

Vorlesungsunterlagen zu Inferenzstatistik 1
und 2

Statistics - Coping with Uncertainty

A modern Introduction to Statistics in times
of Machine Learning

October 15, 2024

Copyright by Göran Kauermann

Contents

1	Introduction	3
1.1	Statistics and Uncertainty	3
1.2	Notation and Technical Requirements	4
2	Uncertainty	7
2.1	What is Uncertainty	7
2.2	Sources of Uncertainty	9
2.3	Classical Statistical Paradigm	12
2.4	Data Generating Process	13
2.5	Statistical Modelling	14
2.6	Exercises	17
3	Learning and Estimating	19
3.1	Kullback-Leibler Divergence	19
3.2	Squared Loss and Likelihood	26
3.3	Likelihood and Bayes	26
3.4	Properties of Estimates	28
3.4.1	Consistency	28
3.4.2	Sufficiency	32
3.4.3	Efficiency	34
3.5	Estimation and Uncertainty	36
3.6	Exercises	47
4	Maximum Likelihood Estimation	51
4.1	Score Equation	51
4.2	Asymptotic Normality	57
4.3	Cramer Rao Bound	61
4.4	Likelihood Ratio	62
4.5	Parameter Transformation	64
4.6	Misspecified Models	66
4.7	Numerical Calculation of the Estimate	68

4.8	The Likelihood Principle	70
4.9	Exercises	72
5	Data Driven Decisions	75
5.1	Significance Tests and p-value	75
5.2	Statistical Hypothesis Tests	84
5.3	Classical Statistical Parameter Tests	87
5.3.1	Likelihood Based Tests	87
5.3.2	Chi-squared-tests	89
5.3.3	Distribution-Free Tests	93
5.3.4	Kolmogorov-Smirnov Test	94
5.4	Power of a Test	96
5.5	Neyman-Pearson Lemma	100
5.6	Testing and Confidence Intervals	102
5.7	Multiple Testing	103
5.8	Universal Inference	106
5.9	Sensitivity, Specificity, Accuracy	110
5.10	Exercises	112
6	Regression	115
6.1	Simple Linear Model	115
6.2	Multiple Linear Regression	119
6.3	Logistic Regression	122
6.4	Generalized Linear Models	126
6.5	Exercises	129
7	Model Selection	133
7.1	Modeling Dependence	133
7.2	Training and Test Data	136
7.3	AIC and BIC	140
7.4	AIC, BIC and Hypothesis Testing	147
7.5	Data Size and Dimension of the Model	149
7.6	Exercises	155
8	Simulation Based Inference	157
8.1	Simulating Uniformly Distributed Random Variables	157
8.2	Simulating from a Distribution Function	161
8.3	Simulating Proportional to a Density (MCMC)	165
8.4	Simulating from a Multivariate Normal Distribution	168
8.5	Gibbs Sampling and Markov Fields	171
8.6	Uncertain Input Parameters	178
8.7	Bootstrapping	181
8.7.1	The Plug-In Principle	181
8.7.2	Bootstrapping in Prediction Models	185
8.8	Permutation	187
8.9	Exercises	189

Contents	vii
9 Bayesian Inference	
9.1 The Bayesian Principle	193
9.2 Prior and Posterior	196
9.3 Hyperparameters and Empirical Bayes	200
9.4 Inference based on MCMC	201
9.5 Approximate Bayes Computation (ABC)	207
9.6 Variational Bayes Reasoning	208
9.7 Bayes and Regularization	212
9.8 Exercises	214
10 Inference in Extreme Data	
10.1 Rare Events	219
10.2 Extreme Data	223
10.3 Exercises	231
11 Multivariate Data	
11.1 Models for contingency tables	233
11.2 Graphical Models	241
11.3 Mutivariate Normal Distribution	249
11.4 Exercises	251
12 Non i.i.d. Data	
12.1 Time Series	255
12.1.1 ARMA Models	261
12.1.2 Autocorrelation and Partial Autocorrelation	272
12.1.3 Forecasting	276
12.2 Multilevel Data and Repeated Measurements	276
12.3 Spatial Data	276
12.4 Network Data	284
12.4.1 Representation and Description	287
12.4.2 Models	288
12.5 Functional Data	289
12.5.1 Analysis and Properties of Functional Data	289
12.5.2 Functional Regression	292
12.6 Exercises	294
13 Latent and Missing Data	
13.1 Types of Missingness	297
13.2 Mixture Models	297
13.3 Expectation Maximization	298
13.4 Complexity through Latent Structures	298
13.5 Exercises	298

14 Relating Machine Learning and Statistics	299
14.1 Machine Learning from a Statistics Perspective	299
14.1.1 Empirical Risk Minimization	300
14.2 Regularization, Penalization and Bayes	301
14.2.1 Regularized Risk Minimization	301
14.2.2 Regularization from a Statistical Perspective	302
14.2.3 Bayesian Approaches	304
14.3 Uncertainty in Machine Learning	305
14.3.1 Prediction Intervals	305
14.3.2 Parameter Uncertainty	306
14.3.3 Distributional Learning	306
14.4 Exercises	306
15 TAKE OUTS	307
15.1 Uncertainty and Prediction	307
15.2 Squared Loss	308
15.3 overparameterized models	309
15.3.1 Overfitting and Generalization	310
15.3.2 Overparameterized Models and Implicit Regularization	310
15.3.3 ABC Extension	313
A Additional results, notions and technical details	317
A.1 Background: Measurement- and Probability Theory	317
A.1.1 Subsection Heading	320
A.2 Notation and Background	321
A.3 Common Distributions	321
A.4 Common Distributions	323
A.5 The Exponential Family	324
A.6 Limiting Distributions and Inequalities	326
A.6.1 Chauchy-Scharz Inequality	326
A.6.2 Markov Inequality	327
A.7 Fisher regular	327
A.8 Exercises	328
References	331
Index	339

Consistency in Notation

- Use " and " as opposed to " and " or ' and ';
- Whenever we are talking about an expected value or variance, check if its clear what the random variable is by capitalizing it;

Chapter 1

Introduction

1.1 Statistics and Uncertainty

Predicting any future event, be it the weather for tomorrow, the safety assessment of a technical system, or the outcome of rolling a die, is invariably subject to uncertainty. The outcome to be predicted is not deterministic, as we are not 100% certain as of the outcome of physical law. If you hold a stone in your hand and let it drop, it will fall down with 100% probability due to the force of gravity. In vacuum conditions, the same would hold for a light feather. However, in non-laboratory conditions, there might be wind as well and the feather will not fall down but fly away. This means though physical laws determine the world, in many situations, additional factors beyond our control influence the setting, and we are faced with uncertainty. In this context, uncertainty signifies that the relationships are too complex to be understood or even described with mathematical or physical models.

With 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 to Natural Language Processing (NLP) utilizing tools like Machine Learning and/or 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. However, this would be a wrong step and a fallacy; because quantifying uncertainty remains as relevant as ever. We will never (or at least not in a foreseeable time) be able to understand all complex structures in the world, which leaves us with uncertainty.

Uncertainty, on the one hand, and ignorance or lack of knowledge, on the other hand, are closely tied together. This point, moreover, demonstrates that uncertainty has a very individual level. Experts know more than laypeople. Hence the question pops up how to quantify knowledge, or reversely, 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.

In other words, 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 but to accept that data is inherently insufficient and can not 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. We will thereby utilize probability theory and its calculus, demonstrating 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 century of statistical developments can 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 follows: "Statistics is 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 can not formulate more elegantly what we consider as the field of statistics and with respect to the mentioned interdisciplinarity, in this book we primarily focus on links to the field of machine learning.

1.2 Notation and Technical Requirements

Before we get into the material, we need to clarify our notation in this book. Uncertainty is apparently closely related to probability theory and we expect some ground knowledge in this field. We provide an Appendix, which lists the main requirements, such as frequently used distribution models and central results from the field of probability. The inexperienced reader is advised to familiarize himself or herself with this material first. We emphasize that this book aims to remain on a technical low level, but still, the one or other notation or result is essential. We therefore list in this section the core results, which also allow to get familiar with the notation.

We define with capital letters the distribution function of a random variable Y , that is

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

where we use the conventional notation that random variables are written 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. The density function $f(\cdot)$ is a positive function, such that

$$P(Y \in [a, b]) = \int_a^b f(y) dy,$$

The density results by $\partial F(y)/\partial y = f(y)$. 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 defining the mean value of a random variable. If Y is continuous, this is obtained by

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

If Y is discrete valued with outcomes $infy < a_1 < a_2 < a_3, \dots$ we get the expected value through

$$E(Y) = \sum_k a_k P(Y = a_k)$$

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

$$E(Y) = \int_{-\infty}^{\infty} y dF(y)$$

We apologize for our notation sloppiness which the mathematically trained reader might not find satisfactory. However, we do not want to get distracted from our main topic, which is statistics. For this reason, we content ourselves with being a little superficial with the notation, wherever suitable.

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, Chi-squared distribution, binomial distribution or Poisson distribution, to name but a few. We give details on these distributions in the Appendix.

Technically, 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 lemma of Glevenko Cantelli. We list these results below:

TO DO

Chapter 2

Uncertainty

This chapter opens the door to a statistical view on data. We will start with defining different sources of uncertainty, which we label as 'aleatoric' and 'epistemic'. We will demonstrate that 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. All we do in any type of data analysis is to suitably approximate this unknown process through 'statistical models'. We will also refer to the classical statistical paradigm which builds on drawing a sample from a population in order to obtain information about a quantity of interest.

2.1 What is Uncertainty

Before we aim to cope with uncertainty, we need to define it. We will see that this is not easy at all. To motivate this we start with a simple thought experiment to prompt 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 upwards is clearly $1/6$. The throw of the die is called a random experiment, i.e. the outcome is random. In this case, a probability model (such as a binomial, multinomial or geometric distribution) 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).

But let's stay with the simple experiment of one die in the cup and modify the experiment 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 is the probability 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 to quantify the probability that the die under the cup - invisible to us - has the number 6 upwards is 1/6, although this is not the probability of a random experiment. This thought experiment reveals two important aspects. First, uncertainty can be represented in the form of probabilities - not always but often. And secondly, a concrete random experiment is not always necessary to quantify uncertainty. We will utilize probability theory to quantify uncertainty even in situations, where no formal random experiment is carried out. Generally, we will make use of probability models which will allow us to quantify knowledge and reversely remaining uncertainty. This view of representing uncertainty through probability distributions is also reflected in Bayesian statistics, which is treated later in the book. But for now we do not conceptually distinguish between the Bayesian and, what is later called, the frequentist's point of view.

We may look at the die experiment also 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 literally a physical process. If we repeat shaking the cup with exactly the same angle, velocity and acceleration etc., the result would be the same assuming that the outcome would be deterministically determined by its physical input parameters. But 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. Hence, if we would equip the dice cup with hundreds of sensors that record the movement of the cup and if we maybe additionally install video cameras that record the person shaking the dice cup, we could in principle circumvent randomness and predict the side the die lands on. Apparently, this is a very complex physical setup, but we could employ ideas of machine learning, build a robot shaking the dice cup randomly, train a machine learner to the so collected data - this would be a classical supervised learning framework - and replace the random experiment with a prediction problem where we predict the number of points on the die based on sensor and other input data. But if this is the case, then the random experiment of rolling a die is no longer random but deterministic in fact. This opens the door to an even more philosophical view which questions, is there randomness at all or is everything (pre-)determined? This leads us to determinism, which does however not directly contribute to the question of quantification of uncertainty. Putting it differently, regardless of whether the world is deterministic or stochastic, our knowledge about things happening in the world or in the near future is limited. That in turn means that we are faced with uncertainty, which again needs to be quantified which we here do with probability models. With respect to the dice cup it means, unless we are not able to record and process all physical quantities which are relevant for the side the dice falls on, we are in the circumstance of uncertainty and we aim to quantify this with means 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, sometimes just vice versa. Consider the example, where we take an image of a horse with a very modern camera that can take pictures with an arbitrary resolution. We could look at a coarse version of the image, that 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 to detect the horse since we focus too much on small things. In other words, we prefer a resolution of the image, that allows for uncertainties - pixels of the image do not show every single hair - but still carry the information of what we want to see.

Quantification of uncertainty is in practice much more complicated and just giving numbers expressing uncertainty does not mean that they are useful or valid. We have all experienced situations, where the weather forecast predicted rain with 100% probability, but in fact, not a single raindrop was falling. While experiencing a dry day instead of a rainy one is for most people tolerable, it shows that even rain probabilities are approximative and not always reliable. In Figure 2.1 we show a different example demonstrating that quantifications of uncertainties can be heavily wrong. This is a famous example of what is called adversarial effects. In machine learning with complex models, it becomes rather complicated to properly quantify uncertainties. Still, based on the architecture of a neural net one can re-express the weights of the last layer as probabilities. This is shown in the example in Figure 2.1 where animal pictures are classified with respect to the animal shown on the picture. A panda is detected, not with a high probability, but with a certainty of about 58%. If now “adversarial noise” is superimposed on the image, that is noise on the image so that classification of the image goes towards a different animal without visually changing the picture, we obtain the right-hand side picture. The latter is just the sum of the first two pictures. Now the animal on the image is classified as a gibbon, and that is almost certain with over 99% probability. Apparently, this shows a pitfall that often occurs in complex neural networks. We bring the example here to motivate, that uncertainties are difficult to calculate and quantifying and coping with uncertainty is one of the major challenges in real data-driven reasoning.

2.2 Sources of Uncertainty

We have seen above, that uncertainties are multifold. In machine learning the different 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 the two. The terms “aleatory” and “epistemology” were already used in 1975 to describe different historical perspectives on the concept of probability.

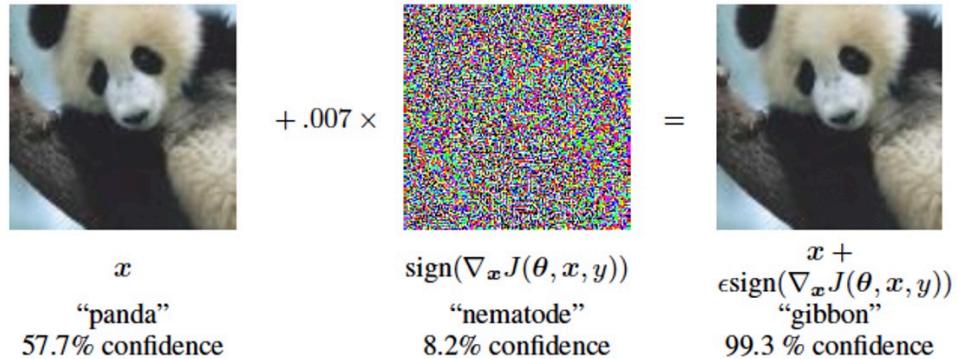


Fig. 2.1 Example for adversarial effects. Adding some “adversarial” noise on the data can lead to wrong outcomes. Source of the image: Goodfellow et al. (2015)

These terms then developed into the nowadays common terms “aleatoric” and “epistemic”. Following Hacking (1975), the aleatoric view considers probability as 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 and 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 uncertainty is thereby defined as uncertainty arising from the inherent randomness of an event, while epistemic uncertainty expresses uncertainty due to a lack of knowledge, which in principle can be diminished by increasing the amount of data or with an excessive number of experiments. We refer to Hüllermeier and Waegeman (2021) for deeper discussion. Conceptually, this decomposition is still not unique and we therefore propose here 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 then called covariates or independent variables while y is the response variable or dependent variable. But this is also the classical setup in supervised machine learning where a neural net is trained for predicting y based on some input data x . The input data itself can be of complex and/or unstructured form, like images or texts. Aleatoric uncertainty is then uncertainty originating from a stochastic/non-deterministic 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

$$Y \sim G(\cdot|x) \quad (2.1)$$

where $G(\cdot|x)$ is some random distribution function which depends on x . The structure of the dependence 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 can not predict y from x

without error. Non-degeneracy means here that the distribution is not concentrated to a single value y_0 . Simply speaking and assuming that the variance exists, aleatoric uncertainty refers to a strictly positive variance $\text{Var}(Y|x)$. The probability model in equation (2.1) serves as definition for aleatoric uncertainty.

With this definition of aleatoric uncertainty we can then also define epistemic uncertainty as all remaining uncertainty which is not aleatoric. In practice this is uncertainty due the fact that we typically do not know $G(\cdot|x)$ in (2.1) and replace it with some numerical model, be it a regression model or a neural network. But epistemic uncertainty also refers to the fact that we have only limited amount of data. Putting it differently, if an infinite amount of data would be available, our prediction models could get arbitrarily complex and data uncertainty vanishes, meaning that epistemic uncertainty decreases to zero. But aleatoric uncertainty remains. We remark that the setup is even more complex, since we also have what is called estimation uncertainty. This means, that whatever model we train or estimate on data, there will be uncertainty due to the finiteness of data. We will go deeper into this later in the book, but refer already at this point to Li and Meng (2021) for a general discussion of this point. Apparently, this is at the time of writing this book a field of active research.

We want to demonstrate the two sources of uncertainties 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 also 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 that are 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 e.g. Bochenek and Ustrnul (2022). For the dice cup this holds as well. If we are at some future point of 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. But for now, dice cups are simple non-technical objects and shaking it is a non-deterministic event so that we are faced with aleatoric uncertainty.

Finally, the terms aleatoric and epistemic uncertainty are not very common in the statistical literature yet. But it does quite easily translate to the statistical jargon. Aleatoric uncertainty is the true underlying data-generating process, the nature that determines y from x with the additional randomness superimposed. If y deterministically follows from x and we are able to observe all x , then there is no uncertainty. Often this will however not be the case in real life. One can therefore state, that aleatoric uncertainty is irreducible and follows the probability model (2.1). Consequently, epistemic uncertainty in the statistical jargon translates to model uncertainty

and estimation uncertainty, both of which are covered excessively later in the book.

We want to add, that in some machine learning applications the aleatoric uncertainty, the irreducible one, is of secondary interest. Consider for instance the classification of images of cats and dogs into one of the two categories. It is quite reasonable to assume that this can be done unambiguously at least by humans, except of blurred images. Hence, given an image x there is no or at least only an ignorable amount of uncertainty whether the animal shown on the image is a dog or a cat. However, this does not apply to all classification tasks. For instance, the classification of medical images is clearly subject to uncertainty. These are the cases where statistical concepts can help to cope with uncertainty.

2.3 Classical Statistical Paradigm

Traditional inferential statistics intends to make a statement about a population based on data, which are typically drawn through a sample. Figure 2.2 gives a sketch. We have a population of individuals for which we want to get information about some quantity. For simplicity, let us consider age as quantity of interest and we are interested in the average age of the population. We draw a sample, that is we randomly select some individuals from the population, which is typically done without replacement so that each individual is drawn only once. Based on the sample we calculate the quantity of interest, hence the average age of the individuals in the sample. The next step is to assess uncertainty. The average age in the sample is apparently 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. Typically, however, through the sampling process, uncertainty about the average age of the population is inevitable. The sample gives us information on both, the average age but also on the heterogeneity in the population, where the latter can be measured through 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 introduce later in the book. For now, consider the confidence interval as an interval in which the true but unknown average age of the population lies with a given degree of certainty. Aleatoric uncertainty in this example comes through heterogeneity of age in the population. But given that we draw without replacement, we can in fact reduce the remaining uncertainty to zero if we include all individuals in the sample. Apparently, a main requirement in this case is that the population is finite.

Sampling and survey design are central areas of statistics and we refer exemplary to Thompson (2010) for more details. In our discussion in this book survey methods are not centrally focused. In other words, we aim to deal with situations where data do not trace from a classical survey sample as sketched in Figure 2.2. Still, we can utilize the main ideas derived but need to extend these to infinite populations, or

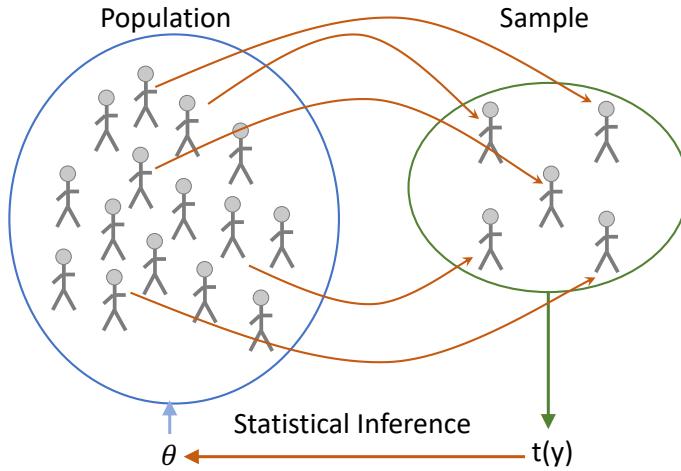


Fig. 2.2 Classical statistical paradigm. A sample is drawn from a population. The sample statistics allow us to draw inference about a population parameter

phrasing more generally, to general data generating processes such as (2.1). The latter term needs to be defined properly which we do subsequently.

2.4 Data Generating Process

We now extend the finite population setup to a general setting. For this we assume that the data are generated by some random process. To keep it simple we ignore the input variables x for the moment and consider a simple quantity Y only. This is assumed to be drawn from some unknown distribution according to

$$Y \sim G(\cdot) \quad (2.2)$$

where $G(\cdot)$ is a distribution function, which is unknown to us. From now on we call (2.2) the **data generating process** which is considered as some natural law. Indeed, this shows similarities to the definition of aleatoric uncertainty in (2.1) ignoring the input variables x for the moment. In fact, aleatoric uncertainty can be explained by the fact that there is irreducible uncertainty which is modelled or better comprehended through a random distribution given in (2.2). This setup will be central throughout the entire book and we want to emphasize right from the start, that (2.2) is of very general structure. We neither assume that $G(\cdot)$ is known nor that it can be learned

with infinite data.

The statistical paradigm shown in Figure 2.2 now extends as follows. We first assume that distribution $G(\cdot)$ is unknown. But we are able to obtain random draws from $G(\cdot)$ which yield the data. Following the conventions of probability theory we denote the data with lower case letters y_1, \dots, y_n and assume that these are **independent and identically drawn** (in short i.i.d) from $G(\cdot)$, i.e.

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

where y_i is the realization of Y_i . Based on the data we now intend to draw inference about general properties of the unknown distribution $G(\cdot)$. These could be quantities like the mean or the variance, i.e.

$$\mu = \int y dG(y) \text{ or } \sigma^2 = \int (y - \mu)^2 dG(y),$$

but it can also be the entire distribution. From the Glevenko-Cantelli theorem ([Reference to Chapter 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 ([Reference to Chapter 2](#)) that

$$\frac{1}{n} \sum_{i=1}^n t(Y_i) \rightarrow \int t(y) dG(y)$$

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

We also emphasize that the data generating process (2.2) is simplified by omitting the input variables x , even though our main 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 to 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 Modelling

We will introduce concepts to approximate $G(\cdot)$ through a parametric distribution function $F(\cdot; \theta)$, where $\theta = (\theta_1, \dots, \theta_p)^\top$ is a parameter vector. For instance, we may assume a continuous distribution for $G(\cdot)$ which we approximate through a normal distribution, where the parameters are the mean μ and the variance σ^2 . Hence, in

this case $\theta = (\mu, \sigma^2)^\top$. Generally, we use the terminology *statistical model* which is defined as a set of distribution functions:

$$\mathcal{F} = \{F(\cdot; \theta), \theta \in \Theta\}. \quad (2.4)$$

In other words, we assume or pretend that the data are generated through an element of set \mathcal{F} for a particular value of the parameters θ_0 . The modelled data-generating process is then

$$Y \sim F(\cdot; \theta_0) \quad (2.5)$$

where θ_0 is sometimes also called the “true” parameter.

In machine learning jargon the choice of \mathcal{F} relates to choosing a loss function. The probability model in fact determines the assumed aleatoric uncertainty. We will come back to this point, but emphasize at this point already, that the choice of \mathcal{F} is apparently crucial - in fact hypothetical - and we will develop tools that assess how well the statistical model serves as an approximation of the unknown data generating process $G(\cdot)$.

Example 1 We want to motivate the idea with a simple example. Assume we have data y_i with related input variables x_i for $i = 1, \dots, n$. We assume that y_i are drawn from a normal distribution with mean μ_i and variance σ_i^2 . We assume further that μ_i is some function dependent on x_i while σ_i^2 is constant for all i . This gives the probability model

$$\mathcal{F} = \left\{ f(y, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}} \sqrt{\sigma^{-2}} \exp \left\{ \frac{(y - \mu)^2}{\sigma} \right\}; \mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}^+ \right\}.$$

Through the shape of the normal distribution, this model relates to square losses, as we will see later. Hence, machine learning based on minimizing the squared loss is comparable to choosing the set \mathcal{F} as the set of normal distributions. \triangleright

The intention is now to choose parameter θ in some optimal way. To do so, we omit the input variables and solely rely on data y_1, \dots, y_n , which we assume to be drawn from (2.3). Classical statistics textbooks are more restrictive and replace the data-generating process (2.3) by

$$Y_i \sim F(\cdot; \theta_0) \text{ i.i.d for } i = 1, \dots, n. \quad (2.6)$$

In other words, one assumes that the true data-generating process comes from \mathcal{F} . The aim is to set parameter θ data-driven to some value $\hat{\theta}$. That is $\hat{\theta}$ can be written as a function of the data y_1, \dots, y_n which could be denoted as

$$\hat{\theta} = t(y_1, \dots, y_n) \quad (2.7)$$

for $t : \mathbb{R}^n \rightarrow \Theta$ being some **statistic**, that is some function of the data. The statistic in (2.7) is called **estimator** and we will discuss the 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. This means it is more plausible to consider $t(\cdot)$ not as some kind of function but as an algorithm applied to data y_1, \dots, y_n yielding the outcome $\hat{\theta}$.

