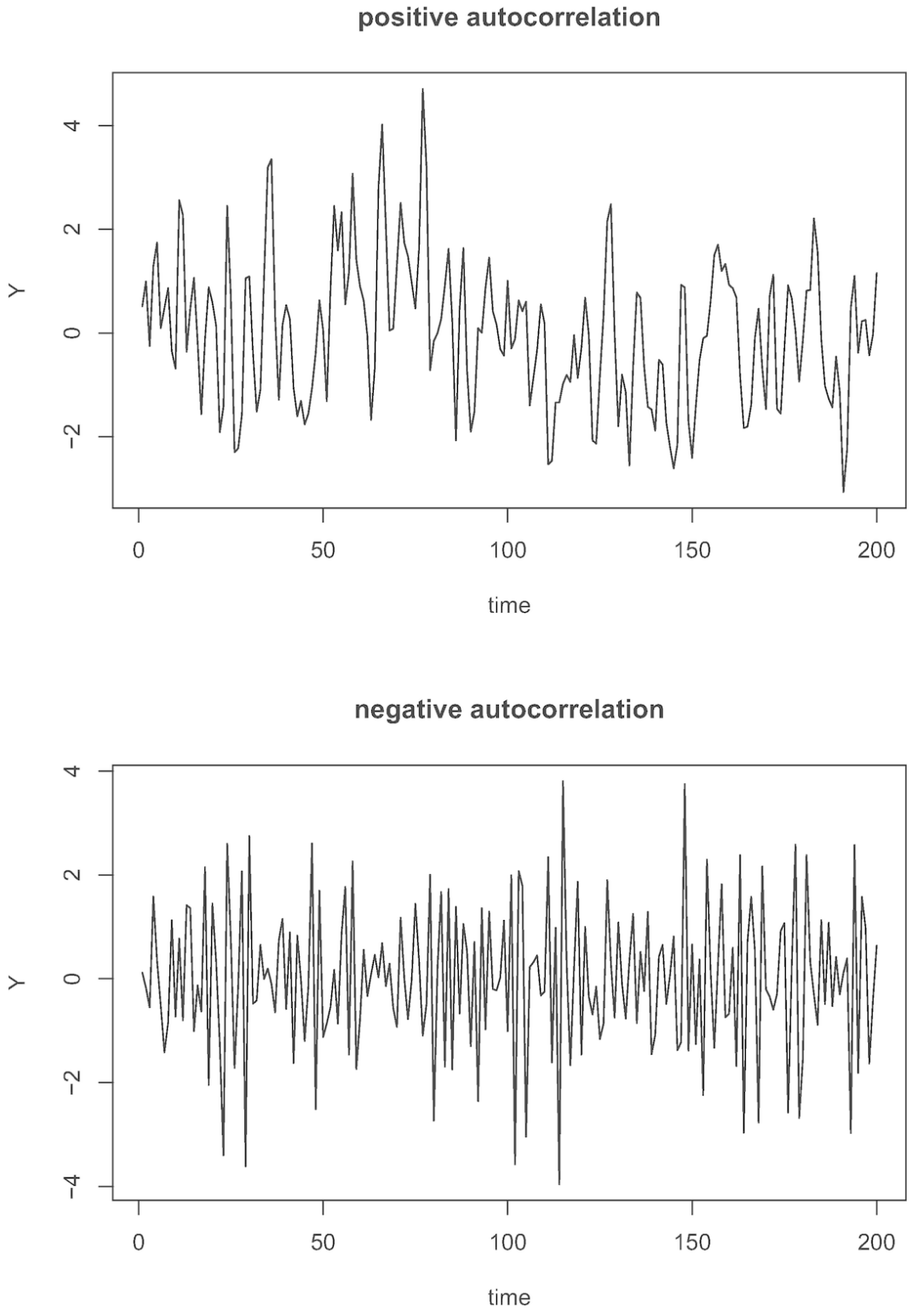


Fig. 12.8 Simulated MA(1) process with positive (top) and negative (bottom) correlation

12.1.6 Autoregressive Moving Average Process (ARMA)

We can now combine the two ideas above, leading to ARMA(p, q) processes.

Definition 12.3 *The Auto Regressive Moving Average process of order p and q, in short ARMA(p, q), is defined as*

$$Y_t = \underbrace{\phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p}}_{\text{AR}} + \underbrace{Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q}}_{\text{MA}}.$$

Using the backshift operator, we can write this as

$$-\sum_{j=0}^{p} \phi_j B^j Y_t = \sum_{i=0}^{q} \theta_i B^i Z_t$$

with $\phi_0 = -1$ and $\theta_0 = 1$.

It is useful to formally define the operators

$$\begin{aligned} \Phi(B) &= 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p \\ \Theta(B) &= 1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q. \end{aligned}$$

This allows us to write an ARMA(p, q) process even more concisely as

$$\Phi(B) Y_t = \Theta(B) Z_t. \tag{12.5}$$

This compact and elegant representation allows us to investigate some properties of ARMA(p, q) processes. Before doing so, it is important to note that the representation is not unique. Taking, for instance, the white noise process

$$Y_t = Z_t,$$

with $Z_t \sim \mathrm{N}(0, \sigma^2)$ $i.i.d.$, we can rewrite this as

$$Y_t - \phi Y_{t-1} = Z_t - \theta Z_{t-1}$$

with $\theta = \phi$. This is the same process but now written as an ARMA(1, 1) process with

$$\Phi(B) = 1 - \phi B \text{ and } \Theta(B) = 1 - \theta B.$$

Next, we will define the "inverse" of $\Phi(B)$, which we denote as $\Phi^{-1}(B)$. Such an "inverse" would be implicitly defined as

$$\Phi^{-1}(B)\Phi(B) = 1,$$

with 1 denoting the identity operator, that is, $1 Y_t = Y_t$. Let us exemplify this with an AR(1) process. In this case, we have

$$\Phi(B) = 1 - \phi B.$$

The "inverse" operator in this case results in

$$\Phi^{-1}(B) = \sum_{j=0}^{\infty} \phi^j B^j = 1 + \phi^1 B^1 + \phi^2 B^2 + \dots$$

To see this, note that we obtain a telescope sum

$$\begin{aligned} \Phi^{-1}(B)\Phi(B) &= (1 + \phi^1 B^1 + \phi^2 B^2 + \dots)(1 - \phi B) \\ &= 1 - \phi B + \phi B - \phi^2 B^2 + \phi^2 B^2 + \dots \\ &= 1. \end{aligned}$$

Such operations are also available for higher-order operators, which are, however, more difficult to construct (see McCullagh, 2018). In any case, we need the convergence of the operator series, which holds for $|\phi| < 1$. But we want to focus on convergence issues in some more depth. For now, we assume that an "inverse" operator exists, which allows us to write the ARMA(p, q) process in (12.5) as

$$Y_t = \Phi^{-1}(B)\Theta(B)Z_t.$$

Taking an ARMA(1, 1) process, we obtain

$$\begin{aligned} Y_t &= (1 + \phi^1 B^1 + \phi^2 B^2 + \dots)(1 - \theta B)Z_t \\ &= Z_t + (\phi + \theta)Z_{t-1} + \phi(\phi + \theta)Z_{t-2} + \dots \\ &= \sum_{i=0}^{\infty} \psi_i B^i Z_{t-i}, \end{aligned}$$

with $\psi_0 = 1$ and $\psi_i = \phi^{i-1}(\phi + \theta)$ for $i \geq 1$. Hence, any AR components can be transferred to an infinite sum of MA components. This holds generally, even for nonstationary processes. But for a stationary process, we can construct the time series from the past, which we call causal.

Definition 12.4 *An ARMA(p, q) process is said to be* **causal** *if the process can be rewritten as*

$$Y_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j} = \Psi(B)Z_t, \tag{12.6}$$

$$\text{with} \Psi(B) = \sum_{j=0}^{\infty} \psi_j B^j \text{and} \sum_{j=0}^{\infty} |\psi_j| < \infty \text{where} \psi_0 = 1.$$

It is useful to interpret formula (12.6) in some more depth. The formula states that the process at time point *t* can be fully explained from the past, that is, from random variables with a time index lower than *t*. This holds for the AR(1) process if $|\phi| < 1$. To see this, note that $\psi_i = \phi^i$ in this case, leading to a converging geometric series $\sum_{i=0}^{\infty} |\phi|^i$. This yields stable process behavior.

But what happens if $|\phi| > 1$? We already saw in Fig. 12.6 that the process "explodes." But let us derive this more rigorously. We write

$$\Phi(B) = (1 - \phi B) = \left(1 - \left(\phi^{-1} B^{-1}\right)^{-1}\right)$$

and expand the series around ϕ^{-1}. The "inverse" of $\Phi(B)$ results through

$$\Phi^{-1}(B) = -\sum\nolimits_{j=1}^{\infty} \phi^{-j} B^{-j},$$

which converges since $|\phi^{-1}| < 1$. This can be seen, again, by telescoping summation leading to

$$\begin{aligned}
\Phi^{-1}(B)\Phi(B) &= \left(-\sum\nolimits_{j=0}^{\infty} \phi^{-j} B^{-j}\right)\left(1 - \left(\phi^{-1} B^{-1}\right)^{-1}\right) \\
&= \left(-\phi^{-1} B^{-1} - \phi^{-2} B^{-2} - \dots\right)\left(1 - \left(\phi^{-1} B^{-1}\right)^{-1}\right) \\
&= -\phi^{-1} B^{-1} + 1 - \phi^{-2} B^{-2} + \phi^{-1} B^{-1} + \dots \\
&= 1
\end{aligned}$$

This shows that the AR(1) process with $|\phi| > 1$ can be written as

$$\begin{aligned}
Y_t &= \Phi^{-1}(B) Z_t \\
&= -\sum\nolimits_{j=0}^{\infty} \phi^{-j} B^{-j} Z_t \\
&= \sum\nolimits_{j=0}^{\infty} \phi^{-j} Z_{t+j}.
\end{aligned}$$

In other words, when $|\phi| > 1$, the AR(1) process is determined by future events rather than the past. Consequently, the process is not causal for $|\phi| > 1$. This justifies why, for most practical applications, we require $|\phi| < 1$.

Finally, we extend the ARMA process toward an ARIMA process, standing for Auto Regressive Integrated Moving Average process. The "integrated" part results from first taking differences. In other words, instead of analyzing Y_t, we look at

$$\nabla Y_t = Y_t - Y_{t-1} = (1 - B)Y_t.$$

Remember that we already applied the idea of taking differences in Fig. 12.2, where we looked at the log-returns, i.e., the difference of the log of stock prices. This leads to ARIMA models, defined as follows.

Definition 12.5 *An ARIMA(p, d, q) model is defined as*

$$\nabla^d Y_t = (1 - B)^d Y_t,$$

with Y_t being an ARMA(p, q) process.

We will not go deeper here, but we emphasize that time series modeling is an extensive and very well-developed field in statistics. A standard textbook is Brockwell and Davis (2016); see also Shumway and Stoffer (2017)

12.1.7 Autocorrelation and Partial Autocorrelation

We already presented the autocorrelation function $\gamma(h)$ and its estimation above. We now extend this toward **partial autocorrelation**, which often provides valuable information about the dependence structure. To motivate the idea, note that the autocorrelation for an AR(1) process equals

$$\rho(h) = \phi^h,$$

such that $\mathrm{Corr}(Y_t, Y_{t-h}) = \phi^h$. However, if we know Y_{t-1}, that is, if we condition on Y_{t-1}, then the conditional correlation of Y_t and Y_{t-2} is zero. This is a Markov property, which we have already utilized in Chap. 8. In fact, the order of the AR process determines the order of the required history, as sketched in Fig. 12.9. To visualize this order in the autocorrelation structure, we define the partial autocorrelation as follows. For an AR(p) process, we have

$$Y_t = \underbrace{\phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p}}_{=:\hat{Y}_{t,p}} + Z_t \ ,$$

and for general $h < p$, we define with $\hat{Y}_{t,h}$ the best linear approximation given $Y_{t-1}, \ldots, Y_{t-h}$. In other words,

$$\hat{Y}_{t,h} = \widetilde{\phi}_{1,h} Y_{t-1} + \cdots + \widetilde{\phi}_{h,h} Y_{t-h},$$

where $\widetilde{\phi}_h = (\widetilde{\phi}_{1,h}, \ldots, \widetilde{\phi}_{h,h})$ minimizes the squared prediction error,

$$\widetilde{\phi}_h = \arg\min \, E\left\{ \left(Y_t - \sum\nolimits_{j=1}^{h} \phi_{h,j} Y_{t-h} \right)^2 \right\}.$$

Note that for $h = p$, we have $\widetilde{\phi}_{p,j} = \phi_j$, and for $h > p$, we have $\widetilde{\phi}_{h,j} = 0$ for $j > p$. This, in turn, leads to $\hat{Y}_{t,h} = \hat{Y}_{t,p}$ for $h > p$. In other words, once *h* reaches the order of the AR process, the best linear approximation has already been achieved, and all additional coefficients vanish. The partial autocorrelation function of a stationary process is then defined as

$$\rho^{(\text{part})}(h) = \text{Corr}(Y_t - \hat{Y}_{t,h-1}, Y_{t-h} - \hat{Y}_{t-h,h-1}),$$

where $\hat{Y}_{t,0} \equiv 0$. This is the correlation of Y_t with Y_{t-h} after conditioning on the intermediate steps $Y_{t-1}, \ldots, Y_{t-h+1}$. Note that for $h > p$, we have

$$\begin{aligned} \rho^{(\text{part})}(h) &= \text{Corr}(Y_t - \hat{Y}_{t,h-1}, Y_{t-h} - \hat{Y}_{t-h,h-1}) \\ &= \text{Cov}(Y_t - \hat{Y}_{t,p}, Y_{t-h} - \hat{Y}_{t-h,p}) \\ &= \text{Cov}(Z_t, Z_{t-h}) \\ &= 0. \end{aligned}$$

Hence, the partial autocorrelation function provides informative insight into the order of the process. We demonstrate this by looking at an AR(3) process as visualized in Fig. 12.10. The top plot shows simulated data from the AR(3) model with $\phi_1 = 0.6$, $\phi_2 = 0.2$ and $\phi_3 = 0.2$. The middle plot gives the autocorrelation function $\rho(h)$. If, instead, we consider the partial correlation function $\rho^{(part)}(h)$, given in the bottom plot, we see that the first three values are nonzero, while all remaining values beyond lag 3 vanish.

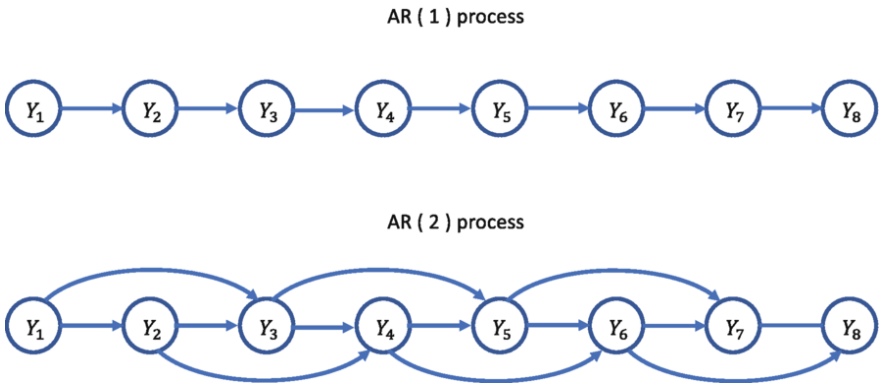
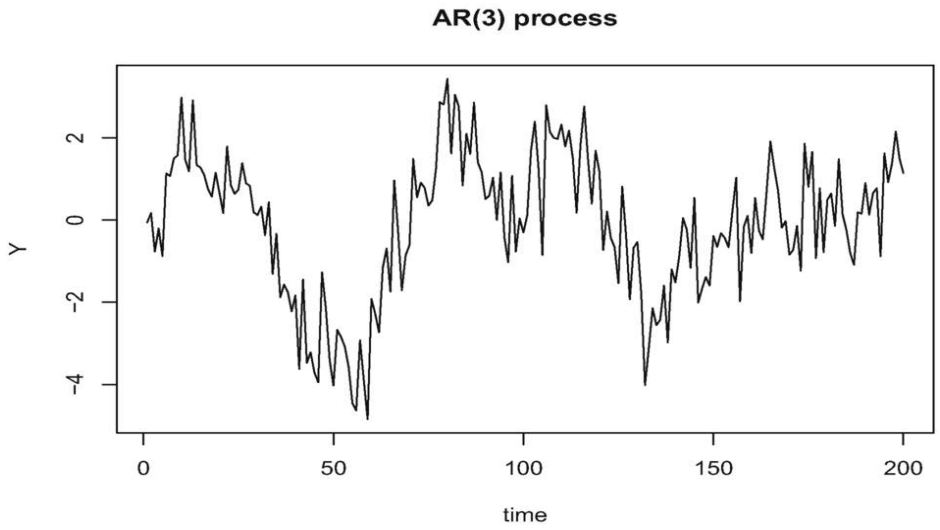
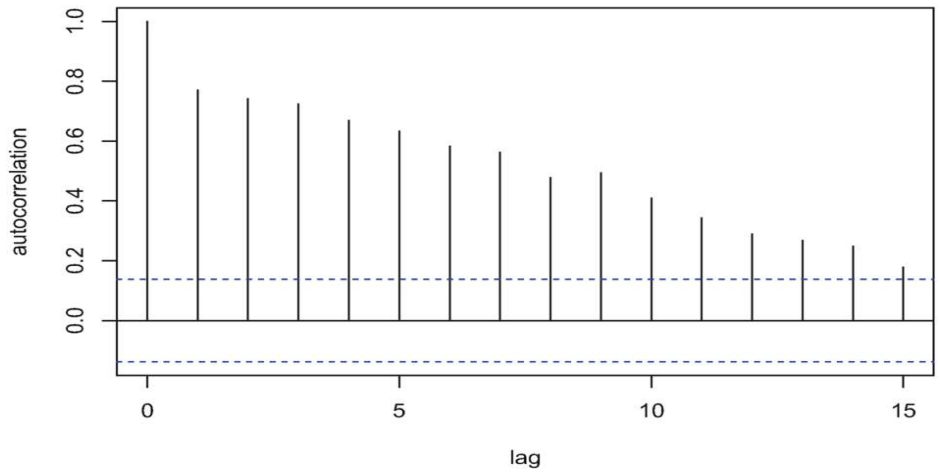


Fig. 12.9 Sketch of AR(1) (top plot) and AR(2) (bottom plot) dependence structure


AR(3) process
Y
2
0
-2
-4
0
50
100
150
200
time


autocorrelation
1.0
0.8
0.6
0.4
0.2
0.0
0
5
10
15
lag

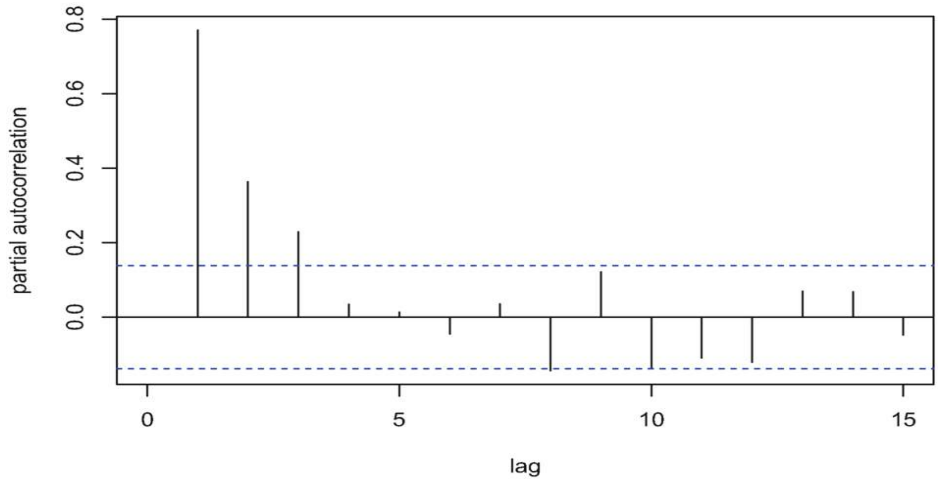

partial autocorrelation
0.8
0.6
0.4
0.2
0.0
0
5
10
15
lag

Fig. 12.10 Simulated AR(3) process (top), estimated autocorrelation function (middle), and estimated partial autocorrelation function (bottom)

12.1.8 Forecasting

One of the main purposes of time series models is to provide reliable and well-performing forecasting tools. In this respect, time series data are well-suited for more complex models trained with machine learning tools, at least if the time series is stationary and observed over a sufficiently long time window *T*. We may now question how well Y_{t+1} can be predicted with $y_t, \ldots, y_{t-p}$ being observed. This can be reformulated to constructing a prediction function $m(y_t, \ldots, y_{t-p})$, such that the mean prediction error

$$E\left\{ \left(Y_{t+1} - m(y_t, \ldots, y_{t-p})\right)^2 \middle| y_t, \ldots, y_{t-p} \right\}$$

is small. The model $m(\cdot)$ may be linear, for example,

$$m(y_t, \ldots, y_{t-p}) = \beta_1 y_t + \beta_2 y_{t-1} + \cdots + \beta_{p-1} y_{t-p},$$

which leads to the autoregressive models discussed above. Of course, we may also use nonlinear models. We do not pursue this here but refer to De Gooijer (2017).

12.1.9 Summarizing Time Series Models

Before turning to other forms of dependence among random variables, let us summarize the main ideas of time series analysis. We observe data sequentially over time, which means that at time *t*, we can look back into the past $t-1, t-2, \ldots$, but we have no data from the future. In this respect, the AR, MA, and ARMA models discussed above play a central role, particularly for those parameter configurations that yield causal models. Closely related questions concern how far dependence extends into the past and whether partial autocorrelations vanish beyond a certain lag. All of these ideas are directly relevant for forecasting future events.

While serial dependence, that is, dependence over time, is certainly one of the most commonly met non-$i.i.d.$ settings, we now turn to other important forms of dependence.

12.2 Spatial Data

12.2.1 Notation and Main Ideas

The renowned geographer Waldo Rudolph Tobler famously stated: "Everything is related to everything else, but near things are more related than distant things," a principle now known as Tobler's First Law of Geography. This idea is directly reflected in the statistical analysis of spatial data, where observations are typically not independent and identically distributed (non-$i.i.d.$). In particular, observations that are spatially close tend to exhibit stronger correlations than those that are further apart from each other.

Spatial data can have different structures and formats, and modeling them requires a wide range of different tools. We provide a short overview and refer to literature for more in-depth treatments. The distinction between different spatial data types is best understood with an example. Figure 12.11 shows data collected in a survey of households in Munich. Households were randomly selected (based on a random sample of the registry office), and points show the location of each household in Munich. The survey focuses on rented apartments; so that, to be more precise, each dot corresponds to the coordinates of a sampled rented apartment. The exact location is slightly coarsened in the data for confidentiality reasons, but that is of no relevance to us now. For each sampled apartment *i*, we have the geo-coordinates $s_i = (s_{i1}, s_{i2})$ given in longitude s_{i1} and latitude s_{i2} for $i = 1, \ldots, n$.

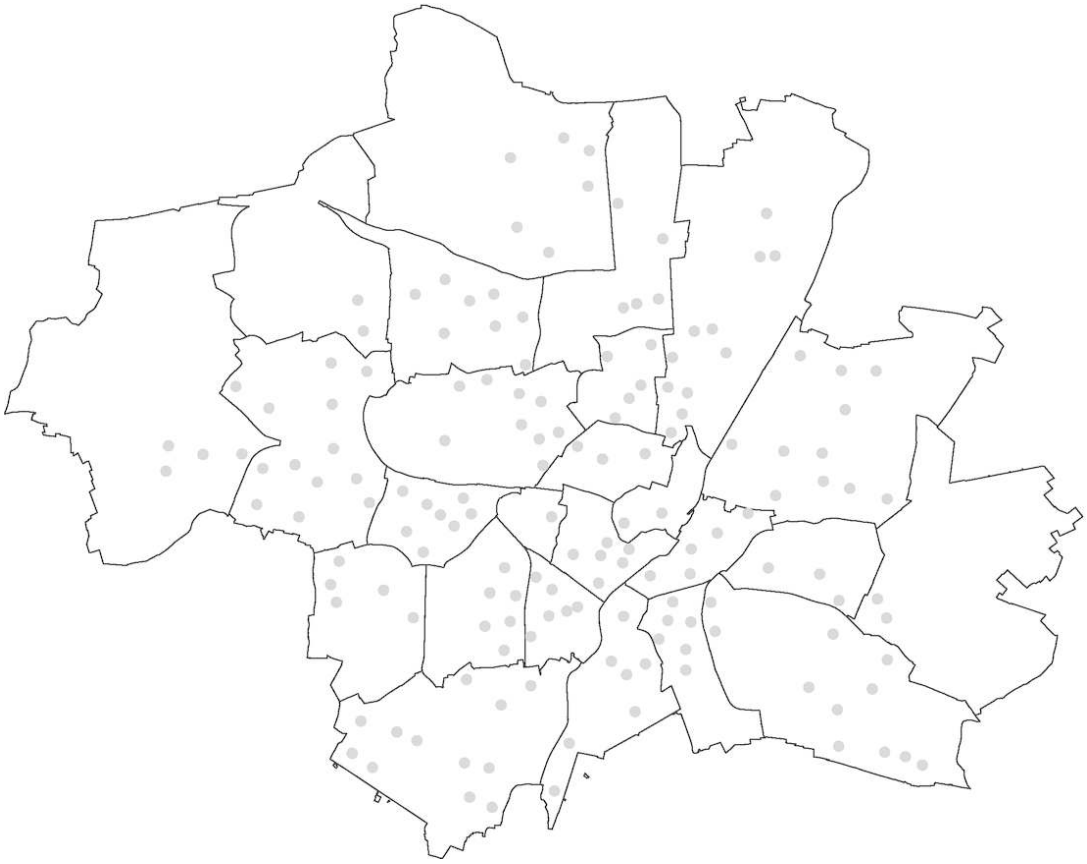


Fig. 12.11 Geolocation of sampled apartments in Munich

In principle, the location s_i itself can be the variable of interest. For instance, we might question: Is the intensity of rented apartments higher in some areas of Munich? Presumably, in the suburbs, we could find more family houses and hence fewer rented apartments. The opposite seems plausible for the city center. In that case, we would consider the location s_i as the variable of interest and aim to model the distribution that generates s_i. We do not pursue this aspect here but refer exemplarily to Baddeley et al. (2007). Instead, throughout this chapter, we will utilize location as information to capture spatial correlation.

We focus on data settings where additional information beyond the locations s_i is available. In the rental example, this entails the rent Y_i of the *i*-th drawn apartment located in s_i. In this setting, Tobler's law suggests that if the rent of two apartments is more correlated, the closer the apartments are. Let Y_i denote the rent per square meter of apartment *i* and Y_j the rent per square meter of apartment *j*. Then Y_i and

Y_j are correlated, and the dependence decreases proportionally to the distance between the apartments increasing, that is, with $\|s_i - s_j\|$ increasing. This is typically modeled with a so-called **isotropic covariance**, which means

$$\mathrm{Cov}(Y_i, Y_j) = c\big(\|s_i - s_j\|\big) \tag{12.7}$$

for some monotonically decreasing function $c(.)$. It is important to note that we need additional assumptions to guarantee valid covariance matrices. We refer to the literature provided at the end of this section for more on this aspect.

In spatial statistics, it is common to work with the **variogram**, which is defined through

$$\gamma\big(\|s_i - s_j\|\big) := \mathrm{Var}(Y_i - Y_j) = \mathrm{Var}(Y_i) + \mathrm{Var}(Y_j) - 2\mathrm{Cov}(Y_i, Y_j),$$

which with (12.7) results to

$$\gamma\big(\|s_i - s_j\|\big) = \mathrm{Var}(Y_i) + \mathrm{Var}(Y_j) - 2c\big(\|s_i - s_j\|\big).$$

Assuming that the variance does not depend on the location itself, we see that

$$c\big(\|s_i - s_j\|\big) = -\gamma\big(\|s_i - s_j\|\big) + \sigma^2$$

where $\sigma^2 = \mathrm{Var}(Y_i) = \mathrm{Var}(Y_j)$.