If the model (2.6) holds, then apparently we hope that $\hat{\theta}$ is close to θ_0 , the true parameter. In fact, if the data size n increases we want that $\hat{\theta} \rightarrow \theta_0$. 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 so that with an increasing data size we obtain a model which is getting closer to reality. We will propose to derive $\hat{\theta}$ by optimizing some target function. In statistics, the likelihood function has achieved the gold standard. We will introduce the function and will discuss the properties of its maximum in depth in Chapter 4.

It is apparent that the true data generating process (2.3) and its approximation (2.6) may differ. If they coincide we have $G(\cdot) \in \mathcal{F}$ and in this case we may rewrite $G(\cdot)$ as

$$G(\cdot) = F(\cdot, \theta_0). \quad (2.8)$$

As remarked above, this setup is mostly assumed in classical statistics textbooks. We remain more general here and emphasize that \mathcal{F} is just a model class to approximate reality. To see this, assume that we are interested in the height of male humans in a particular area. We assume that the height of a person is normally distributed and we aim to estimate the mean μ and variance σ^2 based on some recorded heights. It should be clear that the heights of people are certainly not normally distributed and we have no clue how the true biological height distribution looks like. Still, the normal distribution might work as a good proxy, i.e. the hypothesis that height is normally distributed is a reasonable starting point in the analysis. But if in fact (2.8) holds, we can derive additional mathematical properties for the estimates, some of which do extend to scenarios where (2.8) does not hold but the general setup (2.3) holds.

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 in 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?

Chapter 3

Learning and Estimating

3.1 Kullback-Leibler Divergence

We have defined in Chapter 2 the data generating process, which we will consider now again. Like above we assume that data have been generated through

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

and we aim to model the distribution with the statistical model

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

This scenario can be sketched as in Figure 3.1 in the upper left plot. The true data-generating process $G(\cdot)$ is visualized as a point in the set of all possible data-generating processes. The latter is modelled through a set of distributions \mathcal{F} visualized as a subset that does not contain $G(\cdot)$. The aim is to find the best approximation of the true but unknown data-generating process $G(\cdot)$. In other words, we search for the element in \mathcal{F} which has the smallest distance to $G(\cdot)$. To do so, there are now a number of questions we need to tackle and we go through them step by step. First of all, we need to define some distance measure that describes how far $G(\cdot)$ is away from a distribution $F(\cdot; \theta) \in \mathcal{F}$. Hence, we need to define a measure that allows us to quantify the distance between two distribution functions. A practical and mathematically useful measure is the Kullback-Leibler Divergence, which is formally defined as follows.

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

$$KL(G, F) = \int \log \frac{g(y)}{f(y)} dG(y)$$

where $g(y)$ and $f(y)$ are the densities or probability functions.

If the distribution is continuous we obtain

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

Otherwise, the integration is replaced by a sum.

The term

$$\int \log g(y) g(y) dy$$

is called entropy which for discrete random variables Y is positive unless the distribution is degenerated and has point mass 1 at a single point y_0 .

We sometimes make it explicit in the notation that $G(\cdot)$ and $F(\cdot)$ are in fact distribution functions so that the Kullback-Leibler divergence has two distribution functions as argument. To make this clear we write $KL(G(\cdot), F(\cdot))$. Both notations are used interchangeably.

Note that the Kullback-Leibler divergence is always positive and takes value 0 if and only if $G(\cdot) \equiv F(\cdot)$. This can be easily shown by noting that $\log(y) \leq y - 1$ with equality for $y \equiv 1$. This leads to (for continuous variables)

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

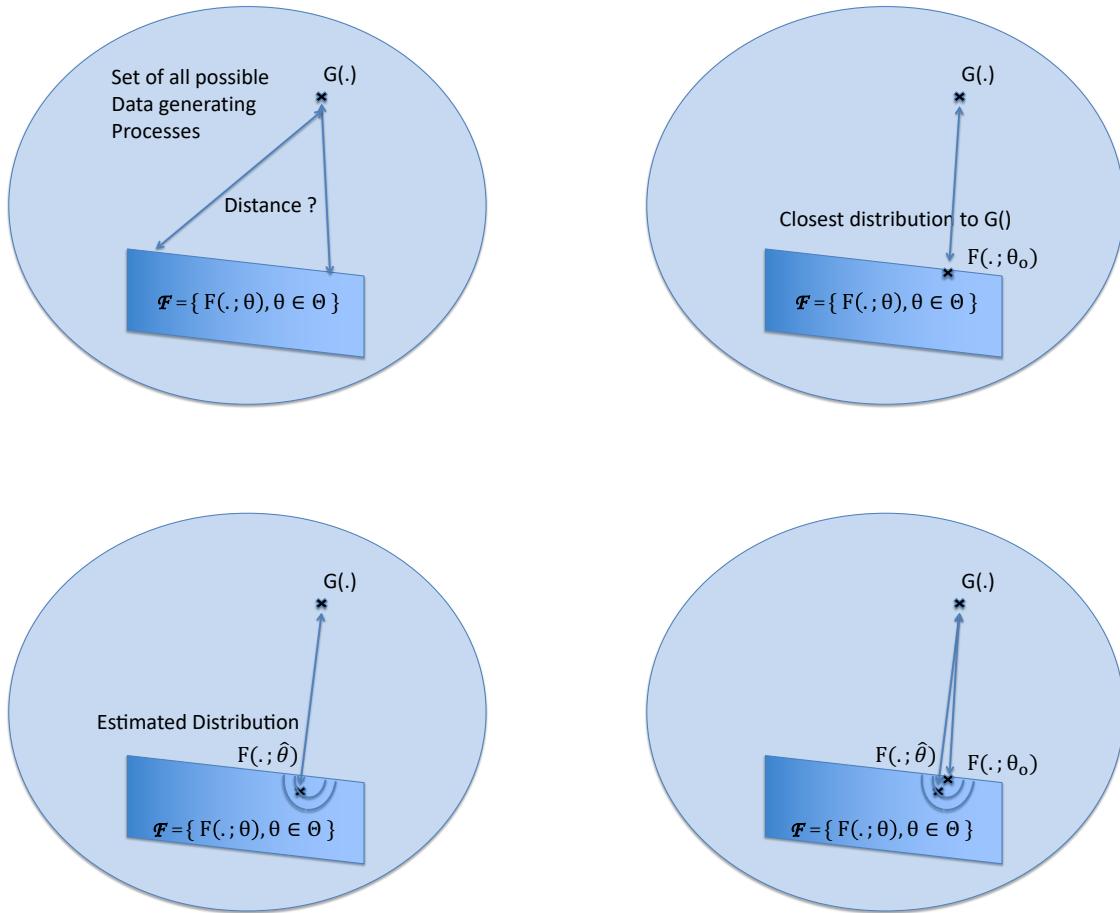


Fig. 3.1 Visualization of the true data generating process and its approximation through statistical models. The top left plot shows the true data generating process $G(\cdot)$ and a statistical model \mathcal{F} . The top right plot demonstrates the optimal parameter value θ_0 . The bottom left gives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

Note also, that the Kullback-Leibler divergence is not symmetric and in general it holds

$$KL(G, F) \neq KL(F, G).$$

This means that the KL divergence is not a metric or a distance measure, since any metric needs to be symmetric by definition. However, we can omit symmetry and

take the true data-generating process as an anchor point for the calculation of the Kullback-Leibler measure.

With this measure, we can now compare different distributions. We exemplify this in Figure 3.2 where we compare two different distributions. In the top plot, we show two exemplary distributions which differ by some amount. The second plot shows the log ratio of the two distributions and we see that this log ratio takes large values for small as well as for large values of y . The density for large and small values of y is however small and by weighting the log ratio in the second plot with the corresponding density of $g(\cdot)$ we obtain the third plot in Figure 3.2. The large discrepancies at the tails of y are now down-weighted and the smaller discrepancies in the centre part of the distribution are becoming dominant. Integrating out the function, as sketched in the bottom plot provides the Kullback-Leibler divergence.

Now that we have defined a (non-symmetric) distance measure we can calculate the distance between two distributions. This implies that for each element in \mathcal{F} we can calculate

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

Apparently, what we need is the distance between the set \mathcal{F} and $G(\cdot)$. This brings us to the second question. How to define the distance between a point - in this case $G(\cdot)$ - and a set - here \mathcal{F} ? We define this as the minimum Kullback-Leibler divergence, that is

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

If we assume \mathcal{F} to be a set of Fisher-regular distributions we can obtain the minimum through differentiation, that is we set

$$\begin{aligned} 0 &= \frac{\partial}{\partial \theta} \int \log \frac{g(y)}{f(y; \theta)} dG(y) \\ &= - \int \frac{\partial}{\partial \theta} \log f(y; \theta) dG(y). \end{aligned}$$

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

$$0 = \int \frac{\partial \log f(y; \theta_0)}{\partial \theta} dG(y) = E \left(\frac{\partial \log f(Y; \theta_0)}{\partial \theta} \right). \quad (3.3)$$

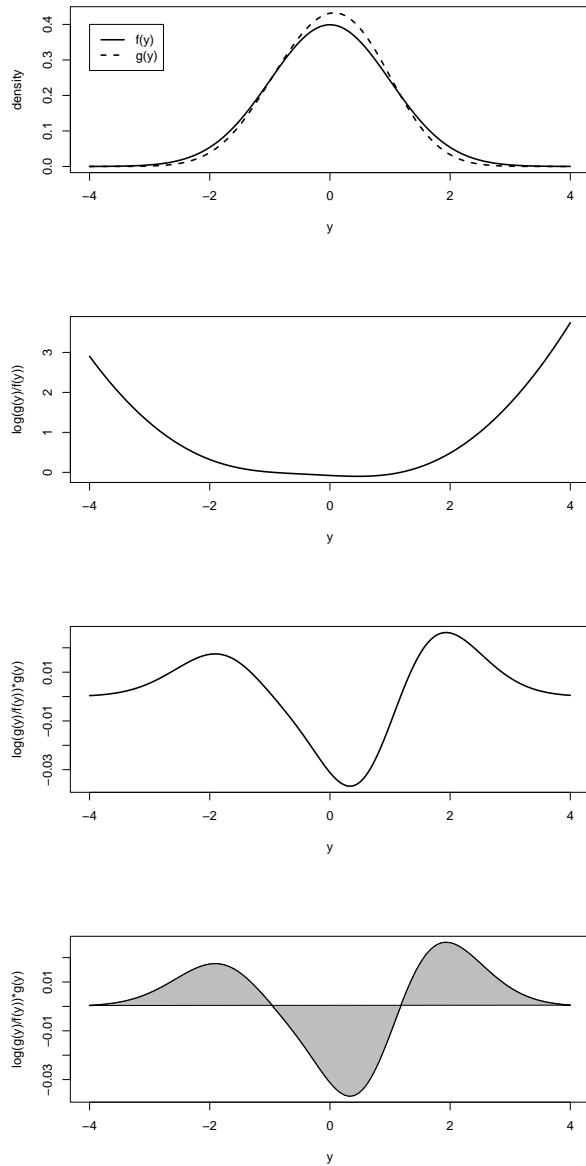


Fig. 3.2 Visualization of Kullback-Leibler divergence. The top row shows two densities $g(\cdot)$ and $f(\cdot)$. The middle row gives the log ratio $\log(g(\cdot)/f(\cdot))$ and the bottom row shows the ratio weighted with $g(\cdot)$. Integration of the bottom plot curve gives the Kullback-Leibler divergence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

We will call θ_0 the "true" parameter, understood here in the sense, that it is the "closest" parameter. This is shown in Figure 3.1 in the top right plot. It should be noted that the above integrals can not be calculated in practice since 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 $F(\cdot; \theta_0)$ which has the smallest distance. For every real-world data example, this will always be the case and the proposed approach looks like a cul-de-sac. However, we have data y_1, \dots, y_n which were generated through $G(\cdot)$ and with the results from Chapter ?? we know that expectations of the form

$$E(t(Y)) = \int t(y) dG(y)$$

for any (measurable) function $t(\cdot)$ can be estimated through the corresponding empirical version, that is

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

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

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

The first component is still not estimable since we do not know $g(\cdot)$ and hence can not 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

$$\frac{1}{n} \sum_{i=1}^n \log f(y_i; \theta) \rightarrow \int \log f(y; \theta) g(y) dy.$$

Our aim will be to maximize this term, which minimizes the Kullback-Leibler divergence. Omitting the $\frac{1}{n}$ term in the sum above leads to 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 likelihood function depends on the observations y_1, \dots, y_n which is suppressed in the notation. Formally however we have

$$l(\theta) = l(\theta; y_1, \dots, y_n).$$

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 was defined through

$$0 = \int \frac{\partial \log f(y; \theta_0)}{\partial \theta} dG(y)$$

which we replace through the empirical version

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

where $\hat{\theta}$ solves (3.4). We define $\hat{\theta}$ as the maximum likelihood estimate. At this point it remains unclear, where the phrase "maximum" comes from since we are "minimizing" a distance. We agree that this appears obscure. We will develop a thorough theoretical framework for maximum likelihood estimation, but the motivation why this is a useful concept is based on the minimum Kullback-Leibler approach. We visualize this in Figure 3.1 in the bottom left. Apparently $\hat{\theta}$ is not necessarily equal to θ_0 so that the distance between $F(\cdot; \theta)$ and $G(\cdot)$ is somewhat larger than the minimal distance between $F(\cdot; \theta_0)$ and $G(\cdot)$. See the bottom right plot in Figure 3.1. We will come back to this point in the next chapter, where we show 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. We will show that with increasing sample size we obtain that

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

or more precisely

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

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

3.2 Squared Loss and Likelihood

The likelihood approach allows to minimize the Kullback-Leibler distance between the data-generating process $G(\cdot)$ and a statistical model $\mathcal{F} = \{F(\cdot; \theta), \theta \in \Theta\}$. It leads to an estimation principle by setting the first-order derivative of the log likelihood function to zero. We want to elaborate on this further and connect this to the familiar squared loss function approach. We initiated this discussion already in Example 1, but we keep it more simple here. Assume we are interested in the mean of data y_1, y_2, \dots, 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

$$\min \sum_{i=1}^n (y_i - \mu)^2 \quad (3.5)$$

We obtain the same loss function if we assume a normal distribution, i.e. \mathcal{F} is the set of all normal distributions with mean μ and variance σ^2 . This leads to the log likelihood function

$$l(\mu, \sigma^2) = \sum_{i=1}^n \left(-\log(\sigma) - \frac{1}{2} \frac{(y_i - \mu)^2}{\sigma^2} \right) \quad (3.6)$$

If we are interested in μ only, i.e. omitting the variance for the moment shows that maximizing (3.6) is equivalent to minimizing (3.5). Phrasing this differently means, using a squared loss is equivalent to assuming a normal distribution for the data y_1, y_2, \dots

3.3 Likelihood and Bayes

The original work by Fisher (1912), in which he introduced the concept of Likelihood, catalyzed the adoption of likelihood-based approaches using Bayesian principles (see also, Aldrich, 1997 or Edwards, 1997). However, the term “Bayesian statistics” wasn’t yet coined during that era; it took nearly 50 more years for this terminology to emerge. What we now recognize as Bayesian reasoning was formerly known as “inverse probability,” as discussed 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 contemplating the probability model for θ given that we have observed y .

We argued before that probability can be used to express uncertainty, and in fact, vice versa as well, to express knowledge. In this way we proceed with the

presumption that prior information concerning the parameter θ is available and we aim to effectively incorporate this into our analysis. This prior information expresses our preliminary beliefs regarding the parameter θ and it is quantified in the form of a probability distribution denoted as $p_\theta(\theta)$. The structure of this distribution may encompass the inclusion of additional parameters, which we ignore for the moment. Building upon a dataset $y = (y_1, \dots, y_n)$, we can deduce the information flow through the posterior distribution. That is to say, what information did we gain through the data about our parameter θ . This is available through the process of "inverting" the distributions. Assuming a continuous random variable Y this leads to

$$\begin{aligned} p(\theta|y) &= \frac{f(y; \theta)p(\theta)}{f(y)} \\ &= \frac{\prod_{i=1}^n f(y_i; \theta)p(\theta)}{f(y)} \\ &= \frac{\exp(l(\theta))p(\theta)}{f(y)}, \end{aligned} \tag{3.7}$$

where the first component in the numerator is the likelihood, that is \exp 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 Chapter 9 again. For now, we concentrate on the numerator in (3.7) and construct an optimization problem from it. For instance, we could be interested in the posterior mode.

Definition 3.3 The posterior mode estimate is defined through

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

There is an interesting relation to the likelihood approach. If $p(\theta)$ is constant, which is also called "flat prior", then the posterior mode is equal to the maximum likelihood estimate, that is

$$\arg \max_{\theta \in \Theta} p(\theta|y) = \arg \max_{\theta \in \Theta} l(\theta). \tag{3.8}$$

In general, however this is not the case.

Throughout this book, we will extensively explore Bayesian concepts on multiple occasions. In Chapter 9, we offer a comprehensive overview of modern Bayesian statistics, setting the stage for our further discussions. A compelling argument we will put forth is that Bayesian ideas lie at the heart of overparameterized models, wherein the number of parameters is often large, and even larger than the number of data points. An example along this line is also deep neural networks. We will extend this discussion in later chapters of the book.

3.4 Properties of Estimates

3.4.1 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 estimates in a statistical model $\mathcal{F} = \{F(\cdot; \theta), \theta \in \Theta\}$. Let us start with consistency. An estimate $\hat{\theta}$ is thereby consistent, if for increasing sample size the range of values for the estimate gets smaller in probability and includes the target parameter θ_0 . Target parameter means here, that it is either the true parameter if the data generating process $G(\cdot)$ is within \mathcal{F} , that is $G(\cdot)$ can be written as $G(\cdot) = F(\cdot; \theta_0)$, or if $G(\cdot) \notin \mathcal{F}$, then the target parameter θ_0 gives the closest distribution in \mathcal{F} based on the Kullback-Leibler divergence, as described above.

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

$$\int \log \frac{f(y; \hat{\theta})}{f(y; \theta_0)} dG(y) \xrightarrow{n \rightarrow \infty} 0. \quad (3.9)$$

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 (3.9) as a probability statement. We want to remain a little sloppy for the moment and get more precise later. Let us look therefore at (3.9) more closely. Simple calculation leads to

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

Hence, the Kullback-Leibler divergences need to converge for an increasing number of data points n . We apply Taylor series expansion to the estimated log density and get

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

Note that with the definition of θ_0 given in (3.3) for Fisher-regular distributions we get

$$2 \int \log \frac{f(y; \hat{\theta})}{f(y; \theta_0)} dG(y) \approx (\hat{\theta} - \theta_0)^\top \underbrace{E\left(-\frac{\partial^2 \log f(y; \theta_0)}{\partial \theta \partial \theta^\top}\right)(\hat{\theta} - \theta_0)}_{:= I(\theta_0)}, \quad (3.10)$$

where the expectation is taken with respect to the data generating process (2.2), that is with respect to $G(\cdot)$. We will later call matrix $I(\theta_0)$ Fisher information or Fisher matrix, which plays an important role in maximum likelihood estimation. As remarked above, (3.9) 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 $\hat{\theta} = t(y_1, \dots, y_n)$. 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 (3.9) by taking expectation over all possible samples we obtain

$$E\left(\int \log \frac{f(y; \hat{\theta})}{f(y; \theta_0)} dG(y)\right) \xrightarrow{n \rightarrow \infty} 0.$$

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

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

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

$$E\left((\hat{\theta} - \theta_0)^\top I(\theta_0)(\hat{\theta} - \theta_0)\right) \xrightarrow{n \rightarrow \infty} 0,$$

where the expectation is taken over the data y_1, \dots, y_n . To simplify this further we restrict the deviations for notational simplicity and without loss of generality to univariate parameters. This allows us to rewrite the above to

$$\begin{aligned}
 E(I(\theta_0)(\hat{\theta} - \theta_0)^2) &\propto E((\hat{\theta} - \theta_0)^2) \\
 &= E((\hat{\theta} - E(\hat{\theta}) + E(\hat{\theta}) - \theta_0)^2) \\
 &= \underbrace{E((\hat{\theta} - E(\hat{\theta}))^2)}_{=Var(\hat{\theta})} + \underbrace{(E(\hat{\theta}) - \theta_0)^2}_{=:bias^2(\hat{\theta}, \theta)} \quad (3.11)
 \end{aligned}$$

where \propto stands for proportionality which allows us to omit components not depending on $\hat{\theta}$. The first term is the variance of the estimate while the second is called bias. The bias thereby quantifies the systematic difference between the mean value of the estimate and the target parameter. The decomposition (3.11) is also called the bias-variance-tradeoff. We sketch the idea in Figure 3.3 when shooting toward the bull's eye. The resulting quantity above is thereby equal to the Mean Squared Error (MSE).

Definition 3.4 The Mean Squared Error of an estimate $\hat{\theta}$ is defined as

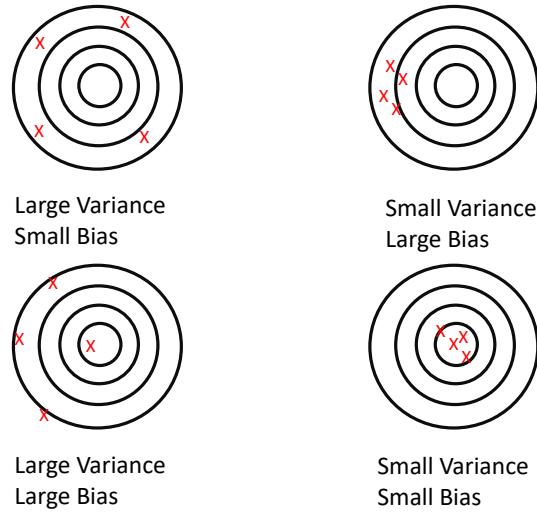


Fig. 3.3 Sketch of Bias and Variance when shooting on bull's eye

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

The bias is defined as

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

and an estimate is called unbiased if

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

The estimate is called asymptotically unbiased if

$$bias(\hat{\theta}, \theta_0) \xrightarrow{n \rightarrow \infty} 0.$$

With these requirements, we can now write the original idea (3.9) more formally and define consistency as follows.

Definition 3.5 The estimate $\hat{\theta}$ is MSE consistent for θ_0 (or in short just consistent estimate) if for increasing data size n with data drawn from the data generating process (2.3) it holds

$$MSE(\hat{\theta}, \theta_0) \xrightarrow{n \rightarrow \infty} 0.$$

Why not just use $\rightarrow 0$ as $n \rightarrow \infty$? I think that the long arrow looks a bit weird. For completeness, we remark that there are various further definitions of consistency. One can also define an estimate $\hat{\theta}$ to be consistent for θ_0 if for every $\delta > 0$ one gets

$$P(|\hat{\theta} - \theta_0| \geq \delta) \xrightarrow{n \rightarrow \infty} 0, \quad (3.12)$$

where the probability refers to the sample Y_1, \dots, Y_n drawn from the data generating process. For the next step, we employ make use of Chebyshev's inequality, stating that for X being a random variable with mean μ and variance $\sigma^2 > 0$ the following

inequality holds for any $k > 0$:

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}.$$

Note that $\hat{\theta}$ is a random variable with mean θ and variance $Var(\hat{\theta}) = E((\hat{\theta} - \theta)^2)$ if $\hat{\theta} = t(Y_1, \dots, Y_n)$ has finite moments. Applying Chebyshev's inequality, we then obtain for arbitrary $\delta > 0$

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

We see that in this case (3.12) is equivalent to MSE consistency, since if $MSE(\hat{\theta}, \theta_0) \rightarrow 0$ for $n \rightarrow \infty$ we obtain (3.12) and vice versa.

3.4.2 Sufficiency

In addition to consistency, we also want to define sufficiency. This concept entails condensing data in such a manner that no relevant information is dropped, and the compressed data encapsulates all the necessary information to estimate the parameter. To motivate the idea, consider a simple example: suppose a binary variable $Y_i \sim \text{Binom}(1, \pi)$, and our focus rests on parameter π . It can be readily derived that the arithmetic mean $\bar{Y} = \sum_{i=1}^n Y_i / n$ is the maximum likelihood estimate for π . Hence, if we distil the data to the mean and the sample size, we have successfully extracted all the essential information. Hence, the following diverse samples, presented as (y_1, \dots, y_n) , all yield the identical maximum likelihood estimate:

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

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 $\bar{y} = 1/4$. We want to formalize this as the concept of sufficiency.

If $t(y_1, \dots, y_n)$ gives sufficient information about parameter θ , it means that all the information concerning θ is in $t(y_1, \dots, y_n)$ and knowing $t(y_1, \dots, y_n)$ is enough to estimate the parameter. Remarkably, the concept of sufficiency permits us to entirely delete our data if we have calculated a sufficient estimate. This can be rather useful,

particularly in the era of big data, as it states that one only needs to calculate and store $t(y_1, \dots, y_n)$ in order to get all information from our sample relevant to our parameter.

To proceed, let us now define sufficiency more formally.

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

This states, once again, that the distribution of the single values y_1, \dots, y_n is non-informative, if we know the value of the statistic $t(y_1, \dots, y_n)$. This idea does look a little confusing at first glance, so an example might help to clarify.

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

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

Apparently, this distribution does not depend on π . ▷

It can be difficult to prove sufficiency in the above defined form, but the **Neyman-factorisation** makes it simple to find a sufficient statistic.

Property 3.1 A statistic $t(Y_1, \dots, Y_n)$ is sufficient for θ if and only if the density or probability function decomposes to

$$f(y_1, \dots, y_n; \theta) = h(y_1, \dots, y_n)g\left(t(y_1, \dots, y_n); \theta\right), \quad (3.13)$$

where $h(\cdot)$ does not depend on θ and $g(\cdot)$ depends on the data only through the statistic $t(y_1, \dots, y_n)$. \square

The proof of this statement is rather simple and given at the end of this section. Note that sufficiency itself is a rather weak statement, as the original sample (y_1, \dots, y_n) itself is already sufficient. We therefore also require some concept of minimality defined as follows:

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

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, \dots, y_n to the value of the minimal sufficient statistic, but we may not reduce it further without losing information about the parameter θ .

3.4.3 Efficiency

Lastly, we want to discuss the concept of efficiency. Here, we're essentially asking a question about which of two possible estimators, which one can view as algorithms, performs better. In this context, "better" refers to having a smaller Mean Squared Error.

Imagine you have one estimator, denoted as $\hat{\theta}_1 = t_1(Y_1, \dots, Y_n)$, and another one, denoted as $\hat{\theta}_2 = t_2(Y_1, \dots, Y_n)$. Both of them are designed to provide an estimate for a certain parameter θ_0 . The comparison between the two boils down to evaluating their performance in terms of Mean Squared Error (MSE). If the MSE of $\hat{\theta}_1$ is less than or equal to the MSE of $\hat{\theta}_2$, then $\hat{\theta}_1$ is considered superior to $\hat{\theta}_2$. It's important to note that this concept is largely theoretical, as explicitly calculating the MSE isn't feasible since it depends on the unknown data generation process.

Example 3 Let us illustrate this concept through a simple example. Assume we have normally distributed random variables $Y_i \sim N(\mu, \sigma^2)$, which are *i.i.d.* We are interested in estimating the mean and propose the estimate

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

for some weights w_i . To obtain unbiasedness it is easy to see that we need to postulate $\sum_{i=1}^n w_i = 1$, because

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

We now question how to choose the weights such that the variance of $t(y)$ is minimized. Note that

$$\text{Var}(t(Y)) = \sum_{i=1}^n w_i^2 \sigma^2.$$

If we set $w_i = \frac{1}{n} + d_i$ we need $\sum_{i=1}^n d_i = 0$ for unbiasedness and hence

$$\begin{aligned} \text{Var}(t(Y)) &= \sum_{i=1}^n \left(\frac{1}{n} + d_i \right)^2 \sigma^2 \\ &= \sum_{i=1}^n \left(\frac{1}{n^2} + d_i^2 \right) \sigma^2 \\ &= \frac{\sigma^2}{n^2} + \sum_{i=1}^n d_i^2 \sigma^2. \end{aligned}$$

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

We will later show that Maximum Likelihood estimation is (approximately) efficient since the resulting estimates are asymptotically unbiased and have the smallest variance.

Proof We here prove Neyman-factorisation. Assume that $t(\cdot)$ is sufficient, then $f(y_1, \dots, y_n | t(y) = t; \theta)$ does not depend on θ . Because $t(\cdot)$ is calculated from y_1, \dots, y_n , with the basic definition of conditional probabilities, we get

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

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 $t(y_1, \dots, y_n)$ and θ . Let us assume now that the density $f(y_1, \dots, y_n)$ is factorized as in (3.13). The marginal density for $t(y_1, \dots, y_n)$ is:

$$\begin{aligned}
f_t(t; \theta) &= \int_{t(y_1, \dots, y_n)=t} f(y_1, \dots, y_n; \theta) dy_1 \dots dy_n \\
&= \int_{t(y_1, \dots, y_n)=t} h(y_1, \dots, y_n) g(t; \theta) dy_1 \dots dy_n \\
&= g(t; \theta) \int_{t(y_1, \dots, y_n)=t} h(y_1, \dots, y_n) dy_1 \dots dy_n.
\end{aligned}$$

The conditional distribution can then be written as

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

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

3.5 Estimation and Uncertainty

So far we developed strategies for estimation, and we emphasized the importance of the Kullback-Leibler divergence. Estimation or learning, respectively, can be thought of as a process or algorithm that gives us a value $\hat{\theta}$ as an output. Note that $\hat{\theta} = t(y_1, \dots, y_n)$ is calculated from data and we assume that the data were random draws from the data-generating process described in (2.3). Consequently, this estimated value $\hat{\theta}$ is not fixed – it has variability due to the data. This means that if we were to use different sets of data, we would get different results of the estimate. While we treat the estimation process as a consistent algorithm (meaning the same inputs yield the same outputs), the inputs themselves are drawn from (2.3). As a result, the outcome of the estimation process can vary due to the inherent randomness in the data. This variability is known as the *estimation variance*, and we aim to come up with 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 by $t_l(y_1, \dots, y_n)$ and $t_r(y_1, \dots, y_n)$, representing the lower and upper bound of the interval. To keep things simple, we are dealing with a univariate parameter.

The interval limits are functions from the data and hence can again be understood as a result of an algorithm applied to the data. Clearly, to achieve a useful interval we need $t_l(Y_1, \dots, Y_n) < t_r(Y_1, \dots, Y_n)$ for all values of Y_1, \dots, Y_n . These statistics can now be defined as follows:

Definition 3.8 The interval $CI = [t_l(Y_1, \dots, Y_n), t_r(Y_1, \dots, Y_n)]$ is called a **confidence interval** for θ with confidence level $(1 - \alpha)$ if

$$P_\theta \left(t_l(Y_1, \dots, Y_n) \leq \theta \leq t_r(Y_1, \dots, Y_n) \right) \geq 1 - \alpha \quad (3.14)$$

for all θ . The value $(1 - \alpha)$ is called the confidence level and α is chosen to be a small value, e.g. $\alpha = 0.01$ or $\alpha = 0.05$.

The probability statement (3.14) can be reformulated as

$$P_\theta \left(\theta \in [t_l(Y_1, \dots, Y_n), t_r(Y_1, \dots, Y_n)] \right) \geq 1 - \alpha,$$

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

It is also clear that defining $t_l(Y_1, \dots, Y_n) = -\infty$ and $t_r(Y_1, \dots, Y_n) = \infty$ automatically gives a $(1 - \alpha)$ confidence interval, as $P(-\infty \leq \theta \leq \infty) = 1$ for any $\theta \in \mathbb{R}$. This is, of course, a useless interval, and hence the intention is to find an interval with the minimal length that still fulfills the definition of a confidence interval. In other words, we preferably choose " $=$ " in (3.14) and not " \geq ".

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

Definition 3.9 A quantity $g(t(Y_1, \dots, Y_n); \theta)$ is called a **pivotal statistic** if its distribution does not depend on θ . The distribution of $g(t(Y_1, \dots, Y_n); \theta)$ 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 estimate $\hat{\theta} = t(Y_1, \dots, Y_n)$ in many cases follows approximately a normal distribution. We denote this as

$$\hat{\theta} = t(Y_1, \dots, Y_n) \xrightarrow{a} N(\theta, Var(\hat{\theta})), \quad (3.15)$$

where \xrightarrow{a} stands for "asymptotically distributed", hence for sufficiently large n . With (3.15) we can construct an approximate pivotal statistic through

$$g(t(Y_1, \dots, Y_n); \theta) = \frac{t(Y_1, \dots, Y_n) - \theta}{\sqrt{Var(\hat{\theta})}} = \frac{\hat{\theta} - \theta}{\sqrt{Var(\hat{\theta})}} \xrightarrow{a} N(0, 1). \quad (3.16)$$

Since $N(0, 1)$ does not depend on θ , the statistic is pivotal. The pivotal distribution can now be used to construct a confidence interval in the following way. With (3.16) we have

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

where $z_{\alpha/2}$ is the $\alpha/2$ quantile of a $N(0, 1)$ distribution and accordingly, $z_{1-\alpha/2}$ the $1 - \alpha/2$ quantile. Because $z_{\alpha/2} = -z_{1-\alpha/2}$, we obtain the confidence interval as

$$CI = \left[\underbrace{\hat{\theta} - z_{1-\alpha/2}\sqrt{Var(\hat{\theta})}}_{t_l(Y_1, \dots, Y_n)}, \underbrace{\hat{\theta} + z_{1-\alpha/2}\sqrt{Var(\hat{\theta})}}_{t_r(Y_1, \dots, Y_n)} \right]. \quad (3.17)$$

Also, remember that if you choose a confidence level of 0.95, the corresponding value $z_{1-\alpha/2}$ is 1.96. Rounding this up to 2 gives the commonly used guideline

$$CI \approx \left[\hat{\theta} \pm 2\sqrt{Var(\hat{\theta})} \right]. \quad (3.18)$$

Figure 3.4 visualizes the construction of the above confidence interval (3.17). In the top left plot, we show the asymptotic normal distribution around the estimate $\hat{\theta}$ with its variance. Standardizing the estimate through

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

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

One should note that the variance $Var(\hat{\theta})$ may itself depend on θ , the unknown parameter. It may also depend on some other parameters, which are unknown. 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.

Since this is a historical problem in statistics we explicitly focus on confidence intervals for the mean of a normal distribution. We assume that data Y_1, \dots, Y_n are normally distributed with

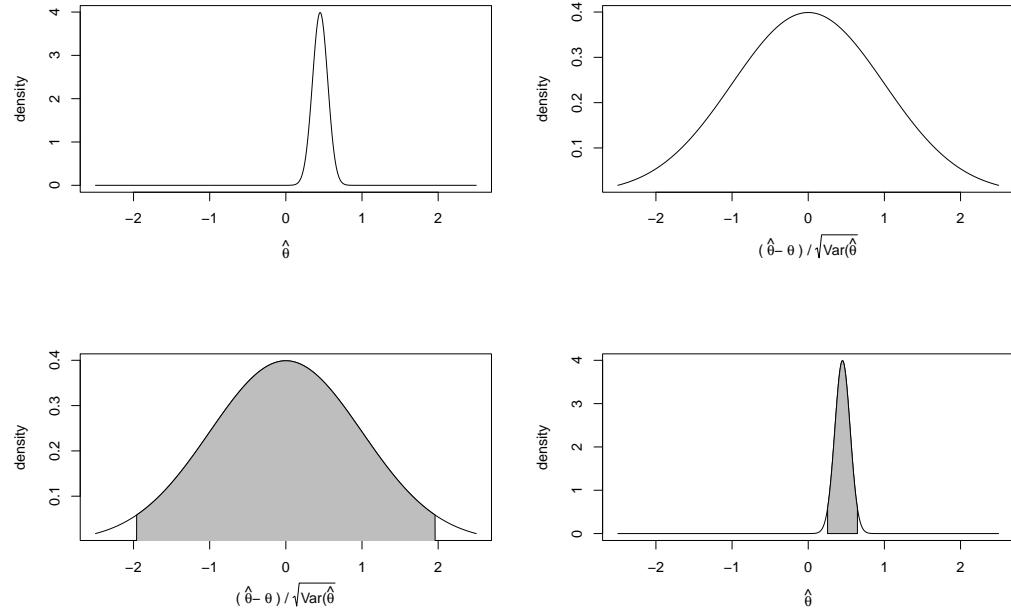


Fig. 3.4 Visualization of constructing a pivotal distribution by standardizing estimate $\hat{\theta}$ (top left plot) through $(\hat{\theta} - \theta_0)/\sqrt{Var(\hat{\theta})}$ (top right plot). This gives a standard normal distribution as pivotal from which confidence intervals can be calculated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

$$Y_i \sim N(\mu, \sigma^2) \text{ i.i.d.} \quad (3.19)$$

The mean is estimated through \bar{Y} and we obtain for the estimate

$$\hat{\mu} = \bar{Y} \sim N(\mu, \sigma^2/n).$$

A confidence interval is then easily constructed through

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

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

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

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

$$\frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

is a pivotal statistic with a known distribution. But if we replace the variance with its estimate we have

$$\frac{\bar{Y} - \mu}{\hat{\sigma}/\sqrt{n}} \not\sim N(0, 1). \quad (3.20)$$

The estimation of the variance leads to additional uncertainty. This phenomenon was identified and quantified by William Sealy Gosset (1876 - 1937). Gosset is thereby a legend in the field of statistics, not only because of his seminal scientific contributions. During his professional life was head brewer at Guinness, the renowned Irish beer company. When he solved the problem shown in (3.20) as an employee at the Guinness brewery, he was unable to publish this because of the possible risk of revealing a trade secret. He therefore shared 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 t-distribution and (3.20) is solved by

$$\frac{\bar{Y} - \mu}{\hat{\sigma}/\sqrt{n}} \sim t_{(n-1)} \quad (3.21)$$

where the index $n - 1$ refers to the degree of freedom. The derivation is given below. Note that (3.21) is again a pivotal statistic as it does neither depend on μ nor σ^2 . This allows to construction of confidence intervals through

$$CI = \left[\bar{y} - t_{(n-1), 1-\alpha/2} \frac{S_n}{\sqrt{n}}, \bar{y} + t_{(n-1), 1-\alpha/2} \frac{S_n}{\sqrt{n}} \right] \quad (3.22)$$

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

The proof that (3.20) is t distributed is given at the end of this section. Here we want to give a graphical motivation why the estimation of the variance in fact leads to increased variation. Let us therefore look at Figure 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 Figure 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 the right are unequal and the left area is larger than the right. This is equivalently shown in the distribution function calculated at the two symmetric quantiles (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 on 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, if we aim for $1 - \alpha$ coverage probability. This motivates the use of a t distribution, shown on the bottom right plot. Note that for increasing n we get the standard normal distribution as the limit.

Before we close this section we want to discuss the interpretation of confidence intervals. Assume that we have calculated a confidence interval with values $[t_l(y_1, \dots, y_n), t_r(y_1, \dots, y_n)]$. We might now pose the question:

$$P(\theta \in [t_l(y_1, \dots, y_n), t_r(y_1, \dots, y_n)]) = ? \quad (3.23)$$

Putting it in words: what is the probability that the true unknown parameter is within the confidence interval, or equivalently, that the confidence interval covers the true parameter? To make a concrete example, assume that we have collected normally distributed data y_1, \dots, y_n with $\bar{y} = 3.2$ and $S_n^2 = 4.2$ where $n = 20$. Then, 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 2.093024, so that the (rounded) confidence interval results to

$$[3.2 - \sqrt{4.2/20} * 2.093024, 3.2 + \sqrt{4.2/20} * 2.093024] = [2.24, 4.16].$$

In order to interpret this we might formulate (3.23) to $P(\mu \in [2.24, 4.16]) = ?$. Formally this is not a question based on a probability model, since data y_1, \dots, y_n are given data and hence there is no random experiment. Putting it differently, the unknown parameter is either in the interval or it is not in the interval and the probability statement (3.23) is not about a random process. This needs clarification. 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". Confidence here translates to the following statement:

Definition 3.10 Confidence principle: If we would repeat the calculation with new data of the same size multiple times, then in $\alpha * 100$ percent of the cases the parameter

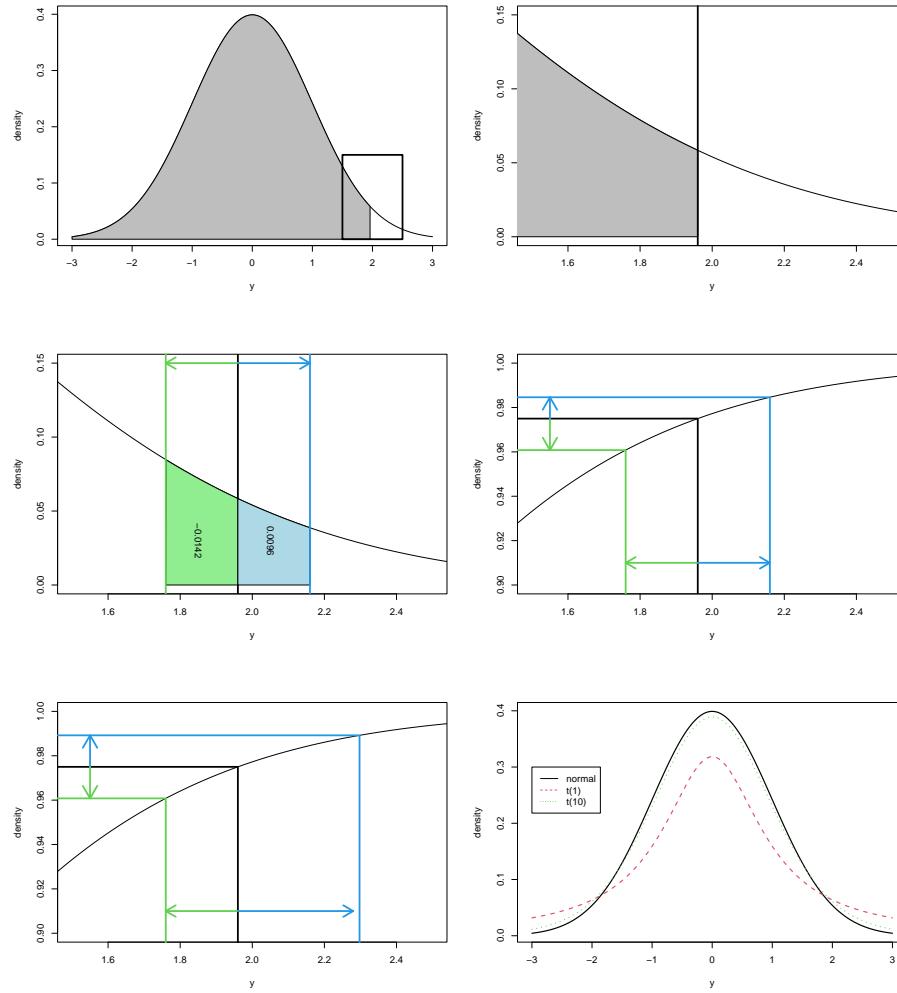


Fig. 3.5 Standard normal distribution with 95% quantile (top left) and zoom-in version (top right). If the variance is estimated unbiasedly, the quantile takes realized values roughly symmetric to the left and the right of the $1 - \alpha$ quantile of the normal distribution (middle left). This leads 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)

would not be within the interval and in $(1 - \alpha) * 100$ percent of the cases it would be.

Though this is formally a correct description, it is certainly awkward since it does not say anything about the concrete sample. But let us look at this from the

perspective of uncertainty and remember the example from Chapter 2.1 with the dice cup not lifted. The situation here is more or less the same, the data are given and we can quantify our uncertainty about the parameter θ . To do so we can 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 $(1 - \alpha) * 100$ percent, bearing in mind though, that “probability” here does not refer to a proper random experiment but to a quantification of our knowledge.

A similar way of interpretation results through Bayesian reasoning. In this case, though, confidence intervals are formally introduced as *credibility intervals* in the literature. In Bayes statistics, we consider the parameter θ to be random so that we can explicitly look at the posterior distribution.

Definition 3.11 A credibility interval for parameter θ is defined as follows. Our knowledge about the parameter θ is given by the posterior distribution $p_\theta(\theta | y_1, \dots, y_n)$. The credibility interval $[t_l(y), t_r(y)]$ is implicitly defined as

$$\begin{aligned} & P_\theta \left(\theta \in [t_l(y_1, \dots, y_n), t_r(y_1, \dots, y_n)] | (y_1, \dots, y_n) \right) \\ &= \int_{t_l(y_1, \dots, y_n)}^{t_r(y_1, \dots, y_n)} p_\theta(\theta | y_1, \dots, y_n) d\theta \geq 1 - \alpha. \end{aligned}$$

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

$$\int_{-\infty}^{t_l(y_1, \dots, y_n)} p_\theta(\theta | y_1, \dots, y_n) d\theta = \int_{t_r(y_1, \dots, y_n)}^{\infty} p_\theta(\theta | y_1, \dots, y_n) d\theta = \frac{\alpha}{2}, \quad (3.24)$$

that is, we cut off a probability mass of $\alpha/2$ in the left and right tail of the posterior probability. We visualize such construction in the top plot of Figure 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 Figure 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 $\theta_1 \notin [t_l(y_1, \dots, y_n), t_r(y_1, \dots, y_n)]$ and for $\theta_2 \in [t_l(y_1, \dots, y_n), t_r(y_1, \dots, y_n)]$ one has

$$p_\theta(\theta_1 | y_1, \dots, y_n) > p_\theta(\theta_2 | y_1, \dots, y_n).$$

This drawback can be avoided by using a **highest posterior density** credibility interval or, in short, the highest density interval. The construction principle is shown in Figure 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

$$HDI(y_1, \dots, y_n) = \{\theta; p_\theta(\theta | y_1, \dots, y_n) \geq c\},$$

where c is chosen such that

$$\int_{\theta \in HDI(y_1, \dots, y_n)} p_\theta(\theta | (y_1, \dots, y_n)) d\theta = 1 - \alpha.$$

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

Proof We now deliver the derivation of the t -distribution for ratio (3.21). We look at the t -statistics

$$T = \frac{\bar{Y} - \mu}{S_n / \sqrt{n}} \sim t(n-1). \quad (3.25)$$

with $S_n^2 = \sum_{i=1}^n (Y_i - \bar{Y})^2 / (n-1)$ and 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 (3.25) can be rewritten as ratio (A.5). To do so let

$$\begin{aligned} Z_1 &= \frac{1}{\sqrt{n}} (Y_1 + Y_2 + \dots + Y_n) = \sqrt{n} \bar{Y} \\ Z_2 &= \frac{1}{\sqrt{2}} (Y_1 - Y_2) \\ Z_3 &= \frac{1}{\sqrt{3}\sqrt{2}} (Y_1 + Y_2 - 2Y_3) \\ Z_4 &= \frac{1}{\sqrt{4}\sqrt{3}} (Y_1 + Y_2 + Y_3 - 3Y_4) \\ &\vdots \\ Z_n &= \frac{1}{\sqrt{n}} \frac{1}{\sqrt{n-1}} (Y_1 + Y_2 + \dots + Y_{n-1} - (n-1)Y_n). \end{aligned}$$

Note that all Z_i are normally distributed as they result as sum of (independent) normal distributed random variables. We will next calculate the corresponding moments. The mean values easily result as

$$\begin{aligned} E(Z_1) &= \sqrt{n}\mu \\ E(Z_2) &= \dots = E(Y_n) = 0. \end{aligned}$$

Accordingly, the pairwise covariances can be calculated through

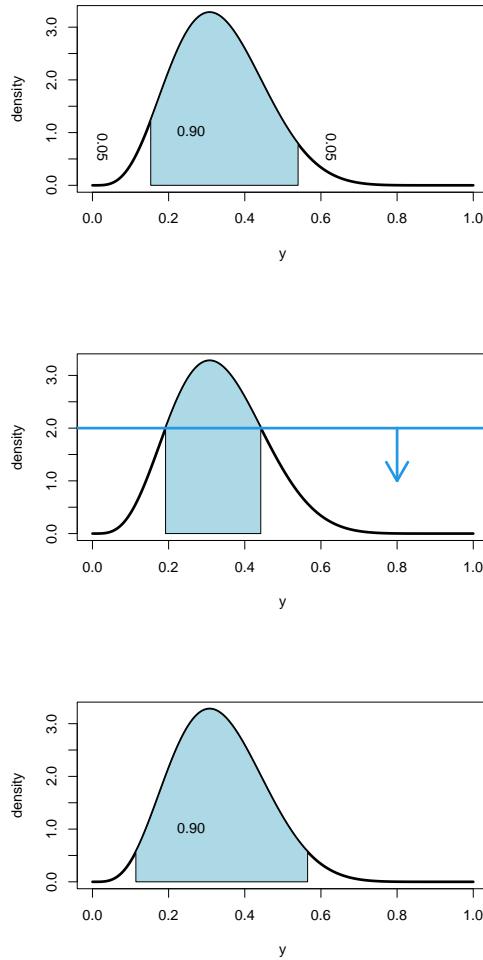


Fig. 3.6 Credibility interval, where we cut off to the left and the Highest density credibility interval for binomial data with $n = 5$ and $y = 0$ (left) and $n = 5$ and $y = 2$ (right).

$$\begin{aligned} \text{Cov}(Z_1, Z_2) &= \frac{1}{\sqrt{n}\sqrt{2}} \text{Cov}(\bar{Y}, Y_1 - Y_2) = \frac{1}{\sqrt{n}\sqrt{2}} \{\text{Cov}(\bar{Y}, Y_1) - \text{Cov}(\bar{Y}, Y_2)\} = 0 \\ &\vdots \\ \text{Cov}(Z_1, Z_i) &= 0, \quad i = 2, \dots, n. \end{aligned}$$

For $n \geq i > j > 1$ we obtain

$$\begin{aligned}
\text{Cov}(Z_i, Z_j) &= \text{Cov}\left(\frac{1}{\sqrt{i}} \frac{1}{\sqrt{(i-1)}} (Y_1 + \dots + (i-1)Y_i), \frac{1}{\sqrt{j}} \frac{1}{\sqrt{(j-1)}} (Y_1 + \dots + (j-1)Y_j)\right) \\
&= \frac{1}{\sqrt{i}} \frac{1}{\sqrt{(i-1)}} \frac{1}{\sqrt{j}} \frac{1}{\sqrt{(j-1)}} \text{Cov}(Y_1 + \dots + Y_j, Y_1 + \dots + (j-1)Y_j) \\
&= \frac{1}{\sqrt{i}} \frac{1}{\sqrt{(i-1)}} \frac{1}{\sqrt{j}} \frac{1}{\sqrt{(j-1)}} \underbrace{\text{Var}(Y_1) + \dots + \text{Var}(Y_{j-1}) - (j-1)\text{Var}(Y_j)}_{=0} = 0.
\end{aligned}$$