For the Munich rental data, the resulting variogram is shown in Fig. 12.12. The variogram shows the variance of the difference of rents per square meter for two apartments which are a given distance apart. The fitted curve shows that there is spatial correlation up to a distance of about 2 km. Beyond that, the variogram flattens out at approximately twice the marginal variance. We include a smooth fitted curve in the plot to account for the estimation variability of the point estimates shown in the plot. These point estimates are derived as follows. We first discretize the distances to $0 < d_1 < d_2 < d_3 \ldots$. For a given distance d_k, we then estimate

$$\widehat{\gamma}(d_k) = \frac{1}{|N(d_k)|} \sum\nolimits_{i,j \in N(d_k)} (Y_i - Y_j)^2$$

where

$$N(d_k) = \left\{ (i,j) : \left| \|s_i - s_j\| - d_k \right| < \Delta \right\}.$$

Hence, $N(d_j)$ is the set of observation pairs with locations that have a distance of about d_k, or, to be more specific, where the distance of the locations differs from d_k less than Δ. The bandwidth Δ needs to be chosen in advance.

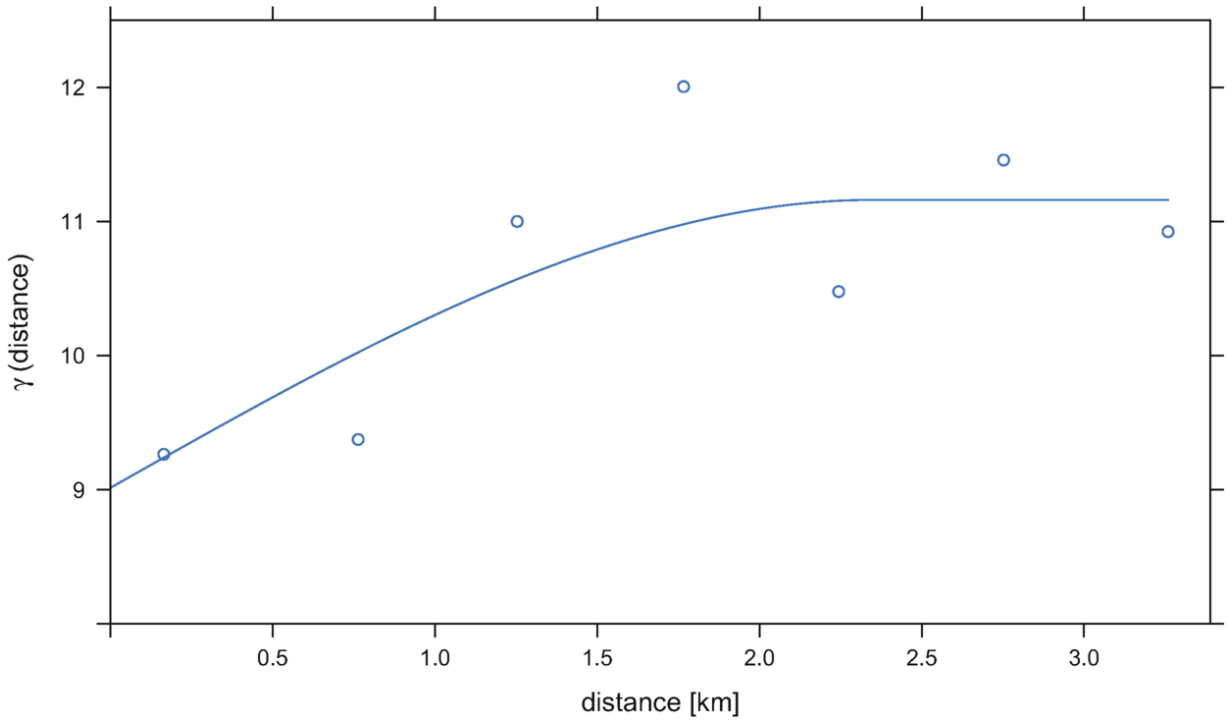


Fig. 12.12 Variogram for rents of apartments in Munich

The definition of a distance between two locations s_i and s_j is itself relevant. At this point, it is worth emphasizing that distance in principle can be any kind of distance. In many cases, the Euclidean distance is a practical choice, that is,

$$d = \|s_i - s_j\|_2,$$

where $\|\cdot\|_2$ refers to the scalar product of the coordinates. However, the definition of the variogram also applies to different distance measures. For example, if the data live on a street network, the relevant distance may be the shortest driving route rather than Euclidean distance. The latter is also known as “Manhattan distance.” Figure 12.13 illustrates this difference.

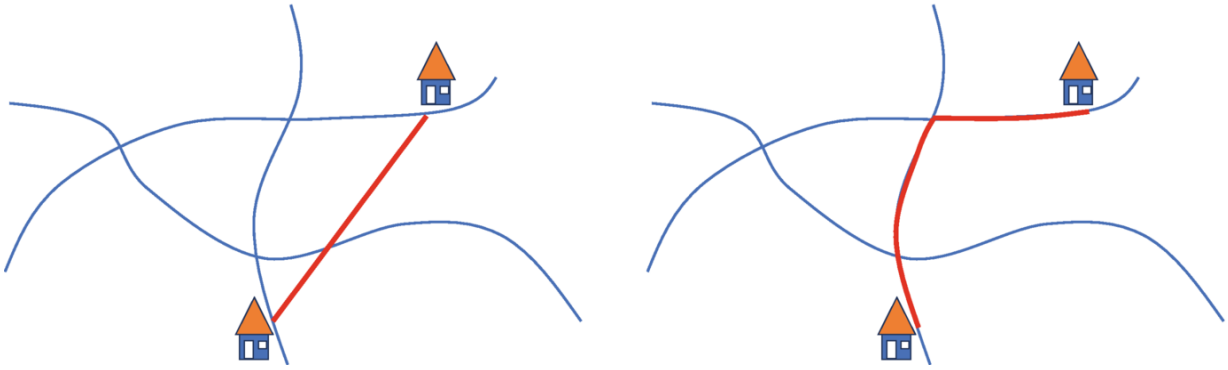


Fig. 12.13 Distance between two locations, Euclidean (left) and based of the shortest route (right)

Conceptually, this changes little; the main requirement is that one can compute pairwise distances d_{ij} under the chosen metric. Ideally, the distance should also be symmetric (implying no one-way streets).

12.2.2 Kriging and Spatial Smoothing

Now that we have defined and estimated the spatial covariance structure, we can approach the next question: Can we model or visualize the spatial effect? The underlying idea is to decompose the observations into a spatially correlated component and an independent noise component, much as in a spatial regression model. Following Bivand et al. (2008, p. 169), a practically useful strategy is to decompose Y into a spatially smooth and an unstructured component, that is,

$$Y_i = h(s_i) + \epsilon_i,$$

where $h(\cdot)$ is a smooth component that is estimated from the data and ϵ_i are independent zero mean variables.

We can consider $h(s)$ as a spatial process with covariance structure $c(\cdot)$ dependent on the distance between two observations. The function $h(s)$ can then be estimated through a concept called **kriging**, which traces back to Krige (1951). The main idea is that we can consider $Y(s)$ as a spatial process with s as the corresponding location. We observe the process at locations $s_1, \ldots, s_n$ leading to data $Y(s_1), \ldots, Y(s_n)$, or in short $Y_1, \ldots, Y_n$. We assume that $Y(s)$ has covariance matrix V constructed from the covariance structure $c\big(\|s_i - s_j\|\big)$, that is, $V_{ij} = \mathrm{Cov}\big(Y(s_i), Y(s_j)\big) = c\big(\|s_i - s_j\|\big)$. We now aim to estimate the

value of the process at location s_0, given the observed values $y = (y(s_1), \dots, y(s_n))^\top$ or, again, in short $y = (y_1, \dots, y_n)^\top$. Utilizing properties of the multivariate normal distribution, we obtain the predictor

$$\hat{Y}(s_0) = wV^{-1}y,$$

where

$$\begin{aligned} w &= \mathrm{Cov}\big(Y(s_0), (Y(s_1), \dots, Y(s_n))\big) \\ &= \big(C(\|s_0 - s_1\|), C(\|s_0 - s_2\|), \dots, C(\|s_0 - s_n\|)\big). \end{aligned}$$

The justification behind the formula is as follows. We assume that $\big(Y(s_0), Y(s_1), \dots, Y(s_n)\big)^\top$ follows a normal distribution with

$$\begin{pmatrix} Y(s_0) \\ Y(s_1) \\ \vdots \\ Y(s_n) \end{pmatrix} \sim \mathrm{N}\left(0, \quad \begin{pmatrix} \sigma^2 & w \\ w^\top & V \end{pmatrix}\right), \tag{12.8}$$

where $\sigma^2 = C(0) = \mathrm{Var}\big(Y(s_0)\big)$. Conditioning on $y = (y_1, \dots y_n)^\top$ leads to

$$Y(s_0)\big|y \sim \mathrm{N}(wV^{-1}y, \sigma^2 - wV^{-1}w), \tag{12.9}$$

which gives the kriging predictor $\hat{Y}(s_0)$ as conditional mean value. It is helpful to note that the formulae above have been utilized before, in Sect. 8.4. In fact, formula (12.9) is a special case of (8.20), while (12.8) is a repetition of (8.18).

We visualize the resulting kriging function in Fig. 12.14. We see that rents in the city center are generally more expensive, which is not surprising.

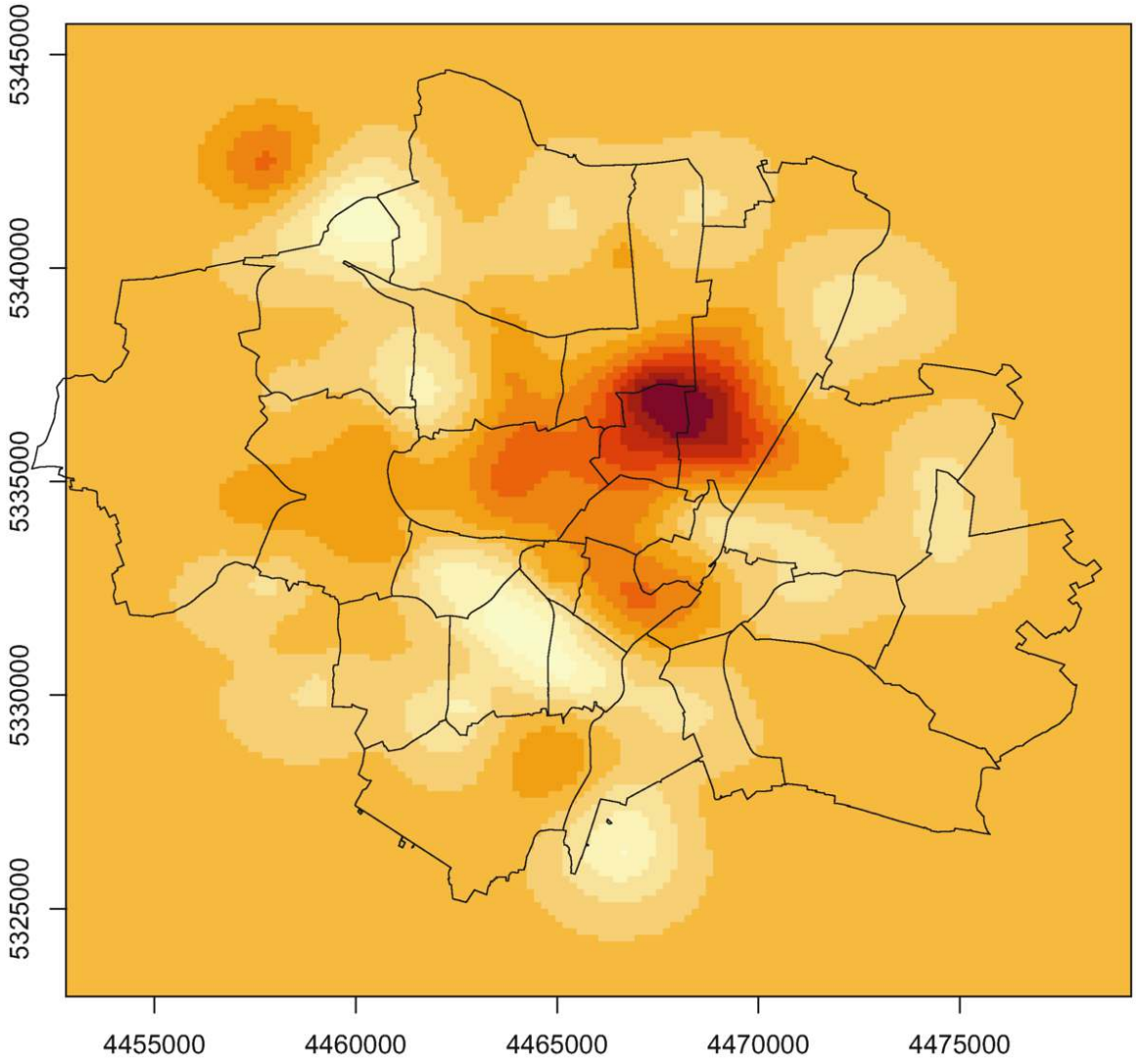


Fig. 12.14 Kriging of rents. Shown is the predicted average rent

An alternative approach results from spatial regression. In this case, we assume

$$Y(s) = h(s) + \epsilon,$$

but now $h(s)$ is considered as a smooth function which can be estimated through, for instance, spline smoothing (Wood, 2017). The residuals ϵ are, on the other hand, taken as white noise, meaning that they are considered independent. This approach builds on the ideas introduced in Chap. 6.

Summing up, when modeling spatial data, two main strategies can be pursued: One focuses on capturing spatial dependence through the

covariance structure, while the other incorporates spatial effects directly into the mean function. This naturally leads to the question of which approach is "correct." There is no universal answer, and maybe the best answer is to quote the prominent spatial statistician Noel Cressie: "One person's mean is another person's covariance" (see Stein & Wikle, 2024). In essence, the choice between the two approaches often comes down to perspective and modeling preference; both are valid and capable of accounting for spatial dependence, and hence, the separation between mean and covariance is model-dependent and often blurred.

12.2.3 Lattice and Grid Data

Often, the location of the data is not observed exactly but only at the level of administrative units such as city districts or counties. In such cases, one can either approximate the location with the centroid of the district or use the resulting lattice to incorporate the spatial structure. We demonstrate this for the city of Munich. In Fig. 12.15, we sketch the lattice structure resulting from the districts and their neighborhood structure. Nodes represent districts, and an edge between them exists if the districts share a boundary. The resulting graph looks like a graphical model, and indeed, one can model the spatial structure this way. To be specific, let $Y(s_i)$ denote the *i*-th observation, where $s_i \in \{1, \ldots, K\}$ denotes now the district in which apartment *i* is located. Tobler's law then suggests that apartments in neighboring districts are more related to each other than apartments in districts that are far away from each other. We can model this by including a spatially correlated district effect, that is,

$$Y(s_i) = U(s_i) + \epsilon,$$

where $U(s)$ is the district effect of district s_i. Given that s_i now refers to the district in which apartment *i* is located, we have $s_i \in \{1, \ldots, K\}$. Consequently, we can write the district effects as $U(1), \ldots, U(K)$ or in short $U_1, \ldots, U_K$ and assume a spatial structure for $U_1, \ldots, U_K$. The latter is incorporated through the neighborhood structure by inducing a graphical model for $U_1, \ldots, U_K$. We thereby postulate

(12.10)

$$U_k \perp\!\!\!\perp U_l \,|\Big(\{U_1, \ldots, U_k\} \backslash \{U_k, U_l\}\Big),$$

i.e., U_k and U_l are independent if and only if they are not neighbors. It is helpful to reformulate (12.10) in relation to Gauss-Markov random fields, discussed in Section 8.5. Indeed, defining with $N_l = N(s_l)$ the neighbors of grid cell s_l and utilizing results on graphical models, we can rewrite (12.10) to

$$U_l \perp\!\!\!\perp \{U_j, \text{ with } s_j \notin \{s_l \cup N_l\} \,|N_l$$

Thus, the spatial structure is reduced to a dependence structure on the district level. We exemplify this using the rental data and show the corresponding estimate for the spatial effect in Fig. 12.16.

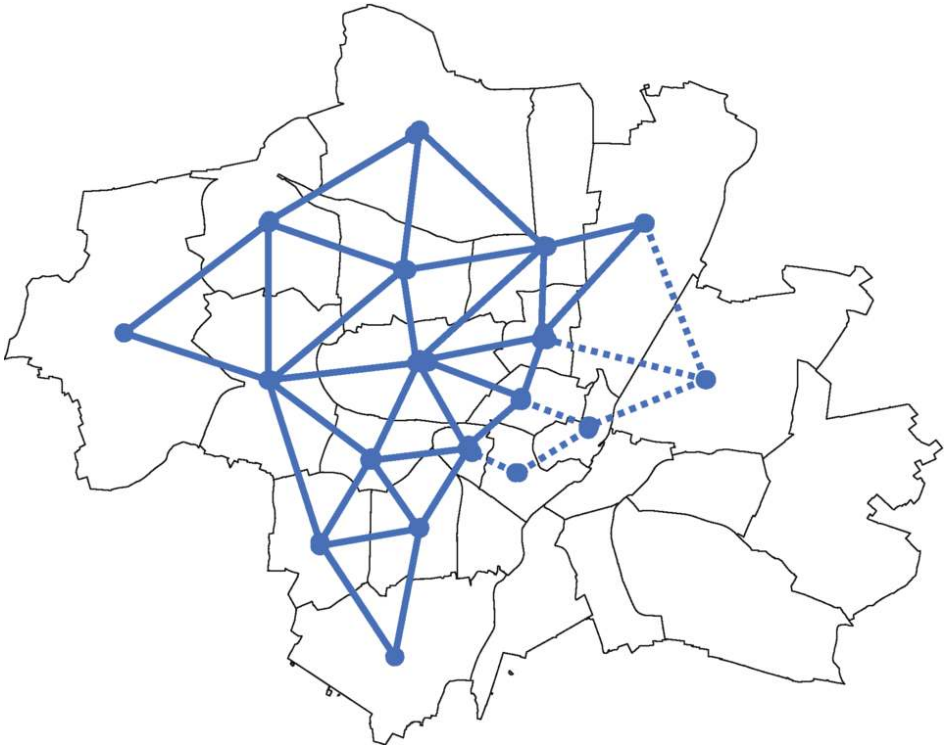


Fig. 12.15 Sketched lattice and neighborhood structure of districts in Munich

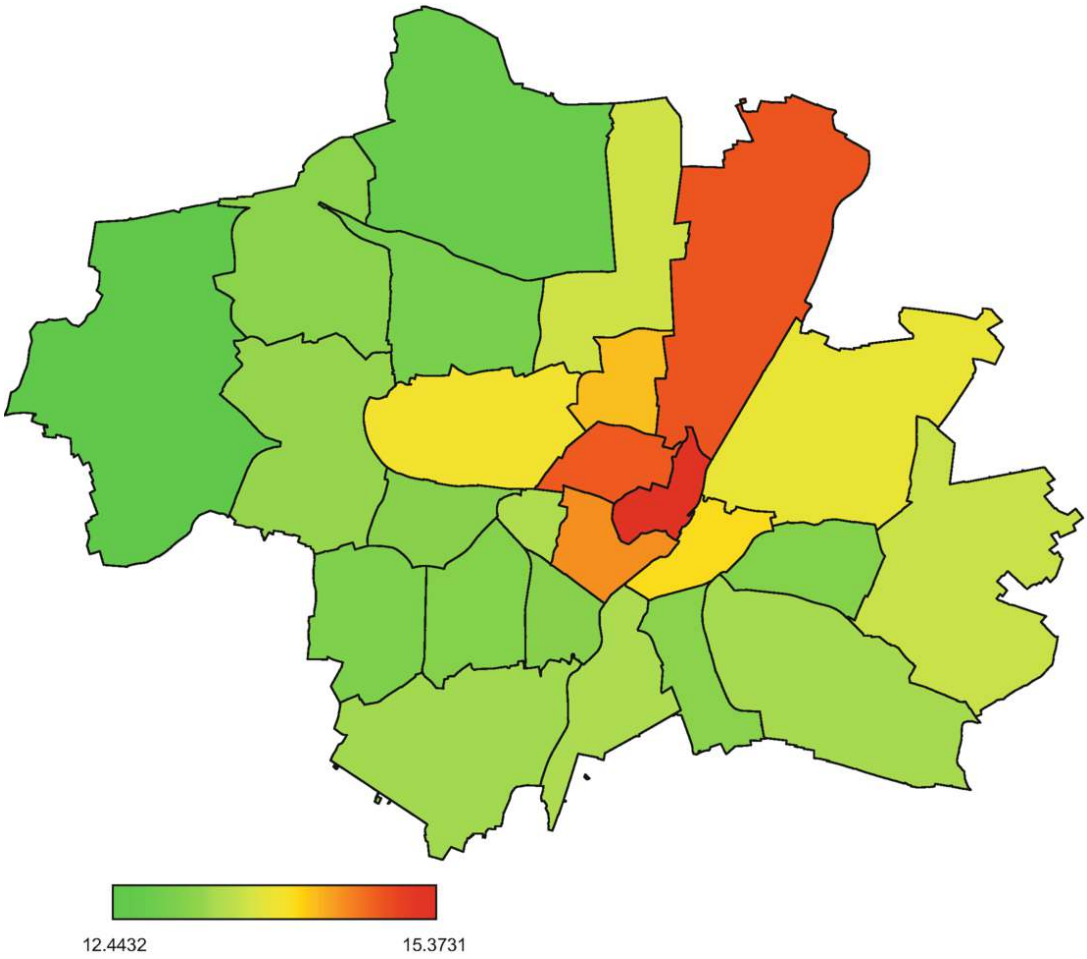


Fig. 12.16 Average rent per district based on spatial lattice data

12.2.4 Summarizing Spatial Data

Spatially correlated data require special attention and, hence, sophisticated statistical treatment. This goes far beyond the scope of this chapter. We refer to the following literature for further reading. One of the most central books in this field is undoubtedly the monograph by Cressie (2015), which first appeared more than three decades ago in 1993. A more applied focus with applications in R is given in Bivand et al. (2008). We also refer to Ripley (2005) for a comprehensive discussion of the field. Parallel to the development of spatial statistics, a substantial literature on spatial econometrics developed over the years; see, e.g., Anselin (1988). A discussion on the different views and aspects of spatial statistics and spatial econometrics is found in Kauermann et al. (2012).

12.3 Network Data

12.3.1 Notation and Main Ideas

Phrases such as "networks are ubiquitous" (see, e.g., Fienberg, 2012; Goldenberg et al., 2010) have themselves become ubiquitous, reflecting the traction network data analysis gained in recent years. In this research area, networks represent relational structures between entities that are generally not conditionally independent of one another. For instance, in a realistic setting, the probability of two pupils to be friends could depend on whether they already share common friends in the class. Contrary to this observation, the bulk of statistical research from the last century is limited to the study of attributes of population members under conditional independence assumptions. In the prior example, this would imply assuming that the probability of becoming friends with someone is independent of existing friendships. For instance, R. Fisher famously tried to investigate the relationship between smoking behavior and cancer (Fisher, 1958a,b) by relying on a sample of around 82 twins to show that smoking habits and cancer can be traced back to inheritance. While numerous studies based on meta-analyses provided evidence of there being a relationship between smoking habits and cancer (Gandini et al., 2008), the social embedding of population members may affect smoking habits as well.

The underlying question in this context is: How is individual behavior affected by the behavior of other actors in the same network and the network itself? In sociology, terms coined for these two mechanisms are "*selection*" and "*influence*." Whereas "*selection*" refers to behavior affecting the relations between actors in the network, the "*influence*" mechanism implies that relations between actors affect their behavior. In the statistical network literature, those phenomena are more frequently called homophily and social contagion, and they are generally not identifiable from observational data (Shalizi & Thomas, 2011). Going further, it is plausible that behaviors affect one another, just like social relations depend on one another. In this context, the focus of this chapter lies on studying networks conditional on behavior and other attributes.

In particular, two types of factors may affect the probability of a tie between two population members:

1. How one relation affects other relation (*endogenous* factors).
2. How relations are affected by variables outside of the network (*exogenous* factors).

Returning to the example on smoking and friendships, an endogenous factor would be the tendency to be friends with friends of your friends, while an exogenous factor would be the tendency to form friendships who share same smoking (or nonsmoking) habits.

Example 12.4 As a running example, we study the friendship network of a university in the UK (Nepusz et al., 2008). The network encompasses 81 students divided into 3 groups based on their affiliation. The network is visualized in Fig. 12.17. A relation in this network indicates a personal relationship between the respective actors. The affiliation attribute for the *i*th student is x_i. A first impression of the network visualization is that many connections exist between actors with the same affiliation and fewer connections if this is not the case. Such visual impressions call for a more rigorous analysis of network data, which we sketch subsequently. ▷

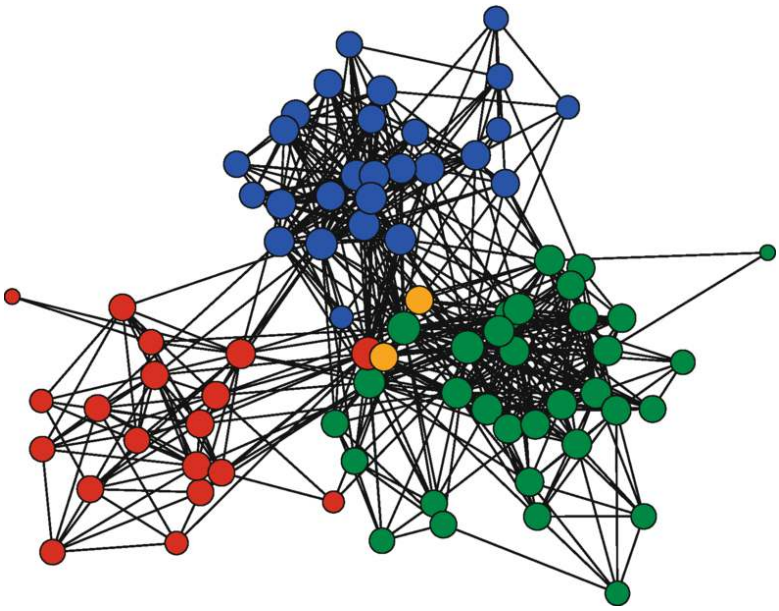


Fig. 12.17 Network of personal friendships of a faculty of a UK university. The color of the nodes represents the affiliation of the academics

The analysis of network data comes up in a wide array of disciplines, ranging from statistics and social sciences to computer science and physics. Therefore, the field is increasingly multidisciplinary, with jargon varying across disciplines. To keep the presentation precise, we commit to the terminology of edges and actors.[1] Networks encompass edges between a set of actors, denoted by $\mathcal{N} := \{1, \dots, N\}$. We represent a network by an adjacency matrix $y = (y_{i,j})$, where $y_{i,j} = 1$ indicates that there is an edge between actors *i* and *j* and $y_{i,j} = 0$ indicates its absence. Generally, $y_{i,j} \neq y_{j,i}$ holds, meaning that there may be an edge from actor *i* to *j* but not vice versa. A network where this is the case is called a *directed* network. Else, the network is *undirected*. Since this section mainly serves introductory purposes, we will focus on the latter case, setting $y_{i,j} = y_{j,i}$. We further assume that there are no self-loops (i.e., ties from an actor to itself) and set $y_{i,i} = 0$ for all $i \in \mathcal{N}$. The set $\mathcal{Y}$ is defined to include all possible networks between *N* actors. In many cases, further information at the actor level, $\boldsymbol{x}_i$ for $i \in \mathcal{N}$, or edges, $\boldsymbol{x}_{i,j}$ for $i \neq j \in \mathcal{N}$, is available.

12.3.2 Descriptive Statistics

Beyond visualizing the network as in Fig. 12.17, we can calculate statistics to understand the structure of a network. As a first step, we define measures at the actor level, which we can then aggregate to the network level. We characterize the position of a particular actor $i \in \mathcal{N}$ through the number of edges it is involved in:

$$\deg_i(y) \quad = \quad \sum_{j \neq i} y_{i,j}. \tag{12.11}$$

These types of statistics refer to a particular kind of centrality of actors in networks, and numerous related statistics were proposed in the literature (Freeman, 1978). In Fig. 12.18, the distribution of the degrees of each actor in the network is visualized. It is often observed in empirical settings that as the number of actors in a network grows, its degree increasingly resembles the form of a power law (Barabási & Albert, 1999), getting more and more skewed to the right.

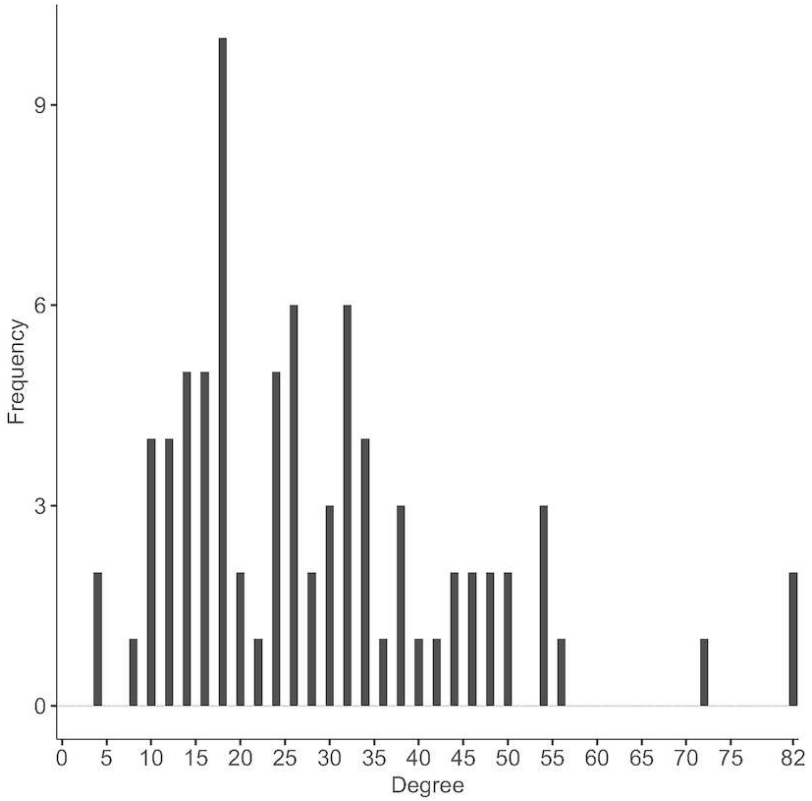


Fig. 12.18 Degree distribution of UK faculty network

While the degrees of actors characterize their local connectivity, researchers are also interested in global connectivity patterns arising from these local structures. A natural approach is to aggregate (12.11), i.e., summing up the degrees of all *N* actors. Since the scale of this sum will depend on the number of involved actors *N*, we divide the result by $N\ (N-1)\ /\ 2$ to obtain a measure, henceforth called the density of $\boldsymbol{y}$, which is independent of *N* and lies between 0 and 1:

$$\text{dens}(\boldsymbol{y}) \quad := \quad \frac{1}{N\ (N-1)\ /\ 2} \sum_{i=1}^{N} \deg_i(\boldsymbol{y}). \tag{12.12}$$

In the running example, the density is at 0.178, implying that around $17.8\,\%$ of all possible friendships in the network are observed. Compared to most networks, this indicates a relatively dense network. In larger networks, one often expects the density to decrease as $N \to \infty$. Of course, many additional summary statistics can be computed, such as clustering coefficients or hub scores, but we do not pursue these here; see Sect. 12.3.4 for further references.

12.3.3 Statistical Model for Networks

Descriptive methods for network data analysis are essential. However, they do not provide readily available uncertainty quantification, nor do they account for the fact that the observed network is often not the whole population network but rather a sample of it. In the following, we introduce two general approaches based on a probabilistic model for $P(Y = y; \theta)$, characterized by some parameter θ, where $Y \in \mathscr{Y}$ is treated as a matrix-valued random variable with random $0/1$ entries, while y denotes the observed network.

The major obstacle in this context is that edges in a network should generally not be assumed to be independent of one another. There are several approaches to tackle this problem (De Nicola et al., 2023):

- **Exponential Random Graph Models** (ERGMs) specify statistics to capture dependencies and account for them in a statistical framework (Lusher et al., 2013).
- **Latent Variable Models** assume that edges are independent of one another when conditioning on some actor-specific latent variables. These latent variables can be discrete or continuous. In the discrete case, they may represent latent categories or group memberships and thereby enable community detection (Snijders & Nowicki, 1997). In the continuous case, they may represent positions in a latent space, where the distance between actors accounts for the dependence (Hoff et al., 2002). Similarly, graphon models rely on a latent, but continuous, structure (Sischka & Kauermann, 2022).

Building on the theory for exponential families (Efron, 2022), ERGMs specify a joint distribution over $y \in \mathscr{Y}$:

$$P(Y = y; \theta) \quad = \quad \frac{\exp\left(\theta^\top s(y)\right)}{\kappa(\theta)}, \tag{12.13}$$

where

- $s : \mathscr{Y} \mapsto \mathbb{R}^p$ is a vector valued function that evaluates p sufficient statistics on any network $y \in \mathscr{Y}$ (examples follow underneath).
- $\theta \in \mathbb{R}^p$ is the coefficient vector weighting all sufficient statistics which needs to be estimated from the observed network.

- $\kappa(\theta) = \sum_{y \in \mathscr{Y}} \exp\left(\theta^\top s(y)\right)$ is a normalizing constant guaranteeing that (12.13) is a proper probability distribution.

The core component of Eq. (12.13) is the vector of sufficient statistics $s(y) = (s_1(y), s_2(y), \ldots, s_p(y))^\top$, which encapsulates the salient global features of the network to be represented in the model. For the UK faculty network in Example 12.4, these may be functions of the network y, such as the number of edges in the network

$$s_{\texttt{edges}}(y) \;\; = \;\; \sum_{i<j} y_{i,j},$$

comparable to (12.12). We can also include statistics such as the sum of edges between actors from a matching faculty

$$s_{\texttt{nodematch}}(y) \;\; = \;\; \sum_{i<j} y_{i,j}\, \mathbb{I}(x_i = x_j).$$

Beyond the assumption of dyadic independence, it is common to include higher-order dependence terms reflecting network transitivity. However, Schweinberger (2011) demonstrated that those types of statistics often induce degeneracy, making the model unsuitable for most applications. To mitigate this, we represent transitivity by counting edges that share a common neighbor:

$$s_{\texttt{transitive}}(\boldsymbol{y}) = \sum_{i<j} y_{i,j}\, \mathbb{I}\Big(\sum_{h \neq (i,j)} y_{i,h}\; y_{h,j} > 0\Big),$$

where the latter component takes the value 1, if student *i* and student *j* have common friends in the network.

Parameter estimation in ERGMs can be carried out via maximum likelihood. This is not as easy as it sounds, since the normalization constant $\kappa(\theta)$ is not available in analytic form. But we have discussed several remedies in Chap. 8 to overcome this problem. Implemented procedures to fit ERGMs are available using the `ergm` package in `R` (Hunter et al., 2008).

To illustrate the role of mutual dependence in the network, we compare this model with a classic regression approach assuming conditional independence, that as such excludes the transitivity term. In other words, we assume $y_{i,j}$ to be independent of $y_{i,k}$ for $j \neq k$. The results, summarized in Table 12.1, demonstrate that incorporating dependence terms substantially improves model fit, as indicated by

lower AIC and BIC values and a higher log-likelihood. The resulting estimated parameters are shown in Table 12.1. The positive and substantial coefficient associated with the "Nodematch Affiliation" statistic provides strong evidence that edges are more likely to form between actors with the same affiliation. Beyond that, the positive coefficient for the "Transitive Ties" statistic indicates an increased probability of edge formation between pairs of actors who share a common neighbor. These results highlight the presence of transitivity, or "triadic closure," in the network.

Table 12.1 Results for the UK Faculty network of the ERGM models with (left column) and without (right column) dependence-inducing terms

12.3.4 Summarizing Network Data

In their most basic form, network data consist of simple binary indicators $y_{i,j} \in \{0, 1\}$. However, these observations cannot be treated as independent: Edges often exhibit intricate and mutually reinforcing dependence structures that may themselves be of substantive interest. That is, the presence of an edge $y_{i,j}$ typically depends on the configuration of other edges in the network. Such dependencies lead to complex modeling frameworks, one of the most important being exponential random graph models (ERGMs). Various alternative models have also been developed and applied in practice, and we refer the reader to the recent literature for further details. A general overview is given in Kolaczyk (2009) and Barbour and Reinert (2026), whereas Lusher et al. (2013) focuses on ERGMs.

12.4 Functional Data

12.4.1 Notation and Main Ideas

Functional data analysis (FDA) uses curves, images, or functions as observations. In the following, we will focus on curves as a prominent example, but most of the presented techniques also apply to different types of functional data. In FDA, observations are usually written as

functions, i.e., $Y(t)$ instead of the previously introduced index notation Y_t. Conceptually, functional data typically involves a large (or, theoretically, an infinite) number of measurements per observational unit. Hence, we obtain data $Y_i(t), t \in \mathscr{T}$ per observation unit *i*, observed on a continuous domain $\mathscr{T}$. Typically, we assume the units to be independent, while the observations within each unit are dependent. In practice, as only a finite number of observed values of each function can be stored, functions $Y_i(t)$ are observed on grid points $t_{i,r} \in \mathscr{T}$, yielding the observations $y_i(t_{i,r}), r = 1, \ldots, R_i, i = 1, \ldots, n$. A common use case is $t_{i,r} \equiv t_r$, i.e., all functions Y_i are observed on the same $R_i \equiv R$ grid points. If, in addition, all observed time points are equidistant, i.e., $t_{j+1} - t_j = t_{k+1} - t_k \; \forall j, k \in \{1, \ldots, R-1\}$, the data can be represented and stored in an $n \times R$-matrix.

12.4.2 Analysis and Properties of Functional Data

Functional data can be analyzed in different ways. The most common approach is to plot all functions as curves over their respective observed time horizons using a line plot. We exemplify this in Fig. 12.19, where we plot the temperature in Montreal, Canada, for years 1972, 1975, and 1992.

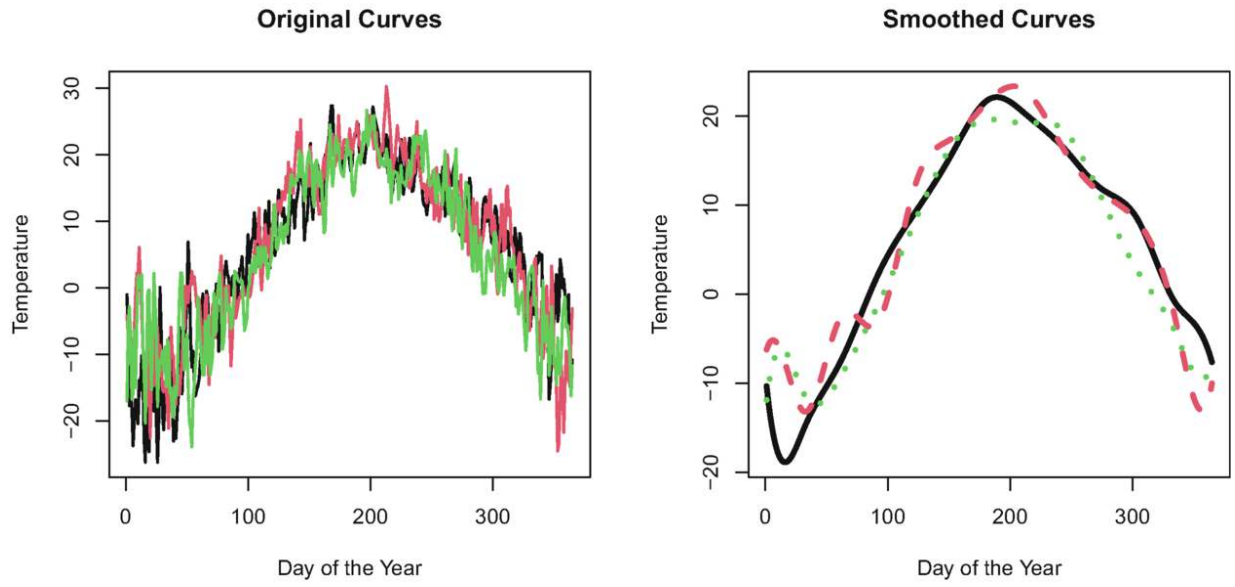


Fig. 12.19 Line plot of original (left) and smoothed functional data showing temperatures in Montreal with different lines corresponding to different years

A possible preprocessing step before further analyzing functional data is to apply pre-smoothing of the curves (Ramsay & Silverman, 2005) to remove observation noise using nonparametric regression or smoothing techniques. We show this in the right panel of Fig. 12.19 for the 3 years considered. Whereas this allows obtaining truly functional observations, it is challenging to account for the uncertainty in estimating the smoothed versions of observations $Y_i(t)$ in subsequent analyses. Newer approaches, therefore, directly deal with the observed unsmoothed data by imposing some form of smoothness requirement in the method of choice (see, e.g., Yao et al., 2005; Scheipl et al., 2015; Greven & Scheipl, 2017; Goldsmith et al., 2020).

Another strategy, akin to pre-smoothing, involves expanding the functions in a basis prior to analysis (see, e.g., Morris & Carroll, 2006). However, instead of subsequently working with the pre-smoothed functions, one continues to work with the multivariate vector of basis coefficients. For splines, this approach closely resembles pre-smoothing and usually also leads to ignoring the resulting uncertainty in curve reconstruction.

12.4.3 Functional Principal Component Analysis

Functional principal component analysis (FPCA) is a core statistical technique in FDA. The principle of FPCA can be derived from the Karhunen-Loève expansion, stating that any stochastic process $Y(t)$ can be represented as an infinite series of uncorrelated random processes. This provides the foundation for dimension reduction in FDA, as the infinite-dimensional functional data can be compressed into a manageable set of finite principal component scores, defined as follows:

$$Y(t) = \mu(t) + \sum_{m=1}^{\infty} \xi_m \phi_m(t), \quad t \in \mathcal{T}. \tag{12.14}$$

In (12.14), μ represents the mean function of the stochastic process $Y(t)$, the $\phi_m(t)$ are orthonormal eigenfunctions, and the ξ_m are uncorrelated random scores. Using this reformulation, the covariance of the stochastic process Y can be represented as

$$\text{Cov}(Y(s), Y(t)) = \sum_{m=1}^{\infty} e_m \phi_m(s) \phi_m(t), \quad t, s \in \mathcal{T} \tag{12.15}$$

where the eigenvalues e_m are by convention sorted as $e_1 \geq e_2 \geq \cdots \geq 0$. As in classical principal component analysis, the eigenvalues can be used to assess the variance explained by only using m^* functional principal components, and, on this basis, the sum in Eq. (12.14) can be truncated to provide an approximation to $Y(t)$ with a finite number of basis functions. This allows for a sample of independent replications of $Y(t)$ to be represented by a set of finite functions, common to all observations, and individual scores that capture function-specific deviations from the mean. We also refer to Sect. 8.4 where we utilized the idea standing behind (12.14) for the simulation of multivariate normally distributed random variables.

In addition to providing a measure of complexity, FPCA is useful in dimension reduction and also allows for the prediction of whole functions from sparse individual data. Figure 12.20 visualizes the functional principal components for the before-shown curves from Fig. 12.19.

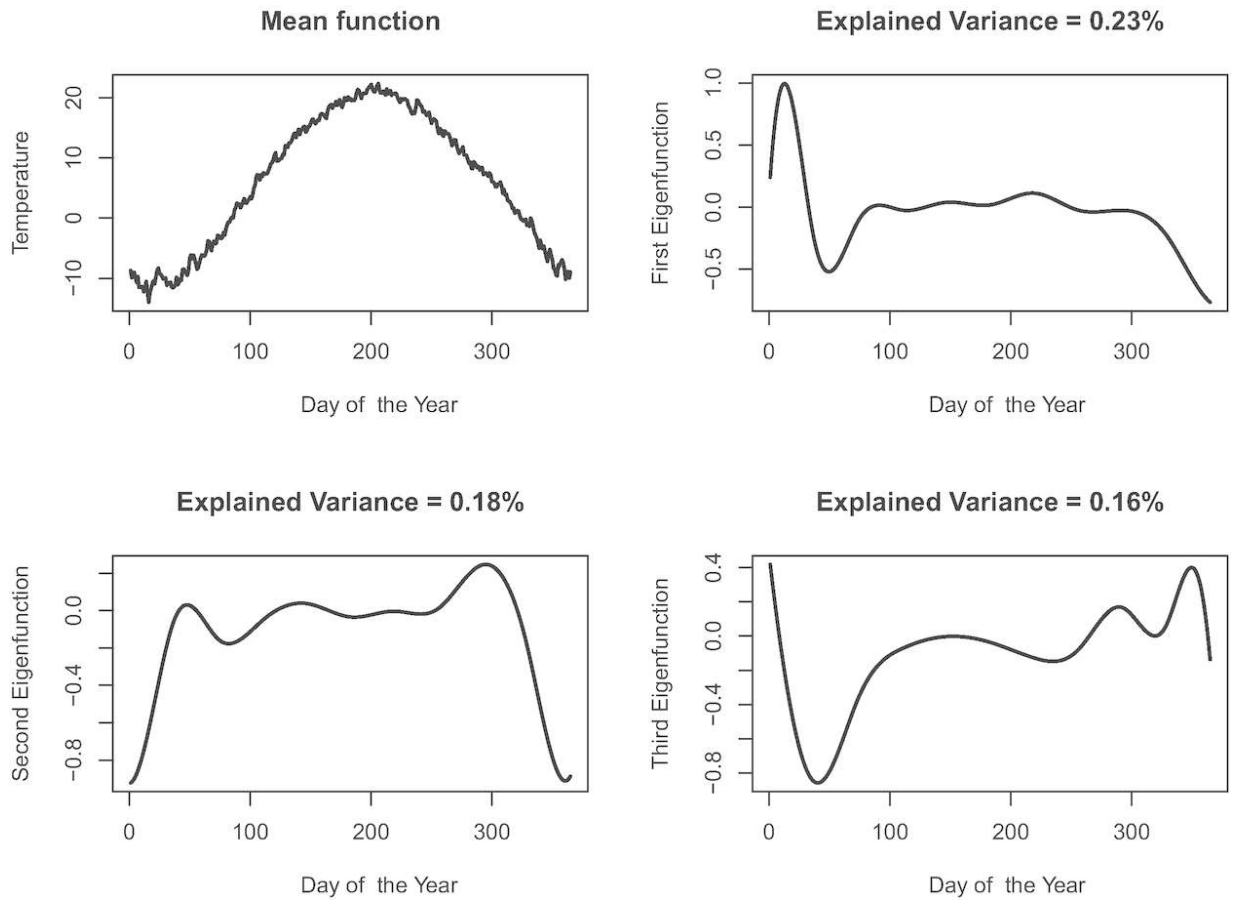


Fig. 12.20 Mean curve and first three functional principal components of the Montreal weather data

12.4.4 Functional Regression

In many settings involving functional data, there is interest in using functions either as covariates, which lead to scalar-on-function regression (SOFR), as responses with scalar covariates, which leads to function-on-scalar regression (FOSR), or in treating both response and covariates as functions, which leads to function-on-function regression (FOFR). We briefly sketch the latter and provide references for further reading.

Assume that both the response $Y(t)$ and the predictor(s) $x(t)$ are functions. To keep notation simple, we postulate that the predictor varies over the same range as the response variable, i.e., $t \in \mathscr{T}$. We assume the regression setup

$$Y_i(t) = \alpha(t) + x_i(t)\beta(t) + \epsilon_i(t), \quad i = 1, \ldots, n, \quad t \in \mathscr{T}. \tag{12.16}$$

The model (12.16) is called the *linear concurrent model* as it provides a concurrent effect for covariates on the response for every point $t \in \mathscr{T}$. Depending on the relationship between covariates and response, various relaxations of Eq. (12.16) exist. For example, one may assume that the regression coefficient is constant, or one may integrate the functional predictors over the entire domain if all points of the covariates are assumed to influence each point of the response.

12.4.5 Summarizing Functional Data

Functional data describe a setting in which the observation themselves are functions. This naturally induces a dependence structure if one assumes continuity of the function. Knowing the value $Y_i(t)$, we obtain that $Y_i(t+h)$ is close to $Y_i(t)$ for h being small. Such a dependence structure needs to be taken into account, which builds the field of functional data analysis.

Based on the short insight above, we see that functional regression models offer a high degree of flexibility and provide a natural way to incorporate non-*i.i.d.* settings. We refer to Greven and Scheipl (2017) for further details on functional regression models and Gertheiss et al. (2024) for a recent tutorial on functional data analysis in general. A general introduction is given in Ramsay and Silverman (2005); see also Kokoszka and Reimherr (2017).

12.5 Exercises

Exercise 1

(a) An analyst is studying three different time series processes. After running a standard diagnostic analysis, they summarize the behavior of the sample autocorrelation function (ACF) and partial autocorrelation function (PACF) for each process.

- **Analysis A:** The ACF has a statistically significant value at lag 1, but all values for lags $h > 1$ are statistically insignificant (i.e., it "cuts off" after lag 1). The PACF shows values that are significant for several lags and appear to decay to zero gradually.

- **Analysis B:** The ACF shows values that are significant for many lags, decaying slowly toward zero (i.e., it "tails off"). The PACF has a significant value at lag 1, but all values for lags $h > 1$ are statistically insignificant.
- **Analysis C:** For both the ACF and PACF, no values are statistically significant for any lag $h > 0$.

Match each analysis description (A, B, C) to the most likely underlying process: **AR(1)**, **MA(1)**, or **White Noise**. Justify your choices by explaining the theoretical properties of the ACF and PACF for each type of process as discussed in the chapter.

(b) Consider the stationary AR(1) process $Y_t = \phi Y_{t-1} + Z_t$, where Z_t is white noise with variance σ^2 and $E[Y_t] = 0$. The chapter states that the autocorrelation function is $\rho(h) = \phi^h$. Show how this result is derived for $h \geq 1$.

(Hint: Start by multiplying the AR(1) equation by Y_{t-h}. Take the expectation and use the fact that $E[Z_t Y_{t-h}] = 0$ for $h > 0$.)

(c) For the MA(q) process, the text states that the autocorrelation function is zero for all lags greater than q ($\rho(h) = 0$ for $h > q$). Using an MA(1) process as an example, explain intuitively why this is true by considering the components of Y_t and Y_{t-h}.

(d) For the AR(p) process, the text states that the partial autocorrelation function (PACF) cuts off after lag p. Explain what this means in terms of conditional correlation and the Markov-like structure of AR processes. Specifically, for an AR(1) process, what is the correlation between Y_t and Y_{t-2} given that you know the value of Y_{t-1}?

Exercise 2

(a) The chapter introduces Tobler's First Law of Geography. In the context of the Munich rental data shown in Fig. 12.11, how would you expect this law to manifest in the rental prices of apartments, even after accounting for other factors like apartment size?

(b) The variogram is defined as $\gamma(d) = \text{Var}(Y_i - Y_j)$ for locations at a distance *d*. If the rental prices Y_i were truly independent and identically distributed (i.e., the i.i.d. assumption held and $\text{Var}(Y_i) = \sigma^2$ for all *i*), what would the theoretical variogram look like as a function of distance *d*? How does this theoretical shape compare to the empirical variogram shown in Fig. 12.12?

(c) Look at the fitted curve in the empirical variogram plot (Fig. 12.12). Identify the approximate **range** and **sill**. What does the range tell you about the distance over which rental prices are spatially correlated? What does the sill represent in terms of the variance of the rental prices?

(d) The chapter discusses two main strategies for handling spatial effects: (i) modeling dependence in the **covariance structure** (e.g., Kriging) or (ii) modeling dependence in the **mean structure** (e.g., spatial regression with a smooth term $h(s)$). If an analyst successfully captures all spatial dependence using approach (ii), what would the variogram of the resulting residuals, $\epsilon_j = Y_j - \hat{h}(s_j)$, look like? Relate your answer to Noel Cressie's quote: "One person's mean is another person's covariance."

(e) Now, imagine the geolocations are aggregated to the administrative districts of Munich, creating lattice data as in Fig. 12.15. This approach models spatial structure on a discrete graph. Describe the conditional independence assumption one makes for the district-level random effects, U_k, in the corresponding Gauss Markov Random Field (GMRF) model.

Exercise 3

(a) The chapter introduces the exponential random graph model (ERGM). In your own words, what is the purpose of the vector of sufficient statistics, $s(y)$? Why is this a more flexible approach for

modeling networks than assuming all potential edges y_{ij} are independent Bernoulli trials?

(b) A researcher studies the collaboration network of computer science faculty. An edge $y_{ij} = 1$ exists if professors *i* and *j* have coauthored a paper. The researcher also has data on each professor's primary research area (e.g., "AI," "Theory," "Systems"). They fit an ERGM and obtain the following results table:

Table

For each of the three statistics, provide a concise interpretation of its coefficient in the context of academic collaboration:

(i) **Edges:** The "Edges" statistic is simply the total number of co-authorships in the network. What does its large negative coefficient imply about the overall density of the collaboration network?

(ii) **Nodematch(Area):** This statistic counts the number of co-authorships between professors in the same research area. What sociological phenomenon (mentioned in the text) does its positive coefficient represent?

(iii) **Transitive Ties:** The "Transitive Ties" statistic counts instances where two professors who share a common collaborator also collaborate with each other. What does the positive coefficient imply about the statement "a collaborator of my collaborator is likely to be my collaborator" in this department?

(c) Suppose the coefficient for "Nodematch(Area)" was instead large and negative. What would this imply about the collaboration patterns in the department? Suggest a reason why such a pattern might emerge in a real-world academic setting.

References

Anselin, L. (1988). *Spatial econometrics: Methods and models* (Vol. 4). Springer.

Baddeley, A., Bárány, I., & Schneider, R. (2007). Spatial point processes and their applications. In *Stochastic geometry: Lectures given at the CIME Summer School Held in Martina Franca, Italy, September 13–18, 2004* (pp. 1–75).

Barabási, A.-L., & Albert, R. (1999). Emergence of scaling in random networks. *Science,286*(5439), 509–512.
[MathSciNet][https://doi.org/10.1126/science.286.5439.509]

Barbour, A. D., & Reinert, G. (2026). *Networks: Probability and statistics* (Cambridge series in statistical and probabilistic mathematics). Cambridge University Press.
[https://doi.org/10.1017/9781009651752]

Bivand, R. S., Pebesma, E. J., & Gómez-Rubio, V. (2008). *Applied spatial data analysis with R*. Springer.

Box, G., & Jenkins, G. (1970). *Time series analysis: Forecasting and control*. Holden-Day.

Box, G. E., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: Forecasting and control*. Wiley.

Brockwell, P. J., & Davis, R. A. (2016). *Introduction to time series and forecasting* (3rd ed.). Springer.
[https://doi.org/10.1007/978-3-319-29854-2]

Cressie, N. (2015). *Statistics for spatial data*. Wiley.

De Gooijer, J. G. (2017). *Elements of nonlinear time series analysis and forecasting* (Vol. 37). Springer.

De Nicola, G., Fritz, C., Mehrl, M., & Kauermann, G. (2023). Dependence matters: Statistical models to identify the drivers of tie formation in economic networks. *Journal of Economic Behavior & Organization,215*, 351–363.
[https://doi.org/10.1016/j.jebo.2023.09.021]

Efron, B. (2022). *Exponential families in theory and practice*. Cambridge University Press.
[https://doi.org/10.1017/9781108773157]

Fienberg, S. E. (2012). A brief history of statistical models for network analysis and open challenges. *Journal of Computational and Graphical Statistics,21*(4), 825–839. Taylor & Francis.