Hence we obtain $\text{Cov}(Z_i, Z_j) = 0$ for $i \neq j$. It remains to derive the univariate variances. These result through

$$\begin{aligned}
\text{Var}(Z_1) &= \sigma^2 \\
\text{Var}(Z_i) &= \text{Var}\left(\frac{1}{\sqrt{i}} \frac{1}{\sqrt{(i-1)}} (Y_1 + \dots + (i-1)Y_i)\right) \\
&= \frac{1}{i} \frac{1}{(i-1)} (\text{Var}(Y_1) + \dots + (i-1)^2 \text{Var}(Y_i)) \\
&= \frac{1}{i} \frac{1}{(i-1)} ((i-1) + (i-1)^2) \sigma^2 \\
&= \sigma^2.
\end{aligned}$$

We conclude that

$$\frac{Z_1 - \sqrt{n}\mu}{\sqrt{\sum_{i=2}^n Z_i^2 / (n-1)}} \sim t_{(n-1)}. \quad (3.26)$$

It remains to show that (3.25) can be expressed as (3.26). We show this by induction. For $n = 2$ we have

$$T = \frac{\bar{Y} - \mu_0}{\sqrt{S^2/2}} = \frac{\frac{Y_1 + Y_2}{2} - \mu_0}{\sqrt{((Y_1 - \bar{Y})^2 + (Y_2 - \bar{Y})^2)/2}} = \frac{Z_1 - \sqrt{2}\mu}{\sqrt{Z_2^2}} \sim t_{(1)}$$

Hence for $n = 2$ we have shown the equivalence between the t -statistics (3.25) and its reformulation (3.26). We can conclude the proof per induction. To do so we define $\bar{Y}_{(n)} = \sum Y_i / n$ and accordingly with $S_{(n)}^2 = \sum_{i=1}^n (Y_i - \bar{Y}_{(n)})^2 / (n-1)$ and $T_{(n)}$, with obvious definition for the latter. Let us therefore assume that the relation between $T_{(n)}$ in (3.25) and (3.26) holds up to n , that is

$$T_{(n)} = \frac{\bar{Y}_{(n)} - \mu}{\sqrt{S_{(n)}^2/n}} = \frac{Z_1 - \sqrt{n}\mu}{\sqrt{\sum_{i=2}^n Z_i^2 / (n-1)}} \sim t_{(n-1)}$$

This implies in particular the equality

$$Z_2^2 + \dots + Z_n^2 = \sum_{i=1}^n (Y_i - \bar{Y})^2 = (n-1)S_{(n)}^2 \quad (3.27)$$

For $n+1$ we get

$$\begin{aligned}
Z_2^2 + \dots + Z_n^2 + Z_{n+1}^2 &= \sum_{i=1}^n (Y_i - \bar{Y}_{(n)})^2 + \frac{1}{n+1} \frac{1}{n} (Y_1 + \dots + Y_n - nY_{n+1})^2 \\
&= \sum_{i=1}^n Y_i^2 - n\bar{Y}_{(n)}^2 + \frac{1}{n+1} \frac{1}{n} (n\bar{Y}_{(n)} - nY_{n+1})^2 \\
&= \sum_{i=1}^n Y_i^2 - n\bar{Y}_{(n)}^2 + \frac{n}{n+1} (\bar{Y}_{(n)} - Y_{n+1})^2 \\
&= \sum_{i=1}^n Y_i^2 - n\bar{Y}_{(n)}^2 + \frac{n}{n+1} \{ \bar{Y}_{(n)}^2 - 2\bar{Y}_{(n)}Y_{n+1} + Y_{n+1}^2 \} \\
&= \sum_{i=1}^{n+1} Y_i^2 - \frac{1}{n+1} \{ n^2\bar{Y}_{(n)}^2 + 2n\bar{Y}_{(n)}Y_{n+1} + Y_{n+1}^2 \} \\
&= \sum_{i=1}^{n+1} Y_i^2 - (n+1)\bar{Y}_{(n+1)}^2
\end{aligned}$$

where the latter follows since

$$\begin{aligned}
\bar{Y}_{(n+1)}^2 &= \left(\frac{1}{n+1} \right)^2 \left(\sum_{i=1}^{n+1} Y_i \right)^2 \\
&= \frac{1}{(n+1)^2} \left(\sum_{i=1}^{n+1} Y_i + Y_{n+1} \right)^2 \\
&= \frac{1}{(n+1)^2} (n\bar{Y}_{(n)} + Y_{n+1})^2 \\
&= \frac{1}{(n+1)^2} (n^2\bar{Y}_{(n)}^2 + 2n\bar{Y}_{(n)}Y_{n+1} + Y_{n+1}^2).
\end{aligned}$$

Consequently

$$T_{(n+1)} = \frac{\bar{Y}_{(n+1)} - \mu}{S_{(n+1)}/\sqrt{n}} = \frac{Z_1}{\sqrt{\sum_{i=2}^{n+1} Z_i^2/(n)}} \sim t(n),$$

which proves that the t -statistics (3.25) can be written as (A.5) and hence follows a t -distribution with $n - 1$ degrees of freedom. \square

3.6 Exercises

Exercise 1

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

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

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

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 $X_i \sim Bin(4, 0.5)$. Calculate the Kullback Leibler Divergence between the empirical distribution of the sample, $g(\cdot)$, and the assumed distribution, $f(\cdot)$.

Exercise 3

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

- 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 \bar{X} - the arithmetic mean - and \bar{X}^{-1} sufficient for λ as well?
- d. Are the statistics t and \bar{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.

Chapter 4

Maximum Likelihood Estimation

4.1 Score Equation

We motivated the use of the likelihood already in Chapter 3.1 so that we can be brief here. In this chapter, we want to derive relevant and very useful properties of the maximum likelihood estimate. Like above we thereby assume the data generating process (3.1) from above, that is

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

We aim to approximate this with the parametric model

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

We will assume that some basic mathematical operations are valid on \mathcal{F} and require that the statistical model \mathcal{F} is Fisher regular as already defined in Chapter A.7. To remind the reader, Fisher regular in particular means that integration with respect to the random variable and differentiation with respect to the parameter are exchangeable. We have derived in Chapter 3.1 that the closest distribution in \mathcal{F} to the true data generating process can be obtained through minimizing the Kullback-Leibler divergence. This defines the "true" parameter θ_0 implicitly through

$$E \left(\frac{\partial f(y; \theta_0)}{\partial \theta} \right) = \int \frac{\partial f(y; \theta_0)}{\partial \theta} dG(y) = 0. \quad (4.2)$$

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 also write $l(\theta; y)$ if we want to stress that the likelihood in fact depends on the data. Considering the data as random implies that the likelihood function itself can be considered random. In this case we have $l(\theta) = \sum_{i=1}^n \log f(Y_i; \theta)$, hence the observed data y_i are replaced by their random version Y_i . Notationally we do not differ between the two versions, but conceptually these are of course different. With this in mind we can reformulate (4.2) to

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

The first order derivative of the likelihood function is called **score function** and (4.3) means 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 equation (4.3) is determined by using the distribution $G(\cdot)$. Now, if we consider the scenario where $G(\cdot)$ belongs to \mathcal{F} , in simpler terms if we assume that the actual process generating the data is one among the set of distributions \mathcal{F} , interesting simplifications arise. In this case there exists a unique parameter θ_0 such that $G(\cdot) \equiv F(\cdot; \theta_0)$. The parameter θ_0 is also denoted as the "true" parameter. Property (4.3) gets a different justification in this case which we want to motivate as follows. 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:

$$1 = \int f(y; \theta) dy. \quad (4.4)$$

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

$$\begin{aligned} 0 &= \frac{\partial 1}{\partial \theta} = \int \frac{\partial f(y; \theta)}{\partial \theta} dy = \int \frac{\partial f(y; \theta)}{\partial \theta} \frac{f(y; \theta)}{f(y; \theta)} dy \\ &= \int \frac{\partial \log f(y; \theta)}{\partial \theta} f(y; \theta) dy = \int s(\theta; y) f(y; \theta) dy, \end{aligned}$$

where $s(\theta; y) = \frac{\partial}{\partial \theta} \log f(y; \theta)$ is the score function. This shows again that the score function has mean zero for the true parameter, i.e.

$$E(s(\theta; Y)) = 0. \quad (4.5)$$

Note that the expectation is calculated with respect to density $f(\cdot; \theta)$, that is the parameter in the density and the parameter in the score are the same. Property (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 θ_0 since the average slope is zero.

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

$$\begin{aligned} 0 &= \frac{\partial 0}{\partial \theta^\top} = \frac{\partial}{\partial \theta^\top} \int \frac{\partial}{\partial \theta} \log f(y; \theta) f(y; \theta) dy \\ &= \int \left(\frac{\partial^2}{\partial \theta \partial \theta^\top} \log f(y; \theta) \right) f(y; \theta) dy + \int \frac{\partial \log f(y; \theta)}{\partial \theta} \frac{\partial f(y; \theta)}{\partial \theta^\top} dy \\ &= E \left(\frac{\partial^2}{\partial \theta \partial \theta^\top} \log f(Y; \theta) \right) + \int \frac{\partial \log f(y; \theta)}{\partial \theta} \frac{\partial \log f(y; \theta)}{\partial \theta^\top} f(y; \theta) dy \\ &\Leftrightarrow E(s(\theta; Y) s^\top(\theta; Y)) = E \left(-\frac{\partial^2}{\partial \theta \partial \theta^\top} \log f(Y; \theta) \right). \end{aligned}$$

Because $E(s(\theta; Y)) = 0$, we obtain the **second order Bartlett identity**

$$Var(s(\theta; Y)) = E \left(-\frac{\partial^2 \log f(Y; \theta)}{\partial \theta \partial \theta^\top} \right). \quad (4.6)$$

Note that this is a matrix-valued component if θ is vector-valued. As before we point out that the expectation is taken over $f(\cdot; \theta)$ so that the parameter in the density and the parameter in the arguments of the expectation and variance in (4.6) is the same. We define the expectation on the right-hand side of equation (4.6) as Fisher information matrix or in short Fisher matrix or Fisher information, respectively. We get that the variance of the score function is equal to the Fisher information

The maximum likelihood estimate is found by maximizing the (log) likelihood function. Because the score function has a non-negative variance, it follows from (4.6) that the second-order derivative is negative in expectation. This in turn guarantees a well-defined optimization problem as the likelihood function is concave and hence the maximum is unique. This property holds in expectation only and for a concrete sample, we might experience local maxima. The Fisher information plays a fundamental role as it mirrors whether the maximization problem is well defined or not. In particular, for overparameterized models, occurring in machine learning,

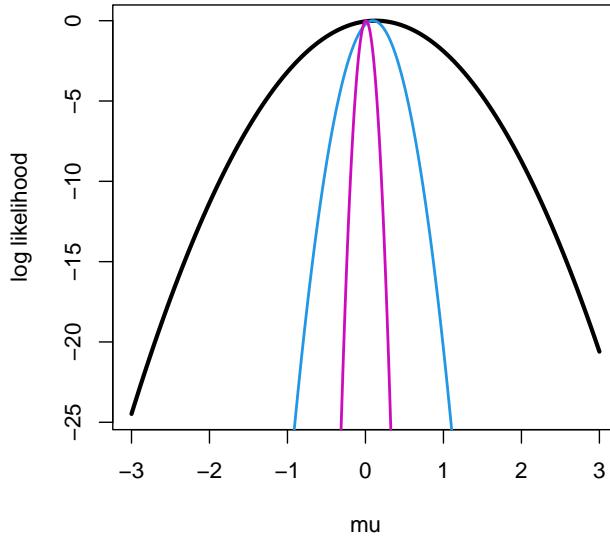


Fig. 4.1 Log likelihood function for μ for normally distributed data with samples sizes $n = 5$ (black outer line), $n = 50$ (blue middle line) and $n = 500$ (purple inner line)

we will pick up the point again.

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 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 Figure 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 with the sample size growing. 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 with the following definition.

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

$$s(\theta; y_1, \dots, y_n) = \frac{\partial l(\theta; y_1, \dots, y_n)}{\partial \theta}. \quad (4.7)$$

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

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

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

$$\text{Var}(s(\theta; Y)) = I(\theta). \quad (4.8)$$

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

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

$$\hat{\theta} = \arg \max l(\theta) = \arg \max l(\theta; y_1, \dots, y_n),$$

which for Fisher-regular distributions occurs when

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

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

Example 4 Assume that Y_i are *i.i.d* uniformly distributed between 0 and θ , that is $Y_i \sim \text{Uniform}[0; \theta]$. Then, the density results through

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

leading to the likelihood function

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

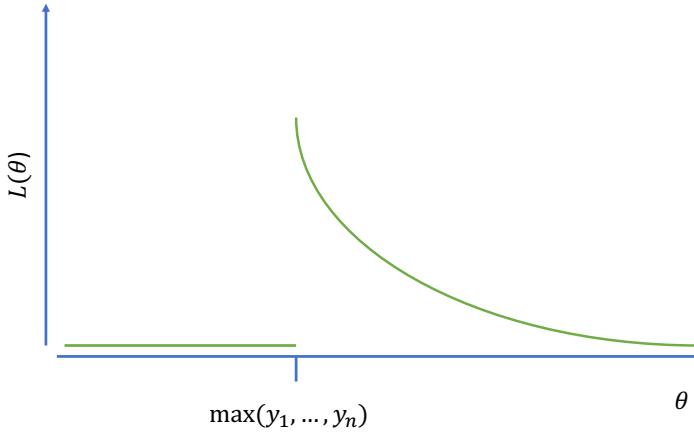


Fig. 4.2 Likelihood function for a uniform distribution on $[0, \theta]$

The log likelihood apparently results through

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

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

We can further explore a possible bias of the estimate. That is we aim to calculate $E(y_{max})$. Note that

$$P(y_{max} \leq y) = P((y_1 \leq y) \wedge (y_2 \leq y) \wedge \dots \wedge (y_n \leq y)) = F^n(y).$$

Hence the density results through

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

For a uniform distribution on $[0, \theta_0]$ with θ_0 as the true parameter, this leads to

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

Calculating the expected value leads to

$$E(y_{max}) = \int_0^{\theta_0} yn \frac{y^{n-1}}{\theta_0^n} dy = \int_0^{\theta_0} n \frac{y^n}{\theta_0^n} dy = \frac{n}{n+1} \frac{y^{n+1}}{\theta_0^n} \Big|_0^{\theta_0} = \frac{n}{n+1} \theta_0.$$

We see that the estimate is biased downwards and we have $E(y_{max}) < \theta_0$ ▷

4.2 Asymptotic Normality

It is important to understand, that the likelihood function itself can be comprehended as random since it is based on data which itself are realizations of a data-generating process. To demonstrate this we consider a simple example and look at the log likelihood for binary data. We assume $Y_i \sim B(1, \pi)$ i.i.d., $i = 1, \dots, n$ with unknown parameter $\pi = P(Y_i = 1)$, which in the simulation is set to $\pi_0 = 0.3$. The different possibly resulting likelihood functions are shown in Figure 4.3 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 $Y = \sum_{i=1}^{10} Y_i$. The width of the functions plotted in Figure 4.3 is proportional to its occurrence probability. The same holds for $n = 100$ on the right-hand side, but now we have 101 possible log likelihood functions. The top plots show the possible maxima of the log likelihood functions and the probability that this value occurs. The true value π_0 is indicated as a vertical line.

We see a number of interesting features. First, as remarked before, the log likelihood function is random. Secondly, the maximum value of the log likelihood function occurs in the vicinity of the true parameter π_0 , and third, with increasing sample size the maximum likelihood estimate appears to follow a normal distribution. This observation can in fact be formally proved and we conclude with the following result.

Property 4.1 Assuming a Fisher-regular distribution with parameter θ_0 from which an i.i.d. sample Y_1, \dots, Y_n is drawn. The maximum likelihood estimate is asymptotically normally distributed with

$$\hat{\theta} \stackrel{a}{\sim} N\left(\theta_0, I^{-1}(\theta_0)\right). \quad (4.9)$$

A formal proof of this statement is a bit lengthy and we therefore only sketch the main ideas. Before doing so, however, we emphasize that (4.9) is *the* central result in maximum likelihood theory and proves very useful in inferring the properties of $\hat{\theta}$. It means in particular, that we can easily derive confidence intervals for the maximum likelihood estimate, at least for sample size n reasonably large, since

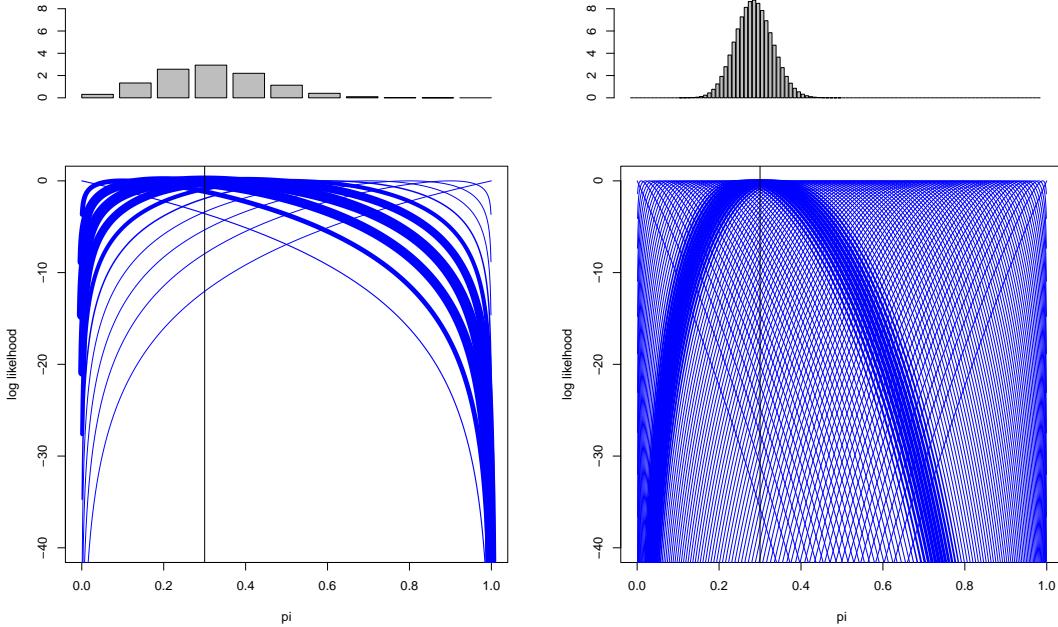


Fig. 4.3 Log likelihood functions for a Binomial distribution with $\pi_0 = 0.3$ for all possible outcomes. Left-hand side is for $n = 10$, right-hand side is for $n = 100$. Lines are weighted by their probability. The top layer shows the probability of different maxima that can be achieved.

a pivotal distribution is directly available from (4.9). This holds since the Fisher matrix $I(\theta)$ can usually be calculated quite easily or it can be approximated through simulations. We will get back to the latter point in a later chapter.

Apparently, the Fisher matrix depends on the true parameter, which in turn is unknown. But we can replace the parameter by its estimate which gives us $I(\hat{\theta})$. Inverting this matrix leads to $I^{-1}(\hat{\theta})$, which is taken as an estimate for the variance matrix of the maximum likelihood estimate, that is $\widehat{Var}(\hat{\theta}) = I^{-1}(\hat{\theta})$. This in turn means that for the k -th component of $\hat{\theta}_k$ we obtain

$$\frac{(\hat{\theta}_k - \theta_{k0})}{\sqrt{\widehat{Var}(\hat{\theta})_{kk}}} \stackrel{a}{\sim} N(0, 1) \quad (4.10)$$

with $\widehat{Var}(\hat{\theta})_{kk}$ as k -th diagonal element of $I^{-1}(\hat{\theta})$ and θ_{k0} referring to the k -th element of the true parameter θ_0 .

The following example demonstrates the above results.

Example 5 Assume that $Y \sim B(n, \pi)$. The log likelihood results to

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

and the derivative leads to the score function

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

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 $\hat{\pi} = \frac{y}{n}$.

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 estimate $\hat{\pi} = \frac{y}{n}$ has variance $\frac{\pi(1-\pi)}{n}$. \triangleright

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

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

Applying simple Taylor series expansion around θ_0 leads to

$$0 = s(\hat{\theta}) = s(\theta_0) + \frac{\partial s(\tilde{\theta})}{\partial \theta^\top} (\hat{\theta} - \theta_0), \quad (4.12)$$

where, based on the mean value theorem, we have $\tilde{\theta}$ located between $\hat{\theta}$ and θ_0 . Solving (4.12) for $\hat{\theta} - \theta_0$ gives

$$\hat{\theta} - \theta_0 = - \left[\frac{\partial s(\tilde{\theta})}{\partial \theta^\top} \right]^{-1} s(\theta_0). \quad (4.13)$$

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

$$s(\theta_0) \stackrel{d}{\sim} N(0, I(\theta_0)). \quad (4.14)$$

Moreover, bear in mind that

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

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 (4.13) we can conclude that with increasing sample size we obtain $\hat{\theta} \rightarrow \theta_0$ and therefore $\tilde{\theta} \rightarrow \theta_0$. Consequently, for increasing n , formula (4.13) can be approximated through

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

The matrix term itself can be written as

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

Note that $U(\theta_0)$ decomposes to

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

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 (4.15) further by dropping $U(\theta_0)$ so that we end up with

$$(\hat{\theta} - \theta_0) \approx I^{-1}(\theta_0) s(\theta_0). \quad (4.16)$$

Utilizing (4.3) directly proves (4.9). \square

4.3 Cramer Rao Bound

The idea of asymptotic normality proves useful because it allows for the construction of confidence intervals. However, we are aiming to delve deeper and establish that the maximum likelihood estimate is not just consistent, but also efficient, meaning that it holds the smallest possible variance. To achieve this, we will examine a central mathematical concept known as the Cramer-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). The result is as follows:

Property 4.2 Let $\hat{\theta} = t(Y_1, \dots, Y_n)$ be an estimate for θ based on the data Y_1, \dots, Y_n i.i.d. with $Y_i \sim F(\cdot; \theta)$, where $F(\cdot; \theta)$ is a Fisher-regular distribution. The mean squared error of $t(\cdot)$ is always larger than the lower **Cramer-Rao bound**

$$MSE(\hat{\theta}, \theta) \geq \text{Bias}^2(\hat{\theta}, \theta) + \frac{(1 + \frac{\partial \text{Bias}(\hat{\theta}, \theta)}{\partial \theta})^2}{I(\theta)}.$$

In particular, if the estimate is unbiased one has

$$MSE(\hat{\theta}, \theta) \geq \frac{1}{I(\theta)}. \quad (4.17)$$

The sketch of the proof 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 Cramer-Rao bound demonstrates why the inverse Fisher information plays a central role. In fact, for unbiased estimates, or at least for asymptotically unbiased estimates like the maximum likelihood estimate, 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 most efficient use of the data".

Proof Here we prove the validity of the Cramer-Rao bound. Note that for unbiased estimates we have

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

Differentiating both sides with respect to θ yields

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

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

$$E(s(\theta; y)) = \int s(\theta; y) f(y; \theta) dy = 0.$$

This implies that

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

With the **Cauchy-Schwarz** inequality (A.7) one obtains

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

Note that

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

using (4.6). As for unbiased estimates we have the equality $\text{Var}_\theta(t(Y)) = \text{MSE}(t(Y), \theta)$, we obtain

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

which is written down here for simplicity of notation for univariate statistics and parameters, respectively. This concludes the proof. \square

4.4 Likelihood Ratio

We have looked above at the maximum likelihood estimate $\hat{\theta}$ and derived its mathematical properties. We will now look at the likelihood function itself. This is sketched

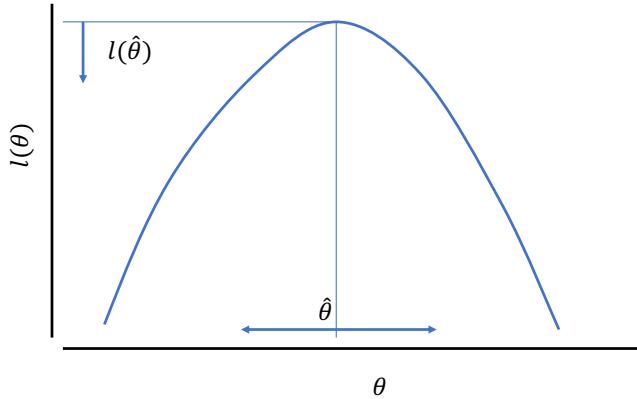


Fig. 4.4 Sketch of log likelihood function. Asymptotic properties of $\hat{\theta}$ refer to "horizontal" statements, while properties of $l(\hat{\theta})$ look at the "vertical" perspective

in Figure 4.4. The asymptotic statements derived above relate to the horizontal axis, i.e. statements relating to $\hat{\theta} - \theta_0$. Statements about the likelihood function $l(\theta)$ refer to the vertical axis. For the latter, we define the log likelihood ratio.

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

$$lr(\hat{\theta}; \theta) = 2\{l(\hat{\theta}) - l(\theta)\} = 2 \log \frac{L(\hat{\theta})}{L(\theta)} \quad (4.18)$$

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

The role of factor 2 above will become obvious in the subsequent property. Note that by definition the log likelihood ratio is always positive since $l(\hat{\theta})$ defines the maximum of the log likelihood function. If we again consider the estimate $\hat{\theta}$ as a random variable, the likelihood ratio itself becomes a random variable, for which we can derive its asymptotic distribution. In other words, besides the asymptotic distribution on the horizontal axis, we can also derive a related statement for the vertical axis.

Property 4.3 Likelihood-Ratio. The likelihood ratio (4.18) for a Fisher regular distribution converges for sample size n increasing to a Chi-squared distribution χ_p^2 with p degrees of freedom, that is

$$2\{l(\hat{\theta}) - l(\theta_0)\} \xrightarrow{a.s.} \chi_p^2, \quad (4.19)$$

where p is the dimension of the parameter θ . \square

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

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

$$l(\theta_0) \approx l(\hat{\theta}) + \underbrace{\frac{\partial l(\hat{\theta})}{\partial \theta^\top}}_{=0} (\theta_0 - \hat{\theta}) + \frac{1}{2} (\theta_0 - \hat{\theta})^\top \left(\frac{\partial^2 l(\hat{\theta})}{\partial \theta \partial \theta^\top} \right) (\theta_0 - \hat{\theta}) \quad (4.20)$$

$$\Leftrightarrow 2\{l(\hat{\theta}) - l(\theta_0)\} \approx -(\theta_0 - \hat{\theta})^\top \left(\frac{\partial^2 l(\hat{\theta})}{\partial \theta \partial \theta^\top} \right) (\theta_0 - \hat{\theta}). \quad (4.21)$$

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 , that is equation (4.21) can be simplified to

$$2\{l(\hat{\theta}) - l(\theta_0)\} \approx -(\theta_0 - \hat{\theta})^\top I(\theta_0)(\theta_0 - \hat{\theta}). \quad (4.22)$$

Using Equation (4.16) we can approximate $\hat{\theta} - \theta_0$ with $I^{-1}(\theta_0)s(\theta_0)$ which allows to simplify (4.22) further to

$$2\{l(\hat{\theta}) - l(\theta_0)\} \approx -s(\theta_0)^\top I^{-1}(\theta_0)s(\theta_0). \quad (4.23)$$

Given (4.14) we can utilize (A.4) to show (4.19). \square

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 clumsy and it is therefore often helpful to transform the parameter and use an unrestricted transformed parameterization of the model instead. In the case of a binomial model, one can use

$$\gamma = \log\left(\frac{\pi}{1-\pi}\right) \Leftrightarrow \pi = \frac{\exp(\gamma)}{1+\exp(\gamma)}.$$

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 now $\gamma = h(\theta)$ for some invertible transformation, so that $\theta = h^{-1}(\gamma)$. The log likelihood for γ is then

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

Differentiation leads to

$$\frac{\partial l_\gamma(\gamma)}{\partial \gamma} = \underbrace{\frac{\partial l_\theta(h^{-1}(\gamma))}{\partial \theta}}_{=0} \underbrace{\frac{\partial h^{-1}(\gamma)}{\partial \gamma}}_{=\theta}.$$

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

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

so that $\hat{\gamma} = h(\hat{\theta})$ results as maximum likelihood estimate 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 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

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

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 Chapter 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 $\mathcal{F} = \{F(\cdot; \theta); \theta \in \Theta\}$ that best matched $G(\cdot)$. This alignment was measured using the Kullback-Leibler divergence as a non-symmetric measure of distance.

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

$$\int \frac{\partial l(\theta_0)}{\partial \theta} dG(y) = E(s(\theta_0)) = 0, \quad (4.24)$$

so that the first Bartlett identity (4.5) remains valid, even though through different reasoning as derived in equation (3.3). However, the second Bartlett identity does not hold. Instead, we obtain the Fisher matrix through

$$I(\theta_0) := - \int \frac{\partial^2 l(\theta_0)}{\partial \theta \partial \theta^\top} dG(y) \quad (4.25)$$

while the variance of the score equals

$$V(\theta_0) := \int s(\theta)s^\top(\theta)dG(y) = \text{Var}(s(\theta_0)). \quad (4.26)$$

In general we have for $G(\cdot) \notin \mathcal{F}$ that $V(\theta_0) \neq I(\theta_0)$. This makes statistical inference slightly more complicated and Property 4.1 extends as follows.

Property 4.4 Assuming a model \mathcal{F} of Fisher-regular distributions and an *i.i.d.* sample Y_1, \dots, Y_n drawn from $Y_i \sim G(\cdot)$. The maximum likelihood estimate is asymptotically normally distributed with

$$\hat{\theta} \stackrel{a}{\sim} N\left(\theta_0, I^{-1}(\theta_0)V(\theta_0)I^{-1}(\theta_0)\right), \quad (4.27)$$

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 (4.25) and the variance $V(\theta_0)$ equals (4.26). \square

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

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

which can then be inserted in (4.27) 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 Chapter 3, the true data-generating process remains unknown in nearly all cases. Hence, it is important to obtain inference on the parameter estimates $\hat{\theta}$ even if $G(\cdot) \notin \mathcal{F}$. 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).

Proof It remains to prove Property 4.4. Taking (4.24) we expand the score around θ_0 which gives

$$\begin{aligned} 0 &= s(\hat{\theta}; y_1, \dots, y_n) \\ &\approx s(\theta_0; y_1, \dots, y_n) - J(\theta_0; y_1, \dots, y_n)(\hat{\theta} - \theta_0) \\ \Leftrightarrow (\hat{\theta} - \theta_0) &\approx J^{-1}(\theta_0; y_1, \dots, y_n)s(\theta_0; y_1, \dots, y_n), \end{aligned} \quad (4.28)$$

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

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

The observed Fisher information can asymptotically be replaced by its expected value given in (4.25) so that (4.28) can be written in short as

$$(\hat{\theta} - \theta_0) \approx I^{-1}(\theta_0)s(\theta_0). \quad (4.29)$$

Reflecting that the score is a sum of independent random variables proves normality following from the central limit theorem where mean and variance directly follow from (4.24), (4.25) and (4.26). \square

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 estimate synergize nicely with a simple method called the Newton-Raphson procedure.

To reiterate, let us outline the Newton-Raphson approach using Figure 4.5. 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 Figure 4.5. 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 $\hat{\theta}$.

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

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

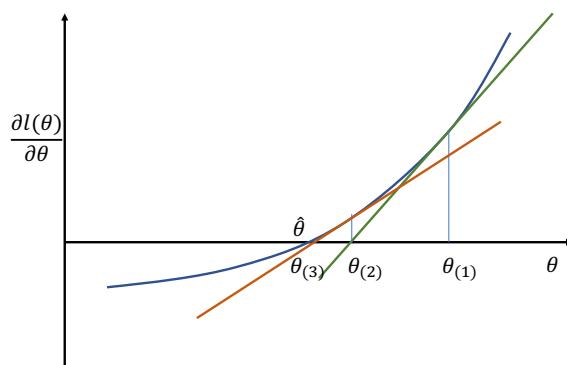
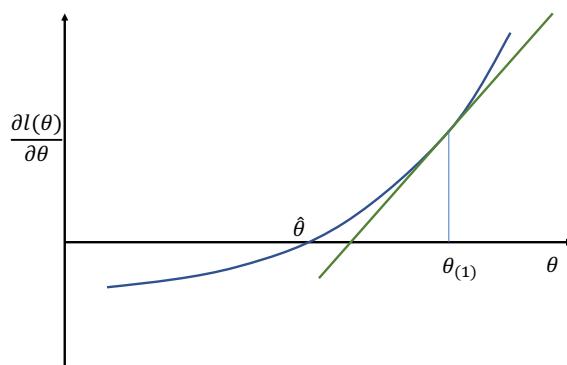
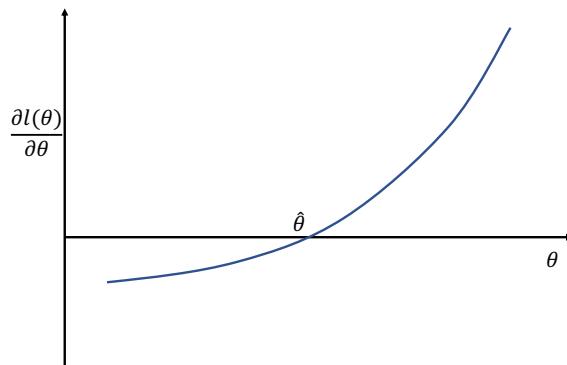


Fig. 4.5 Sketch of Newton-Raphson procedure

Rewriting this gives

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

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 $\theta_{(t+1)} := \theta_{(t)} + I^{-1}(\theta_{(t)})s(\theta_{(t)}; y)$
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 $0 < \delta < 1$:

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

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. Following that, 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 parameter θ , all 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 Figure 4.6 we show some simulated data. The measurements y are thereby generated through $y_i = \exp(z_i)$ with $z_i \sim N(-1, 1)$. Hence, the measurements

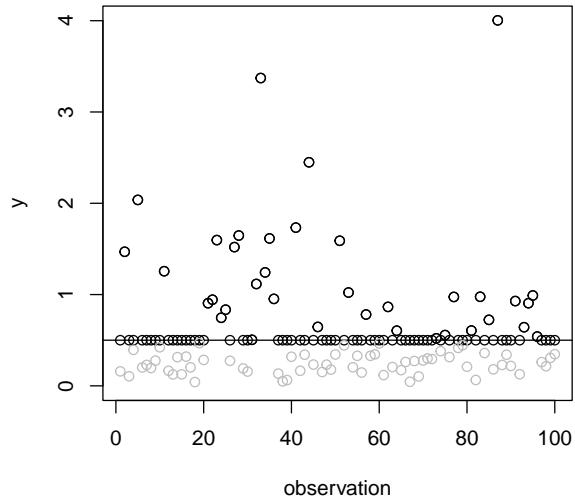


Fig. 4.6 Simulated data with precision threshold. Values below this threshold are set to the threshold. The true simulated values below the threshold are shown as grey points.

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 grey. Note that these values remain unobservable and they are shifted to the threshold level, which in this case is $c = 0.05$.

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 original data and just calculating their mean value, which corresponds to applying an L2 loss for the mean, would lead to the optimization problem

$$\sum_{i=1}^n (z_i - \mu)^2 \rightarrow \min.$$

The resulting mean estimate will be biased upwards 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 the true air pollution. Instead of the squared loss, the likelihood principle would lead to a more elaborated extraction of information. For observations with $y_i > c$ we obtain from the density of a normal distribution the component $-1/2(z_i - \mu)^2$, which corresponds to the L2 loss. 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

$$l(\mu, \sigma^2) = \sum_{i=1}^n \mathbb{I}\{Y_i > c\} \left[-\frac{1}{2} \log \sigma^2 - \frac{1}{2} \frac{(\log(y_i) - \mu)^2}{\sigma^2} \right] \quad (4.30)$$

$$+ \sum_{i=1}^n \mathbb{I}\{Y_i \leq c\} \Phi\left(\frac{\log(c) - \mu}{\sigma}\right), \quad (4.31)$$

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

$$P(Y_i \leq c; \mu, \sigma^2) = P(\underbrace{\log(Y_i)}_{\sim N(\mu, \sigma^2)} \leq \log(c); \mu, \sigma^2) \quad (4.32)$$

$$= P\left(\underbrace{\frac{\log(Y_i) - \mu}{\sigma}}_{\sim N(0,1)} \leq \frac{\log(c) - \mu}{\sigma}\right) \quad (4.33)$$

$$= \Phi\left(\frac{\log(c) - \mu}{\sigma}\right). \quad (4.34)$$

Applying this to the data shown in Figure 4.6 we get the following results. If we use a simple L2 loss for the log measurements z_i we end up with estimates $\hat{\mu} = -0.42$ and $\hat{\sigma}^2 = 0.22$ which clearly does not match the simulation setup. If in fact we use the likelihood as derived above we get $\hat{\mu} = -1.04$ and $\hat{\sigma}^2 = 1.04$ 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 confidently assert 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:

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

This distribution, sometimes also called 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, \dots, X_n) from this distribution. Write the log-likelihood function of the sample.
- b. Find the maximum likelihood estimator (MLE) \hat{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 \hat{p} .

Exercise 3

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

$$f(x_i; \theta) = \frac{1}{\sqrt{2\pi}} \theta^{-\frac{3}{2}} x_i^2 \exp\left(-\frac{x_i^2}{2\theta}\right), \quad \theta > 0, \quad i = 1, \dots, n. \quad (4.35)$$

Further, $\mathbb{E}(X_i) = 0$ and $\text{Var}(X_i) = 3\theta$ 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 fulfills the criteria for Fisher regularity. Show that the expected Fisher information is $I_X(\theta) = \frac{1.5n}{\theta^2}$.
- 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:

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

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

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

Chapter 5

Data Driven Decisions

Statistical inference is commonly understood as the process of decision making based on statistical tests. This is a central pillar-stone in statistical reasoning which boils down to hypotheses tests and/or p -values. We will provide a gentle introduction to the topic which follows the chronological developments of the different ideas. This means we will start with the idea of a p -value before we formally introduce hypotheses tests. Our main intention is not to give a massive collection of different test procedures, but instead to motivate the (simple) main ideas of statistical tests. From there, different test strategies follow immediately. We will thereby not ignore central mathematical results which provide the reasoning for optimal tests. Finally, tests and confidence intervals, as discussed in 3.5, are directly related. To some extent, these are two sides of the same coin, as we will demonstrate.

5.1 Significance Tests and p-value

Our current focus shifts to making decisions based on data, starting with examining the validity of a specific hypothesis. One should therefore bear in mind that we can rarely prove that a hypothesis holds, but we can collect evidence against it. When the evidence opposing the hypothesis becomes compelling, it undermines its validity, leading us to reject it. To simplify this concept, consider the hypothesis that the Earth is flat. We might hold onto this idea until sufficient evidence occurs to contradict it. Linking this idea to the broader theme of measuring uncertainty, our objective is to create a measure that quantifies the evidence against a given hypothesis.

We motivate the idea with a historical example. Pierre-Simon Laplace (1749 - 1827) was certainly a genius. He not only proved the central limit theorem but also made multiple enduring contributions to mathematics, with his findings and methods still influential today. He might even be considered one of the early pioneers in statistics, delving into the exact question we are discussing here: How much evidence is required to reject a hypothesis? Laplace directed his attention to the gender ratio

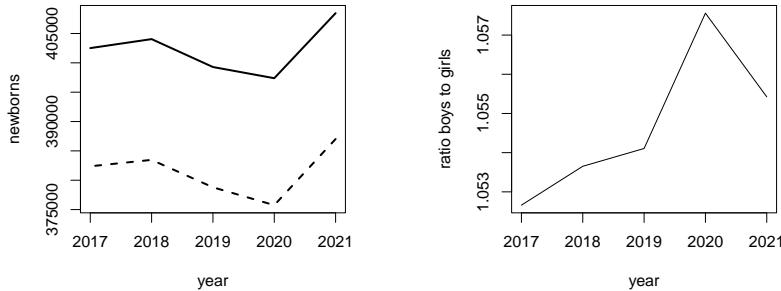


Fig. 5.1 Left-hand side: Number of male (solid line) and female (dashed line) newborns. Right-hand side: proportion of males in the last years.

of newborns, a topic already observed in his time due to the more frequent birth of males. The question arose whether this was mere chance or indicative of a natural law governing human births, suggesting a higher likelihood of male newborns. In those days, natural laws were often tied to religious beliefs and considered as outcomes of divine providence, although Laplace disagreed with this connection. Intriguingly, Laplace compiled data from various parts of Europe, consistently uncovering similar trends. Yet, he hesitated to assert the existence of the aforementioned natural law until a substantial volume of data—considering the standards of that era—was available. During that time, proper methods to quantify evidence against a hypothesis remained unknown, and computational calculations were burdensome. This situation compelled Laplace to derive a range of approximations to address the issue.

To make this more concrete, let us look at the question with current data. Imagine we have a hypothesis suggesting that boys and girls are born in equal proportions. However, in 2021, in Germany, we observed 408,490 boys being born compared to 387,027 girls. How much evidence is in these data against the hypothesis? Refer to Figure 5.1, where we present the count of newborns by gender on the left side and the ratio of male newborns on the right side. Much like Laplace did, we might ask whether there is sufficient evidence to reject the idea of equal gender proportions. But we may apparently also formulate a different hypothesis, namely is the gender ratio constant over time? We see some variation over time in the right-hand side plot, with a peak in the year 2020. Is this peak random or evidence for time variation? Bear in mind, 2020 was the year when the COVID-19 pandemic started. Is there evidence, that this had an influence on the gender ratio of newborns?

To answer these questions more formally we need to define our hypothesis which we subsequently denote as follows.

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

In statistical significance tests, which we develop subsequently, the hypothesis can often be formulated as a hypothesis on a parameter. For instance, we can reformulate the gender question of newborns by assuming a Bernoulli distribution. Let Y_i be the coded gender for a newborn and let $Y_i = 1$ denote a male child and accordingly $Y_i = 0$ for a female child (for mathematical simplicity we ignore diverse newborns 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. It is certainly plausible to assume independence - at least approximately - so that

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

This brings us now in the position to formulate the hypothesis as

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

Next, we need to construct evidence against the hypothesis. It is not difficult to see that $\hat{\pi} = Y/n$ is the maximum likelihood estimate and the further away $\hat{\pi}$ is from $1/2$, the more evidence there is against the hypothesis. While this sounds plausible, the real challenge is now to quantify this evidence. This can be done with the so-called *p*-value, which traces back to the above mentioned analyses of Laplace, but which was formally introduced by Pearson (1900) and heavily promoted by Fisher, see e.g. 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 the particular year. To be concrete, in 2021 we observed $y_{obs} = 408.490$ male babies among all together $n = 795478$ newborns. We denote the observed data with subscript *obs*. If we assume an equal gender share we obtain the *p*-value through

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

where the conditioning in the hypothesis means that we assume $Y \sim B(795478, \pi = 0.5)$. This apparently looks in particular in the direction of excess male births. Exact calculation would require heavy computation, but the central limit theorem simplifies the computation tremendously. Indeed, once Laplace derived the central

limit theorem he could have calculated the p -value for the gender example at that time. Bear in mind that the variance of Y under H_0 equals $n * 0.5^2$ so that simple calculus yields

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

where $\Phi(\cdot)$ is the distribution function of the standard normal distribution. The p -value is more or less zero. Hence, the probability of obtaining new data that contradicts the hypothesis even more than the observed data is equal to zero. Consequently, the numbers that we observed carry strong evidence against the hypothesis. We can therefore reject the hypothesis and 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:

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

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 Figure 5.1. Let Y_t be observed male newborns in year t and $\bar{Y}_t = Y_t / n_t$ the corresponding observed proportion. The hypothesis is now that

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

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 two years t and t' we look at the quantity

$$D_{t,t'} := \bar{Y}_t - \bar{Y}_{t'} \quad (5.2)$$

The larger the absolute difference, that is the larger the absolute value of $D_{t,t'}$, 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 two years t and t' can then be calculated as follows. Let $d_{t,t',obs}$ be the observed values. For instance comparing years 2020 and 2021 we have $\bar{y}_{21} = 0.5134825$ and $\bar{y}_{20} = 0.5139873$ so that $d_{21,20,obs} = -0.0005048$. Note that if H_0 holds then $D_{t,t'}$ should have zero mean since π is the same in both years. This can be utilized and the p -value results through

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

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

$$\widehat{Var}(D_{t,t'}) = \frac{\bar{Y}_t * (1 - \bar{Y}_t)}{n_t} + \frac{\bar{Y}_{t'} * (1 - \bar{Y}_{t'})}{n_{t'}},$$

where the sum follows due to assumed independence between different years. In Figure 5.2 we show the resulting standardized values of $d_{t,t',obs}$ with an over-imposed standard normal distribution for the 10 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 Figure 5.1 can be explained by pure random fluctuation and there is no evidence, that the proportion of male newborns varies over the years. We see that there is no evidence, that the year 2020 was special and the peak visible in Figure 5.1 is fully explainable through randomness.

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 which 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 $Y_i \sim N(\mu, \sigma^2)$, *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 $\bar{Y} = \sum_{i=1}^n Y_i/n$, where apparently large values speak against the hypothesis. Let \bar{y}_{obs} be the observed arithmetic mean. The p -value then results through

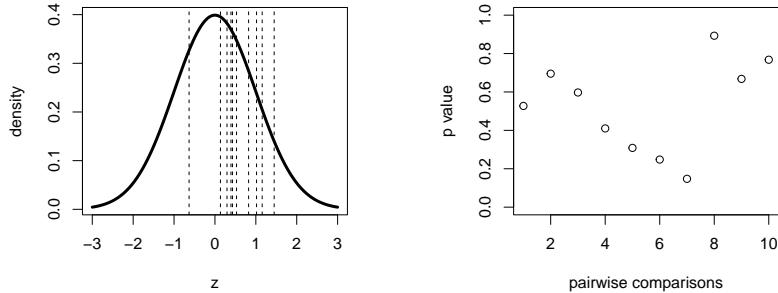


Fig. 5.2 Left-hand side: Standard normal distribution and standardized values of $d_{t,t',obs}$. Right-hand side: Resulting p -values

$$p\text{-value} = P(\bar{Y} > \bar{y}_{obs} | H_0)$$

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

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

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

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

Hence, we find that under the hypothesis the p -value has a uniform distribution. This is an important property and a necessary requirement for the p -value to be unbiased. But what happens if the hypothesis does not hold? In principle, we can pursue the above calculations in a comparable manner. Let μ be the true mean value and assume now that $\mu > \mu_0$. In other words, the hypothesis is void. In this case $Z_{obs} \sim N(0, 1)$. But we can still calculate the distribution of the p -value. Note that

$$\begin{aligned}
\frac{\bar{Y} - \mu}{\sqrt{\sigma^2/n}} &= \underbrace{\frac{\bar{Y} - \mu_0}{\sqrt{\sigma^2/n}}}_{\sim N(0, 1)} - \underbrace{\frac{\mu - \mu_0}{\sqrt{\sigma^2/n}}}_{\sim N\left(\frac{\mu - \mu_0}{\sqrt{\sigma^2/n}}, 1\right)}
\end{aligned}$$

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

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

where $c = (\mu - \mu_0)/\sqrt{\sigma^2/n}$ and z_{1-p} is the $1 - p$ quantile of the $N(0, 1)$ distribution. We show the distribution of the p -value in Figure 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 that small p -values occur if the hypothesis is in fact void.

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

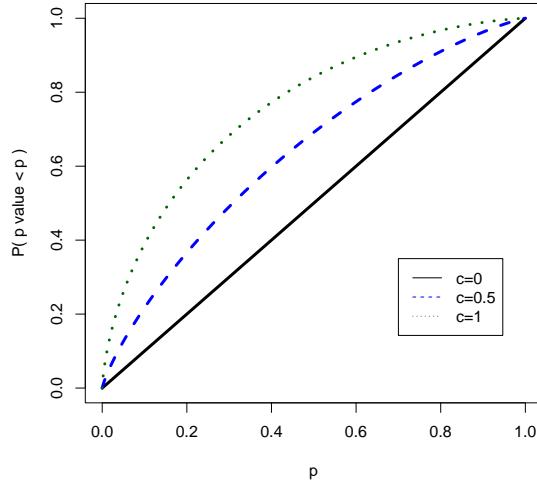


Fig. 5.3 Distribution of p -value for different setting of c

that numerous misconceptions of the approach exist. Goodman (2008) lists twelve typical misunderstandings. For instance: “*if p -value < 0.05 , then there is only a 5% chance that the hypothesis holds*”. This is clearly a wrong statement since no probability statement has been formulated about the validity of the hypothesis. Similarly, and quite a typical error results through “*if the p -value > 0.05 then the hypothesis could be confirmed*”. This is wrong as well, since we can never prove the validity of a hypothesis but can only reject it if there is sufficient evidence against it. Hence, it is important to keep the correct definition of the p -value in mind in order to interpret it properly. The p -value is a concept to quantify evidence against the hypothesis, not more but also not less.