Fisher, R. (1958a). Cigarettes, cancer, and statistics. *The Centennial Review of Arts & Science,2*, 151–166. Michigan State University Press.

Fisher, R. A. (1958b). Lung cancer and cigarettes? *Nature,182*(4628), 108–108.
[https://doi.org/10.1038/182108a0]

Freeman, L. C. (1978). Centrality in social networks conceptual clarification. *Social Networks,1*(3), 215–239.

[https://doi.org/10.1016/0378-8733(78)90021-7]

Gandini, S., Botteri, E., Iodice, S., Boniol, M., Lowenfels, A. B., Maisonneuve, P., & Boyle, P. (2008). Tobacco smoking and cancer: A meta-analysis. *International Journal of Cancer,122*(1), 155–164. [https://doi.org/10.1002/ijc.23033]

Gertheiss, J., Rügamer, D., Liew, B. X., & Greven, S. (2024). Functional data analysis: An introduction and recent developments. *Biometrical Journal,66*(7), e202300363. [MathSciNet][https://doi.org/10.1002/bimj.202300363]

Goldenberg, A., Zheng, A. X., Fienberg, S. E., Airoldi, E. M., & Others (2010). A survey of statistical network models. *Foundations and Trends in Machine Learning,2*(2), 129–233. Now Publishers, Inc.

Goldsmith, J., Scheipl, F. Huang, L., Wrobel, J., Di, C., Gellar, J., Harezlak, J., McLean, M. W., Swihart, B., Xiao, L., Crainiceanu, C., & Reiss, P. T. (2020). *Refund: Regression with functional data*. R package version 0.1-23.

Greven, S., & Scheipl, F. (2017). A general framework for functional regression modelling. *Statistical Modelling,17*(1–2), 1–35. [MathSciNet][https://doi.org/10.1177/1471082X16681317]

Hoff, P. D., Raftery, A. E., & Handcock, M. S. (2002). Latent space approaches to social network analysis. *Journal of the American Statistical Association,97*(460), 1090–1098. [MathSciNet][https://doi.org/10.1198/016214502388618906]

Hunter, D. R., Handcock, M. S., Butts, C. T., Goodreau, S. M., & Morris, M. (2008). Ergm: A package to fit, simulate and diagnose exponential-family models for networks. *Journal of Statistical Software,24*, 1–29. [https://doi.org/10.18637/jss.v024.i03]

Kauermann, G., Haupt, H., & Kaufmann, N. (2012). A hitchhiker's view on spatial statistics and spatial econometrics for lattice data. *Statistical Modelling,12*(5), 419–440. [MathSciNet][https://doi.org/10.1177/1471082X12457493]

Kokoszka, P., & Reimherr, M. (2017). *Introduction to functional data analysis*. Chapman and Hall/CRC. [https://doi.org/10.1201/9781315117416]

Kolaczyk, E. D. (2009). *Statistical analysis of network data*. Springer. [https://doi.org/10.1007/978-0-387-88146-1]

Krige, D. G. (1951). A statistical approach to some basic mine valuation problems on the witwatersrand. *Journal of the Southern African Institute of Mining and Metallurgy,52*(6), 119–139.

Lusher, D., Koskinen, J., & Robins, G. (2013). *Exponential random graph models for social networks*. Cambridge University Press.

McCullagh, P. (2018). *Tensor methods in statistics*. Courier Dover Publications. [https://doi.org/10.1201/9781351077118]

Morris, J. S., & Carroll, R. J. (2006). Wavelet-based functional mixed models. *Journal of the Royal*

Statistical Society: Series B (Statistical Methodology),68(2), 179–199.
[MathSciNet][https://doi.org/10.1111/j.1467-9868.2006.00539.x]

Nepusz, T., Petróczi, A., Négyessy, L., & Bazsó F. (2008). Fuzzy communities and the concept of bridgeness in complex networks. *Physical Review E,77*(1), 1–12.
[MathSciNet][https://doi.org/10.1103/PhysRevE.77.016107]

Ramsay, J. O., & Silverman, B. W. (2005). *Functional data analysis* (Springer series in statistics). Springer.
[https://doi.org/10.1007/b98888]

Ripley, B. D. (2005). *Spatial statistics*. Wiley.

Scheipl, F., Staicu, A.-M., & Greven, S. (2015). Functional additive mixed models. *Journal of Computational and Graphical Statistics,24*(2), 477–501.
[MathSciNet][https://doi.org/10.1080/10618600.2014.901914]

Schweinberger, M. (2011). Instability, sensitivity, and degeneracy of discrete exponential families. *Journal of the American Statistical Association,106*, 1361–1370.
[MathSciNet][https://doi.org/10.1198/jasa.2011.tm10747]

Shalizi, C. R., & Thomas, A. C. (2011). Homophily and contagion are generically confounded in observational social network studies. *Sociological Methods & Research,40*(2), 211–239.
[MathSciNet][https://doi.org/10.1177/0049124111404820]

Shumway, R. H., & Stoffer, D. S. (2017). *Time series analysis and its applications: with R examples* (4th ed.). Springer.
[https://doi.org/10.1007/978-3-319-52452-8]

Sischka, B., & Kauermann, G. (2022). Em-based smooth graphon estimation using mcmc and spline-based approaches. *Social Networks,68*, 279–295.
[https://doi.org/10.1016/j.socnet.2021.08.007]

Snijders, T. A., & Nowicki, K. (1997). Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *Journal of Classification,14*(1), 75–100.
[MathSciNet][https://doi.org/10.1007/s003579900004]

Stein, A., & Wikle, C. K. (2024). Space, uncertainty, and the environment: Honoring the distinguished career of noel cressie. *Spatial Statistics,61*, 100835.
[MathSciNet][https://doi.org/10.1016/j.spasta.2024.100835]

Tsay, R. S. (2000). Time series and forecasting: Brief history and future research. *Journal of the American Statistical Association,95*(450), 638–643.
[https://doi.org/10.1080/01621459.2000.10474241]

Whittle, P. (1951). *Hypothesis testing in time series analysis* (Vol. 4). Almqvist & Wiksells boktr.

Wood, S. N. (2017). *Generalized additive models: An introduction with R*. CRC Press.
[https://doi.org/10.1201/9781315370279]

Yao, F., Müller, H.-G., & Wang, J.-L. (2005). Functional data analysis for sparse longitudinal data.

Journal of the American Statistical Association,100(470), 577–590.
[MathSciNet][https://doi.org/10.1198/016214504000001745]

Footnotes

1 Alternative names in the literature for edges are connections, relations, or ties, and for actors, nodes, agents, or vertices.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
G. Kauermann et al., *Statistics—Coping with Uncertainty*, Textbooks in Statistical Science
https://doi.org/10.1007/978-3-032-33658-3_13

13. Relating Machine Learning and Statistics

Göran Kauermann[1] ✉, Giacomo De Nicola[2], Cornelius Fritz[3] and David Rügamer[4]
(1) Department of Statistics, LMU Munich, Munich, Germany
(2) Harvard T.H. Chan School of Public Health, Harvard University, Boston, MA, USA
(3) School for Computer Science and Statistics, Trinity College Dublin, Dublin, Ireland
(4) Munich Uncertainty Quantification AI Lab, LMU Munich, Munich, Germany

In this chapter, we change our focus from traditional statistics to machine learning, emphasizing strong connections and shared principles between the two fields. The chapter aims to show how statistics and machine learning complement and strengthen each other. We start by outlining essential machine learning concepts and approaches and then highlight and explain the statistical theory underlying machine learning. The emphasis lies on viewing machine learning algorithms as natural extensions of statistical models built to learn from data. This approach connects theoretical concepts with practical use, helping readers of previous chapters to understand machine learning through the lens of statistical principles. We also make use of the jargon used in the field of machine learning and refer, where necessary, to the terminology used synonymously in statistics.

13.1 Primer on Machine Learning and Neural Networks

13.1.1 Machine Learning

Machine learning is, in general, more concerned with making accurate predictions than with uncovering or understanding the true data-generating process $G(\cdot \mid x)$. This focus distinguishes machine learning from much of classical statistics, which often emphasizes interpretability and explicit modeling of $G(\cdot \mid x)$. In many applications, the machine learning paradigm is primarily driven by empirical performance on predictive tasks, where the main objective is to

minimize prediction error on unseen data rather than to recover and/or interpret the exact underlying stochastic mechanism.

Over the past decades, many different machine models have been proposed and have become popular. Among the earlier widely adopted methods are *support vector machines* (SVMs; Cortes & Vapnik, 1995), which achieve classification and regression by maximizing the margin between classes in a high-dimensional feature space. Subsequent years saw the rise of *random forests* (Breiman, 2001), an ensemble of decision trees combined via bagging to improve predictive accuracy and robustness. Another influential family of models are *boosting* methods (Freund & Schapire, 1997), which build strong predictors by iteratively combining so-called weak learners, i.e., simple models that perform slightly better than guessing. Modern incarnations of boosting, such as *gradient boosting machines* (Friedman, 2001) and their highly optimized implementations, including CatBoost (Prokhorenkova et al., 2018), LightGBM (Ke et al., 2017), and XGBoost (Chen & Guestrin, 2016), are among the most widely used ML tools today. More details on classic machine learning approaches can be found, for example, in Bishop and Nasrabadi (2006) and Hastie et al. (2009).

In recent years, much research has shifted toward *neural networks* (NNs) due to their flexibility and scalability, as well as their success as the foundation of powerful generative models for images, text, and other modalities. Recent advancements in tabular foundation models, e.g., prior-fitted networks (Müller et al., 2025), have made neural networks also a competitive choice for classic machine learning tasks, achieving state-of-the-art predictive performance without the need for tuning, in contrast to previously used methods such as boosting. While an introduction to modern (generative) architectures such as prior-fitted networks, diffusion models, or large language models is beyond the scope of this book, we will briefly introduce the basic neural network architecture, the *multilayer perceptron* (MLP), in the following.

13.1.2 Multilayer Perceptrons

A multilayer perceptron (MLP) is a function $h : \mathbb{R}^p \to \mathbb{R}^v$ composed of layers of interconnected *neurons*. Each neuron takes an input vector $z \in \mathbb{R}^m$ and computes an *affine transformation*, which means a linear calculation $w^\top z + b$, where $w \in \mathbb{R}^m$ are the neuron's weights and $b \in \mathbb{R}$ is called *bias*, which is simply a different name for "intercept." This is used as input for nonlinear *activation function*τ yielding the neuron output

$$u = \tau\big(w^\top z + b\big).$$

The activation function τ can, for example, be a rectified linear unit (ReLU) $\tau(\cdot) = \max(0, \cdot)$ or a logistic sigmoid function; see Eq. (6.14). A neural network layer consists of several such neurons operating in parallel, taking the same input z but with different weights and biases, producing a vector of outputs. Mathematically, this is defined by a matrix W of combined weights for each neuron and a vector of stacked biases b. These transformations can be applied in a sequence of transformations:

$$z^{(1)} = \tau\big(W^{(1)}x + b^{(1)}\big), \quad z^{(2)} = \tau\big(W^{(2)}z^{(1)} + b^{(2)}\big), \quad \dots$$

The final layer (often called the *output layer* or *last layer*) computes a linear combination of the learned features from the last hidden layer:

$$\hat{y} = W^{(L)}z^{(L-1)} + b^{(L)},$$

which is equivalent to a linear regression on these learned features $z^{(L-1)}$. Depending on the domain of the outcome values, the last layer might also include an additional nonlinear activation, such as a sigmoid function for mapping the predictions into the space of probabilities. Formally, an L-layer feedforward neural network with input $x \in \mathbb{R}^p$ can be written as

$$h(x;\theta) = W^{(L)}\,\tau\left(W^{(L-1)}\,\tau\left(\dots\tau\left(W^{(1)}x + b^{(1)}\right)\dots\right) + b^{(L-1)}\right) + b^{(L)}, \tag{13.1}$$

where $W^{(L)} \in \mathbb{R}^{v \times m_{L-1}}$ and $b^{(L)} \in \mathbb{R}^v$ are the weights and biases of the Lth layer and $z^{(L-1)} \in \mathbb{R}^{m_{L-1}}$ are the outputs of the $(L-1)$-th hidden layer with $z^{(0)} := x$, that is, the first input is the original input variables x. The parameters of the network are collectively denoted by $\theta = \{W^{(\ell)}, b^{(\ell)}\}_{\ell=1}^{L}$.

Figure 13.1 illustrates the structure of an MLP with an input layer, two hidden layers, and an output layer. The hidden layers learn nonlinear feature transformations of the input, while the last layer forms the final prediction by linearly combining these features. For a more detailed exposition of neural networks and related concepts, see Goodfellow et al. (2016).

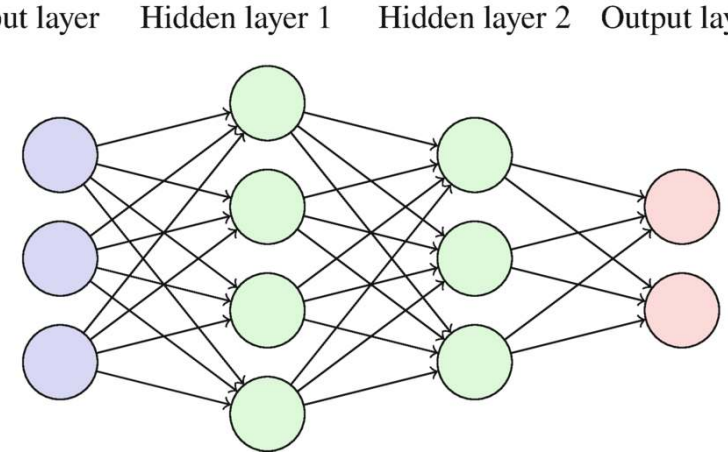


Fig. 13.1 A multilayer perceptron (MLP) with three input features, two hidden layers with four and three neurons, and two outputs. Each hidden layer neuron applies a nonlinear transformation to a weighted sum of its inputs. The biases are omitted for easier visualization. The last layer forms a linear regression on the learned features from the last hidden layer

13.2 Machine Learning from a Statistics Perspective

13.2.1 Hypothesis Space

From a statistical perspective, the goal in supervised machine learning is to approximate the data-generating process $G(\cdot \mid x)$ in (2.1) using a machine learning model from the so-called **hypothesis space** (or model space) $\mathscr{H}$, introduced earlier in Sect. 7.1 as *structural model*; see also Eq. (7.3). Instead of approximating or modeling $G(\cdot \mid x)$ with a set of distribution functions, the hypothesis space is usually defined by a set of functions

$$\mathscr{H} = \{h(\cdot;\theta), \theta \in \Theta\}. \tag{13.2}$$

The machine learning *model* $h(\cdot;\theta) : \mathscr{X} \to \Gamma$ approximates the unknown mapping from the covariate or feature space $\mathscr{X}$ to a space of predictions Γ, where $\Gamma \subseteq \mathbb{R}^v$ with $v \in \mathbb{N}$. The parameter (vector) θ defines a single instance (a machine learning model) of $\mathscr{H}$. For example, if $h(x;\theta) = x^\top\theta$, $\mathscr{H}$ is the space of linear regression models with covariates (features) x, coefficients $\theta \in \Theta = \mathbb{R}^p$, and $\Gamma = \mathbb{R}$.

In many of the classic machine learning applications, this hypothesis space is usually not paired with a distribution assumption or model $\mathscr{F}$. Instead, in classic machine learning, we only distinguish between a *regression* and a *classification task*. More contemporary machine learning models, in contrast, build on the same principle as statistical models and assume an underlying probability model $\mathscr{F}$ (see, e.g., Murphy, 2022).

13.2.2 Empirical Risk Minimization

Learning in machine learning is the combination of a hypothesis space $\mathscr{H}$, a *risk* $\mathscr{R}$ and an optimization routine. To define the risk, we first need to define a loss function

$$\ell : \mathscr{Y} \times \mathscr{H} \to \mathbb{R} : \ (y, h(x;\theta)) \to \ell(y, h(x;\theta)), \tag{13.3}$$

where $\mathscr{Y}$ is the set of possible outcomes. We want to minimize the loss with respect to $h \in \mathscr{H}$. An appropriate loss is usually chosen for each particular application, for instance, to find a model h such that the squared deviations of $h(x;\theta)$ and y are minimized. Acknowledging that the true data-generating process $G(\cdot \mid x)$ is (in most cases) not deterministic, the overall objective of machine learning is to find the model h which minimizes the expected loss, the so-called **risk**

$$\mathcal{R}(h) := \int \left(\int \ell(y, h(x; \theta)) dG(y \mid x) \right) dG(x) \tag{13.4}$$

with $G(x)$ as a known or unknown distribution over the feature set $\mathcal{X}$. As (13.4) involves the unknown true data-generating process, we approximate the theoretical risk by its empirical analog, the **empirical risk**, which for a given data set of $i.i.d.$ observations $(x_i, y_i)_{i=1,\ldots,n}$ is

$$\mathcal{R}_{\text{emp}}(h) := \sum_{i=1}^{n} \ell(y_i, h(x_i; \theta)). \tag{13.5}$$

Learning is thus the search for the best model that minimizes the empirical risk using a suitable optimization technique (often a gradient-based method). For a parametrized family of models $\mathcal{H}$, this implies the Empirical Risk Minimizer

$$\hat{\theta}_{\text{ERM}} = \underset{\theta \in \Theta}{\text{argmin}} \; \mathcal{R}_{\text{emp}}(h(\cdot; \theta)). \tag{13.6}$$

Putting (13.5) and (13.6) together, we can directly see the correspondence between *estimating* the parameter θ using maximum likelihood estimation and *learning* the parameter θ using empirical risk minimization: Choosing ℓ to be the negative log-likelihood of a parametric model

$$\mathcal{F} = \{f(y; \gamma); \gamma \in \Gamma\} \tag{13.7}$$

yields $\hat{\theta}_{\text{ERM}} \equiv \hat{\theta}_{\text{ML}}$. In this case, the model h (implicitly) describes the parametrized transformation from x to the distribution parameters γ of some distribution $F(\cdot \mid x; \gamma)$. For example, in a regression setting with $\mathcal{H}$ being the class of linear models, one typically defines the loss to be the mean squared error, and thus $h(x; \theta) = \mu(x; \theta) = x^\top \theta$ (implicitly) defines the mean μ of a normal distribution F.

In fact, many loss functions in machine learning have a one-to-one correspondence with a (negative) log-likelihood function in statistics. For example, optimizing the log-loss in machine learning is equivalent to optimizing the negative log-likelihood based on a Bernoulli distribution assumption. Moreover, minimizing the mean squared error is equivalent to maximum likelihood estimation of the mean of a Gaussian distribution. And, a multinomial distribution assumption for likelihood-based estimation results in the same estimates as a corresponding model in machine learning, minimizing the so-called cross-entropy. On the other hand, there are examples exist where there is no direct correspondence to a distribution assumption, e.g., the Hinge loss (Cortes & Vapnik, 1995). The rationale, in this case, is not to directly approximate the data-generating process $G(\cdot \mid x)$ but to optimize a predefined metric that is of particular interest and which is not implicitly optimized by finding the best data-generating process approximation. Yet even these metrics often have a statistical

origin. The Huber loss, for example, is used in machine learning as a loss function, but its motivation goes back to ideas in robust statistics (Huber, 1964).

In machine learning, a large portion of models deal with the classification of observations into two or more classes. If Γ is of discrete nature, i.e., has a finite number of elements, *h* is usually called a *classifier*. A further distinction can be made regarding whether *h* returns a probability for every possible value in $\mathscr{Y}$. If it does so, we call *h* a probabilistic classifier; otherwise, if a score for every value in $\mathscr{Y}$ is returned, *h* is called a scoring function or scoring classifier. A logistic regression, for example, is a probabilistic classifier, returning probabilities $\pi = P(Y = 1 \mid x)$ and $1 - \pi = P(Y = 0 \mid x)$ for $\mathscr{Y} = \{0, 1\}$. The additive predictor in logistic regression, in turn, is a scoring function that produces a score (the logits) that can be normalized to probabilities.

We can summarize that *learning* in machine learning follows conceptually statistical ideas as derived in the book. Depending on the chosen loss function, it can, however, go beyond some classical statistical concepts.

13.3 Regularization, Penalization, and Bayes

13.3.1 Regularized Risk Minimization

One of the most crucial parts of machine learning is regularization. Being often highly parameterized and expressive, there is a great risk of overfitting the given dataset with a machine learning model, as described in more detail in the following subsections.

The most common way to avoid overfitting in machine learning is to regularize or penalize models depending on their flexibility. More specifically, we can extend (13.5) to the so-called regularized risk

$$\mathscr{R}_{\text{reg}}(h) := \sum_{i=1}^{n} \ell(y_i, h(x_i; \theta)) + \lambda \cdot \text{pen}(\theta), \tag{13.8}$$

where $\text{pen} : \Theta \to \mathbb{R}^+$ is a function assigning larger values (penalties) to model parameters θ that induce very or too flexible hypotheses and $\lambda > 0$ is a tuning parameter controlling the trade-off between model fit and flexibility. A common way of searching for the optimal λ value is by using cross-validation. The explicit form of the penalty depends on the model class. Information criteria introduced in Sect. 7 are some form of regularization, for example, with the penalty function of the AIC defined as $\text{pen}(\theta) = \|\theta\|_0$, where $\|\theta\|_0$ is the L_0-norm[1] which counts the number of nonzero elements in θ. If we choose the empirical risk to be twice the negative log-likelihood and set $\lambda = 2$, we can then consider the optimization of the AIC as a way of performing regularized risk minimization.

13.3.2 Regularization from a Statistical Perspective

Many regularization approaches in machine learning can be related to concepts that are also prominent in the (high-dimensional) statistical literature. Starting from the previous analogy and model selection criteria in Sect. 7, we could try to find the AIC-optimal model directly:

$$\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmin}} \; \mathcal{R}_{\text{emp}}(h(\cdot; \theta)) + \|\theta\|_0. \tag{13.9}$$

This problem, optimizing a model with the so-called L_0 penalty, is, however, known to be NP-hard from optimization theory. One of the key problems with the L_0-norm is its non-differentiability at 0. The "next best" convex relaxation of the L_0-norm is the L_1-norm $\|\theta\|_1$. Instead of counting the nonzero entries, the l_1-norm is defined as

$$\text{pen}(\theta) = \|\theta\|_1 = \sum_{j=1}^{p} |\theta_j|, \tag{13.10}$$

i.e., sums up the absolute values of all entries in θ. This penalty, also known as the lasso penalty (Tibshirani, 1996), can be solved using specialized routines (depending on the model class). Similar to information criteria, the regularizing effect of the lasso penalty induces sparser models, i.e., models with fewer parameters. As the L_1-norm is not differentiable at 0, the penalty is less common in machine learning than in statistics and optimization, where specialized optimization routines for it are available. Recent findings, however, show how to optimize these non-differentiable penalties also in deep neural networks (Kolb et al., 2023), thereby making it available to a much larger machine learning community.

Another common penalty in both machine learning and statistics is the L_2- or ridge penalty:

$$\text{pen}(\theta) = \|\theta\|_2^2 = \sum_{j=1}^{p} \theta_j^2, \tag{13.11}$$

which, contrary to the name, is a squared L_2-norm penalty.

We visualize the effects of both the L_1 and the L_2 penalties in Fig. 13.2. The two plots show the typical performance of penalized parameter estimates, with the amount of penalization on the horizontal axis. While the L_1 penalty, shown in the left plot, can shrink parameter estimates to zero, the ridge penalty, right-hand side plot, cannot enforce sparsity in that parameters are forced to zero but instead regularizes the model by "only" shrinking the coefficients toward zero.

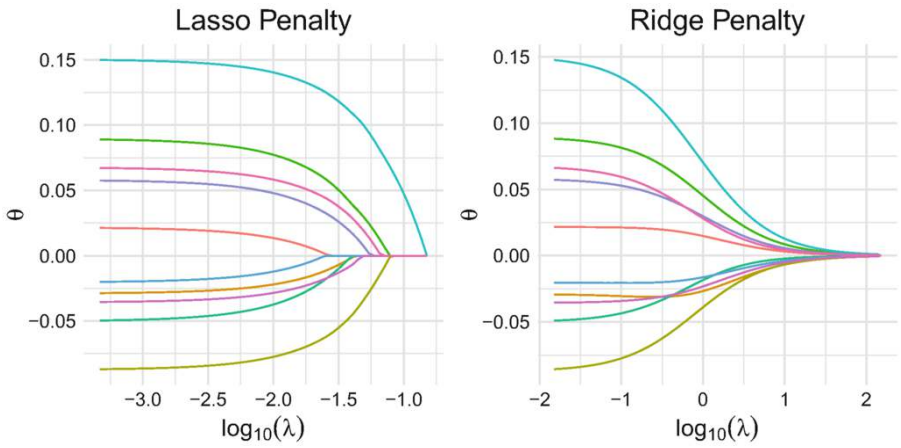


Fig. 13.2 Paths of coefficients of a linear model (in different colors) with lasso (left) and ridge regularization (right) depending on the strength of the penalty λ

For linear models and the L_2-penalty, for example, the regularized risk minimizer $\hat{\theta}_{\text{RRM}}$ has a closed form solution

$$\hat{\theta}_{\text{RRM}} = \left(X^\top X + \lambda I_p\right)^{-1} X^\top y. \tag{13.12}$$

For models without an explicit analytical solution, a common strategy in machine learning is to use gradient descent. The gradient of a regularized risk objective with a ridge penalty (divided by 2 for convenience) is

$$\begin{aligned} \frac{\partial}{\partial \theta}\{\mathscr{R}_{\text{reg}}(h(\cdot;\theta))\} &= \frac{\partial}{\partial \theta}\left\{\sum_{i=1}^{n} \ell(y_i, h(x_i;\theta)) + \tfrac{\lambda}{2}\|\theta\|_2^2\right\} \\ &= \frac{\partial}{\partial \theta}\{\mathscr{R}_{\text{emp}}(h(\cdot;\theta))\} + \lambda\theta. \end{aligned} \tag{13.13}$$

The weights of the model are then updated via gradient descent as follows:

$$\theta^{[t+1]} = \theta^{[t]} - \nu \frac{\partial}{\partial \theta}\{\mathscr{R}_{\text{reg}}(h(\cdot;\theta^{[t]}))\} \tag{13.14}$$

with learning rate $\nu \in (0, 1]$ and iteration $t + 1$. Expanding (13.14) using (13.13), we see that the ridge penalty induces a *decay* of the current parameter value defined by λ:

$$\begin{aligned} \theta^{[t+1]} &= \theta^{[t]} - \nu \frac{\partial}{\partial \theta}\{\mathscr{R}_{\text{reg}}(h(\cdot;\theta^{[t]}))\} \\ &= \theta^{[t]} - \nu\left(\frac{\partial}{\partial \theta}\{\mathscr{R}_{\text{emp}}(h(\cdot;\theta^{[t]}))\} + \lambda\theta^{[t]}\right) \\ &= \theta^{[t]}(1 - \nu\lambda) - \nu \frac{\partial}{\partial \theta}\{\mathscr{R}_{\text{emp}}(h(\cdot;\theta^{[t]}))\}. \end{aligned} \tag{13.15}$$

In machine learning and particularly deep learning, it is common to apply *weight decay* during optimization. Therefore, one would impose a ridge penalty not explicitly through a regularized objective as in (13.11) but implicitly by altering the optimization routine as in (13.15).

Similarly, the so-called *early stopping* of iterative optimization routines is common practice and can be shown to implicitly regularize the model. For the simple case of a linear model with L_2-loss optimized by gradient descent as

defined above, there exists a direct relationship between the number of iterations T the algorithm is run (i.e., how often the old parameter is updated) and the induced ridge penalty (see, e.g., Goodfellow et al., 2016):

$$\lambda \approx \frac{1}{\nu T}. \tag{13.16}$$

Put differently, early stopping the routine imposes a ridge regularization, controlled by the number of iterations T and the learning rate ν. In particular, the earlier we stop the routine, the larger the induced ridge penalization and vice versa.

13.3.3 Bayesian Approaches

Machine learning also incorporates many different Bayesian approaches. Apart from offering a flexible framework to incorporate prior knowledge into the learning process, priors in Bayesian approaches act as a form of regularization. The connection between Bayesian methods and regularized risk minimization can be understood by examining how priors influence the optimization of model parameters, particularly through the lens of normal priors and their relationship to ridge regularization. The posterior distribution in a Bayesian model is obtained via Bayes' theorem (see Sect. 9):

$$p(\theta \mid x, y) \propto p(y \mid x, \theta) p(\theta), \tag{13.17}$$

where $p(\theta \mid x, y)$ is the posterior distribution, $p(y \mid x, \theta)$ is the likelihood, and $p(\theta)$ is the prior distribution. Let's consider a linear regression model where $h(x; \theta) = x^\top \theta$. In a Bayesian framework, we might impose a normal prior on the parameter vector θ:

$$\theta \sim \mathrm{N}(0, \tau^2 I), \tag{13.18}$$

to express the belief that the parameters θ are normally distributed with mean zero and variance τ^2. The likelihood of the data given the parameters, assuming Gaussian noise with variance σ^2, is

$$p(y \mid x, \theta) \propto \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y_i - x_i^\top \theta\right)^2\right). \tag{13.19}$$

The posterior distribution is then given by

$$p(\theta \mid x, y) \propto \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y_i - x_i^\top \theta\right)^2\right) \exp\left(-\frac{1}{2\tau^2} \parallel \theta \parallel_2^2\right). \tag{13.20}$$

Taking the negative log of the posterior results in the following objective function:

$$-\log p(\theta \mid x, y) \propto \sum_{i=1}^{n} \left(y_i - x_i^\top \theta\right)^2 + \frac{\sigma^2}{\tau^2} \parallel \theta \parallel_2^2 . \tag{13.21}$$

We can see that minimizing this negative log-posterior is equivalent to minimizing the regularized risk function with a ridge penalty:

(13.22)

$$\mathscr{R}_{\text{reg}}(h) = \sum_{i=1}^{n} \left(y_i - x_i^\top \theta\right)^2 + \lambda \parallel \theta \parallel_2^2,$$

where $\lambda = \frac{\sigma^2}{\tau^2}$.

The equivalence between maximizing the posterior distribution under a normal prior and performing ridge regression illustrates a direct link between Bayesian methods and regularized risk minimization. A similar link can be shown for various other prior distribution assumptions.

The following example highlights the connection of the Bayesian formulation to a recently discovered phenomenon called *double descent*, which occurs in overparametrized models where $p > n$.

Example 13.1 Consider again the Bayesian linear regression model with Gaussian prior $\theta \sim \mathrm{N}(0, \tau^2 I)$ and Gaussian likelihood. We investigate the behavior of the model in an overparameterized regime, where the number of parameters p exceeds the sample size n.

To illustrate this, we simulate $\theta_j \sim \mathrm{N}(0,1) i.i.d.$ for $j = 1, \ldots, p$ and sort the coefficients such that $|\theta_1| \geq |\theta_2| \geq \cdots \geq |\theta_p|$. We fix $n = 100$ and consider p up to 200, hence covering both the classical setting ($p < n$) and the overparameterized case ($p > n$). Data are generated from

$$y_i = x_i^\top \theta + \varepsilon_i, \quad \varepsilon_i \sim \mathrm{N}(0, \rho^2),$$

with $\rho^2 = 0.5$ and covariates x_i drawn i.i.d. from a standard normal distribution.

For each p, we fit the model using ridge regularization (equivalently, the Bayesian posterior mean) and evaluate predictive performance via the Kullback–Leibler divergence

$$\mathrm{KL}\big(G(\cdot), F(\cdot\,; \hat{\theta})\big),$$

where G denotes the true data-generating distribution and $F(\cdot\,; \hat{\theta})$ the fitted model. The results, shown in Fig. 13.3, reveal two minima of the divergence: one in the underparameterized regime ($p < n$) and a second in the overparameterized regime ($p > n$). The latter occurs because the prior (or equivalently, the ridge penalty) effectively regularizes the model even when the parameter dimension exceeds the sample size. This highlights that with suitable priors, Bayesian approaches can yield competitive performance well beyond the classical $p < n$ paradigm.▷

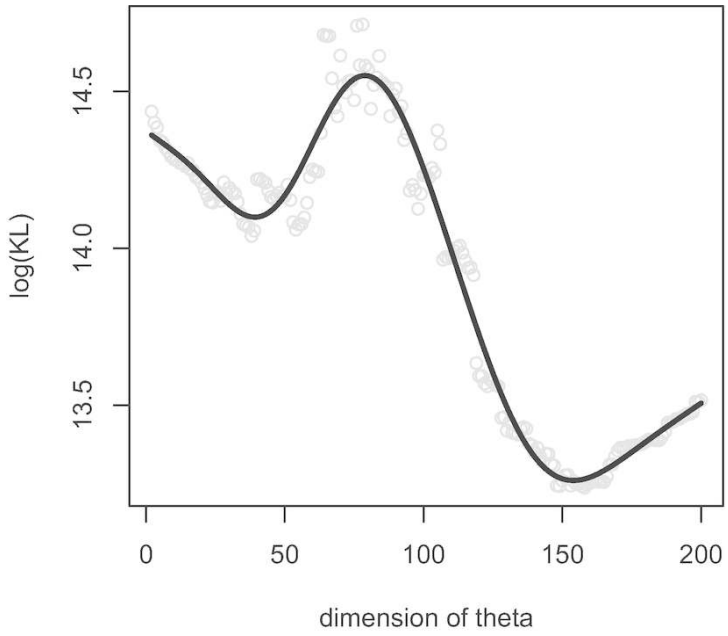


Fig. 13.3 Kullback–Leibler divergence of the fitted model for varying parameter dimension p in a ridge-regularized linear model with $n = 100$ and $\rho^2 = 0.5$

13.4 Uncertainty in Machine Learning

13.4.1 Prediction and Parameter Uncertainty

Although there is ongoing debate about whether a universally valid categorization of uncertainties in machine learning exists, three areas of uncertainty quantification (UQ) are of particular practical interest:

1. uncertainty in model predictions,
2. uncertainty in model parameters,
3. and uncertainty in the data-generating process.

In the following, we will discuss these three areas, outlining their goals and practical considerations. For simplicity, we only model the mean μ of outcome Y given features x and use $\widehat{\mu}(\widetilde{x}) := \hat{E}(\widetilde{Y} \mid \widetilde{x}) = h(\widetilde{x}; \hat{\theta})$ to make predictions about a new, unobserved observation $\widetilde{Y}$ with covariates $\widetilde{x}$. We will first consider approaches that focus on the uncertainty $\widehat{\mu}$ and, related to this, the uncertainty of $\hat{\theta}$ in h.

13.4.2 Predictive Uncertainty

For the uncertainty of $\widehat{\mu}$, one can use either model-specific or model-agnostic approaches. Model-agnostic methods include bootstrap (Sect. 8.7.2) and cross-validation (Sect. 7.2), where repeated training on resampled data yields intervals for $\widehat{\mu}$. Conformal prediction, which is formally introduced in Sect. 13.4.6, offers another way to construct such intervals. Model-specific methods often rely on the model's parametric distribution assumption. An example is a Gaussian process, which quantifies predictive uncertainty using a Bayesian framework.

13.4.3 Parameter Uncertainty

In contrast to the previously discussed prediction uncertainty, quantifying the uncertainty of parameters in machine learning models is typically more complicated. As machine learning models are often "black boxes" where single parameter states do not provide meaningful insights, quantifying the parameter uncertainty is often merely a means to an end to quantify the uncertainty of the prediction function.

Consider, for example, a neural network $h = h_{\text{lin}} \circ h_{\text{nlin}}$, meaning that we decompose the network in a nonlinear (hidden layer) part and the last layer, which is linear, with parameters $\theta = (\beta, \xi)$ being a decomposition of a linear h_{lin} part, parameterized by β, and a nonlinear function h_{nlin}, parameterized by ξ. Such a decomposition is the basis to quantify uncertainty using a *last-layer approach*, providing the uncertainty of the last layer's weights β in the function $h_{\text{lin}}(z, \beta) = z^\top \beta$, where $z = h_{\text{nlin}}(x; \xi)$ are learned representation of the original data x. Ignoring the uncertainty in ξ underestimates the model's overall uncertainty. However, focusing solely on β is relevant, e.g., in *transfer learning*, where ξ is learned on a separate dataset and then fixed when applying the model to new data. In this setting, only β remains uncertain, allowing statistical inference to be applied under suitable conditions.

13.4.4 Data Uncertainty

While classic machine learning approaches have focused on risk minimization, the links to statistics and awareness of the need for uncertainty quantification have sparked much interest in probabilistic methods recently. Many probabilistic models allow to directly model the *aleatoric uncertainty*, i.e., learn the distribution F with all its parameters.

Distributional Regression Models In statistics, *distributional regression* refers to modeling more than just the mean of a distribution using an additive predictor. Jointly estimating multiple parameters, such as μ and σ^2 of a normal distribution,

is substantially more challenging, since even this simple case leads to a non-convex optimization problem.

In machine learning and especially in deep learning, convexity is usually not assumed, and global optima are rarely feasible nor the goal, as this often refers to overfitting. It is therefore natural to define a separate model h_j for each distribution parameter γ_j and optimize them jointly (see, e.g., Rügamer et al., 2024). To avoid restrictive parametric assumptions, more flexible approaches employ monotonic transformations to represent complex distributions. A prominent example is *normalizing flows* (see, e.g., Kook et al., 2024).

13.4.5 Calibration

The simplest case of a distributional regression model is a parametric distribution with a single parameter, e.g., $Y \sim \text{Binomial}(1, \pi)$. While this example is not a very complex data-generating process, it is instructive to understand a common problem in machine learning when modeling distributions related to the model's calibration.

Example 13.2 Consider a binary classification task where $y_1, \ldots, y_{n-1} = 0$ and $y_n = 1$. If our data is almost linearly separability given some feature *x*, e.g., for $y_i = 1$, we have $x_i = c$, and for almost all $y_i = 0$, we have $x_i = -c$ for some $c \in \mathbb{R}$, the coefficient of the simple logistic regression model $\pi_i = P(Y_i = 1 \mid x_i) = 1 \, / \, (1 + \exp\,(x_i \beta))$ will become larger the bigger the constant *c*.[2] However, if the model is also subject to regularization such that large values for the coefficients β are penalized, the model is incentivized to return a small β value. In the case of exact linear separability, the coefficient is

$$\widehat{\beta} \approx -c^{-1} W(c^2 \lambda^{-1}),$$

where *W* is the Lambert function. For example, for $c = 1$ and $\lambda = 1$, we obtain $\widehat{\beta} = -0.5671433$. Due to the relatively small size of $\widehat{\beta}$, the predicted probabilities will be close to 0.5 (see also Firth (1993) for an earlier reference for this matter). More precisely, we get $P(Y_i = 1 \mid x_i = c) = 0.6381037$ and $P(Y_i = 1 \mid x_i = -c) = 0.3618963$. The model is thus clearly *miscalibrated* for large *n* as the above-computed probabilities do not represent the observed frequencies of 0s and 1s in any way. Put differently, for a calibrated model, we expect probabilities that are (closer to) $(n-1) \, / \, n$ and $1 \, / \, n$. ▷

The previous example shows that a perfect model does not necessarily have a correct representation of the underlying aleatoric uncertainty, i.e., the data-generating distribution. It is therefore common practice to calibrate predictions

after optimizing the model. If the distribution of the model is known, a second model or procedure can be applied on top of the first model's prediction to ensure proper calibration. One method with increased popularity is *conformal prediction* or *conformal inference*.

13.4.6 Conformal Prediction

Predictive uncertainty (Sect. 13.4.1) quantifies how uncertain a model's output is. Conformal prediction instead calibrates these uncertainties to produce prediction sets or intervals that contain the true outcome with a specified probability, under minimal assumptions (often exchangeability). Its main goal is guaranteed coverage, not just uncertainty quantification.

We introduce conformal prediction using the earlier miscalibration example, where we assumed binary labels[3] $Y \in \{0, 1\}$ modeled via a logistic regression with linear predictor η:

$$\pi(x) := P(Y = 1 \mid x) = h(x; \theta) = \tfrac{1}{1+\exp(-\eta(x))}.$$

We define a *nonconformity score* using the predicted probability

$$s(x, y) = \begin{cases} \pi(x), & \text{if } y = 0, \\ 1 - \pi(x), & \text{if } y = 1. \end{cases} \tag{13.23}$$

When $y = 1$, a high π indicates high confidence and leads to a small score. For $y = 0$, the reverse holds. We can now generate a conformal prediction set with miscoverage level α in two steps: the calibration and the prediction step.

Calibration For each data point $i = 1, \ldots, n$ used for estimation, we compute $s_i = s(x_i, y_i)$, i.e., the score quantifying to what extent our model would predict "something wrong." Using all these scores for $i = 1, \ldots, n$, we compute the empirical $1 - \alpha$ quantile $\hat{q}$ of $\{s_1, \ldots, s_n\}$.

Prediction For a new input $\tilde{x}$, compute $\hat{\pi} = h(\tilde{x}, \hat{\theta})$ and the score

$$s(\tilde{x}, \tilde{y}) = \begin{cases} \hat{\pi}, & \tilde{y} = 0, \\ 1 - \hat{\pi}, & \tilde{y} = 1. \end{cases}$$

Our conformal prediction set C can now be constructed as

$$C(\tilde{x}) = \{\tilde{y} \in \{0, 1\} : s(\tilde{x}, \tilde{y}) \leq \hat{q}\}.$$

In other words, our prediction set includes the label 0 if $\hat{\pi} \leq \hat{q}$ and the label 1 if $1 - \hat{\pi} \leq \hat{q}$. High confidence usually yields a single label, but if there is large nonconformity (large values s_i) in our calibration data, $C = \{0, 1\}$ if the empirical quantile $\hat{q} \geq \max(\hat{\pi}, 1 - \hat{\pi})$.

Example 13.3 Assume $n = 10$ calibration pairs (x_i, y_i) with fitted probabilities $\widehat{\pi}_i$:

i	1	2	3	4	5	6	7	8	9	10
y_i	0	1	0	1	0	1	0	1	0	1
$\hat{\pi}_i$	0.12	0.84	0.20	0.55	0.33	0.40	0.18	0.77	0.25	0.32

The nonconformity scores from (13.23)

$$s_i = \begin{cases} \widehat{\pi}_i & \text{if } y_i = 0, \\ 1 - \widehat{\pi}_i & \text{if } y_i = 1, \end{cases}$$

are

$$(0.12,\ 0.16,\ 0.20,\ 0.45,\ 0.33,\ 0.60,\ 0.18,\ 0.23,\ 0.25,\ 0.68).$$

If we choose $1 - \alpha = 0.9$, $\hat{q}$ is the tenth order statistic, i.e., the largest nonconformity score, which is $\hat{q} = 0.68$.

For a new input $\widetilde{x}$ with predicted probability $\widehat{\pi} = h(\widetilde{x}, \hat{\theta})$, different sets C can manifest, for example:

- Let $\widehat{\pi} = 0.20$. Then, $s(0) = 0.20 \leq 0.68$, $s(1) = 0.80 > 0.68$, and therefore $C = \{0\}$.
- Let $\widehat{\pi} = 0.90$. Then, $s(0) = 0.90 > 0.68$, $s(1) = 0.10 \leq 0.68$, and therefore $C = \{1\}$.
- Let $\widehat{\pi} = 0.50$. Then, $s(0) = 0.50 \leq 0.68$, $s(1) = 0.50 \leq 0.68$, and therefore $C = \{0, 1\}$.

The conformal guarantee (Vovk et al., 1999) ensures

$$P(\tilde{y} \in C(\widetilde{x})) \geq 1 - \alpha$$

and produces well-calibrated prediction sets even when model probabilities are miscalibrated. The nonconformity score can, e.g., be based on logits, ensemble uncertainty, or other measures. However, the score should reflect true model uncertainty to avoid unnecessarily wide prediction sets. Further note that the method guarantees *marginal* coverage: The coverage probability holds over the data distribution not for each individual prediction. For some inputs—particularly in imbalanced or adversarial settings—the true label may fall outside the set despite the overall guarantee. ▷

13.5 Exercises

Exercise 1

(a) Consider a standard linear regression model where the goal is to predict a

continuous outcome y_i from features x_i. The most common approach in machine learning is to find the parameters θ that minimize the **mean squared error (MSE)** loss

$$\mathscr{R}_{\text{emp}}(\theta) = \sum_{i=1}^{n} \left(y_i - h(x_i; \theta)\right)^2 \quad \text{where} \quad h(x_i; \theta) = x_i^\top \theta.$$

Show that minimizing this loss function is equivalent to finding the maximum likelihood estimate (MLE) for θ under the assumption that the data follows a specific probability distribution. What is this distribution for the conditional outcome, $Y_i \mid x_i$?

(b) Now, consider a binary classification problem where $y_i \in \{0, 1\}$. A common loss function is the **binary cross-entropy** or **log-loss**:

$$\mathscr{R}_{\text{emp}}(\theta) = -\sum_{i=1}^{n} [y_i \log\left(\pi_i\right) + (1 - y_i) \log\left(1 - \pi_i\right)],$$

where $\pi_i = h(x_i; \theta)$ is the model's predicted probability that $y_i = 1$. What is the statistical distribution (i.e., the likelihood) that corresponds to this loss function? What common statistical model, discussed in a previous chapter, uses this likelihood?

(c) A multilayer perceptron (MLP) can be used for classification by adding a sigmoid activation function to its final layer, making its output a probability. Explain how such a complex neural network can be viewed from a statistical perspective as a powerful generalization of the model you identified in part (b). What part of the MLP architecture allows it to learn more complex relationships than the simpler statistical model?

Exercise 2

(a) In machine learning, **Ridge regression** is used to fit a linear model $h(x; \theta) = x^\top \theta$. The objective is to minimize the sum of squared errors plus an L_2 penalty on the parameters

$$\mathscr{R}_{\text{reg}}(\theta) = \sum_{i=1}^{n} \left(y_i - x_i^\top \theta\right)^2 + \lambda \sum_{j=1}^{p} \theta_j^2.$$

What is the purpose of the penalty term $\lambda \sum \theta_j^2$? What problem, common in both statistics and machine learning, is it intended to solve, especially when the number of features p is large?

(b) From a statistical viewpoint, we can create a **Bayesian linear regression** model. We assume a Gaussian likelihood for the data and place a Gaussian prior on the parameters θ:

$$Y_i \mid x_i, \theta \sim \mathrm{N}(x_i^\top \theta, \sigma^2) \quad \text{and} \quad \theta_j \sim \mathrm{N}(0, \tau^2) \quad \text{for } j = 1, \ldots, p.$$

The posterior distribution is $p(\theta \mid y, X) \propto p(y \mid X, \theta) p(\theta)$. Write down the expression for the **negative log-posterior** distribution, $-\log p(\theta \mid y, X)$.

(c) By comparing your results from (a) and (b), show that minimizing the Ridge regression objective is equivalent to maximizing the posterior probability (i.e., finding the MAP estimate). What is the relationship between the machine learning regularization parameter λ and the statistical parameters σ^2 (noise variance) and τ^2 (prior variance)?

(d) Using the relationship you found in (c), what happens to the optimal θ in the Bayesian model as the prior variance $\tau^2 \to \infty$? What does this imply about our "prior belief" regarding the parameters, and to which simpler model does Ridge regression converge in this case?

References

Bishop, C. M., & Nasrabadi, N. M. (2006). *Pattern recognition and machine learning* (Vol. 4). Springer.

Breiman, L. (2001). Random forests. *Machine Learning,45*(1), 5–32. [https://doi.org/10.1023/A:1010933404324]

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM sigkdd international conference on knowledge discovery and data mining* (pp. 785–794).

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning,20*(3), 273–297. [https://doi.org/10.1023/A:1022627411411]

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika, 80*, 27–38. [MathSciNet][https://doi.org/10.1093/biomet/80.1.27]

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences,55*(1), 119–139. [MathSciNet][https://doi.org/10.1006/jcss.1997.1504]

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics, 29*, 1189–1232. [MathSciNet][https://doi.org/10.1214/aos/1013203451]

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press. http://www.deeplearningbook.org

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. Springer. [https://doi.org/10.1007/978-0-387-84858-7]

Huber, P. J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics,35*(1), 73–101.

[MathSciNet][https://doi.org/10.1214/aoms/1177703732]

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In *Advances in neural information processing systems,30*.

Kolb, C., Müller, C. L., Bischl, B., & Rügamer, D. (2023). Smoothing the edges: A general framework for smooth optimization in sparse regularization using hadamard overparametrization.

Kook, L., Kolb, C., Schiele, P., Dold, D., Arpogaus, M., Fritz, C., Baumann, P. F. M., Kopper, P., Pielok, T., Dorigatti, E., & Rügamer, D. (2024). How inverse conditional flows can serve as a substitute for distributional regression. In *The 40th conference on uncertainty in artificial intelligence*.

Müller, S., Reuter, A., Hollmann, N., Rügamer, D., & Hutter, F. (2025). Position: The future of Bayesian prediction is prior-fitted. In *Forty-second international conference on machine learning position paper track*.

Murphy, K. P. (2022). *Probabilistic machine learning: An introduction*. MIT Press.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. In *Advances in neural information processing systems,31*.

Rügamer, D., Kolb, C., & Klein, N. (2024). Semi-structured distributional regression. *The American Statistician,78*(1), 88–99.
[MathSciNet][https://doi.org/10.1080/00031305.2022.2164054]

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology,58*(1), 267–288.
[MathSciNet][https://doi.org/10.1111/j.2517-6161.1996.tb02080.x]

Vovk, V., Gammerman, A., & Saunders, C. (1999). Machine-learning applications of algorithmic randomness.

Footnotes

1 Technically, the L_0-norm is not an actual norm as it violates the absolute homogeneity property.

2 Note that in the case of actual linear separability, the solution is $\beta = -\infty$.

3 Conformal predictions also exist for other multivariate, count and continuous outputs.

Appendix A

Background in Probability Theory

This chapter provides background on probability theory, including important technical notions and results for which we did not have space to develop in full detail in the main text. In particular, we introduce more formal notation to anchor later derivations without ambiguity. For more thorough treatments of the area, we refer to the classic books by Feller (1968) and Billingsley (2012), while Kauermann et al. (2021) provide a more accessible introduction in Chap. 2. Note that Chap. 1 already introduces the concepts of distribution functions, densities, expectations, and some of the main asymptotic results. In this Appendix, we do not repeat but complement them. We start in Sect. A.1 by introducing how probabilities for events are defined. In Sect. A.2, we introduce conditional probabilities. In Sect. A.3, we provide the mathematical definition of a random variable via a function that transforms stochastic noise to an observed outcome in the real numbers. Within this framework, we define fundamental ways to describe random variables in Sect. A.4 and summarize them in Sect. A.5. Finally, Sects. A.6 and A.7 define different ways random variables can converge and state the most often used inequalities used throughout the book. While this mathematical background is not strictly required for practical application and understanding the main takeaways from this book, it provides valuable intuition and more rigorous understanding of the underlying statistical mechanics.

A.1 Defining Probability

This book builds on understanding statistics from the perspective of data-generating processes. Generally, this process is stochastic rather than deterministic. The outcome of stochastic behavior is known only probabilistically; that is, some events or outcomes are more probable than others. A natural question arising in this context is then how to characterize this stochasticity in a mathematically rigorous framework. Therefore, we introduce probabilities to describe how probable particular events are. The basic building block of probability is a random experiment. Imagine a toy like a magic 8-ball, but instead of giving answers to yes-or-no questions, it changes color when you shake it. The toy can take four different colors that we collect in the set $\Omega = \{\text{“red”}, \text{“blue”}, \text{“orange”}, \text{“purple”}\}$. Each shake is one *random experiment*, and the *outcome* is the color the box takes after the shake. For instance, possible random outcomes from three consecutive experiments could be “orange”, “blue”,

and “orange”. Assuming that every shake of the toy yields independent outcomes, meaning that every outcome is at all times equally likely, we can describe the entire stochastic behavior of the toy box via four probabilities of the box turning “red”, “blue”, “orange”, or “purple”. These are the probabilities of *elementary events*, which need to abide a set of axioms derived by Kolmogorov (1933):

1. All probabilities must be greater than zero, $P(A) \geq 0$.
2. The probability of getting any outcome must be 1, $P(\Omega) = 1$.
3. The probability of a union of two disjoint events equals the sum of their individual probabilities. For A, B with $A \cap B = \emptyset$, we have $P(A \cup B) = P(A) + P(B)$.

These axioms are rather abstract and do not provide an interpretation of probability itself, which sparked a long philosophical debate on how probabilities should be interpreted. We refer to Gillies (2010) for an overview of this topic.

Against this background, we can describe any stochastic mechanism by a *probability space* through a tuple with three components:

1. Sample space, Ω, defined as the set encompassing all possible “atomic” or “elementary” outcomes of the random process; these elementary outcomes are denoted by $\omega \in \Omega$;
2. σ-algebra, $\mathscr{F}$, defined as the set of all compound events to which we want to be able to assign a probability. For a discrete Ω, $\mathscr{F} = 2^{\Omega}$ is commonly defined as the power set with $2^{|\Omega|}$ entries, which is the set of all possible subsets of Ω, including the empty set, $\emptyset$, and Ω. For continuous Ω, $\mathscr{F} = \mathscr{B}(\Omega)$, which is the Borel σ-algebra, defined as the smallest σ-algebra containing all open sets of Ω;
3. Probability measure, *P*, being defined as a function that assigns all elements in $\mathscr{F}$ a probability that needs to abide by Kolmogorov’s Axioms.

To make these abstract definitions more concrete, we illustrate two examples, one for discrete and one for continuous outcomes.

Example A.1 We flip a coin twice and, keeping track of the order, record for each throw separately whether the coin lands on heads (*H*) or tails (*T*).

1. Sample space: $\Omega = \{(H, H), (H, T), (T, H), (T, T)\}$;

2. σ-algebra:

$$\begin{aligned}\mathscr{F} = 2^{\Omega} = \Big\{ \quad & \emptyset, \{(H,H)\}, \{(H,T)\}, \{(T,H)\}, \{(T,T)\}, \\ & \{(H,H),(H,T)\}, \{(H,H),(T,H)\}, \{(H,H),(T,T)\}, \\ & \{(H,T),(T,H)\}, \{(H,T),(T,T)\}, \{(T,H),(T,T)\}, \\ & \{(H,H),(H,T),(T,H)\}, \{(H,H),(H,T),(T,T)\}, \\ & \{(H,H),(T,H),(T,T)\}, \{(H,T),(T,H),(T,T)\}, \\ & \{(H,H),(H,T),(T,H),(T,T)\} \Big\};\end{aligned}$$

3. Probability measure, *P*: Assuming a fair coin, we assign equal probability mass to each outcome in Ω. Any event $A \in \mathscr{F}$ can now be decomposed into elementary elements, e.g., $\{(H,H),(H,T),(T,H),(T,T)\}$ into the four separate outcomes (H,H), (H,T), (T,H), and (T,T). For any event $A \in \mathscr{F}$, we then obtain the probability by summing up over these elementary elements: $P(A) = \sum_{\omega \in A} P(\omega)$. For instance, the probability of $A = \{(H,H),(H,T),(T,H),(T,T)\}$ is $P(A) = 1/4 + 1/4 + 1/4 + 1/4 = 1$. This equates to 1 because $A = \Omega$. ▷

Example A.2 Assume that for one night, we track the percentage spent in REM (rapid eye movement) sleep.

1. Sample Space: $\Omega = (0,1)$ represents the percentage of sleep spent in the REM state;
2. σ-algebra: $\mathscr{F} = \mathscr{B}((0,1))$.
3. Probability measure, *P*: For any interval $(a,b) \subseteq (0,1)$, we assume $P((a,b)) = b - a$. The measure of multiple disjoint intervals is the sum of the respective intervals. For instance, for $(a_1,b_1) \subseteq (0,1)$ and $(a_2,b_2) \subseteq (0,1)$ with $b_1 < a_2$, we get $P((a_1,b_1) \cup (a_2,b_2)) = P((a_1,b_1)) + P((a_2,b_2))$. ▷

The definition of probabilities through a probability space is implicitly assumed in the ensuing sections. Note also that for sets with just one element, such as $\{a\}$, we employ the shortened notation $P(\{a\}) = P(a)$.