Example 6 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 colours and shapes. To be specific, he mated purebred yellow and round peas with purebred green and wrinkled peas. Both, the yellow colour 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 colour 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 would apply. The original numbers of the experiment that he reported are given in Table 6

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

Table 5.1 Mendel's data

Mendel argued, that the numbers are close to his hypothesis which he used as evidence for the validity of his inheritance law. Apparently, we know that a hypothesis can not be proved but we can calculate the p -value, which is the evidence against the hypothesis. We will demonstrate this later in the chapter when we introduce the Chi-squared test. For now, we make use of the log likelihood function and calculate the log likelihood ratio. To do so we number the possible outcomes from 1 to 4 and assume that each bred pea is characterized by $Y_i \in \{1, 2, 3, 4\}$ for $i = 1, \dots, n$ where $n = 556$ is the total number of peas in the experiment. We assume that Y_i are i.i.d. drawn from a multinomial distribution with cell probabilities $\pi_0 = (9/16, 3/16, 3/16, 1/16)$. This is the hypothesis. We can take the log likelihood function

$$l(\pi) = \sum_{i=1}^n \sum_{k=1}^4 1\{Y_i = k\} \log(\pi_k)$$

with π_k as the k -th element of vector π and $1\{Y_i = k\}$ as 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

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

so that the log likelihood ratio is calculated through $2\{l(\hat{\pi}) - l(\pi_0)\}$. The larger this value, the more evidence exists against the hypothesis. We have derived the asymptotic distribution of the likelihood ratio in Chapter 4.3 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

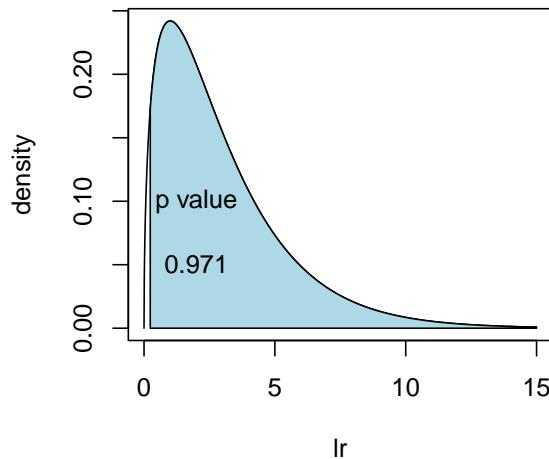


Fig. 5.4 Likelihood ratio of Mendel data with corresponding p -value

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

which is also shown in 5.4. Apparently, the p -value is rather large, hence there is hardly any evidence against the hypothesis. One should bear in mind that at the time when Mendel conducted his experiments, it was unknown how far the data may differ from the hypothesis. Given the large p -value though it might be quite likely, that Mendel reported data from one of his possibly many experiments, which led to the highest concordance with his theory. \triangleright

5.2 Statistical Hypothesis Tests

So far we have questioned the validity of a hypothesis and quantified the evidence against it through the p -value. We want to take a different view now and consider the problem from a decision-theoretic point of view. In this case, we want to decide in favour or against the hypothesis, which requires us to formulate the opposite of the hypothesis.

Definition 5.3 We denote the hypothesis with H_0 and the alternative with H_1 . We assume that one of the two conditions holds.

A data-driven decision now requires that based on data we decide either for H_0 or alternatively for H_1 . Notationally, we put the corresponding decisions in quotes, that is " H_0 " stands for a decision for the hypothesis while " H_1 " gives a decision for the alternative. As in each decision problem, we can make mistakes. If the hypothesis holds but based on the data we conclude with " H_1 ", we are making a mistake. Similarly, if H_1 holds but we decide for " H_0 " we also conduct a mistake. The problem can be written as a decision matrix as shown in Table 5.2. We label the two types of errors as type I and type II errors, respectively. Alternative notation refers to the α and β error, respectively.

truth	decision	
	" H_0 "	" H_1 "
H_0	correct decision	type I error (α error)
H_1	type II error (β error)	correct decision

Table 5.2 Error matrix in significance tests

It is clearly impossible to simultaneously avoid both types of errors. For instance, if we aim to avoid the type I error, the only suitable strategy is to generally decide for " H_0 ". While this would be the correct decision if the hypothesis holds, we will always conduct the type II error if the alternative is valid. 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 random variables. 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 conduct 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

$$P("H_1" | H_0) \leq \alpha \quad (5.3)$$

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

This definition is sufficient to define a test and everything else which is required can be derived conceptually. To demonstrate this, we formulate in general, how a statistical hypothesis test can be carried out by pursuing the following steps.

1. Construct or select a test statistics $t(Y_1, \dots, Y_n)$, which can discriminate between H_0 and H_1 . For instance, large values of $t(Y_1, \dots, Y_n)$ speak in favour of H_1 while small values speak in favour of H_0 . We assume the latter in the following.
2. Derive the distribution of $t(Y_1, \dots, Y_n)$ under H_0 . This step is important and it might even influence the previous step. If the distribution of $t(Y_1, \dots, Y_n)$ is known under H_0 , then a decision rule can be easily derived. We assume this for the moment, that is we know the distribution of $t(Y_1, \dots, Y_n)$ if H_0 holds.
3. The decision rule then results through "reject the hypothesis if $t(Y_1, \dots, Y_n)$ exceeds a threshold c ", that is " H_1 " : $t(Y_1, \dots, Y_n) > c$.
4. It remains to specify the critical value c . Based on (5.3) this is implicitly defined such that

$$P("H_1" | H_0) = P(t(Y_1, \dots, Y_n) > c | H_0) \leq \alpha$$

5. Take the observed value $t_{obs} = t(y_1, \dots, y_n)$ and compare this with c . If $t_{obs} \leq c$ the test decision is " H_0 ", otherwise the test 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 5.1 Assume a statistical hypothesis test with significance level α , then it holds

$$\text{p-value} \leq \alpha \Leftrightarrow "H_1". \quad (5.4)$$

The proof of this statement is rather simple.

Proof Assume we have a test statistic $t()$ such that we decide in favour of H_1 if $t(y_{obs}) > c$, where y_{obs} are the observed data and the critical value c is determined such that

$$P(t(Y) > c | H_0) \leq \alpha.$$

Because

$$\text{p-value} = P(t(Y) > t(y_{obs}) | H_0)$$

it is directly clear that if $t(y_{obs}) > c$, then the p-value $\leq \alpha$ and vice versa. \square

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 favour 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 can not provide and complete list here. There are however some tests which are commonly used which will be discussed in the following. We thereby consider the following setup. We assume that the data are generated through a Fisher regular distribution

$$Y_i \sim f(y; \theta) \quad i.i.d. \quad (5.5)$$

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 cases where $\Theta_0 = \{\theta_0\}$ and $\Theta_1 = \Theta \setminus \{\theta_0\}$ or $\Theta_0 = \{\theta : \theta \leq \theta_0\}$ and $\Theta_1 = \Theta \setminus \Theta_0$.

5.3.1 Likelihood Based Tests

We derived in Chapter 4 asymptotic properties of the maximum likelihood estimate which we will now turn into three test principles. We visualize the idea in Figure 5.5 where we show the likelihood function which takes its maximum at $l(\hat{\theta})$. We investigate the hypothesis $H_0 : \theta = \theta_0$ against the alternative $H_1 : \theta \neq \theta_0$. There are now three possible quantities we can look at to question the validity of the hypothesis. First, we could look at the difference between θ_0 and $\hat{\theta}$. Apparently, if this difference is large it speaks against the hypothesis. This is called the **Wald-test**. Secondly, we could look at the vertical difference of the likelihood at its maximum, i.e. $l(\hat{\theta})$, and the likelihood evaluated under H_0 , i.e. $l(\theta_0)$. Again, if this difference is large it speaks against the hypothesis. This is called the **likelihood-ratio-test**. Finally, we could calculate 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 is large (in absolute terms) it again speaks against the hy-

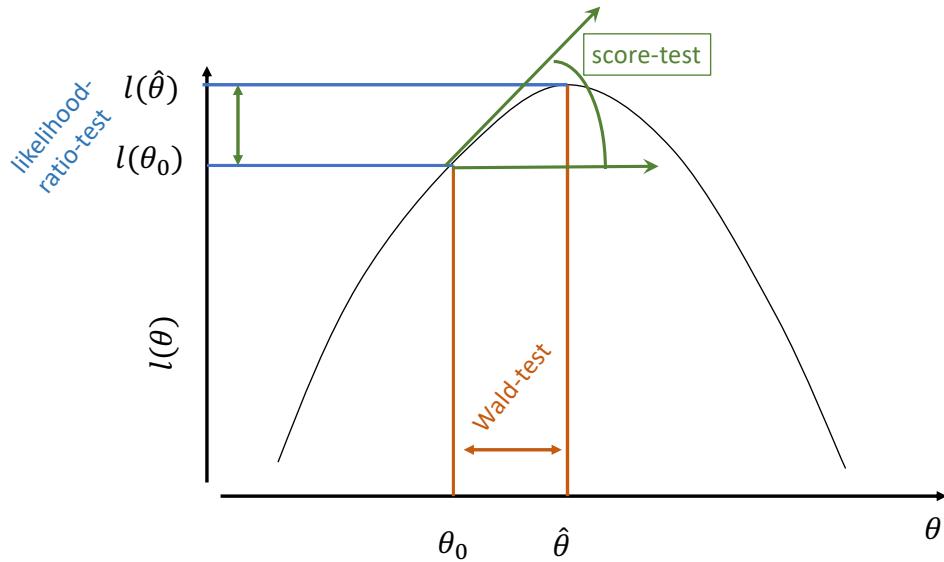


Fig. 5.5 Sketch of Wald-, score- and likelihood-ratio-test

pothesis. This is called the **score-test**. We go through the three tests now step by step.

Note that the maximum likelihood estimate is asymptotically normally distributed, i.e.,

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

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

$$\text{"}H_1\text{"} \Leftrightarrow |\hat{\theta} - \theta_0| > c,$$

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 have estimated the parameter as well as its variance.

Often, the variance is calculated with the Fisher information not at its hypothetical value θ_0 but at the estimated value $\hat{\theta}$. In this case, the critical value is derived from $c = z_{1-\alpha/2} \sqrt{I^{-1}(\hat{\theta})}$. Because under H_0 we know that $\hat{\theta}$ converges to θ_0 , it is plausible to use $I(\hat{\theta})$ instead of $I(\theta_0)$. Given that standard software packages also give $I(\hat{\theta})$ as an output, it is also more convenient to work with the Fisher information calculated at $\hat{\theta}$ and not at θ_0 .

In Chapter 4.4 we derived the asymptotic distribution of the likelihood ratio. This can now be utilized to construct a test from it. In fact, 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 is the dimension of the parameter. Hence, large values of the likelihood ratio $lr(\hat{\theta}; \theta)$ calculated at θ_0 speak against the hypothesis $H_0 : \theta = \theta_0$. This allows us to derive 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 can thereby be calculated 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 is the most powerful test. In other words, it is the best test with respect to having the sharpest decision rule. More details are given in Chapter 5.4.

Finally, from (4.14) we can directly formulate a test on the score. In fact, the score has mean value zero for the true parameter, so we reformulate the hypothesis $H_0 : E(s(\theta_0)) = 0$. This leads to the decision rule

$$\text{"}H_1\text{"} \Leftrightarrow |s(\theta_0; y_1, \dots, y_n)| > c$$

where $c = z_{1-\alpha/2} \sqrt{I(\theta_0)}$. The **score test** is comparable to the Wald test, but there is a subtle difference. For the Wald test, we first need to calculate the maximum likelihood estimate $\hat{\theta}$ to test whether H_0 holds. This is not the case for the score test. Here we simply have to calculate the score function at the hypothetical parameter θ_0 , but the calculation of the maximum likelihood estimate is not required. The score test is therefore particularly useful if the calculation of the maximum likelihood estimate is cumbersome or numerically demanding.

5.3.2 Chi-squared-tests

We focus now on a discrete random variable $V \in \{1, \dots, K\}$ with probability $P(V = k) = \pi_k$, where $\sum_{k=1}^K \pi_k = 1$. Assuming an *i.i.d.* setting for variables V_1, \dots, V_n we obtain a multinomial distribution. Let therefore

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

with $1\{\cdot\}$ as indicator function. We define with $Y = (Y_1, \dots, Y_K)$ the vector with the number of counts in the K categories. Then Y has the distribution

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

A possible hypothesis could be that we assume a particular proportion of the probabilities, like in the Mendel data above, see Example 6. Let therefore π_0 be the hypothetical value and consider the hypothesis $H_0 : \pi = \pi_0$ where $\pi = (\pi_1, \dots, \pi_K)$ and $\pi_0 = (\pi_{01}, \dots, \pi_{0K})$. This allows to calculate the so-called **Chi-squared statistics**

$$X^2 = \sum_{k=1}^K \frac{(y_k - e_k)^2}{e_k}. \quad (5.6)$$

where $e_k = n\pi_{0k}$ are the expected case under the hypothesis H_0 which are set in relation to the observed cases y_k . It can be shown that X^2 is Chi-squared distributed with $K - 1$ degrees of freedom. We provide a sketch of the proof at the end of this section.

Tests, which are based on the Chi-squared statistics X^2 are generally known as Chi-squared tests. The test statistic X^2 given in (5.6) has an intuitive structure. We compare the expected counts e_k , which we would expect if the hypothesis holds, with the observed counts y_k . We take the square to make the difference positive and we divide it by the expected number in that cell. The latter makes sense since a small difference between observed and expected matters more in cells where we don't have many observations, compared to cells with many expected observations. As a result, the division by the expected counts can be seen as a form of standardization.

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

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

and want to test the hypothesis of independence, that is

$$H_0 : P(Z_{i1} = k, Z_{i2} = l) = P(Z_{i1} = k) * P(Z_{i2} = l)$$

To do so we apply the Chi-squared-test to the data which can be written as a contingency table:

		Z_2						
		1	2	\dots	K_2			
Z_1	1	n_{11}	n_{12}		n_{1K_2}	$n_{1\cdot}$		
	2	n_{21}	n_{22}		n_{2K_2}	$n_{2\cdot}$		
	\vdots					\vdots		
	K_1	$n_{K_1 1}$	$n_{K_1 2}$		$n_{K_1 K_2}$	$n_{K_1 \cdot}$		
		$n_{\cdot 1}$	$n_{\cdot 2}$	\dots	$n_{\cdot K_2}$	$n_{\cdot \cdot}$		

With n_{kl} we define the number of observations with $Z_{i1} = k$ and $Z_{i2} = l$, i.e.

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

Moreover, we define $n_{\cdot l} = \sum_k n_{kl}$ as the column sum, $n_{k\cdot}$ as the row sum and $n_{\cdot \cdot} = n$ as the sample size. We now compare the observed counts to their expected versions under independence. Let $\pi_{kl} = P(Z_{i1} = k, Z_{i2} = l)$, which in case of independence decomposes to $\pi_{k\cdot} \pi_{\cdot l} = P(Z_{i1} = k) \cdot P(Z_{i2} = l)$. It is not difficult to see that the maximum likelihood estimates under independence equal

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

such that the expected counts under independence are

$$e_{kl} = n_{\cdot \cdot} \hat{\pi}_{k\cdot} \hat{\pi}_{\cdot l} = \frac{n_{k\cdot} n_{\cdot l}}{n_{\cdot \cdot}}$$

This leads to the Chi-squared statistic

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

The distribution of X^2 can again be approximated with a Chi-squared distribution, but to obtain the correct number for the degrees of freedom we need to be a little bit careful. Indeed, when the test ideas were introduced by Pearson (1904) and further developed by Fisher (1922), this initiated a harsh debate on the correct specification of the degree of freedom, see Baird (1983). Based on Fisher's results the correct calculation of the degree of freedom follows the rule

$$df = \text{Number of cells} - \text{number of constraints} - \text{number of estimated parameters}.$$

For the Chi-squared independence test, we have $K_1 * K_2$ cells. The number of side constraints is 1, since all cell probabilities need to sum up to 1. We estimated the $K_1 - 1$ probabilities $\hat{\pi}_k$, for $k = 1, \dots, K_1 - 1$, which defines the estimate for the last category through $\hat{\pi}_{K_1} = 1 - \sum_{k=1}^{K_1} \hat{\pi}_k$. Analogously we estimated $K_2 - 1$ parameters $\hat{\pi}_{l,l}$ for $l = 1, \dots, K_2 - 1$. This cumulates to $K_1 + K_2 - 2$ estimated parameters. Hence, following the rule to calculate the degrees of freedom, we obtain

$$K_1 * K_1 - 1 - (K_1 + K_2 - 2) = (K_1 - 1) * (K_2 - 1)$$

as the correct degree of freedom.

Proof We sketch the proof that X^2 defined in (5.6) follows a Chi-squared statistics with $K - 1$ degrees of freedom. Bear in mind that the hypothesis relates to the first $K - 1$ parameters since $\pi_K = 1 - \pi_1 - \dots - \pi_{K-1}$. We collect these $K - 1$ components in parameter vector θ . A test can then be constructed utilizing the above results on the likelihood based tests. In fact, we can directly utilize the likelihood-ratio-test and take the test statistics

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

where

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

The likelihood ratio can be approximated with expansions for the log as follows.

$$\begin{aligned} 2 \sum_{k=1}^K y_k \log \left(\frac{\hat{\pi}_k}{\pi_{0k}} \right) &= -2 \sum_{k=1}^K y_k \log \left(\frac{\pi_{0k}}{\hat{\pi}_k} \right) \\ &= -2 \sum_{k=1}^K y_k \log \left(1 + \frac{\pi_{0k} - \hat{\pi}_k}{\hat{\pi}_k} \right) \\ &\approx -2 \sum_{k=1}^K y_k \left(\frac{\pi_{0k} - \hat{\pi}_k}{\hat{\pi}_k} \right) + \sum_{k=1}^K y_k \left(\frac{\pi_{0k} - \hat{\pi}_k}{\hat{\pi}_k} \right)^2. \end{aligned}$$

Note that $\hat{\pi}_k = \frac{y_k}{n}$, so that the first term above results in zero since both, π_{0k} as well as $\hat{\pi}_k$ sum up to one. Defining the expected counts with $e_k = n\pi_{0k}$ simplifies the second term above, by again using $\hat{\pi}_k = \frac{y_k}{n}$

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

If we replace the denominator y_k by its expectation under the hypothesis we get the Chi-squared statistics (5.6). Given that approximation relation of X^2 to the likelihood ratio we can state that if π_0 are true cell probabilities then X^2 follows a Chi-squared distribution with $K - 1$ degree of freedom \square

5.3.3 Distribution-Free Tests

A large class of tests are so-called distribution-free tests, occasionally also labelled as nonparametric tests. The terminology is not unique and in fact confusing, since nonparametric concepts in statistics also relate to functional estimation. We will therefore motivate the ideas here using the terminology distribution free tests. It implies, that we do not assume a concrete distributional model for the data at hand but rely on a general data-generating process (2.3). The focus of interest is now on a hypothesis related to the distribution $G(\cdot)$, e.g. questioning whether the median takes a particular value. We present the idea of distribution-free tests here for a more general setup where we assume two samples

$$Y_i \sim G_Y(\cdot), i = 1, \dots, n \text{ and } X_j \sim G_X(\cdot), j = 1, \dots, m.$$

We question whether the two distributions are the same, that is

$$H_0 : G_Y(\cdot) \equiv G_X(\cdot).$$

This is quite a general hypothesis that simplify further and instead, we question whether the two distributions have the same median. To construct a test statistics we take the pooled sample $z = (y_1, \dots, y_n, x_1, \dots, x_m)$ and order this sample such that

$$z_1 \leq z_2 \leq z_3 \leq \dots \leq z_{n+m}$$

where z_j is either from the Y or the X sample. Let $rg(x_j)$ denote the rank of the j th observation in the above z data. That is $rg(x_j) = k$ if and only if $z_k = x_j$, assuming that there are no ties, i.e. all observations are pairwise different. Figure 5.6 motivates the idea. The test statistics then result through

$$T = \sum_{j=1}^m rg(x_j).$$

It is easy to see that large and small values of T speak against the hypothesis that the median of the $G_Y(\cdot)$ coincides with the median of the $G_X(\cdot)$. The distribution under H_0 is known as the Wilcoxon distribution and the test carries the name **Wilcoxon-test**, referring to Wilcoxon (1945). The decision rule is

$$\text{"}H_1\text{"} \Leftrightarrow (T < w_{\alpha/2}) \text{ or } (T > w_{1-\alpha/2})$$

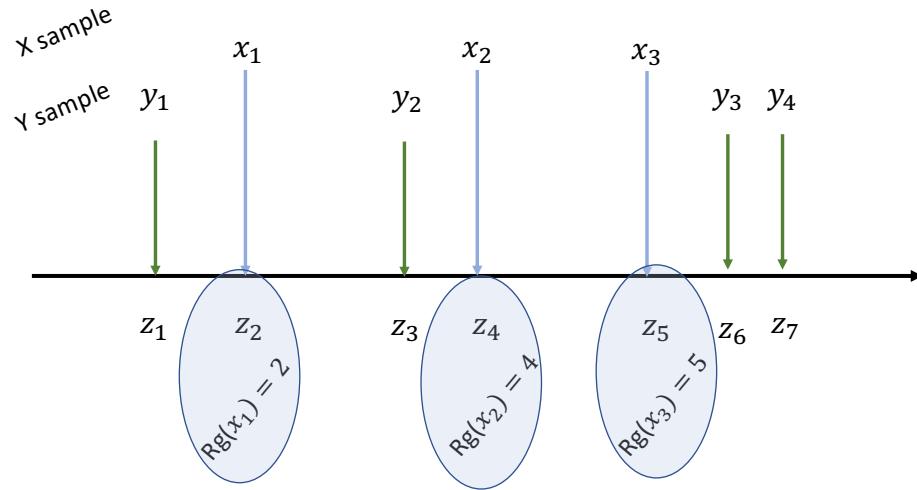


Fig. 5.6 Sketch of Wilcoxon test statistics

where w_α is the α Quantile of the matching Wilcoxon distribution. The test is implemented in most statistics software packages and hence can be easily carried out practically.

5.3.4 Kolmogorov-Smirnov Test

The Kolmogorov-Smirnov test serves as another example of a distribution-free test. This test enables a direct examination of whether two distributions are identical or not. In this context, we are introducing the test for situations where our goal is to assess if the data originates from a predetermined model. With the general setup (2.3) we aim to test the hypothesis

$$H_0 : G(\cdot) = F(\cdot; \theta).$$

For the moment we assume that θ is given and known.

The test builds upon the difference between the empirical distribution function and the hypothetical distribution. With $F_n(y) = \frac{1}{n} \sum_{i=1}^n 1_{\{Y_i \leq y\}}$ as empirical cumulative distribution function we question whether the distance between $F(y; \theta)$ and $F_n(y)$ is small, at least when n is large. To do so we construct a test statistic by looking at the difference between $F(y; \theta)$ and $F_n(y)$. To be specific, we consider the supremum of the absolute difference

$$D_n := \sup_y |F_n(y) - F(y; \theta)|. \quad (5.7)$$

This quantity is also called the Kolmogorov-Smirnov-Distance. Figure 5.7 illustrates the construction of D_n with an example. The empirical distribution function of a sample is given as $F_n(\cdot)$ and the standard normal distribution as hypothetical distribution $F(\cdot)$. This is shown in the top plot. The bottom plot gives the difference $F_n(y) - F(y)$ for all values of y , which takes its maximum (in absolute terms) at the locations indicated by the vertical line. This defines the concrete value of D_n .

The next step is to find the distribution for D_n under H_0 , i.e. if $F(y; \theta)$ is the true distribution. It can be shown that the distribution of D_n does not depend on the concrete distribution $F(\cdot; \theta)$ but in fact has a general distribution. This in turn means that D_n follows some pivotal distribution.

Property 5.2 The distribution of D_n given in (5.7) (asymptotically) follows a Kolmogorov-Smirnov distribution. In particular, the distribution is pivotal, that is it does not depend on $F(\cdot; \theta)$. \square

The proof is given at the end of this section. The property states that the test decision is now given by

$$\text{“}H_1\text{”} \Leftrightarrow D_n \geq KS_{1-\alpha},$$

where $KS_{1-\alpha}$ denotes the $1 - \alpha$ quantile of the Kolmogorov-Smirnov distribution.

The above test assumes that the parameter θ of the distribution is known, which in most practical settings is not the case. Hence, the parameter needs to be estimated and the derivation of the exact distribution of the test statistic with estimated parameter

$$\hat{D}_n = \sup_y |F_n(y) - F(y; \hat{\theta})|$$

is thereby complicated and has only been determined for specific distributions, such as the normal or exponential distributions. For an unspecified distribution, the Kolmogorov-Smirnov distribution only holds asymptotically for D_n . For more details, we refer to Berger and Zhou (2014).

Proof We show now that the distribution of D_n is pivotal. Let therefore $F^{-1}(y; \theta)$ be the inverse of the distribution function, or more formally for $x \in [0, 1]$

$$F^{-1}(x; \theta) = \min\{y : F(y; \theta) \geq x\}.$$

Then for $y = F^{-1}(x; \theta)$ we get $F(F^{-1}(x; \theta); \theta) = x$ such that

$$P(\sup_y |F_n(y) - F(y; \theta)| \leq t) = P(\sup_x |F_n(F^{-1}(x; \theta)) - x| \leq t) \quad (5.8)$$

where

$$F_n(F^{-1}(x; \theta)) = \frac{1}{n} \sum_{i=1}^n 1_{\{Y_i \leq F^{-1}(x; \theta)\}} = \frac{1}{n} \sum_{i=1}^n 1_{\{F(Y_i; \theta) \leq x\}}.$$

Note that if $F(y; \theta)$ is the true distribution, then $F(Y_i; \theta)$ has a uniform distribution on $[0, 1]$. This is easily seen as

$$P(F(Y_i; \theta) \leq x) = P(Y_i \leq F^{-1}(x; \theta)) = F(F^{-1}(x; \theta); \theta) = x.$$

Hence, we define $U_i = F(Y_i; \theta)$, which is uniform on $[0, 1]$, which in turn allows us to rewrite (5.8) as

$$P(\sup_x \left| \frac{1}{n} \sum_{i=1}^n 1_{\{U_i \leq x\}} - x \right| \leq t).$$

Note that this probability statement does not depend on the hypothetical distribution and thus we have formulated a pivotal distribution. \square

5.4 Power of a Test

In Figure 5.3 we showed the distribution of the p -value in constellations where the hypothesis is not true. We want to elaborate on this further now. So far we have only looked at the performance of the test under the validity of the hypothesis, that is if H_0 holds. This led to the p -value defined in 5.2 and the hypothesis tests introduced in 5.4. The view of only looking at the hypothesis can be seen as a restriction, which is easily demonstrated with the following pathological test. Consider any hypothesis, no matter what, which is supposed to be tested by analysing data. Instead of looking at the data at all, we base the test decision by drawing 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, it holds $P("H_1" | H_0) = P("H_1") = P(U \leq \alpha) = \alpha$ so this yields a valid statistical hypothesis test based on definition (5.3). Needless to say the test is useless. Hence, condition (5.3) is not sufficient to define a test that is good, useful, or both. In general, we need to also look at the type II error. In fact, a test is good, if the type II error occurs only with a small probability. To completely understand this setup we need a couple of technical definitions. First, we define the power of a test.