A.2 Conditional Probability and Bayes' Theorem

In many settings, we want to refine probabilities given some additional knowledge. For instance, the probability of having cervical cancer is around $1\,\%$ for a woman over her lifetime. If we focus on the subgroup of women with a positive screening result, already around $46\,\%$ of the women have some abnormal cell changes. In this setting, it becomes apparent that the knowledge of a positive screening result should change one's probability of having cervical cancer. This notion of updating one's prior beliefs is fundamental to statistical learning and lets us formulate conditional distributions, which come up repeatedly in this book:

1. Most prominently, all Bayesian models covered in Sect. 3.5 and Chap. 9 are fundamentally based on updating one's knowledge about the parameters with observed data. We start with the prior distribution and learn the posterior distribution.
2. In Chap. 5, we studied statistical tests, which are based on conditional probabilities and defined for a particular $\alpha \in (0, 1)$ by $P(\text{“H}_1\text{”} \mid H_0) \leq \alpha$.
3. Sampling from a multivariate complex distribution can be simplified by breaking it into several samplers from conditional distributions via the Gibbs sampler, illustrated in Sect. 8.5.

We therefore introduce conditional probabilities for measurable events more formally and show how this relates to Bayes' theorem, which serves as the foundation of Bayesian methods.

Definition A.1 (Conditional Probability for Measurable Events) *For two events A, B where the conditioning event B has a strictly positive probability, such that $P(B) > 0$, the conditional probability of event A given that event B has occurred is denoted by $P(A \mid B)$ and defined by*

$$P(A \mid B) := \frac{P(A \cap B)}{P(B)}.$$

With Definition A.1, we can define conditional random variables based on conditional probability measures, which, in turn, give rise to conditional expectations and so on. In many settings, the joint probability of events *A* and *B*, $P(A \cap B)$, is not known, but its symmetry, $P(A \cap B) = P(B \cap A)$, allows us to rewrite the conditional into a form that is often called Bayes' theorem.

Theorem A.1 (Bayes' Theorem) *For two events A, B where the conditioning event B has a strictly positive probability, such that $P(B) > 0$, the conditional probability of event A given that event B has occurred is denoted by $P(A \mid B)$ and defined by*

$$P(A \mid B) = \frac{P(B|A)\ P(A)}{P(B)}. \tag{A.1}$$

Assuming that we have a partition of the sample space into disjoint sets, i.e., $\Omega = \cup_{k=1}^{K} A_k$ with $A_h \cap A_l = \emptyset$ for all $h \neq l$ in the set $\{1, \ldots, K\}$, the *total law of probability* states

$$P(B) = \sum_{k=1}^{K} P(B \mid A_k)\ P(A_k). \tag{A.2}$$

Plugging (A.2) into (A.1) yields

$$P(A_i \mid B) = \frac{P(B|A_i)\ P(A_i)}{\sum_{k=1}^{K} P(B|A_k)\ P(A_k)} \tag{A.3}$$

for any $i \in \{1, \ldots, K\}$. This identity illustrates why Bayesian statistics is often referred to as "probability in reverse": The roles of the conditioning and random elements in (A.3) are swapped between its left- and right-hand sides, $P(A_i \mid B)$ vs. $P(B \mid A_i)$.

Example A.3 In Sect. 6.3, we introduced logistic regression to model binary outcomes ($Y_i \in \{0, 1\}$) with the following model:

$$\pi_i = P(Y_i = 1 \mid \eta_i) = \frac{\exp(\eta_i)}{1+\exp(\eta_i)},$$

with linear predictor

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip},$$

which is a function of β. In a Bayesian context, the parameter vector $\beta = (\beta_0, \ldots, \beta_p)$ is now assumed to be a random variable with a standard normal prior distribution

$$P(\beta) = \prod_{j=0}^{p} \frac{1}{\sqrt{2\pi}\sigma}\ \exp\left(-\frac{\beta_j^2}{2\sigma^2}\right) \tag{A.4}$$

for some $\sigma > 0$. Excluding the covariates in the conditioning argument for simplicity, the distributional assumption that $Y_i \mid \beta \sim \mathrm{B}(n = 1, p = \pi_i)$ for all $i \in \{1, \ldots, n\}$, we can construct the conditional probability of all binary outcomes:

$$P(Y \mid \beta) = \prod_{i=1}^{n} \pi_i^{Y_i}\ (1 - \pi_i)^{1-Y_i}. \tag{A.5}$$

Next, we can make use of Theorem A.1 to derive the posterior $\beta \mid Y$ by combining (A.4) and (A.5):

$$P(\beta \mid Y) = \frac{P(Y|\beta)\ P(\beta)}{P(Y)} \propto P(Y \mid \beta)\ P(\beta).$$

▷

A.3 Real Random Variables

Working with probabilities defined on sets is often inconvenient. Take clinical trials, where we are primarily interested in determining whether a specific medical intervention is safe and effective: Wilding et al. (2021) tests the efficacy of the medication Semaglutide for the treatment of obesity using data from 1961 patients randomized to a test or placebo group. Since the body weight of all patients was recorded weekly for 68 weeks, the sample space Ω is the Cartesian product of all possible weight trajectories for all patients. At the same time, we are mainly interested in the probability of the event that the intervention was successful, which seems tedious to articulate given this complicated sample space. To simplify, we can introduce a function that maps observations in the sample space to a real number, commonly called a *real random variable* or just *random variable*. A meaningful function for assessing the intervention's efficiency is the difference in mean weight loss between baseline (week 0) and the final week (68) for the test and placebo groups. The probability of the original observations in Ω then implies a probability measure, so-called *push-forward measure*, defined on the real line, which is a lot easier to deal with than the original probability measure defined over Ω. Statistical analysis of this induced measure on the real line bypasses the complexities of the original sample space while preserving the stochastic properties needed to test the question at hand: Is the medication Semaglutide an effective obesity treatment?

Definition A.2 (Random Variable) *If $(\Omega, \mathscr{F}, P)$ is a probability space, the measurable mapping*

$$X : \Omega \to \mathbb{R}$$

is called a (real) random variable.

For every event $E \in \mathscr{F}$ in the probability space $(\Omega, \mathscr{F}, P)$, the measure *P* assigns a specific probability $P(E)$. Since *X* is a measurable map, $X^{-1} : \mathbb{R} \to \Omega$, which is called a *preimage*, exists and is given by

$$X^{-1}(A) = \{\omega \in \Omega : X(\omega) \in A\}$$

for any $A \subseteq \mathbb{R}$. The probability measure P_X is then given by

$$P_X(A) = P(X^{-1}(A))$$

for any $A \subseteq \mathbb{R}$. Put differently, the probability of observing event $A \subseteq \mathbb{R}$ in the random variable X is given by first searching for $\omega \in \Omega$ that are projected to A, i.e., $\omega \in \Omega$ such that $X(\omega) \in A$ holds. Then, the induced probability of A results from applying the original probability measure $P(\omega)$ with ω such that $X(\omega) \in A$ holds.

Example A.4 Continuing Example A.2, we are now studying the waiting time until a patient enters their first REM (rapid eye movement) sleep cycle. The unobservable probability space is defined just like in Example A.2:

1. Sample Space: $\Omega = (0, 1)$.
2. σ-algebra: $\mathscr{F} = \mathscr{B}((0, 1))$.
3. Probability Measure: For any subinterval $(a, b] \subseteq (0, 1)$, $P((a, b]) = b - a$.

Different from Example A.2, an element $\omega \in \Omega$ represents now uniform stochastic noise. Since a uniform distribution on the open interval $(0, 1)$ is not a realistic representation of the waiting time until the first REM, we define a random variable to transform this uniform noise to a more reasonable distribution. Neurological waiting times are frequently assumed to follow an exponential distribution with a rate parameter of 1. We construct a random variable $X : (0, 1) \to \mathbb{R}^+$ that deterministically transforms the uniform noise $\omega \in (0, 1)$ into this distribution $X(\omega) = -\log(1 - \omega) \in \mathbb{R}^+$. As derived in Example A.7, one can then show that the induced image measure P_X is a standard exponential distribution.▷

A.4 Characterizing the Distribution of a Random Variable

Having established the random variable X and its probability measure P_X, we next define different ways to characterize the unique stochastic behavior of X. The most fundamental characterization of P_X is via distribution functions.

Definition A.3 (Cumulative Distribution Function) *For random variable X with induced probability measure P_X, the Cumulative Distribution Function (CDF), denoted by F_X, is the function $F_X : \mathbb{R} \to [0, 1]$ defined as the distribution function of P_X:*

$$F_X(x) := P_X((-\infty, x]) = P(\{\omega \in \Omega : X(\omega) \leq x\})$$

for any $x \in \mathbb{R}$ with the following properties:

1. *Nondecreasing: For any $x_1, x_2 \in \mathbb{R}$ with $x_1 \leq x_2$, $F_X(x_1) \leq F_X(x_2)$.*

2. *Right-continuous: For any* $x \in \mathbb{R}$, $\lim_{y \downarrow x} F_X(y) = F_X(x)$.

3. *Boundary Limits:* $\lim_{x \to -\infty} F_X(x) = 0$ *and* $\lim_{x \to \infty} F_X(x) = 1$.

Example A.5 Continuing Example A.4, we study the random variable X representing the waiting time until a patient enters their first REM (rapid eye movement) sleep cycle. Since the waiting time is assumed to follow a standard exponential distribution, the induced probability measure P_X is

$$F_X(x) = 1 - \exp(-x).$$

▷

Since $F_X(x)$ is a nondecreasing function, we can define it as a Lebesgue-Stieltjes integral

$$F_X(x) = \int_{(-\infty, x]} \mathrm{d}\, F_X(u), \tag{A.6}$$

which evaluates as either a discrete summation over mass points, a standard Riemann-Lebesgue integral over a density function, or a linear combination of both, depending on whether X is a discrete, continuous, or mixed random variable:

1. For discrete random variables X with support $\mathscr{X}$, the integral in (A.6) denotes a summation

$$\int_{(-\infty, x]} \mathrm{d}\, F_X(u) = \sum_{k \in \mathscr{X} : k \leq x} P_X(k),$$

   where the probability mass function P_X (PMF) is covered in Definition A.4. The PMF evaluates the probability that X is a specific outcome $f_X(x) = P_X(x)$.
2. For absolutely continuous random variables, (A.6) becomes a standard integral

$$\int_{(-\infty, x]} \mathrm{d}\, F_X(u) = \int_{-\infty}^{x} f_X(u) \mathrm{d}\, u,$$

   where $f_X(x)$ denotes the probability density function, defined as

$$f_X(x) = \frac{\mathrm{d}}{\mathrm{d}x} F_X(x)$$

   at all points x where the cumulative distribution function $F_X(x)$ is differentiable. The function f_X is called the probability density function (PDF) and covered in Definition A.5. The PDF defines $P_X(B)$ implicitly: For any set B, we integrate $f_X(x)$ over $x \in B$.

B, we integrate $f_X(x)$ over $x \in B$.

3. For mixed random variables X, the integral decomposes into a linear combination of a summation over the discrete mass points and an integration over the continuous support via the Lebesgue-Stieltjes integral.

Definition A.4 (Probability Mass Function) *For a discrete random variable X with support $\mathscr{X}$, the function $f_X : \mathscr{X} \to [0, \infty)$ is called the probability mass function (PMF) of X if, for any $x \in \mathscr{X}$, the probability $P_X(x)$ is given by*

$$P_X(x) = f_X(x).$$

Example A.6 Let X be a discrete random variable representing the outcome of rolling a fair six-sided die. The target space of possible outcomes is the finite, discrete set $\mathscr{X} = \{1, 2, 3, 4, 5, 6\}$. Because X only takes values within this discrete set, its stochastic behavior is characterized by a probability mass function (PMF), $f_X(x)$, which assigns an exact probability mass to each specific outcome. Assuming the die is perfectly fair, this mass is distributed equally:

$$f_X(x) = \begin{cases} \frac{1}{6} & \text{for } x \in \{1, 2, 3, 4, 5, 6\} \\ 0 & \text{otherwise} \end{cases}$$

To compute the probability of any event, such as rolling an even number (defined by the subset $B = \{2, 4, 6\}$), we sum the discrete probability masses evaluated at those specific points

$$\begin{aligned} P_X(B) &= \sum\nolimits_{x \in B} f_X(x) \\ &= f_X(2) + f_X(4) + f_X(6) \\ &= \tfrac{1}{6} + \tfrac{1}{6} + \tfrac{1}{6} = \tfrac{1}{2}. \end{aligned}$$

▷

Definition A.5 (Probability Density Function) *For an absolutely continuous random variable X with support $\mathscr{X}$, the function $f_X : \mathscr{X} \to [0, \infty)$ is called the probability density function (PDF) or just density function of X if, for any set $B \in \mathscr{B}(\mathscr{X})$, the probability measure $P_X(B)$ is given by*

$$P_X(B) = \int_B f_X(x) \, d\, x.$$

Example A.7 Continuing Example A.5, X is a continuous random variable representing the waiting time (in hours) until a sleeping patient enters their first

REM cycle. We model this neurological waiting time using a standard exponential distribution with rate parameter $\lambda = 1$. First, we repeat the setup of Example A.4:

1. Sample Space: $\Omega = (0, 1)$.
2. σ-algebra: $\mathscr{F} = \mathscr{B}((0, 1))$.
3. Probability Measure: For any sub-interval $(a, b] \subseteq (0, 1)$, $P((a, b]) = b - a$.

The probability density function for a uniform random variable between 0 and 1 is known to be $f(\omega) = 1$ for any $\omega \in (0, 1)$. We can then derive the density of $X = -\log(1 - \omega) = G(\omega)$, where ω has a uniform density, by transforming the density

$$f_X(x) = f(G^{-1}(x)) \left| \tfrac{\mathrm{d}}{\mathrm{d}\, x} G^{-1}(x) \right|.$$

Calculating

$$-\log(1 - \omega) = G(\omega) = x \;\Rightarrow\; G^{-1}(x) = 1 - \exp(-x),$$

and

$$\tfrac{\mathrm{d}}{\mathrm{d}\, x} G^{-1}(x) = \tfrac{\mathrm{d}}{\mathrm{d}\, x}(1 - \exp(-x)) = \exp(-x),$$

we can characterize the stochastic behavior of X with the following probability density function (PDF), $f_X(x)$:

$$f_X(x) = \begin{cases} \exp(-x) & \text{for } x \geq 0 \\ 0 & \text{for } x < 0 \end{cases}$$

To compute the actual probability that the patient enters their first REM cycle between hour 1 and hour 2 of observation, we have to integrate the density function over the set $B = (1, 2]$:

$$\begin{aligned} P(1 < X \leq 2) &= \int_1^2 f_X(x)\, \mathrm{d}\, x = \int_1^2 \exp(-x)\, \mathrm{d}\, x = [-\exp(-x)]_1^2 \\ &= -\exp(-2) - \Big(-\exp(-1)\Big) \approx 0.368 - 0.135 = 0.233. \end{aligned}$$

▷

A.5 Expected Values

The most common way to summarize a random variable is by its typical or expected values. We call this value the expected value of a random variable, which is the average value we would get if we drew the random variable an infinite number of times.

Definition A.6 (Expected Value) *Let $X : \Omega \to \mathbb{R}$ be a real random variable, then*

$$E(X) := \int x \; d \; P_X(x)$$

is the expected value of X.

Example A.8 Recall the continuous random variable *X* representing the waiting time (in hours) until a patient enters their first REM sleep cycle. We employed an exponential distribution with rate parameter $\lambda = 1$ in this setting. To obtain the expected value, we evaluate the integral over the target space $\mathbb{R}$ using the induced image measure P_X. The differential $\mathrm{d}\ P_X(x)$ is replaced by $f_X(x)\mathrm{d}\ x$, where $f_X(x) = \exp\ (-x)$ for $x \geq 0$:

$$E(X) = \int_{\mathbb{R}} x\ \mathrm{d}\ P_X(x) = \int_0^\infty x f_X(x)\ \mathrm{d}\ x = \int_0^\infty x\ \exp\ (-x)\ \mathrm{d}\ x.$$

With substitution $u = x$ and $\mathrm{d}\ v = \exp\ (-x)\mathrm{d}\ x$, $\mathrm{d}\ u = \mathrm{d}\ x$ and $v = -\exp\ (-x)$ hold

$$\begin{aligned} E(X) &= [-x\ \exp\ (-x)]_0^\infty - \int_0^\infty -\exp\ (-x)\ \mathrm{d}\ x \\ &= (0-0) + \int_0^\infty \exp\ (-x)\ \mathrm{d}\ x \\ &= [-\exp\ (-x)]_0^\infty = 0 - (-1) = 1 \end{aligned}$$

Thus, the expected waiting time until the first REM cycle is precisely 1 hour. ▷

In particular, it holds for a measurable function *t* and distribution P_Y of random variable *Y* that

$$\int t(x)\ \mathrm{d}\ P_X(x) = E(t(X)). \tag{A.7}$$

With $t(x) = (x - \mathbb{E}(X))^2$, we define the variance of the random variable *X*, which describes the volatility of the random variable.

Definition A.7 (Variance) *For random variable X with $\int X^2\ dP < \infty$, the variance of X, denoted by $Var(X)$ or σ_X^2, is given by*

$$Var(X) := E\Big((X - E(X))^2\Big) = \int_{\mathbb{R}} (x - E(X))^2\ dP_X(x).$$

A.6 Stochastic Convergence

For completeness, we define here different ways in which a series of random variables can converge.

Definition A.8 (Convergence in Probability) *The sequence of random variables* $(X_n)_{n\in\mathbb{N}}$ *converges in probability to* $X : \Omega \to \mathbb{R}$*, if*

$$P(|X_n - X| > \varepsilon) \to 0 \; as \; n \to \infty \; for \; all \; \varepsilon > 0.$$

We write $X_n \xrightarrow{P} X$ *and also speak of stochastic convergence.*

Definition A.9 (Convergence in Distribution) *The sequence of random variables* $(X_n)_{n\in\mathbb{N}}$ *with distribution functions* F_i*,* $i = 1, 2, \ldots,$ *n converges in distribution to* $X : \Omega \to \mathbb{R}$ *with distribution function* F*, if*

$$\lim_{n\to\infty} F_n(x) = F(x)$$

for all x *where* F *is continuous. We write* $X_n \xrightarrow{D} X$.

A.7 Probabilistic Inequalities

In many settings encountered throughout this book, we need to bound probabilities or expectations. This section provides the analytic tools required to bound these objects:

1. When proving the Cramér-Rao Bound from Theorem 4.3, we bound the covariance between an unbiased estimator and the score function via the Cauchy–Schwarz inequality (Proposition A.1);
2. To find an upper bound on the split likelihood ratio statistics, proving the validity of universal inference in Sect. 5.8, we make use of the Markov inequality (Proposition A.2);
3. To show that MSE consistency of an estimator implies weak consistency, we bound the probability of the estimator deviating from the true parameter in Sect. 3.3.1 via the Chebyshev inequality (Proposition A.3).

Next, we state these fundamental results.

Proposition A.1 (Cauchy–Schwarz Inequality) *Let U and V be two random variables with finite variances (see Definition A.7), then*

$$\left|E(UV)\right| \leq \sqrt{E(U^2)}\,\sqrt{E(V^2)}.$$

As a direct consequence, for any random variables X and Y with finite variances, we obtain

$$Cov(X, Y)^2 \leq Var(X)\; Var(Y). \tag{A.8}$$

Proposition A.2 (Markov Inequality) *For a nonnegative random variable Y, we have*

$$P(Y > a) \leq \frac{E(Y)}{a}.$$

Proof Note that $E(Y) > 0$. Then

$$\begin{aligned} E(Y) &= \int_0^\infty y f(y) \mathrm{d}\ y \\ &= \int_0^a y f(y) \mathrm{d}\ y + \int_a^\infty y f(y) \mathrm{d}\ y \\ &\geq \int_a^\infty y f(y) \mathrm{d}\ y \\ &\geq a \int_a^\infty f(y) \mathrm{d}\ y = a P(Y \geq a). \end{aligned}$$

□

Proposition A.3 (Chebyshev Inequality) *Let Y be a random variable with finite mean $E(Y) = \mu$ and finite variance $\mathrm{Var}(Y) = \sigma^2 < \infty$. Then, for any $\varepsilon > 0$,*

$$P(|Y - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2}.$$

Proof Chebyshev's inequality follows directly from Markov's inequality applied to the nonnegative random variable $(Y - \mu)^2$. □

Appendix B

Parametric Distribution Families

We now make the abstract definitions from Appendix A more concrete. First, we state the distributions most frequently employed for discrete, countable, and uncountable outcomes. Second, introduce the exponential family as a class of distributions for random variables that covers many realistic settings. Third, we state regularity assumptions that hold for those distributions.

B.1 Frequently Used Distributions

Some distributions appear so often in statistics that familiarity with their parameters and features can be greatly beneficial. The distributions themselves depend on one or more parameters, denoted with Greek letters. We follow the convention of listing the distribution's parameters after the semicolon (Fig. B.1).

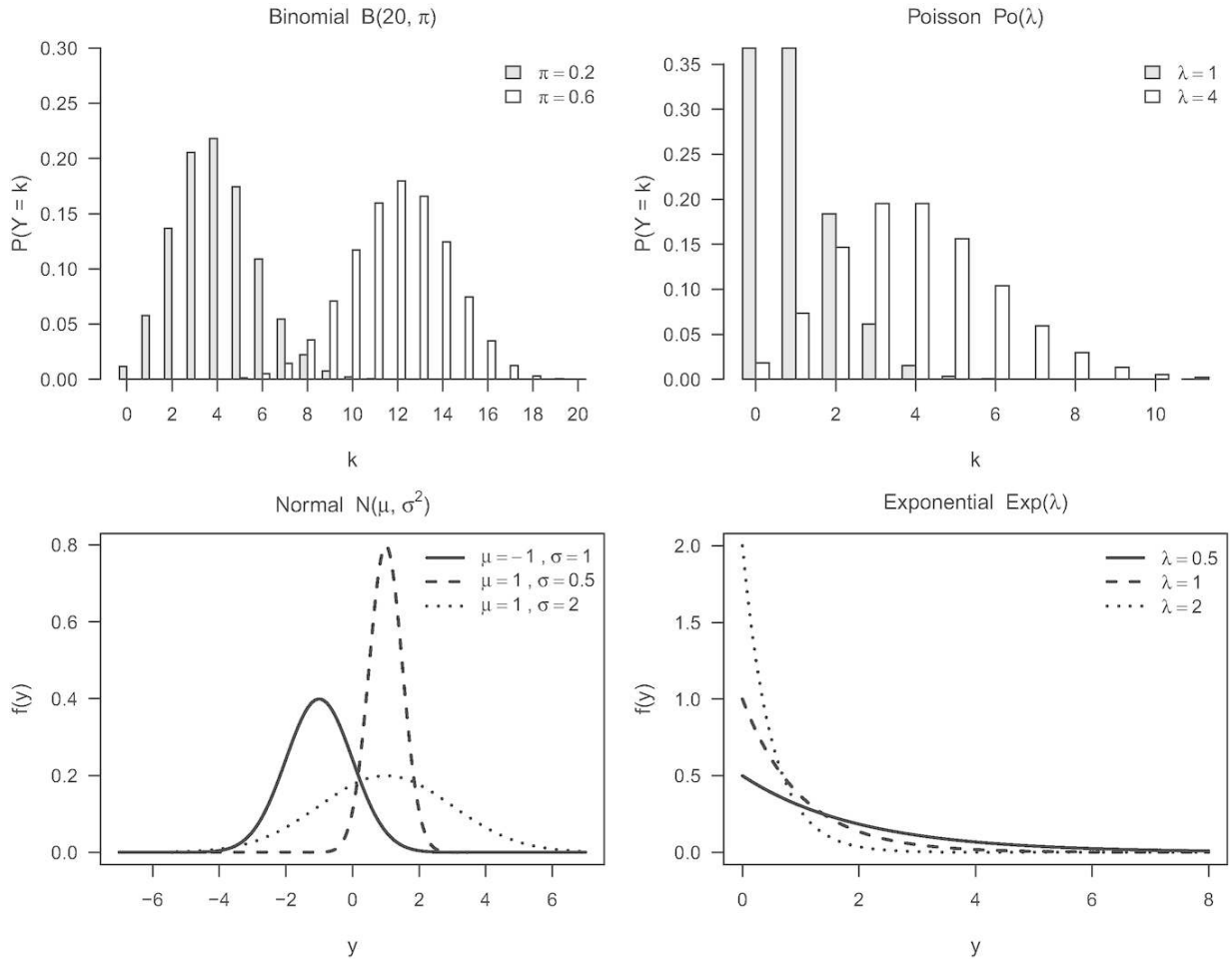


Fig. B.1 Common distributions used throughout the book. The panels show the binomial, Poisson, normal, and exponential distributions for different parameter values

Definition B.1 (Binomial Distribution $B(n, \pi)$)

- *Application: Two outcomes—success or failure with a chance of success of* $\pi \in [0, 1]$*. We count the number of successes in n independent trials.*
- *Probability function:* $P(Y = k; \pi) = \binom{n}{k} \pi^k (1 - \pi)^{n-k}$*, for* $k = 0, \ldots, n$
- *Expectation:* $E(Y) = n\pi$
- *Variance:* $\mathrm{Var}(Y) = n\pi(1 - \pi)$

Definition B.2 (Poisson Distribution $Po(\lambda)$)

- *Application: Count of events in a certain period of time with events occurring at an average rate of* $\lambda > 0$*.*
- *Probability function:* $P(Y = k; \lambda) = \frac{\lambda^k}{k!} \exp(-\lambda)$ *for* $k = 0, 1, 2, \ldots$

- *Expectation:*$E(Y) = \lambda$
- *Variance:*$Var(Y) = \lambda$

Definition B.3 (Normal Distribution $\mathrm{N}(\mu, \sigma^2)$)

- *Application: Metric symmetrically distributed random variables with mean*$\mu \in \mathbb{R}$*and variance*$\sigma^2 > 0$.
- *Density function:*$f(y; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$*for*$y \in \mathbb{R}$
- *Expectation:*$E(Y) = \mu$
- *Variance:*$Var(Y) = \sigma^2$

Definition B.4 (Exponential Distribution $\mathrm{Exp}(\lambda)$)

- *Application: Time between two events following a Poisson process, where the waiting time for the next event is on average*$1 \ / \ \lambda$, $\lambda > 0$.
- *Density function:*$f(y; \lambda) = \lambda \exp(-\lambda y)$*for*$y \geq 0$
- *Expectation*$E(Y) = \frac{1}{\lambda}$
- *Variance*$Var(Y) = \frac{1}{\lambda^2}$

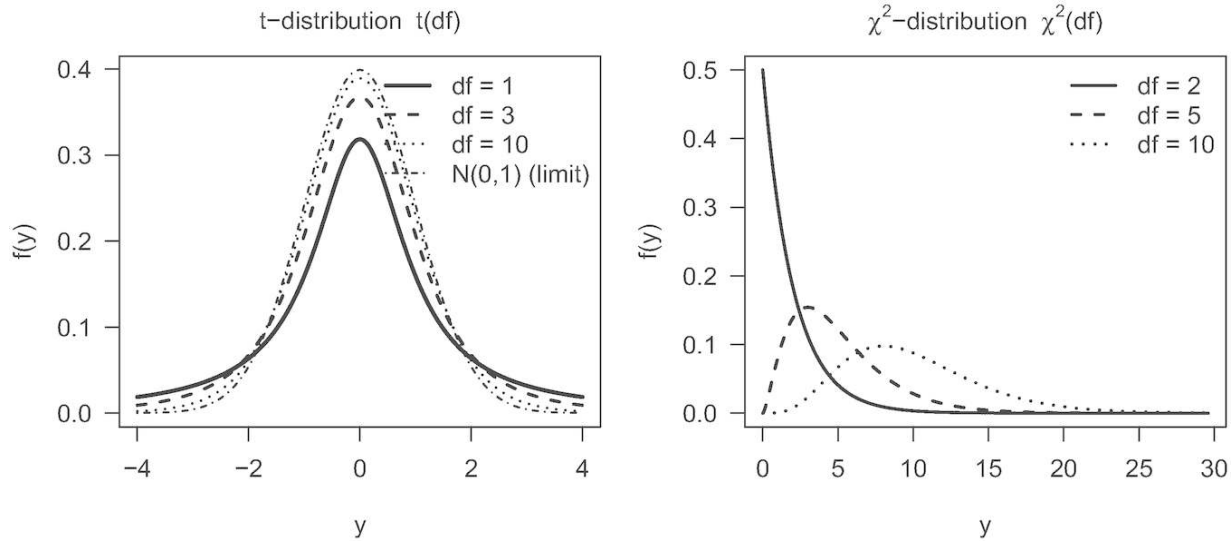


Fig. B.2 Distributions that occur frequently in statistical inference. The panels show the *t*-distribution and the chi-squared distribution for different degrees of freedom

Definition B.5 (t-Distribution t_n)

- *Application: Metric symmetrically distributed random variable with a considerable probability of extreme outcomes ("heavy tails"). Also used for statistical tests of the mean of normally distributed variables when the variance is unknown and estimated from the data and the degrees of freedom* (n) *are small.*
- *Density function:*

$$f(y;n) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})}\left(1+\frac{y^2}{n}\right)^{-\frac{n+1}{2}}$$

  where $\Gamma(\cdot)$ *is the gamma function and* $n = 1, 2, \ldots$ *is the degree of freedom (which in principle can take any positive value, but we only work with positive discrete values here)*
- *Expectation:* $E(Y) = 0$ *for* $n \geq 2$
- *Variance:* $Var(Y) = \frac{n}{n-2}$ *for* $n \geq 3$
- *Properties:*
  - *The t-distribution resembles the normal distribution but has heavier tails, that is, more probability mass for extreme values. We show different t-distributions and their limit, the normal distribution, in Fig.*B.2
  - *We can combine the normal distribution and the Chi-squared distribution to derive the t-distribution: Let* $Z_1, \ldots, Z_n$ *be n independently* $N(0,1)$ *distributed random variables. Then the ratio*

$$\frac{Z_1}{\sqrt{\sum_{i=2}^{n} Z_i^2 \Big/ (n-1)}} \sim t_{(n-1)} \tag{B.1}$$

    follows a t-distribution with $n-1$ *degrees of freedom.*
  - *Property* (B.1) *can be used to prove* (3.20).

Proof Subsequently, we provide the distribution of the *t*-distribution

$$T = \frac{\overline{Y}-\mu}{S_n/\sqrt{n}} \sim t_{(n-1)}, \tag{B.2}$$

with $S_n^2 = \sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2 \Big/ (n-1)$. We aim to show that this follows a *t*-distribution with $n-1$ degrees of freedom. The idea of the proof is to

show that the ratio (B.2) can be rewritten as the ratio (B.1). To do so, we define the following constructive variables:

$$
\begin{aligned}
Z_1 &:= \frac{1}{\sqrt{n}} \left(Y_1 + Y_2 + \ldots + Y_n\right) = \overline{Y}\,\sqrt{n} \\
Z_2 &:= \frac{1}{\sqrt{2}} \left(Y_1 - Y_2\right) \\
Z_3 &:= \frac{1}{\sqrt{3}\sqrt{2}} \left(Y_1 + Y_2 - 2Y_3\right) \\
Z_4 &:= \frac{1}{\sqrt{4}\sqrt{3}} \left(Y_1 + Y_2 + Y_3 - 3Y_4\right) \\
&\vdots \\
Z_n &:= \frac{1}{\sqrt{n}} \frac{1}{\sqrt{n-1}} \left(Y_1 + Y_2 + \ldots + Y_{n-1} - (n-1)Y_n\right).
\end{aligned}
$$

All Z_i with $i = 1, \ldots, n$ are normally distributed as they are sums of (independent) normally distributed random variables. Their moments are

$$
\begin{aligned}
E(Z_1) &= \sqrt{n}\mu \\
E(Z_2) &= \ldots = E(Z_n) = 0.
\end{aligned}
$$

with the following pairwise covariances

$$
\begin{aligned}
\mathrm{Cov}(Z_1, Z_2) &= \frac{1}{\sqrt{n}\sqrt{2}} \mathrm{Cov}(\overline{Y}, Y_1 - Y_2) \\
&= \frac{1}{\sqrt{n}\sqrt{2}} \left\{\mathrm{Cov}(\overline{Y}, Y_1) - \mathrm{Cov}(\overline{Y}, Y_2)\right\} = 0 \\
\mathrm{Cov}(Z_1, Z_2) &= 0 \\
&\vdots \\
\mathrm{Cov}(Z_1, Z_n) &= 0.
\end{aligned}
$$

For $n \geq i > j > 1$, we obtain

$$\begin{aligned}\mathrm{Cov}(Z_i, Z_j) &= \mathrm{Cov}\left(\frac{1}{\sqrt{i}}\frac{1}{\sqrt{(i-1)}}\left(Y_1 + \ldots - (i-1)Y_i\right), \frac{1}{\sqrt{j}}\frac{1}{\sqrt{(j-1)}}\left(Y_1 + \ldots - (j-1)Y_j\right.\right.\\ &= \frac{1}{\sqrt{i}}\frac{1}{\sqrt{(i-1)}}\frac{1}{\sqrt{j}}\frac{1}{\sqrt{(j-1)}}\mathrm{Cov}\left(Y_1 + \ldots + Y_j, Y_1 + \ldots - (j-1)Y_j\right)\\ &= \frac{1}{\sqrt{i}}\frac{1}{\sqrt{(i-1)}}\frac{1}{\sqrt{j}}\frac{1}{\sqrt{(j-1)}}\left(\underbrace{\mathrm{Var}(Y_1) + \ldots + \mathrm{Var}(Y_{j-1}) - (j-1)\mathrm{Var}(Y_j)}_{=0}\right) =\end{aligned}$$

Thus, $\mathrm{Cov}(Z_i, Z_j) = 0$ holds for $i \neq j$. It remains to derive the univariate variances

$$\begin{aligned}\mathrm{Var}(Z_1) &= \sigma^2\\ \mathrm{Var}(Z_i) &= \mathrm{Var}\left(\frac{1}{\sqrt{i}}\frac{1}{\sqrt{(i-1)}}\left(Y_1 + \ldots - (i-1)Y_i\right)\right)\\ &= \frac{1}{i}\frac{1}{(i-1)}\left(\mathrm{Var}(Y_1) + \ldots + (i-1)^2\,\mathrm{Var}(Y_i)\right)\\ &= \frac{1}{i}\frac{1}{(i-1)}\left((i-1) + (i-1)^2\right)\sigma^2\\ &= \sigma^2.\end{aligned}$$

We can conclude that with (B.1) we obtain

$$\frac{Z_1 - \sqrt{n}\mu}{\sqrt{\sum_{i=2}^{n} Z_i^2 \Big/ (n-1)}} \sim t_{(n-1)}. \tag{B.3}$$

It remains to show that (B.2) can be expressed as (B.3), which we will show by induction. For $n = 2$, we have

$$T = \frac{\overline{Y} - \mu_0}{\sqrt{S^2/2}} = \frac{\frac{Y_1+Y_2}{2} - \mu_0}{\sqrt{\left((Y_1 - \overline{Y})^2 + (Y_2 - \overline{Y})^2\right)/2}} = \frac{Z_1 - \sqrt{2}\mu}{\sqrt{Z_2^2}} \sim t_{(1)}$$

Hence, for $n = 2$, we have shown the equivalence between the t-statistic (B.2) and its reformulation (B.3). We can conclude the proof now by induction. To do so, we define $\overline{Y}_{(n)} = \sum Y_i \Big/ n$ and accordingly with $S^2_{(n)} = \sum_{i=1}^{n} \left(Y_i - \overline{Y}_{(n)}\right)^2 \Big/ (n-1)$ and $T_{(n)}$, with obvious definition for

the latter. Let us therefore assume that the relation between $T_{(n)}$ in (B.2) and (B.3) holds up to n, that is,

$$T_{(n)} = \frac{\overline{Y}_{(n)} - \mu}{\sqrt{S^2_{(n)}/n}} = \frac{Z_1 - \sqrt{n}\mu}{\sqrt{\sum_{i=2}^{n} Z_i^2 \Big/ (n-1)}} \sim t_{(n-1)}$$

This implies in particular the equality

$$Z_2^2 + \ldots + Z_n^2 = \sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2 = (n-1) S^2_{(n)} \tag{B.4}$$

For $n+1$, we get

$$\begin{aligned}
Z_2^2 + \ldots + Z_n^2 + Z_{n+1}^2 &= \sum_{i=1}^{n} \left(Y_i - \overline{Y}_{(n)}\right)^2 + \frac{1}{n+1}\frac{1}{n}(Y_1 + \ldots + Y_n - nY_{n+1})^2 \\
&= \sum_{i=1}^{n} Y_i^2 - n\overline{Y}^2_{(n)} + \frac{1}{n+1}\frac{1}{n}(n\overline{Y}_{(n)} - nY_{n+1})^2 \\
&= \sum_{i=1}^{n} Y_i^2 - n\overline{Y}^2_{(n)} + \frac{n}{n+1}(\overline{Y}_{(n)} - Y_{n+1})^2 \\
&= \sum_{i=1}^{n} Y_i^2 - n\overline{Y}^2_{(n)} + \frac{n}{n+1}\left(\overline{Y}^2_{(n)} - 2\overline{Y}_{(n)}Y_{n+1} + Y_{n+1}^2\right) \\
&= \sum_{i=1}^{n+1} Y_i^2 - \frac{1}{n+1}\left(n^2\overline{Y}^2_{(n)} + 2n\overline{Y}_{(n)}Y_{n+1} + Y_{n+1}^2\right) \\
&= \sum_{i=1}^{n+1} Y_i^2 - (n+1)\overline{Y}^2_{(n+1)}
\end{aligned}$$

where the latter follows since

$$\begin{aligned}\overline{Y}_{(n+1)}^2 &= \left(\tfrac{1}{n+1}\right)^2\left(\sum_{i=1}^{n+1} Y_i\right)^2 \\ &= \tfrac{1}{(n+1)^2}\left(\sum_{i=1}^{n+1} Y_i + Y_{n+1}\right)^2 \\ &= \tfrac{1}{(n+1)^2}\left(n\overline{Y}_{(n)} + Y_{n+1}\right)^2 \\ &= \tfrac{1}{(n+1)^2}\left(n^2\overline{Y}_{(n)}^2 + 2n\overline{Y}_{(n)}Y_{n+1} + Y_{n+1}^2\right).\end{aligned}$$

Consequently

$$T_{(n+1)} = \frac{\overline{Y}_{(n+1)}-\mu}{S_{(n+1)}/\sqrt{n}} = \frac{Z_1}{\sqrt{\sum_{i=2}^{n+1} Z_2^2 \Big/ (n)}} \sim t_{(n)},$$

which proves that the *t*-statistics (B.2) can be written as (B.1) and hence follows a *t*-distribution with $n-1$ degrees of freedom. □

Definition B.6 (Chi-squared Distribution $\mathscr{X}^2(n)$)

- *Application: Squared quantities like sample variances, sum of n independent squared normal distributed random variables.*
- *Density function:* $f(y;n) = y^{\frac{n}{2}-1}\,\exp\left(-\frac{y}{2}\right)\Big/ 2^{\frac{n}{2}}\Gamma(\frac{n}{2})$ *where* $n = 1, 2, 3, \ldots$
- *Expectation:* $E(Y) = n$
- *Variance:* $\mathrm{Var}(Y) = 2n$
- *Properties:*
  - *Let* $Z_1, \ldots, Z_n$ *be n independently* $N(0,1)$ *distributed random variables. Then*

    $$X^2 := \sum_{i=1}^{n} Z_i^2$$

    follows a Chi-squared distribution with n degrees of freedom.
  - *Assume that Y is vector-valued normally distributed, that is,* $Y \sim N(\mu, \Sigma)$ *where* $\mu \in \mathbb{R}^p$ *is the vector of mean values and* $\Sigma \in \mathbb{R}^{p\times p}$ *is the positive definite variance matrix. Then*

    $$(Y-\mu)^\top \Sigma^{-1}(Y-\mu) \sim \mathscr{X}_p \tag{B.5}$$

with p as dimension of Y.

B.2 The Exponential Family

Several of the above distributions admit a mathematical generalization, which in turn enables the development of theories and models that apply to the entire model class. One such class of distributions, central to statistics, is the exponential family. This class covers many of the most common distributions in statistics, including the normal, Poisson, and binomial distributions. Although these distributions are very different, they share structural properties.

Definition B.7 *A class of distributions for a random variable Y is called an***exponential family***, if the density (or probability function) can be factorized in the following form*

$$f_Y(y;\theta) = \exp\left(t(y)^\top \theta - \kappa(\theta)\right) h(y), \tag{B.6}$$

where $h(y) \geq 0$ and $t(y) = (t_1(y), \ldots, t_p(y))^\top$ are a vector of known functions and $\theta = (\theta_1, \ldots, \theta_p)^\top$ is a parameter vector.

The density (or probability function) given in (B.6) is very general and thus appears complicated, but we will see in subsequent examples that the different terms simplify substantially in most cases. Consider the quantities in (B.6) in a little more depth. First of all, the function $t(y)$ is a function of the random variable and will later be called a statistic. The term θ is the parameter of the distribution, also called the natural parameter, and the quantity $\kappa(\theta)$ serves as a normalization constant, such that the density integrates out to 1. Hence, $\kappa(\theta)$ is defined by the following:

$$\begin{aligned} 1 &= \int \exp\left(t(y)^\top \theta\right) h(y) \mathrm{d}\, y \exp\left(-\kappa(\theta)\right) \\ \Leftrightarrow \quad \kappa(\theta) &= \log \int \exp\left(t(y)^\top \theta\right) h(y) \mathrm{d}\, y. \end{aligned}$$

As a consequence, by differentiating $\kappa(\theta)$ with respect to θ, we get

$$\frac{\partial \kappa(\theta)}{\partial \theta} = \frac{\int t(y)^\top \exp(t(y)^\top \theta) h(y) \mathrm{d}\, y}{\int \exp(t(y)^\top \theta) h(y) \mathrm{d}\, y} = \int t(y)^\top f_Y(y;\theta) \mathrm{d}\, y = E\Big(t(Y)^\top\Big).$$

Finally, the quantity $h(y)$ depends on the random variable, not on the parameter, and we will see later that this quantity will not be of particular interest.

We will now demonstrate how some of the distributions introduced above can be written in the form of an exponential family. We thereby start with the normal distribution, which can be rewritten as

$$\begin{aligned} f(y) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(y-\mu)^2}{\sigma^2}\right) \\ &= \exp\left(-\frac{1}{2}\frac{y^2 - 2y\mu + \mu^2}{\sigma^2} - \frac{1}{2}\log(\sigma^2)\right)\frac{1}{\sqrt{2\pi}} \\ &= \exp\left(\underbrace{\left(-y^2/2, y\right)}_{t(y)^\top}\underbrace{\begin{pmatrix}1/(\sigma^2)\\ \mu/(\sigma^2)\end{pmatrix}}_{\theta} - \underbrace{\frac{1}{2}\left(-\log\frac{1}{\sigma^2} + \frac{\mu^2}{\sigma^2}\right)}_{\kappa(\theta)}\right)\underbrace{\frac{1}{\sqrt{2\pi}}}_{h(y)}. \end{aligned}$$

where $\theta_1 = \dfrac{1}{\sigma^2}$ and $\theta_2 = \frac{\mu}{\sigma^2}$ such that

$$\kappa(\theta) = \tfrac{1}{2}\left(-\log(\theta_1) + \tfrac{\theta_2^2}{\theta_1}\right) = -\tfrac{1}{2}\left(\log(\theta_1) - \tfrac{\theta_2^2}{\theta_1}\right).$$

Note that

$$\begin{aligned} \tfrac{\partial\kappa(\theta)}{\partial\theta_1} &= -\tfrac{1}{2}\left(\tfrac{1}{\theta_1} + \tfrac{\theta_2^2}{\theta_1^2}\right) = -\tfrac{1}{2}(\sigma^2 + \mu^2) = E\left(-\tfrac{Y^2}{2}\right) = E\Big(t_1(Y)\Big) \\ \tfrac{\partial\kappa(\theta)}{\partial\theta_2} &= \tfrac{\theta_2}{\theta_1} = \mu = E(Y) = E\Big(t_2(Y)\Big). \end{aligned}$$

Hence, we see that the normal distribution is an exponential family distribution.

The same holds for the Binomial distribution, because

$$
\begin{aligned}
f(y) = P(Y = y) &= \binom{n}{y} \pi^y (1-\pi)^{n-y} \\
&= \left(\frac{\pi}{1-\pi}\right)^y (1-\pi)^n \binom{n}{y} \\
&= \exp\left(\underbrace{y}_{t(y)} \underbrace{\log\left(\frac{\pi}{1-\pi}\right)}_{\theta} - \underbrace{n \log\left(\frac{1}{1-\pi}\right)}_{\kappa(\theta)} \right) \underbrace{\binom{n}{y}}_{h(y)}.
\end{aligned}
$$

where $\theta = \log\left(\frac{\pi}{(1-\pi)}\right)$ is also known as log-odds ratio and $\kappa(\theta)$ results in

$$\kappa(\theta) = n \log\left(1 + \exp\left(\theta\right)\right).$$

As $\pi = \exp\left(\theta\right) / \left(1 + \exp\left(\theta\right)\right)$, one gets

$$\frac{\partial \kappa(\theta)}{\partial \theta} = n \frac{\exp(\theta)}{1+\exp(\theta)} = n\pi = E(Y) = E\Big(t(Y)\Big).$$

Among many other distributions, the Poisson distribution is also part of the exponential family, the proof of which we leave as an exercise for the reader.

B.3 Fisher Regularity

We will work with distributions that have special properties that allow optimization through differentiation. Such distributions are defined as Fisher-regular below, where the eponym refers to Sir Ronald Aylmer Fisher, one of the most famous and productive statisticians in the twentieth century (see Box, 1978).

Definition B.8 *A distribution $f(y;\theta)$ is* **Fisher-regular** *if the following properties hold the following:*

1. *The support of Y is not dependent upon θ, i.e., the set $\{y : f(y;\theta) > 0\}$ does not depend on θ.*
2. *The possible parameter space Θ is open; i.e., if θ is univariate, it has the form $\Theta = (a, b)$ with $a < b$.*

3.

   The probability function $f(y;\theta)$ *can be differentiated twice with respect to* θ.

4. *Integration with respect to some suitable measure and differentiation with respect to the parameter* θ *are exchangeable.*

The latter property means that, for instance, for continuous random variables, we have the equality

$$\int \frac{\partial}{\partial\theta} f(y;\theta)\mathrm{d}\ y = \frac{\partial}{\partial\theta} \int f(y;\theta)\mathrm{d}\ y.$$

We emphasize that Fisher-regularity is not a strong assumption and is met by most common distributions. Still, it can be violated, and the standard counterexample is the uniform distribution on $[0,\theta]$, i.e.,

$$f(y;\theta) = \begin{cases} \frac{1}{\theta} & \text{for } y \in [0,\theta] \\ 0 & \text{otherwise.} \end{cases}$$

Here, the support of the distribution depends explicitly on the parameter, and hence, property 1 above is violated.

References

Agresti, A. (2012). *Categorical data analysis* (Vol. 792). Wiley.

Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle. In *Proceedings of the 2nd international symposium on information theory* (pp. 267–281).

Aldrich, J. (1997). R.A. Fisher and the making of maximum likelihood 1912–1922. *Statistical Science,12*(3), 162–176.
[MathSciNet][https://doi.org/10.1214/ss/1030037906]

Andrieu, C., & Thoms, J. (2008). A tutorial on adaptive MCMC. *Statistics and Computing,18*, 343–373.
[MathSciNet]

Anselin, L. (1988). *Spatial econometrics: Methods and models* (Vol. 4). Springer.

Baddeley, A., Bárány, I., & Schneider, R. (2007). Spatial point processes and their applications. In *Stochastic geometry: Lectures given at the CIME Summer School Held in Martina Franca, Italy, September 13–18, 2004* (pp. 1–75).

Baird, D. (1983). The Fisher/Pearson chi-squared controversy: A turning point for inductive inference. *The British Journal for the Philosophy of Science,34*(2), 105–118.
[MathSciNet]

Barabási, A.-L., & Albert, R. (1999). Emergence of scaling in random networks. *Science,286*(5439), 509–512.
[MathSciNet]

Barbour, A. D., & Reinert, G. (2026). *Networks: Probability and statistics* (Cambridge series in statistical and probabilistic mathematics). Cambridge University Press.

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, FRS communicated by Mr. Price, in a letter to John Canton, AMFR S. *Philosophical Transactions of the Royal Society of London,* (53), 370–418.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society,57*, 289–300.
[MathSciNet]

Berger, V. W., & Zhou, Y. (2014). *Kolmogorov–Smirnov test: Overview*. Wiley statsref: Statistics reference online.

Berry, K. J., Johnston, J. E., & Mielke, P. W. (2019). *A primer of permutation statistical methods*. Springer.

Berry, K. J., Johnston, J. E., & Mielke, P. W. Jr. (2011). Permutation methods. *Wiley Interdisciplinary Reviews: Computational Statistics,3*(6), 527–542.

Billingsley, P. (2012). *Probability and measure* (Wiley series in probability and statistics). Wiley.

Bishop, C. M., & Nasrabadi, N. M. (2006). *Pattern recognition and machine learning* (Vol. 4). Springer.

Bivand, R. S., Pebesma, E. J., & Gómez-Rubio, V. (2008). *Applied spatial data analysis with R*. Springer.

Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association,112*(518), 859–877. [MathSciNet]

Bochenek, B., & Z. Ustrnul (2022). Machine learning in weather prediction and climate analyses—applications and perspectives. *Atmosphere,13*(2), 180. [https://doi.org/10.3390/atmos13020180]

Boland, P. J. (1984). A biographical glimpse of William Sealy Gosset. *The American Statistician,38*(3), 179–183. [https://doi.org/10.1080/00031305.1984.10483195]

Bonferroni (1936). *Teoria statistica delle classi e calcolo delle probabilità*. Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze.

Box, G., & Jenkins, G. (1970). *Time series analysis: Forecasting and control*. Holden-Day.

Box, G. E. (1976). Science and statistics. *Journal of the American Statistical Association,71*(356), 791–799. [MathSciNet][https://doi.org/10.1080/01621459.1976.10480949]

Box, G. E., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: Forecasting and control*. Wiley.

Box, J. F. (1978). *R.A. Fisher, the life of a scientist*. Wiley.

Breiman, L. (2001). Random forests. *Machine Learning,45*(1), 5–32. [https://doi.org/10.1023/A:1010933404324]

Brockwell, P. J., & Davis, R. A. (2016). *Introduction to time series and forecasting* (3rd ed.). Springer.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM sigkdd international conference on knowledge discovery and data mining* (pp. 785–794).

Claeskens, G., & Hjort, N. L. (2008). *Model selection and model averaging*. Cambridge Books.

Clopper, C. J., & Pearson, E. S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika,26*(4), 404–413. [https://doi.org/10.1093/biomet/26.4.404]

Coles, S., Bawa, J., Trenner, L., & Dorazio, P. (2001). *An introduction to statistical modeling of extreme values* (Vol. 208). Springer.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning,20*(3), 273–297. [https://doi.org/10.1023/A:1022627411411]

Cowell, R. G., Dawid, P., Lauritzen, S. L., & Spiegelhalter, D. J. (2007). *Probabilistic networks and expert systems: Exact computational methods for Bayesian networks*. Springer.

Cramér, H. (1946). *Mathematical methods of statistics*. Princeton University Press.

Cressie, N. (2015). *Statistics for spatial data*. Wiley.

Dale, A. I. (1982). Bayes or laplace? An examination of the origin and early applications of Bayes' theorem. *Archive for History of Exact Sciences, 27*, 23–47.
[MathSciNet]

Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their application* (Cambridge series in statistical and probabilistic mathematics). Cambridge University Press.

de Carvalho, M., & Marques, F. (2012). Jackknife Euclidean likelihood-based inference for spearman's RHO. *North American Actuarial Journal,16*(4), 487–492.
[MathSciNet][https://doi.org/10.1080/10920277.2012.10597644]

De Gooijer, J. G. (2017). *Elements of nonlinear time series analysis and forecasting* (Vol. 37). Springer.

de Laplace, P. S. (1820). *Théorie analytique des probabilités* (Vol. 7). Courcier.

De Nicola, G., Fritz, C., Mehrl, M., & Kauermann, G. (2023). Dependence matters: Statistical models to identify the drivers of tie formation in economic networks. *Journal of Economic Behavior & Organization,215*, 351–363.

Dean, A., Voss, D., & Draguljić, D. (2017). Response surface methodology. In *Design and analysis of experiments* (pp. 565–614). Springer.

Dekking, F. M., Kraaikamp, C., Lopuhaä, H. P., & Meester, L. E. (2006). *A modern introduction to probability and statistics: Understanding why and how*. Springer.

DelSole, T., & Tippett, M. K. (2021). Correcting the corrected aic. *Statistics & Probability Letters,173*, 109064.
[MathSciNet]

Devlin, K. (2010). The pascal-fermat correspondence: How mathematics is really done. *The Mathematics Teacher,103*(8), 579–582.
[https://doi.org/10.5951/MT.103.8.0579]

Dikta, G., & Scheer, M. (2021). *Bootstrap methods*. Springer.

Edwards, A. W. (1997). What did fisher mean by "inverse probability" in 1912–1922? *Statistical*

Science,12(3), 177–184.
[MathSciNet][https://doi.org/10.1214/ss/1030037907]

Efron, B. (1979). Bootstrap methods: Another look at the Jackknife. *The Annals of Statistics,7*(1), 1–26.
[MathSciNet]

Efron, B. (2010). *Large-scale inference: Empirical Bayes methods for estimation, testing, and prediction*. Cambridge University Press.

Efron, B. (2022). *Exponential families in theory and practice*. Cambridge University Press.

Eicker, F. (1967). Limit theorems for regressions with unequal and dependent errors. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, Berkeley, CA, 1965/66 (Vol. I: Statistics, pp. 59–82). University California Press.

Fahrmeir, L., Kneib, T., Lang, S., & Marx, B. (2013). *Regression—models, methods and applications*. Springer.
[https://doi.org/10.1007/978-3-642-34333-9]

Feller, W. (1968). *An introduction to probability theory and its applications* (Wiley series in probability and mathematical statistics, 3rd ed., rev ed.). Wiley.

Fienberg, S. E. (2006). When did Bayesian inference become "Bayesian"? *Bayesian Analysis,1*(1), 1–40.
[MathSciNet][https://doi.org/10.1214/06-BA101]

Fienberg, S. E. (2012). A brief history of statistical models for network analysis and open challenges. *Journal of Computational and Graphical Statistics,21*(4), 825–839. Taylor & Francis.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika, 80*, 27–38.
[MathSciNet][https://doi.org/10.1093/biomet/80.1.27]

Fisher, R. (1922). On the interpretation of chi-squared from contingency tables, and the calculation of p. *Journal of the Royal Statistical Society, 85*, 87–94.

Fisher, R. (1955). Statistical methods and scientific induction. *Journal of the Royal Statistical Society: Series B (Methodological),17*(1), 69–78.
[MathSciNet]

Fisher, R. (1958a). Cigarettes, cancer, and statistics. *The Centennial Review of Arts & Science,2*, 151–166. Michigan State University Press.

Fisher, R. A. (1912). On an absolute criterion for fitting frequency curves. *Messenger of Mathematics,41*, 155–156.