Definition 5.5 The power of a statistical hypothesis test is defined as $P("H_1" | H_1)$.

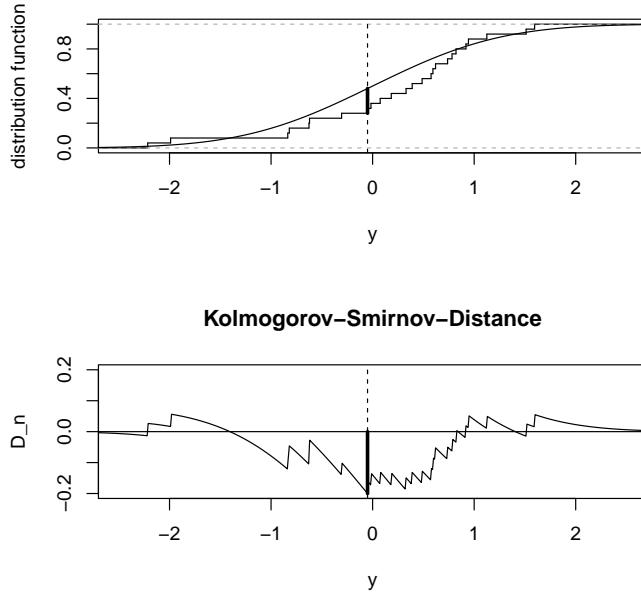


Fig. 5.7 Empirical distribution function and hypothetical distribution (top plot). The difference between these two distributions (bottom plot) with the maximum (supreme) value is indicated as a vertical line.

The power of a test quantifies the probability of correctly rejecting the null hypothesis if H_1 holds. This is best understood with a simple example. Assume *i.i.d.* normally distributed data $Y_i \sim N(\mu, \sigma^2)$ with σ^2 known and consider the hypothesis $H_0 : \mu \leq \mu_0$, where for simplicity we set $\mu_0 = 0$. With the results derived above it is easy to obtain the decision rule:

$$\text{“}H_1\text{”} \Leftrightarrow \bar{Y} \geq \mu_0 + z_{1-\alpha} \frac{\sigma}{\sqrt{n}}.$$

The power of the test is now easily calculated through

$$\begin{aligned}
P("H_1"|\mu) &= P\left(\bar{Y} \geq \mu_0 + z_{1-\alpha} \frac{\sigma}{\sqrt{n}} |\mu\right) \\
&= P\left(\frac{\bar{Y} - \mu_0}{\sigma/\sqrt{n}} \geq z_{1-\alpha} |\mu\right) \\
&= P\left(\frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \geq z_{1-\alpha} + \frac{\mu_0 - \mu}{\sigma/\sqrt{n}} |\mu\right) \\
&= 1 - \Phi\left(z_{1-\alpha} + \frac{\mu_0 - \mu}{\sigma/\sqrt{n}}\right).
\end{aligned}$$

We can plot $P("H_1"|\mu)$ against μ , which is shown in Figure 5.8. We show the function for values $-1 \leq \mu \leq 1$ and two data sizes, namely $n = 30$ and $n = 60$. The top plot shows the entire function. Note that for values $\mu \leq \mu_0$ the power is less or equal to α , which follows from definition (5.3). The maximum is achieved for $\mu = \mu_0$. Hence, we may ignore the left-hand side of the plot since we know that for all values $\mu \leq \mu_0$ we have that $P("H_1"|\mu) \leq P("H_1"|\mu_0) \leq \alpha$. We visualize this in the middle plot in Figure 5.8 by erasing the function values for $\mu \leq \mu_0$. We also see that for $\mu > \mu_0$ we obtain $\min P("H_1"|\mu) = \alpha$, or putting it differently the maximum probability for a type II error is $1 - \alpha$. This occurs at $\mu = \mu_0 + \delta$ with δ infinitely small but positive and it holds for both sample sizes. But if we set δ to some value bounded away from zero, we see that $P("H_1"|\mu = \mu_0 + \delta)$ increases with increasing sample size. Hence, the probability for the type II error $P("H_0"|"H_1")$ decreases. It gets largest at $\mu = \mu_0 + \delta$. If we set $\delta = 0.25$, we see a difference in the power function for the different sample sizes. In fact, the only relevant point of the power function on the x axis is $\mu = \mu_0 + \delta$. This is visualized in the bottom plot in Figure 5.8. The quantity δ determines the relevant difference that one wants to detect. Once this is fixed we can determine the required data size in order to limit the type II error. This is in particular used in experiments and the calculations are sketched below following the normal distribution example.

In Figure 5.9 we focus on the relevant quantities from Figure 5.8 and show $P("H_1"|\mu = \mu_0 + \delta)$ for different data sizes, namely $n = 60$, $n = 90$ and $n = 120$. If we postulate the type II error should be less than $\beta = 0.2$, we need the power function to exceed value $1 - \beta$ at $\mu = \mu_0 + \delta$. Setting $\beta = 0.2$ this is visualized in Figure 5.9 with the blue shade in the top right corner. Now we can control both, the type I, as well as the type II error, and apparently, the required sample size is between $n = 90$ and $n = 120$. In fact, we can calculate the required data size exactly by solving the following implicit equation with respect to n :

$$n_0 = \arg \min \{n : 1 - \beta \leq P("H_1"|\mu = \mu_0 + \delta)\} \quad (5.9)$$

$$= \arg \min \left\{n : 1 - \beta \leq 1 - \Phi\left(z_{1-\alpha} + \frac{\mu_0 - (\mu_0 + \delta)}{\sigma/\sqrt{n}}\right)\right\} \quad (5.10)$$

$$= \arg \min \left\{n : 1 - \beta \leq 1 - \Phi\left(z_{1-\alpha} + \frac{-\delta}{\sigma/\sqrt{n}}\right)\right\}. \quad (5.11)$$

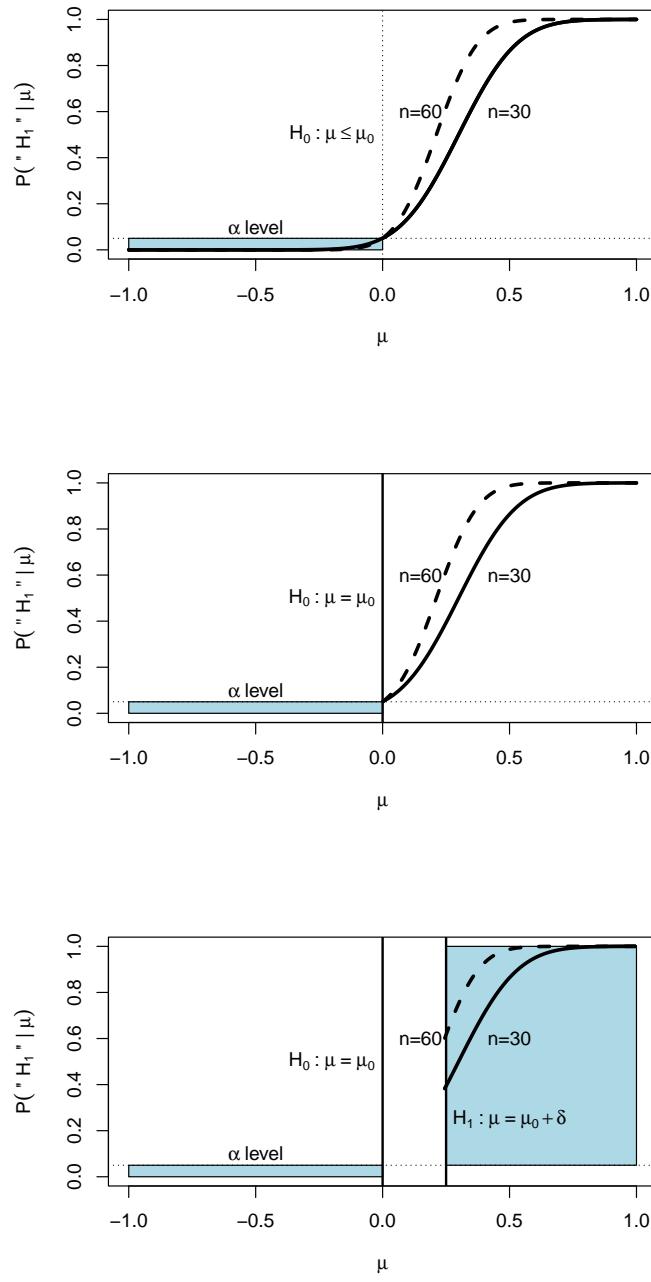


Fig. 5.8 Power of a test on $H_0 : \mu \leq \mu_0$ with $\mu_0 = 0$. Top plot shows $P(H_1 | \mu)$ for all values of μ . The middle plot shows $P(H_1 | H_1)$ and the bottom plot shows the relevant power $P(H_1 | \mu = \mu_0 + \delta)$, demonstrating the effect of the sample size.

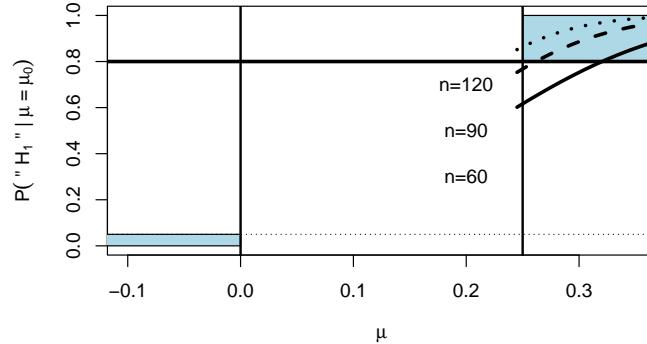


Fig. 5.9 Power of a test on $H_0 : \mu \leq \mu_0$ with $\mu_0 = 0$. Top plot shows $P(H_1 | \mu)$ for all values of μ . Middle plot shows $P(H_1 | H_1)$ and bottom plot shows the relevant power $P(H_1 | \mu = \mu + \delta)$, demonstrating the effect of the sample size.

In the concrete example, we get $n_0 = 99$ for $\beta = 0.2$.

5.5 Neyman-Pearson Lemma

The next question that we want to tackle is, whether there is an optimal test. Looking at Figure 5.8 this means we are looking for a test which fulfils (5.3) and has the largest power, i.e. which maximizes $P(H_0 | \mu = \mu + \delta)$. This is derived from the following lemma.

Property 5.3 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 $P(\text{“}H_1\text{”}|H_0) \leq \alpha$. This test is called the **Neyman-Pearson test**. □

The Lemma was published by Neyman and Pearson (1933) and the result demonstrates once more the central role of the likelihood. Utilizing the distribution of the likelihood ratio as in (4.19) we can easily quantify the threshold value c using a Chi-

squared distribution. The above results motivate the use of the likelihood ratio test defined in Section 5.3.1, as this is the most powerful test. The proof of the Lemma above is not complicated 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

$$P(\psi(Y) = 1; \theta_1) \leq P(\varphi(Y) = 1; \theta_1),$$

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

$$P(\varphi(Y) = 1; \theta_1) - P(\psi(Y) = 1; \theta_1) = \int \{\varphi(y) - \psi(y)\} f(y; \theta_1) dy.$$

We have to prove that this integral is greater or equal to zero. The above integral can be labelled over three regions as

$$\begin{aligned} R_1 &= \left\{ y : \frac{f(y; \theta_0)}{f(y; \theta_1)} < k \right\} \\ R_2 &= \left\{ y : \frac{f(y; \theta_0)}{f(y; \theta_1)} > k \right\} \\ R_3 &= \left\{ y : \frac{f(y; \theta_0)}{f(y; \theta_1)} = k \right\}. \end{aligned}$$

For region R_1 we have $\varphi(y) \equiv 1$ and $f(y; \theta_1) > f(y; \theta_0)/k$, such that

$$\int_{y \in R_1} [\varphi(y) - \psi(y)] f(y; \theta_1) dy \geq \frac{1}{k} \int_{y \in R_1} [\varphi(y) - \psi(y)] f(y; \theta_0) dy.$$

For region R_2 we have $\varphi(y) \equiv 0$ and $-f(y; \theta_1) > -f(y; \theta_0)/k$, such that

$$\begin{aligned} \int_{y \in R_2} [\varphi(y) - \psi(y)] f(y; \theta_1) dy &= - \int_{y \in R_2} \psi(y) f(y; \theta_1) dy \\ &\geq -\frac{1}{k} \int_{y \in R_2} \psi(y) f(y; \theta_0) dy = \frac{1}{k} \int_{y \in R_2} [\varphi(y) - \psi(y)] f(y; \theta_0) dy. \end{aligned}$$

And finally for $y \in R_3$ we have $f(y; \theta_1) = f(y; \theta_0)/k$, such that

$$\int_{y \in R_3} [\varphi(y) - \psi(y)] f(y; \theta_1) dy = \frac{1}{k} \int_{y \in R_3} [\varphi(y) - \psi(y)] f(y; \theta_0) dy.$$

Collecting the right-hand sides of the three regions we can conclude

$$\int [\varphi(y) - \psi(y)] f(y; \theta_1) dy \geq \frac{1}{k} \int [\varphi(y) - \psi(y)] f(y; \theta_0) dy. \quad (5.12)$$

As both tests have a significance level of α one obtains

$$\alpha = P(\text{"H}_1\text{"} | H_0) = \int \varphi(y) f(y; \theta_0) dy = \int \psi(y) f(y; \theta_0) dy,$$

such that the right-hand side in (5.12) is equal to zero and the proof is completed. \square

5.6 Testing and Confidence Intervals

We have now introduced two concepts for statistical inference, namely confidence intervals and statistical hypothesis tests. This raises the question of whether the two ideas are related. In fact, we can show that the two approaches are not only related but they can also be interchanged. This means, that given a confidence interval, one can directly construct a corresponding test and vice versa.

Let us motivate this and reconsider the definition of a confidence interval given in (3.14) in definition 3.8. The confidence interval was defined as $[t_l(y), t_r(y)]$ such that

$$P_\theta(t_l(Y) \leq \theta \leq t_r(Y)) \geq 1 - \alpha,$$

where $Y = (Y_1, \dots, Y_n)$. Taking this interval we define a statistical hypothesis test on the hypothesis $H_0 : \theta = \theta_0$ through the decision rule

$$\text{"H}_1\text{"} \Leftrightarrow \theta_0 \notin [t_l(y), t_r(y)].$$

Hence, for all parameter values which are not in the confidence interval, we reject the hypothesis while parameter values lying in the confidence interval are not rejected. This test construction defines a statistical hypothesis test with significance level α . To see this, we can 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

$$\begin{aligned}
P(\text{"H}_1\text{"}|H_0) &= 1 - P(\text{"H}_0\text{"}|H_0) \\
&= 1 - P(\varphi(Y) = 0|\theta = \theta_0) \\
&= 1 - \underbrace{P(t_l(Y) \leq \theta_0 \leq t_r(Y))}_{\geq 1-\alpha} \leq \alpha.
\end{aligned}$$

The above construction principle takes a confidence interval and constructs a hypothesis test from it. We can also go in the other direction and take a hypothesis test and construct a confidence interval based on the test. Assume we have a test, which we can define as

$$\varphi_\theta(Y) = \begin{cases} 0 & \text{if "H}_0\text{"} \\ 1 & \text{if "H}_1\text{"}. \end{cases}$$

The corresponding confidence interval can then be calculated with

$$CI(y) = \{\theta : \varphi_\theta(y) = 0\}.$$

Since $P(\text{"H}_1\text{"}|H_0) = P(\varphi_\theta(Y) = 1|\theta) \leq \alpha$ we have

$$\begin{aligned}
P(\theta \in CI(Y)|\theta) &= 1 - P(\theta \notin CI(Y)|\theta) \\
&= 1 - P(\varphi_\theta(Y) = 1|\theta) \\
&= 1 - P(\text{"H}_1\text{"}|H_0) \geq 1 - \alpha.
\end{aligned}$$

We see that confidence intervals and statistical hypothesis tests are related and are two sides of the same coin.

5.7 Multiple Testing

So far we have looked at univariate or at least low-dimensional parameters and a single hypothesis. But what happens if we have a multivariate parameter $\theta = (\theta_1, \dots, \theta_p)$ and multiple hypotheses of the form

$$H_{0j} : \theta_j = \theta_{0j}$$

with $j \in \mathcal{J} \subset \{1, \dots, p\}$ and $\theta_0 = (\theta_{01}, \dots, \theta_{0p})$. Assume that $|\mathcal{J}| = m$, that is we aim to test m components out of θ at the same time with m hypotheses. Instead of conducting a test on all m components simultaneously, we test each of the m

hypotheses separately. Let p_1, \dots, p_m be the p -values of each of these m different tests leading to the m test decision rules

$$\text{"}H_{1j}\text{"} \Leftrightarrow p_j \leq \alpha \quad j = 1, \dots, m.$$

For every single test, we have a statistical hypothesis test in that

$$P(\text{"}H_{1j}\text{"}|H_{0j}) \leq \alpha. \quad (5.13)$$

But what can be said about the rejection probability for all tests simultaneously?

Let us look at the problem from a different angle. For each test we have an error matrix as in Table 5.3. Putting all m tests together leads to an error matrix as shown in Table 5.3.

	non rejected hypotheses i.e. " H_{0j} "	rejected hypotheses i.e. " H_{1j} "	
true hypotheses H_{0j}	U	V	M_0
false hypotheses, i.e. H_{1j}	T	S	M_1
	$m - r$	r	m

Table 5.3 Error matrix for multiple testing

Note that the entries in the table are unknown since we do not know whether the hypotheses are true or false. This justifies using capital letters. The same holds for the margins. The only number we know is m , the number of tests, and once the tests are conducted we know r , the number of rejected hypotheses out of the m tested hypothesis. Apparently, if $r = m$ or $r = 0$, that is all hypotheses are rejected or not rejected, respectively, we obtain knowledge about further entries in the table. In general, however, the inner part of the table remains unknown.

Multiple testing relates now to the problem of looking at the column " H_{1j} " and focuses on the question of what can be said about S and V . To do so we define the family-wise error rate.

Definition 5.6 The **family-wise error rate** (FWER) is defined as

$$\alpha_{FWER} := P(V > 0 | H_{0j} \text{ for } j = 1, \dots, m), \quad (5.14)$$

with V as defined in the Table 5.3.

We can interpret the FWER as the probability of falsely rejecting at least one of the m hypotheses if all hypotheses hold. It is important to relate the FWER to the decision rules (5.13). In fact, we will show that $\alpha_{FWER} \geq \alpha$, and even worse, if the separate tests are not completely dependent we will have $\alpha_{FWER} \rightarrow 1$ as m is increasing. To see this note that

$$\begin{aligned}\alpha_{FWER} &= P(V > 0 | H_0) \text{ for } j = 1, \dots, m \\ &= P((p_1 \leq \alpha) \vee \dots \vee (p_m \leq \alpha) | H_0) \text{ for } j = 1, \dots, m \\ &= 1 - P((p_1 > \alpha) \wedge \dots \wedge (p_m > \alpha) | H_0) \text{ for } j = 1, \dots, m\end{aligned}\quad (5.15)$$

where “ \vee ” denotes the logical “or” and “ \wedge ” is the logical “and”. If we assume independence of the tests, the last row (5.15) simplifies further to

$$1 - \prod_{j=1}^m P((p_j > \alpha) | H_0) = 1 - (1 - \alpha)^m \quad (5.16)$$

utilizing the fact that the p -value has a uniform distribution if the hypothesis holds. Now it is easy to see that as m increases, the FWER α_{FWER} converges to 1. In particular, we find that $\alpha_{FWER} > \alpha$. This is to say, that if m independent tests are carried out with significance level α , then the FWER α_{FWER} is always larger than the significance level α used for each single test in (5.13).

We may also postulate that the FWER α_{FWER} should not exceed a given significance level and question how to adjust the significance levels for the individual tests so that an overall level α_{FWER} remains small. It can be shown that α_{FWER} gets the largest if the separate tests are independent. Hence we obtain in general

$$\alpha_{FWER} \geq 1 - (1 - \alpha)^m \Leftrightarrow 1 - \alpha_{FWER} \leq (1 - \alpha)^m \quad (5.17)$$

It can further be shown that for α small we have $(1 - \alpha)^m > (1 - m\alpha)$ so that if we adjust the significance level for the separate tests to $\alpha_{adjust} = \alpha/m$ we obtain that the FWER is smaller or equal to a general level α . This adjustment is called Bonferroni adjustment, which goes back to Bonferroni (1936).

Property 5.4 Bonferroni adjustment. *The FWER α_{FWER} is limited to α , if we set the significance level of each individual test in (5.13) to α/m , where m is the number of tests. The adjusted significance level is defined by $\alpha_{adjust} = \alpha/m$.* \square

Though the Bonferroni adjustment is simple, it is not always recommendable. In particular, if m is large other procedures are more advisable. We refer to Efron

(2010) for further details. An alternative is to look at the ratio V/r in the table above, which is called the *False Discovery Rate*. The idea is that instead of forcing V to be positive only with small probability α , it might be more useful to allow V to be positive as long as V/r remains small. Approaches in this direction have been proposed by Benjamini and Hochberg (1995), see also Storey (2011).

5.8 Universal Inference

Tests and confidence intervals proposed above rely on asymptotic (or correct) distributions. These are not always available, for instance, we might not be able to derive the asymptotic distribution. For this setup Wasserman et al. (2020) recently proposed a general framework, labeled as universal inference. The procedure works for large data and the idea is to split the data into two parts, which for ease of notation are of equal size. Let $2n$ here define the size of the data so that we can denote the data as $Y_1, \dots, Y_n, Y_{n+1}, \dots, Y_{n+n}$. 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 likelihood for some model calculated from data $D_{(0)}$. Accordingly, we define

$$L_{(1)}(\theta) = \prod_{i=n+1}^{2n} \log f(y_i; \theta)$$

as likelihood of the second part. The latter is maximized leading to $\hat{\theta}_{(1)}$ as the maximum likelihood estimate. With this notation, we can now define the universal confidence interval.

Definition 5.7 The universal confidence interval is defined as

$$CI_{\text{universal}} = \left\{ \theta : \frac{L_{(0)}(\hat{\theta}_{(1)})}{L_{(0)}(\theta)} \leq \frac{1}{\alpha} \right\}$$

It can be shown that this confidence interval has a coverage of at least $1 - \alpha$, that is

$$P(\theta \in CI_{\text{universal}}) \geq 1 - \alpha \quad (5.18)$$

Bear in mind that the likelihood is calculated for observation 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, \dots, Y_n . This is an important property which is utilized in the proof. We sketch the proof at the end of this section. It is rather simple as it relies on the Markov inequality.