Fisher, R. A. (1958b). Lung cancer and cigarettes? *Nature,182*(4628), 108–108.
[https://doi.org/10.1038/182108a0]

Fox, C. W., & Roberts, S. J. (2012). A tutorial on variational Bayesian inference. *Artificial*

Intelligence Review,38, 85–95.

Fréchet, M. (1943). Sur l'extension de certaines évaluations statistiques au cas de petits échantillons. *Revue de l'Institut International de Statistique,11*, 182–205. [https://doi.org/10.2307/1401114]

Freeman, L. C. (1978). Centrality in social networks conceptual clarification. *Social Networks,1*(3), 215–239.

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences,55*(1), 119–139. [MathSciNet][https://doi.org/10.1006/jcss.1997.1504]

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics, 29*, 1189–1232. [MathSciNet][https://doi.org/10.1214/aos/1013203451]

Gandini, S., Botteri, E., Iodice, S., Boniol, M., Lowenfels, A. B., Maisonneuve, P., & Boyle, P. (2008). Tobacco smoking and cancer: A meta-analysis. *International Journal of Cancer,122*(1), 155–164.

Garwood, F. (1936). Fiducial limits for the poisson distribution. *Biometrika,28*(3/4), 437–442. [https://doi.org/10.2307/2333958]

Gawlikowski, J., Tassi, C. R. N., Ali, M., Lee, J., Humt, M., Feng, J., Kruspe, A., Triebel, R., Jung, P., Roscher, R., et al. (2023). A survey of uncertainty in deep neural networks. *Artificial Intelligence Review,56*(Suppl 1), 1513–1589. [https://doi.org/10.1007/s10462-023-10562-9]

Gelfand, A. E. (2000). Gibbs sampling. *Journal of the American Statistical Association,95*(452), 1300–1304. [MathSciNet]

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence,PAMI-6*(6), 721–741.

Gertheiss, J., Rügamer, D., Liew, B. X., & Greven, S. (2024). Functional data analysis: An introduction and recent developments. *Biometrical Journal,66*(7), e202300363. [MathSciNet]

Gillies, D. (2010). *Philosophical theories of probability* (Philosophical issues in science, Reprinted ed.). Routledge.

Goldenberg, A., Zheng, A. X., Fienberg, S. E., Airoldi, E. M., & Others (2010). A survey of statistical network models. *Foundations and Trends in Machine Learning,2*(2), 129–233. Now Publishers, Inc.

Goldsmith, J., Scheipl, F. Huang, L., Wrobel, J., Di, C., Gellar, J., Harezlak, J., McLean, M. W., Swihart, B., Xiao, L., Crainiceanu, C., & Reiss, P. T. (2020). *Refund: Regression with functional data*. R package version 0.1-23.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press. http://www.deeplearningbook.org

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. In *ICLR proceedings*.

Goodman, S. (2008). A dirty dozen: Twelve p-value misconceptions. *Semin Hematol, 45*, 235–240.

Green, P. J. (2001). A primer on Markov chain Monte Carlo. *Monographs on Statistics and Applied Probability,87*, 1–62.
[MathSciNet]

Greven, S., & Scheipl, F. (2017). A general framework for functional regression modelling. *Statistical Modelling,17*(1–2), 1–35.
[MathSciNet]

Grimmett, G., & Stirzaker, D. (2001). *Probability and random processes* (3rd ed.). Oxford University Press.

Grimmett, G. R. (1973). A theorem about random fields. *Bulletin of the London Mathematical Society,5*(1), 81–84.
[MathSciNet]

Gruber, C., Schenk, P. O., Schierholz, M., Kreuter, F., & Kauermann, G. (2026). Sources of uncertainty in machine learning—a statisticians' view. *Statistical Science*, *41*(3), 671–696.
[MathSciNet][https://doi.org/10.1214/25-STS1003]

Grünwald, P. (1999). Viewing all models as "probabilistic". In *Proceedings of the twelfth annual conference on Computational learning theory* (pp. 171–182).

Gumbel, E. J. (1958). *Statistics of extremes*. Columbia University Press.
[https://doi.org/10.7312/gumb92958]

Haan, L., & Ferreira, A. (2006). *Extreme value theory: An introduction* (Vol. 3). Springer.

Hacking, I. (1975). *The emergence of probability: A philosophical study of early ideas about probability, induction and statistical inference*. Cambridge University Press.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. Springer.
[https://doi.org/10.1007/978-0-387-84858-7]

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika,57*(1), 97–109.
[MathSciNet]

Hoff, P. D., Raftery, A. E., & Handcock, M. S. (2002). Latent space approaches to social network analysis. *Journal of the American Statistical Association,97*(460), 1090–1098.
[MathSciNet]

Hoffman, M. D., Blei, D. M., Wang, C., & Paisley, J. (2013). Stochastic variational inference. *Journal of Machine Learning Research, 14*, 1303–1347.
[MathSciNet]

Hora, S. C. (1996). Aleatory and epistemic uncertainty in probability elicitation with an example from hazardous waste management. *Reliability Engineering & System Safety,54*(2), 217–223. Treatment of Aleatory and Epistemic Uncertainty.

Huber, P. J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics,35*(1), 73–101.
[MathSciNet][https://doi.org/10.1214/aoms/1177703732]

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, Berkeley, CA, 1965/66 (Vol. I: Statistics, pp. 221–233). University California Press.

Hüllermeier, E., & Waegeman, W. (2021). Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning,110*(3), 457–506.
[MathSciNet][https://doi.org/10.1007/s10994-021-05946-3]

Hunter, D. R., Handcock, M. S., Butts, C. T., Goodreau, S. M., & Morris, M. (2008). Ergm: A package to fit, simulate and diagnose exponential-family models for networks. *Journal of Statistical Software,24*, 1–29.

Hurvich, C. M., & Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika,76*(2), 297–307.
[MathSciNet]

Kauermann, G., Haupt, H., & Kaufmann, N. (2012). A hitchhiker's view on spatial statistics and spatial econometrics for lattice data. *Statistical Modelling,12*(5), 419–440.
[MathSciNet]

Kauermann, G., Krivobokova, T., & Fahrmeir, L. (2009). Some asymptotic results on generalized penalized spline smoothing. *Journal of the Royal Statistical Society Series B: Statistical Methodology,71*(2), 487–503.
[MathSciNet]

Kauermann, G., Küchenhoff, H., & Heumann, C. (2021). *Statistical foundations, reasoning and inference: For science and data science*. Springer International Publishing.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In *Advances in neural information processing systems,30*.

Kingma, D. P., & Welling, M. (2014). Auto-encoding variational Bayes. In Bengio & Lecun (Eds.), *2nd international conference on learning representations*.

Kiureghian, A. D., & Ditlevsen, O. (2009). Aleatory or epistemic? Does it matter? *Structural Safety,31*(2), 105–112.
[https://doi.org/10.1016/j.strusafe.2008.06.020]

Kokoszka, P., & Reimherr, M. (2017). *Introduction to functional data analysis*. Chapman and Hall/CRC.

Kolaczyk, E. D. (2009). *Statistical analysis of network data*. Springer.

Kolb, C., Müller, C. L., Bischl, B., & Rügamer, D. (2023). Smoothing the edges: A general framework for smooth optimization in sparse regularization using hadamard overparametrization.

Kolmogorov, A. N. (1933). *Grundbegriffe Der Wahrscheinlichkeitsrechnung*. Springer.

Kook, L., Kolb, C., Schiele, P., Dold, D., Arpogaus, M., Fritz, C., Baumann, P. F. M., Kopper, P., Pielok, T., Dorigatti, E., & Rügamer, D. (2024). How inverse conditional flows can serve as a substitute for distributional regression. In *The 40th conference on uncertainty in artificial intelligence*.

Krige, D. G. (1951). A statistical approach to some basic mine valuation problems on the witwatersrand. *Journal of the Southern African Institute of Mining and Metallurgy,52*(6), 119–139.

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., & Blei, D. M. (2017). Automatic differentiation variational inference. *Journal of Machine Learning Research*, *18*, 1–45.
[MathSciNet]

Lang, J. B. (1996). On the comparison of multinomial and poisson log-linear models. *Journal of the Royal Statistical Society: Series B (Methodological),58*(1), 253–266.
[MathSciNet][https://doi.org/10.1111/j.2517-6161.1996.tb02079.x]

Lauritzen, S. L. (1996). *Graphical models* (Vol. 17). Clarendon Press.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature,521*(7553), 436–444.
[https://doi.org/10.1038/nature14539]

Lehmer, D. H. (1951). Mathematical methods in large-scale computing units. *Annals of the Computation Laboratory of Harvard University,26*, 141–146.
[MathSciNet]

Li, X., & Meng, X.-L. (2021). A multi-resolution theory for approximating infinite-p-zero-n: Transitional inference, individualized predictions, and a world without bias-variance tradeoff. *Journal of the American Statistical Association,116*(533), 353–367.
[MathSciNet][https://doi.org/10.1080/01621459.2020.1844210]

Lindsay, B. G., Kettenring, J., & Siegmund, D. O. (2004). A report on the future of statistics.
[https://doi.org/10.1214/088342304000000404]

Lusher, D., Koskinen, J., & Robins, G. (2013). *Exponential random graph models for social networks*. Cambridge University Press.

Maathuis, M., Drton, M., Lauritzen, S., & Wainwright, M. (2018). *Handbook of graphical models*. CRC Press.
[https://doi.org/10.1201/9780429463976]

Marin, J.-M., Pudlo, P., Robert, C. P., & Ryder, R. J. (2012). Approximate bayesian computational methods. *Statistics and Computing,22*(6), 1167–1180.
[MathSciNet]

McCullagh, P. (2018). *Tensor methods in statistics*. Courier Dover Publications.

Mendel, G. (1865). Versuche über Pflanzenhybriden. *Verhandlungen des Naturforschenden Vereins Brünn* (4), 3–47.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & E. Teller (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics,21*(6), 1087–1092.

Morris, J. S., & Carroll, R. J. (2006). Wavelet-based functional mixed models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology),68*(2), 179–199.
[MathSciNet]

Müller, S., Reuter, A., Hollmann, N., Rügamer, D., & Hutter, F. (2025). Position: The future of Bayesian prediction is prior-fitted. In *Forty-second international conference on machine learning position paper track*.

Murphy, K. P. (2022). *Probabilistic machine learning: An introduction*. MIT Press.

Neal, P., & Roberts, G. (2006). Optimal scaling for partially updating MCMC algorithms. *The Annals of Applied Probability,16*(2), 475–515.
[MathSciNet]

Nelder, J. A., & Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society Series A: Statistics in Society,135*(3), 370–384.
[https://doi.org/10.2307/2344614]

Nepusz, T., Petróczi, A., Négyessy, L., & Bazsó F. (2008). Fuzzy communities and the concept of bridgeness in complex networks. *Physical Review E,77*(1), 1–12.
[MathSciNet]

Neyman, J., & Pearson, E. S. (1933). Ix. on the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character,231*(694–706), 289–337.

Pearl, J. (2009). *Causality*. Cambridge University Press.
[https://doi.org/10.1017/CBO9780511803161]

Pearson, K. (1904). On the theory of contingency and its relation to association and normal correlation. In *Mathematical contributions to the theory of evolution XIII*. Dulau and Co.

Pearson, K. F. (1900). X. 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. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science,50*(302), 157–175.

Peters, J., Janzing, D., & Schölkopf, B. (2017). *Elements of causal inference: Foundations and*

learning algorithms. MIT Press.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. In *Advances in neural information processing systems,31*.

Ramsay, J. O., & Silverman, B. W. (2005). *Functional data analysis* (Springer series in statistics). Springer.

Rao, C. R. (1945). Information and the accuracy attainable in the estimation of statistical parameters. *Bulletin of the Calcutta Mathematical Society,37*, 81–89.
[MathSciNet]

Ravishanker, N., Chi, Z., & Dey, D. K. (2021). *A first course in linear model theory*. Chapman and Hall/CRC.
[https://doi.org/10.1201/9781315156651]

Rendtel, U., Wasmuht, U. C., & Wilrich, P.-T. (2021). Emil Julius Gumbel: An innovative statistician and a dedicated pacifist. *AStA Wirtschafts-und Sozialstatistisches Archiv,15*, 273–291.
[https://doi.org/10.1007/s11943-021-00293-0]

Riabiz, M., Chen, W. Y., Cockayne, J., Swietach, P., Niederer, S. A., Mackey, L., & Oates, C. J. (2022). Optimal thinning of MCMC output. *Journal of the Royal Statistical Society Series B: Statistical Methodology, 84*(4), 1059–1081.
[MathSciNet]

Ripley, B. D. (2005). *Spatial statistics*. Wiley.

Robbins, H., & Monro, S. (1951). A stochastic approximation method. *The Annals of Mathematical Statistics, 22*, 400–407.
[MathSciNet][https://doi.org/10.1214/aoms/1177729586]

Robert, C., & Casella, G. (2004). *Monte Carlo statistical methods*. Springer.

Ross, S. M. (2022). *Simulation*. Academic.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics, 12*, 1151–1172.
[MathSciNet]

Rügamer, D., & Greven, S. (2018). Selective inference after likelihood-or test-based model selection in linear models. *Statistics & Probability Letters,140*, 7–12.
[MathSciNet]

Rügamer, D., Kolb, C., & Klein, N. (2024). Semi-structured distributional regression. *The American Statistician,78*(1), 88–99.
[MathSciNet][https://doi.org/10.1080/00031305.2022.2164054]

Scheipl, F., Kneib, T., & Fahrmeir, L. (2013). Penalized likelihood and bayesian function selection in regression models. *AStA Advances in Statistical Analysis,97*(4), 349–385.
[MathSciNet]

Scheipl, F., Staicu, A.-M., & Greven, S. (2015). Functional additive mixed models. *Journal of Computational and Graphical Statistics,24*(2), 477–501.
[MathSciNet]

Schweinberger, M. (2011). Instability, sensitivity, and degeneracy of discrete exponential families. *Journal of the American Statistical Association,106*, 1361–1370.
[MathSciNet]

Shalizi, C. R., & Thomas, A. C. (2011). Homophily and contagion are generically confounded in observational social network studies. *Sociological Methods & Research,40*(2), 211–239.
[MathSciNet]

Shumway, R. H., & Stoffer, D. S. (2017). *Time series analysis and its applications: with R examples* (4th ed.). Springer.

Sischka, B., & Kauermann, G. (2022). Em-based smooth graphon estimation using mcmc and spline-based approaches. *Social Networks,68*, 279–295.

Snijders, T. A., & Nowicki, K. (1997). Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *Journal of Classification,14*(1), 75–100.
[MathSciNet]

Somerville, M. C., & Brown, R. S. (2013). Exact likelihood ratio and score confidence intervals for the binomial proportion. *Pharmaceutical Statistics,12*(3), 120–128.
[https://doi.org/10.1002/pst.1560]

Spiegelhalter, D. (2024). Does probability exist? *Nature,636*, 19.
[https://doi.org/10.1038/d41586-024-03807-2]

Stein, A., & Wikle, C. K. (2024). Space, uncertainty, and the environment: Honoring the distinguished career of noel cressie. *Spatial Statistics,61*, 100835.
[MathSciNet]

Stigler, S. M. (1975). Studies in the history of probability and statistics. xxxiv: Napoleonic statistics: The work of laplace. *Biometrika,62*(2), 503–517.
[MathSciNet]

Stigler, S. M. (1980). Stigler's law of eponymy. *Transactions of the New York Academy of Sciences,39*(1), 147–157.
[https://doi.org/10.1111/j.2164-0947.1980.tb02775.x]

Stigler, S. M. (2018). Richard price, the first Bayesian. *Statistical Science,33*(1), 117–125.
[MathSciNet]

Storey, J. D. (2011). False discovery rate. *International Encyclopedia of Statistical Science,1*, 504–508.

Student (1908). The probable error of a mean. *Biometrika,6*(1), 1–25.
[https://doi.org/10.2307/2331554]

Suess, E. A., & Trumbo, B. E. (2010). *Introduction to probability simulation and Gibbs sampling with R* (Vol. 307). Springer.

Tager, I. B., Weiss, S. T., Rosner, B., & Speizer, F. E. (1979). Effect of parental cigarette smoking on the pulmonary function of children. *American Journal of Epidemiology,110*(1), 15–26.

Taleb, N. N. (2007). *The Black Swan : The impact of the highly improbable*. Random House.

Tavaré, S., Balding, D. J., Griffiths, R. C., & Donnelly, P. (1997). Inferring coalescence times from dna sequence data. *Genetics,145*(2), 505–518.

Thompson, S. K. (2010). *The unfinished game: Pascal, fermat, and the seventeenth-century letter that made the world modern*. Basic Books.

Thomson, W. E. (1958). A modified congruence method of generating pseudo-random numbers. *The Computer Journal,1*(2), 83–83.

Thulin, M. (2014). The cost of using exact confidence intervals for a binomial proportion. *Electronic Journal of Statistics,8*(1), 817–840. [MathSciNet][https://doi.org/10.1214/14-EJS909]

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology,58*(1), 267–288. [MathSciNet][https://doi.org/10.1111/j.2517-6161.1996.tb02080.x]

Tibshirani, R. J., & Efron, B. (1993). *An introduction to the bootstrap* (Monographs on statistics and applied probability, Vol. 57, pp. 1–436). Chapman & Hall/CRC.

Tsay, R. S. (2000). Time series and forecasting: Brief history and future research. *Journal of the American Statistical Association,95*(450), 638–643.

Tutz, G. (2011). *Regression for categorical data* (Vol. 34). Cambridge University Press.

van de Schoot, R., Depaoli, S., King, R., Kramer, B., Märtens, K., Tadesse, M. G., Vannucci, M., Gelman, A., Veen, D., Willemsen, J., et al. (2021). Bayesian statistics and modelling. *Nature Reviews Methods Primers,1*(1), 1.

van den Hout, A., & van der Heijden P. G. (2002). Randomized response, statistical disclosure control and misclassificatio: A review. *International Statistical Review,70*(2), 269–288.

Vovk, V., Gammerman, A., & Saunders, C. (1999). Machine-learning applications of algorithmic randomness.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transaction of American Mathematical Society,54*, 426–482. [MathSciNet]

Wang, Q., Ma, Y., Zhao, K., & Tian, Y. (2022). A comprehensive survey of loss functions in machine learning. *Annals of Data Science,9*(2), 187–212.

Wasserman, L., Ramdas, A., & Balakrishnan, S. (2020). Universal inference. *Proceedings of the National Academy of Sciences,117*(29), 16880–16890.
[MathSciNet]

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica,48*, 817–838.
[MathSciNet][https://doi.org/10.2307/1912934]

Whittaker, J. (2009). *Graphical models in applied multivariate statistics*. Wiley Publishing.

Whittle, P. (1951). *Hypothesis testing in time series analysis* (Vol. 4). Almqvist & Wiksells boktr.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin,1*(6), 80–83.

Wilding, J. P. H., Batterham, R. L., Calanna, S., Davies, M., Gaal, L. F. V., Lingvay, I., McGowan, B. M., Rosenstock, J., Tran, M. T. D., Wadden, T. A., Wharton, S., Yokote, K., Zeuthen, N., & Kushner, R. F. (2021). Once-weekly semaglutide in adults with overweight or obesity. *New England Journal of Medicine,384*(11), 989–1002.

Wood, S. N. (2017). *Generalized additive models: An introduction with R*. CRC Press.

Yao, F., Müller, H.-G., & Wang, J.-L. (2005). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association,100*(470), 577–590.
[MathSciNet]

Ziliak, S. T. (2008). Retrospectives: Guinnessometrics: the economic foundation of "student's" t. *Journal of Economic Perspectives,22*(4), 199–216.
[https://doi.org/10.1257/jep.22.4.199]

Index

A

Accuracy 110
Akaike information criterion (AIC) 138
Aleatoric uncertainty 7, 10
Autocorrelation function 256
Autocovariance function 256

B

Bayesian inference 42
Bayesian information criterion (BIC) 139, 144
Bernstein-von Mises theorem 197
Bias 28
Bias-variance-trade-off 28
Bonferroni adjustment 104
Bootstrapping 181

C

Cauchy–Schwarz inequality 322
Central limit theorem 4
Chebyshev inequality 323
Chi-squared statistic 90
Concentration matrix 246
Confidence interval 36
Consistency 27
Cramér-Rao bound 59
Credibility interval 43
Cross-validation 135

D

Data-generating process 14
Directed acyclic graphs (DAG) 243

E

Efficiency 35
Empirical Bayes 198, 199
Entropy 21
Epistemic uncertainty 7, 10
Estimator 16
Exact confidence interval 218
Extreme value distribution 222

F

False discovery rate 110
Family-wise error rate 103
Fisher information 50, 51
Fisher-Neyman-factorization 32
Fisher-regular 334
Forward selection 148
Frequentist inference 42
F score 110

G

Gaussian Markov Random Field (GMRF) 177
Generalized extreme value 223
Generalized extreme value distribution (GEV) 223
Gibbs sampling 171
Glivenko-Cantelli theorem 4
Graphical model 237

H

Hammersley Clifford theorem 175
Highest density interval 45
Hyperparameter 194
Hypothesis test 83

I

i.i.d. 14
Importance sampling 163

K

Karhunen-Loève expansion 287
k-fold cross-validation 137
Kolmogorov-Smirnov test 94
Kriging 276
Kullback-Leibler divergence 21

L

Law of large numbers 5
Likelihood function 25
Likelihood principle 69
Likelihood ratio 62
Log-linear model 231

M

Markov Chain Monte Carlo (MCMC) 165
Markov fields 175
Markov inequality 322
Maximum likelihood 52
Maximum likelihood estimator 25
Mean squared error 28
Metropolis-Hastings 164

N

Nested models 132
Neyman-Pearson test 99
Null hypothesis 74

P

Pairwise conditional independence 239
Partial autocorrelation 269
Pivotal statistic 36
Plug-in principle 183
Posterior distribution 194
Posterior mode 43
Power of test 96

Prevalence 110
Prior distribution 194
p-value 75

R

Rejection sampling 161
Residual 114
Risk 296

S

Score function 49, 51
Sensitivity 109
Specificity 109
Stationarity 256
Statistical model 15
Sufficiency 31
Sufficient 31

T

Training and test data 135

U

Umbrella density 161

V

Variogram 274

W

Wald test 88
Weak law of large numbers 5
Wilcoxon test 93

`flight`	Flight number
`temp`	Temperature in $^\circ$F
`td`	Thermal overstress (1 = Yes/0 = No)

Year	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990
Max inter-arrival time	11	12	10	22	16	19	14	8	12	16	9

Today	Tomorrow	S	C	R
S		0.50	0.25	0.25
C		0.20	0.40	0.40
R		0.20	0.50	0.30

Table 11.2 Munich rental guide survey, tabulated with respect to new contract and district

New	**District**						
Contract	**Allach**	**Altstadt**	**Au-Haidhausen**	**Aubing**	**Berg am Laim**	**Bogenhausen**	···
No	17	38	131	18	50	111	···
Yes	12	20	75	36	33	73	···

	Estimate	**Std. error**	***p*-Value**
$\widehat{\beta}_0$ **(intercept)**	-7.05	32.97	0.83
$\widehat{\beta}_1$ **(height)**	0.42	0.19	0.04
$\widehat{\beta}_2$ **(age)**	-0.06	0.45	0.89
$\widehat{\beta}_3$ **(gender)**	13.69	4.06	< 0.01

Variable	Description
Gamble	Money spent on gambling in pounds per year
Status	Socioeconomic status (based on the job of the parents)
Income	Income (in pounds per week)
Verbal	Test score based on the number of properly defined words (max. 12)
Sex	Sex (0 = male, 1 = female)

Table 5.1 Mendel's data

Outcome				
number	**Characteristics**	**Assumed proportions**	**Assumed numbers**	**Observed numbers**
1	Round/yellow	9/16	312.75	315
2	Wizened/yellow	3/16	104.25	101
3	Round/green	3/16	104.25	108
4	Wizened/green	1/16	34.75	32

Statistic	Coefficient	Std. Error
Edges	−4.20	(0.25)
Nodematch(Area)	+2.10	(0.30)
Transitive Ties	+1.50	(0.20)

Table 12.1 Results for the UK Faculty network of the ERGM models with (left column) and without (right column) dependence-inducing terms

	Coefficient	**Coefficient**
Statistic	**(Standard error)**	**(Standard error)**
	Dependence	**Independence**
Edges	-4.70	-2.84
	$(.44)$	$(.09)$
Nodematch affiliation	2.58	2.57
	$(.11)$	$(.11)$
Transitive ties	1.75	
	$(.42)$	
AIC	2343.31	2378.43
BIC	2361.56	2390.60
Log likelihood	-1168.66	-1187.22

Table 10.1 Parameter estimates for the two time periods

	Before 1985	**After 1985**
μ	22.88 (0.30)	24.68 (0.32)
σ	1.57 (0.20)	1.78 (0.22)
γ	− 0.19 (0.10)	− 0.37 (0.09)

Table 6.2 Coefficients, standard errors, and *p*-values for the regression model in Example 6.4

	Estimate	**Std. error**	***p*-Value**
$\hat{\beta}_0$ **(intercept)**	0.642	0.173	< 0.01
$\hat{\beta}_1$ **(review_score)**	0.473	0.034	< 0.01
$\hat{\beta}_2$ **(price)**	0.001	0.002	0.69
$\hat{\beta}_3$ **(promotion)**	0.166	0.037	< 0.01

Table 5.3 Error matrix for multiple testing

	Nonrejected hypotheses	**Rejected hypotheses**	
	i.e., "H_{0j}"	**i.e.,** "H_1"	
True hypotheses H_{0j}	U	V	M_0
False hypotheses, i.e., H_{1j}	T	S	M_1
	$m - r$	r	m

		Z_2				
Z_1		**1**	**2**	**...**	K_2	
	1	n_{11}	n_{12}		n_{1K_2}	$n_{1\cdot}$
	2	n_{21}	n_{22}		n_{2K_2}	$n_{2\cdot}$
	⋮					⋮
	K_1	n_{K_11}	n_{K_12}		$n_{K_1K_2}$	$n_{K_1\cdot}$
		$n_{\cdot 1}$	$n_{\cdot 2}$	...	$n_{\cdot K_2}$	$n_{\cdot\cdot}$

Table 9.1 Examples for conjugate distributions

Parameter	$\mathscr{F}$	$\mathscr{P}$
π	Binomial distribution	Beta distribution
λ	Poisson distribution	Gamma distribution
μ	Normal distribution	Normal distribution
λ	Exponential distribution	Gamma distribution

Table 8.1 Error matrices for two classifiers. Classifier 1 on the left and classifier 2 on the right, applied to the same test data

	Classifier 1				**Classifier 2**		
true	**decision**			**true**	**decision**		
class	**“cat”**	**“dog”**		**class**	**“cat”**	**“dog”**	
cat	235	57	292	cat	250	42	292
dog	82	626	708	dog	78	630	708
	317	683	1000		328	672	1000

Table 11.1 Munich rental guide survey, tabulated with respect to floor space and center

Floor	**Center**	
Space	**No**	**Yes**
< **50**	228	227
50–100	1455	803
> **100**	207	205

Table 6.1 Coefficients, standard errors, and *p*-values for the regression model in Example 6.3

	Estimate	**Std. error**	***p*-Value**
$\widehat{\beta}_0$ **(intercept)**	-5.53	1.67	< 0.01
$\widehat{\beta}_1$ **(age)**	0.06	0.02	< 0.01
$\widehat{\beta}_2$ **(gender)**	1.45	0.85	0.09

Sales	Score	Price	Promotion
19	4.1	27€	No
6	3.8	43€	No
29	4.6	85€	Yes
21	4.0	60€	No
...	...	...	...

Table 5.2 Error matrix in significance tests

	Decision	
Truth	"H_0"	"H_1"
H_0	Correct	Type I
	decision	error
		(α error)
H_1	Type II	Correct
	error	decision
	(β error)	

Firm	α	β	γ	δ	ϵ	ζ	η	θ	ι	κ	λ	μ	ν	ξ	o	π	ρ	σ	τ	υ
Breaches	0	0	0	0	0	1	0	0	0	0	0	0	0	2	0	0	0	1	0	0