With the link between confidence intervals and testing, we can directly extend this to universal testing. We will not give details here but we want to emphasize that the approach quite easily extends to more complex confidence intervals and more complex testing problems. Assume for instance one has fitted a model $\hat{f}_{(1)}(y)$ which may or may not be parametric using the data $D_{(1)}$. If we want to test whether an alternative model $\tilde{f}(y)$ would hold as well we can check whether $\tilde{f}(y)$ is in the universal confidence interval. To be specific, let \mathcal{F} be a set of densities. The universal confidence interval is then

$$CI_{\text{universal}} = \left\{ f \in \mathcal{F} : \frac{\prod_{i=1}^n \hat{f}_{(1)}(y_i)}{\prod_{i=1}^n f(y_i)} \leq \frac{1}{\alpha} \right\} \quad (5.19)$$

Putting it differently, a test of a specific hypothesis $H_0 : f(\cdot) = \tilde{f}(\cdot)$ can be done through checking

$$\frac{\prod_{i=1}^n \hat{f}_{(1)}(y_i)}{\prod_{i=1}^n \tilde{f}(y_i)} \leq \frac{1}{\alpha} \Leftrightarrow "H_0"$$

Let us look at universal inference in a simple example. Let $Y_i \sim N(\mu, 1)$ and we aim to construct a confidence interval for μ . As before, we split the data $Y_1, \dots, Y_n, Y_{n+1}, \dots, Y_{2n}$ and take Y_{n+1}, \dots, Y_{2n} to obtain the mean estimate

$$\hat{\mu}_{(1)} = \frac{1}{n} \sum_{i=n+1}^{2n} Y_i =: \bar{y}_{(1)}$$

The likelihood ratio split statistics results to

$$T_n(\mu) = \frac{\exp\left(-\frac{1}{2} \sum_{i=1}^n (y_i - \bar{y}_{(1)})^2\right)}{\exp\left(-\frac{1}{2} \sum_{i=1}^n (y_i - \mu)^2\right)}.$$

The universal confidence interval then results to

$$CI_{\text{universal}} = \left\{ \mu : T_n(\mu) \leq \frac{1}{\alpha} \right\}.$$

Simple calculation gives

$$\begin{aligned} T_n(\mu) \leq \frac{1}{\alpha} &\Leftrightarrow -\sum_{i=1}^n (y_i - \bar{y}_{(1)})^2 + \sum_{i=1}^n (y_i - \mu)^2 \leq 2 \log\left(\frac{1}{\alpha}\right) \\ &\Leftrightarrow -\sum_{i=1}^n (y_i - \mu + \mu - \bar{y}_{(1)})^2 + \sum_{i=1}^n (y_i - \mu)^2 \leq 2 \log\left(\frac{1}{\alpha}\right) \\ &\Leftrightarrow 2n(\bar{y}_{(0)} - \mu)(\mu - \bar{y}_{(1)}) + n(\mu - \bar{y}_{(1)})^2 \geq 2 \log\left(\frac{1}{\alpha}\right) \end{aligned}$$

with $\bar{y}_{(0)} = \frac{1}{n} \sum_{i=1}^n y_i$. The universal confidence interval takes the form

$$CI_{\text{universal}} = \left\{ \mu : 2n(\bar{y}_{(0)} - \mu)(\bar{y}_{(1)} - \mu) + n(\bar{y}_{(1)} - \mu)^2 \geq 2 \log(\alpha) \right\}$$

Let us compare this with the exact confidence interval, which in this case would be available. Note that

$$\sqrt{n}(\bar{y}_{(0)} - \mu) \sim N(0, 1) \text{ and } \sqrt{n}(\bar{y}_{(1)} - \mu) \sim N(0, 1).$$

Hence the confidence interval for $\bar{y}_{(1)}$ results to

$$CI = \left\{ \mu : n(\bar{y}_{(1)} - \mu)^2 \leq X_{1,1-\alpha}^2 \right\}.$$

With the exact distribution, we can calculate the coverage of the universal interval. Let therefore Z_0 and Z_1 be independent $N(0, 1)$ variables. Then the coverage probability of the universal interval equals

$$P(\mu \in CI_{\text{universal}}) = P(2Z_0 Z_1 + Z_1^2 \geq 2 \log(\alpha)).$$

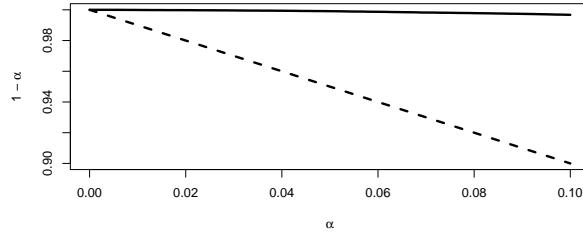


Fig. 5.10 Coverage probability of the universe confidence interval (solid line) compared to the nominal coverage (dashed line)

We compare the coverage probabilities of the true interval with its universal counterpart in Figure 5.10. We recognize that the coverage is by far larger than the nominal coverage. Hence, the universal approach is not recommendable in cases where an exact distribution is available. The general form, however, allows us to draw universal inferences, even in situations without known limiting distributions or in more complex testing constellations, like the general interval on distributions given in (5.19). We just have to pay the price of increased coverage.

We can investigate this loss of information also with a different plot. Omitting the data splitting we can consider $T_n(\mu)$ as likelihood ratio by replacing $\hat{\mu}_{(1)}$ by $\hat{\mu}_{(0)} = \bar{y}_{(0)}$. In this case, the threshold for a confidence interval would result through the asymptotics of the log-likelihood ratio and in fact, we would get

$$CI = \left\{ \mu : \frac{\exp\left(-\frac{1}{2} \sum_{i=1}^n (y_i - \bar{y}_{(0)})^2\right)}{\exp\left(-\frac{1}{2} \sum_{i=1}^n (y_i - \mu)^2\right)} \leq \exp\left(\frac{1}{2} \chi^2_{1,1-\alpha}\right) \right\}.$$

In Figure 5.11 we plot the two thresholds, namely based on the asymptotic Chi-squared approach as well as based on the universal approach. To be specific, we plot the log threshold. The asymptotic threshold, shown as a dotted line, is clearly smaller than the universal threshold. The shapes of the two curves are however comparable.

Proof To prove the validity of (5.18) 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

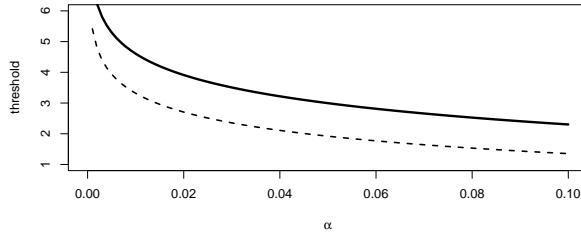


Fig. 5.11 Comparison of threshold based on the correct Chi-squared approximation (dashed line) and the universal threshold (solid line)

$$P_\theta(\theta \notin CI_{\text{universal}}) = P_\theta\left(T_n(\theta) > \frac{1}{\alpha}\right) \leq \alpha E_\theta(T_n(\theta)),$$

where the inequality results from the Markov inequality (see Chapter A.6). Calculating the expectation is easy since

$$\begin{aligned} E_\theta(T_n(\theta)) &= E\left(\frac{L_{(0)}(\hat{\theta}_{(1)})}{L_{(0)}(\theta)}\right) \\ &= \int \frac{\prod_{i=1}^n f(y_i; \hat{\theta}_{(1)})}{\prod_{i=1}^n f(y_i; \theta)} \prod_{i=1}^n f(y_i; \theta) dy_i \\ &= \int \prod_{i=1}^n f(y_i; \hat{\theta}_{(1)}) dy_i = 1, \end{aligned}$$

which concludes the proof. \square

5.9 Sensitivity, Specificity, Accuracy

Before concluding this chapter, we aim to relate the results to the field of classification using alternative terminology. This does not introduce new material but rather presents the same concepts using different wording and expressions. To achieve this, we will direct our attention to the table labelled 5.3. In the context of classification terminology, a decision H_1 is labelled as "positive," while a decision as H_0 is termed "negative." For instance, consider a medical test where a decision of " H_1 " indicates a positive test result, meaning a disease is detected, while " H_0 " signifies a negative test result. By adopting this binary perspective, one obtains:

- S = true positives (TP)
- U = true negatives (TN)
- V = false positive (FP)

- $T = \text{false negative. (FN)}$

With these labels, one defines a number of quantities that express the performance of the test. First, with **sensitivity** one denotes the proportion of correctly rejected hypotheses, or phrased differently, the proportion of correct positive results. Synonyms for the term are "recall" as well as "True Positive Rate (TPR)". Hence

$$\text{Sensitivity} = \text{Recall} = \text{TPR} = \frac{TP}{P} = \frac{S}{M_1} \quad (5.20)$$

where $P = FN + TP$ denotes the number of positives, denoted as M_1 in Table 5.3. Roughly, the sensitivity is the power of a test, that is $P("H_1" | H_1)$.

The **specificity** of a test is the proportion of correctly classified negatives. Here, again, there are frequently used synonyms, namely "selectivity" and "True Negative Rate (TNR)" so that

$$\text{Specificity} = \text{Selectivity} = \text{TNR} = \frac{TN}{N} = \frac{U}{M_0}, \quad (5.21)$$

where $N = TN + FP$ denotes overall number of negatives which is M_0 in Table 5.3. Note that the specificity can roughly be understood as $P("H_0" | H_0)$.

The **prevalence** is the proportion of positives, that is

$$\text{Prevalence} = \frac{P}{P+N} = \frac{M_1}{m}, \quad (5.22)$$

We mentioned above already the *False Discovery Rate (FDR)* which we repeat here for completeness. This is defined as

$$\text{FDR} = \frac{FP}{FP+TP} = \frac{V}{r}, \quad (5.23)$$

which in turn defines the *precision*, also called *Positive Predictive Value (PPV)* through

$$\text{Precision} = \text{PPV} = \frac{TP}{FP+TP} = \frac{S}{r} = 1 - \text{FDR}. \quad (5.24)$$

This gives the proportion of correctly found positive results.

In classification one aims to combine these measures into a single measure of accuracy. The first is called "accuracy" and it is defined as

$$\text{Accuracy} = \frac{TP+TN}{P+N} = \frac{U+S}{m}. \quad (5.25)$$

The measure gives the proportion of correct results amongst all tests.

The second occasionally used measure is the so-called *F score*. The idea is to combine sensitivity with the false discovery rate, or equivalently the precision. To be specific, the F score is defined as

$$\text{F score} = 2 \frac{\text{sensitivity} * \text{precision}}{\text{sensitivity} + \text{precision}} = \frac{2TP}{2TP+FP+FN} = \frac{2S}{r+M_1} \quad (5.26)$$

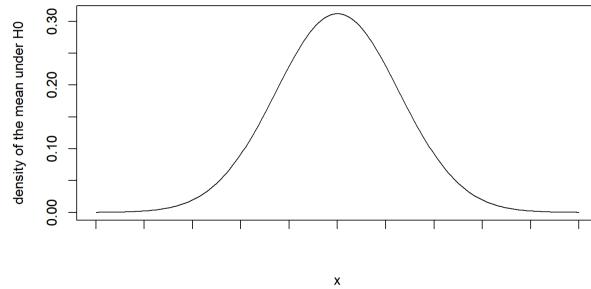
The above quantities have been interpreted in the context of testing already. Hence, the above definitions are more meant as a guideline to the reader to relate the different terms, both in mathematical definition as well as terminology.

5.10 Exercises

Exercise 1

The best seller of Edumm cheese dairy is their 1000g 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 1000g, 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.

- Formulate the hypotheses for this question. Which test would you use and why?
- Which test decision do you arrive at for a measured mean cheese weight of 1003g? Perform the test for $n = 10$ and $n = 30$. Interpret the results. Also calculate the *p*-values and rejection region of the tests.
- 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, \dots, X_n from a normal distribution $X_i \sim N(\mu, \sigma^2)$, 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.

- 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.
- Derive the score function $s(\mu; x_1, \dots, x_n)$. 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.
- State the idea of the Wald test, the general test statistic and 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?
- 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 $\chi^2_{k, 1-\alpha} = z_{1-\frac{\alpha}{2}}^2$)

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 one 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 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 χ^2 test to test if two variables in a contingency table are independent, i.e. derive the χ^2 test on Independence on your own.
Hint: Derive the distribution of the number in each cell under the independence assumption, then test if the data as a whole follows this distribution.
- d. Use the χ^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.

Chapter 6

Regression

The previous chapters discussed general ideas of statistical inference. We restricted the presentation to the case where we observe data y_1, \dots, y_n and omitted any input variables for simplicity of notation. We will change this now and assume that we observe the outcome variable Y (which may also be multivariate) given the input variable x , which may also be vector-valued. In this context, Y is often termed as “outcome”, “target” or “dependent variable”, as its distribution depends on x , often called “inputs”, “features” or “independent variables”. Hence, we observe pairs of variables (y_i, x_i) and aim to model the distribution of Y given x . In statistics this is usually referred to a regression, which can be seen as one of the most central fields in statistics. We touch on the main ideas in this chapter and emphasize right from the start, that the model class has been widely explored, and the possibilities for extensions are almost endless, and go far beyond what is discussed in this chapter. We here 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. In investigating questions regarding heredity, Galton collected data on the body height 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 also be taller than the average person, while at the same being shorter than their parents on average. Galton described this tendency towards the population average as *regression* towards the mean, and termed the linear trend line “regression line”. Nowadays, the term regression analysis is broadly used for describing statistical processes for estimating the relationships between a dependent variable and one or more independent ones. The simplest type of regression models are linear models, 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:

$$Y = \beta_0 + \beta_1 x + \epsilon \quad (6.1)$$

Here, β_0 is the intercept, while β_1 is the slope of the regression line. The term ϵ is called the error term and consists of random deviations with mean 0. Because of that, the expect value of Y is assumed to be a linear function of x , that is

$$E(Y|x) = \beta_0 + \beta_1 x.$$

In other words, we assume (a) a linear relationship between x and Y , and (b) that the relationship is only disturbed by the random perturbation ϵ . These are two fundamental assumptions of a linear regression model. Note that the relationship does not necessarily need to be causal in nature, but can also be a simple (linear) association. In addition, the error term is often assumed to be normally distributed and independent of the covariate, that is

$$\epsilon_i \sim N(0, \sigma^2), \quad i = 1, \dots, n$$

for each subject i , where σ^2 is known as the error variance. In this case, it follows that the single observations on the response variable follow a (conditional) normal distribution, with

$$E(Y_i|x_i) = \beta_0 + \beta_1 x_i, \quad \text{Var}(Y_i|x_i) = \sigma^2$$

for each subject $i = 1, \dots, n$. The property of constant variance σ^2 is also termed *homoscedasticity*. In particular, this implies that the errors do not depend on the value of the covariate. In contrast, if the variance does depend on x , we are in the presence of *heteroscedasticity*. An empirical regression application under homoscedasticity conditions is given in example 7.

Example 7 In a class of 30 statistics student, we have information on their height and weight and are interested in learning about the relationship between the two quantities. We here consider weight as the response (or dependent) variable, and height as the single covariate (or independent variable). We therefore estimate a simple linear regression model, which takes the form:

$$\text{weight}_i = \beta_0 + \beta_1 \text{height}_i + \epsilon_i$$

Estimating the coefficients gives an estimated slope of $\beta_{\text{height}} = 0.89$, meaning that students that are taller tend to also weigh more. More specifically, within our sample, an increase of 1cm in height is associated with an increase of 0.89 kg in body mass on average. A graphical representation of the data and the corresponding regression line is given in Figure 6.1.

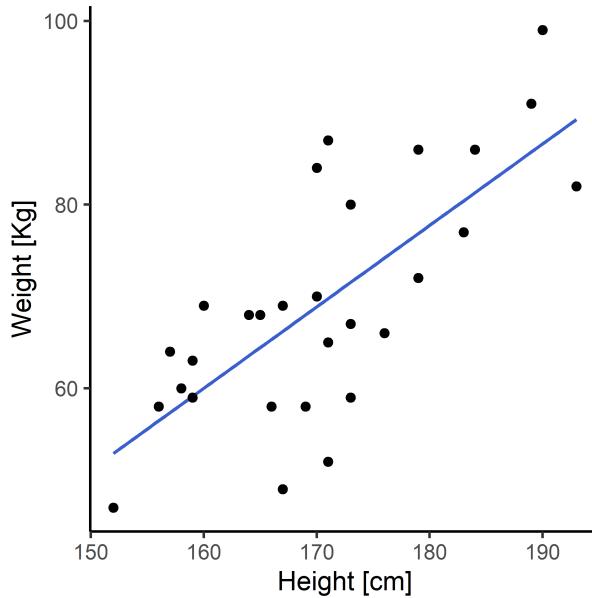


Fig. 6.1 Scatterplot of weight and height data, with regression line.

Having defined the theoretical aspects of the regression model, the natural question arises: how can we estimate its unknown parameters? Galton only drew the regression line visually, as his mathematical capabilities were limited. His successors, however, were able to progressively formalize his work. Nowadays, the parameters β_0 and β_1 can be estimated through maximum likelihood. Assuming normality and independence of the error terms ϵ_i results in the log-likelihood

$$l(\beta_0, \beta_1, \sigma^2) = \sum_{i=1}^n \left\{ -\frac{1}{2} \log \sigma^2 - \frac{1}{2} \frac{(y_i - \beta_0 - \beta_1 x_i)^2}{\sigma^2} \right\}. \quad (6.2)$$

Deriving with respect to the three parameters returns the maximum likelihood estimates:

$$\hat{\beta}_0 = \frac{1}{n} \sum_{i=1}^n y_i - \frac{1}{n} \sum_{i=1}^n x_i \hat{\beta}_1$$

$$\hat{\beta}_1 = \sum_{i=1}^n x_i (y_i - \hat{\beta}_0) / \sum_{i=1}^n x_i^2 \quad (6.3)$$

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{\beta}_0 - x_i \hat{\beta}_1)^2$$

The estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are also known as “least squares estimates”, as they can also be obtained through the method of *least squares*. The method, which pre-dates maximum likelihood and traces back to Gauss (1777-1855), consists in finding the parameters that minimize the sum of the squared deviations between predicted and observed values, $\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$. The equivalence between the two is apparent from equation 6.2, 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 e.g. in Figure 6.1.

The calculation of the estimates can also be carried out in matrix notation. To do so, let

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \in \mathbb{R}^{nx1}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \in \mathbb{R}^{nx2}, \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix} \in \mathbb{R}^{nx1} \quad (6.4)$$

We can then write the regression model (6.1) in the form

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \quad (6.5)$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1)^\top$ and $\mathbf{Y} = (Y_1, \dots, Y_n)^\top$. The matrix \mathbf{X} is often referred to as the design matrix, and by default, it contains a column of 1s, which implies that the model includes an intercept (which would otherwise be implicitly assumed to be 0). Given the data \mathbf{y} , the likelihood is

$$l(\boldsymbol{\beta}, \sigma^2) = -\frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \quad (6.6)$$

Differentiation with respect to $\boldsymbol{\beta}$ using matrix algebra and setting the derivative to 0 yields the maximum likelihood estimate

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} \quad (6.7)$$

which is identical to the estimates in (6.3), but has a more convenient form. It is easy to show that $\hat{\boldsymbol{\beta}}$ is exactly unbiased, as

$$E(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top E(\mathbf{Y} | \mathbf{X}) = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}.$$

Moreover, as $\text{Var}(\boldsymbol{\epsilon}) = \sigma^2 I_n$, where I_n is a diagonal matrix of dimension n , it follows that

$$\text{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}.$$

It is also easy to show that $\sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}$ is equal to the inverse Fisher information matrix, implying that the estimator variance is minimal. Furthermore, $\hat{\boldsymbol{\beta}}$ is the best linear unbiased estimator (BLUE) for $\boldsymbol{\beta}$, which is also known as Gauss-Markov property.

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 non-linear dependencies that interact with each other. Luckily, it is fairly straightforward to extend the simple, single-variable regression setting seen in the previous section to the case of multiple covariates. Let us introduce the *multiple linear regression model*:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \epsilon \quad (6.8)$$

where x_1, x_2, \dots, x_k are k different covariates assumed to influence the response variable Y , $\beta_1, \beta_2, \dots, \beta_k$ are the coefficients associated with each covariate, and ϵ is the random error term. As before, the 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 is given by the structural component of the model, i.e.

$$E(Y|X) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k,$$

as in the case of simple linear regression. We can rewrite the model again in matrix form. Let therefore x_{ij} denote the i -th observation on covariate j , then for observation i the multiple regression formula (6.8) becomes

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik} + \epsilon_i.$$

Defining the design X as

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix} \in \mathbb{R}^{n \times (k+1)}$$

allows to write the regression model as (6.5), with y and ϵ as defined in (6.4) and $\beta = (\beta_0, \beta_1, \dots, \beta_k)^\top$. Accordingly, the maximum likelihood estimate results as in (6.7).

Example 8 Let's consider once again the 30 statistics students from Example 7. In addition to their height and weight, we now have additional information on their gender and age. We are interested in learning how height, gender and age relate to the students' weight. We thus consider weight as the response variable, with height, age and gender as covariates. We estimate the multiple linear regression model:

$$\text{weight}_i = \beta_0 + \beta_1 \text{height}_i + \beta_2 \text{age}_i + \beta_3 \text{gender}_i + \epsilon_i,$$

where the variable "gender" takes value 1 if the student identifies as male, and 0 otherwise. Estimates of the unknown parameters $\beta_0, \beta_1, \beta_2$ and β_3 were obtained via maximum likelihood, and are shown in the following table:

	Estimate	Std. error	p-value
$\hat{\beta}_0$ (intercept)	-7.05	32.97	0.83
$\hat{\beta}_1$ (height)	0.42	0.19	0.04
$\hat{\beta}_2$ (age)	-0.06	0.45	0.89
$\hat{\beta}_3$ (gender)	13.69	4.06	<0.01

In the table, the p-values are obtained by performing a Wald test for the null hypotheses of the corresponding coefficient being equal to 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 real coefficient might be zero.

Plugging the estimated coefficients into the model equation yields:

$$\widehat{\text{weight}} = -7.05 + 0.42\text{height}_i - 0.06\text{age}_i + 13.69\text{gender}_i.$$

The single coefficients can be interpreted similarly as in the simple regression model, with the additional caveat that the interpretation is valid by keeping all other coefficients fixed. 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:

- $\hat{\beta}_0$: 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 0cm is very much out of the sample and also does not make sense from a substantive standpoint.
- $\hat{\beta}_1$: An increase of 1cm in height is associated, *ceteris paribus*, with an average increase of 0.42Kg in weight.
- $\hat{\beta}_2$: Keeping height and gender the same, an age increase of one year is associated with an average reduction of 0.07 in weight. Note that the variance of this coefficient is much higher than the absolute value of the coefficient itself, leading to a high p-value. This means that the data at hand does not provide us with enough evidence to say that age has any effect on weight for people comparable to those within our sample.
- $\hat{\beta}_3$: The average weight for a male student is about 13.69kg higher than that of a female student with the same height and age.

You may have noticed in (8) 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 on the nature of the covariates x_1, \dots, x_k . 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 non-numeric variables and Y ;
- Variable transformations, such as $\log(x)$ or $\frac{1}{x}$, to model non-linear relationships;
- Using polynomials to model more complex relationships, e.g. including variables $x_1 = z, x_2 = z^2$ and $x_3 = z^3$ 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 $x_3 = x_1x_2$ 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 non-linearity 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 non-linearly.

6.3 Logistic Regression

The linear regression model is suitable for response variables which approximately follow a normal distribution. In the real world, however, there are many cases of non-gaussian variables of interest. Some variables are not even continuous, but rather categorical, i.e. taking a finite number of non-numeric 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 getting through an exam (pass/fail). In general, the two possible outcomes of a binary variable Y can be characterized as *success*, i.e. $Y = 1$, and *failure*, i.e. $Y = 0$. We first show that modelling a binary response Y with a linear model as presented in the previous chapter 6.2, i.e. as

$$Y_i = E(Y_i | x_{i1}, x_{i2}, \dots, x_{ik}) + \epsilon_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik} + \epsilon_i \quad (6.9)$$

with $\epsilon_i \sim N(0, \sigma^2)$ isn't feasible. Note that here $E(Y_i)$ is equivalent to $P(Y_i = 1)$, as for binary variables we have that

$$E(Y_i) = P(Y_i = 0) \cdot 0 + P(Y_i = 1) \cdot 1 = P(Y_i = 1).$$

Given this, it is easy to see that the linear model breaks down in this case, as:

- Any probability can only take values between 0 and 1, while the predicted values on the right side of (6.9) are unbounded;
- The errors ϵ_i would not be Gaussian, and even relaxing the normality assumption they could not be homoscedastic with $\text{Var}(\epsilon_i) = \sigma^2$, given that the response Y_i would follow a Bernoulli distribution with $\pi_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik}$, and thus we would have that

$$\text{Var}(Y_i) = \pi_i(1 - \pi_i),$$

which depends on the value of the covariates, and thus could not be the same (and equal to σ^2) for every observation i .

These issues can be solved by limiting ourselves to modelling the conditional expectation $E(Y_i|x_{i1}, x_{i2}, \dots, x_{ik})$ (i.e. $P(Y_i = 1|x_{i1}, x_{i2}, \dots, x_{ik})$, or more shortly π_i) instead of predicting observations Y_i , and by assuming the model:

$$\pi_i = P(Y_i = 1|x_{i1}, x_{i2}, \dots, x_{ik}) = h(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik}),$$

where $h()$ is a function mapping the real axis from $-\infty$ to ∞ uniquely to values in the interval $[0, 1]$. By modelling the expectation we do not need to make any assumption on the errors; further, by using the function h to transform the output we ensure that the predicted probabilities will only take acceptable values. The argument of the h function is termed *linear predictor*, and usually indicated with the letter η_i , i.e.

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik}. \quad (6.10)$$

For reasons of interpretability, the function h is typically chosen 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:

$$h(\eta) = \frac{\exp(\eta)}{1 + \exp(\eta)} \quad (6.11)$$

The function is represented in Figure 6.2, from which we can appreciate its characteristic S-shape. The use of the logistic function results in the *logistic regression model*, or, more shortly, the *logit model*:

$$\pi_i = P(Y_i = 1|\eta_i) = \frac{\exp(\eta_i)}{1 + \exp(\eta_i)}. \quad (6.12)$$

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 $\frac{\exp(\eta)}{1 + \exp(\eta)}$ to $\frac{\exp(\eta+1)}{1 + \exp(\eta+1)}$. An alternative interpretation can be obtained by solving equation 6.12 in η_i , obtaining:

$$\eta_i = \log\left(\frac{\pi_i}{1 - \pi_i}\right) = \log\left(\frac{P(Y_i = 1)}{1 - P(Y = 1)}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik} \quad (6.13)$$

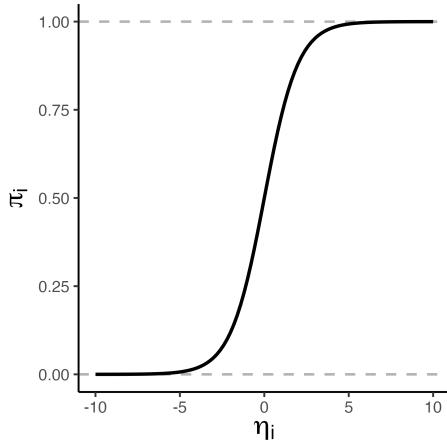


Fig. 6.2 Logistic distribution function, showing the probability $P(Y_i = 1)$, π_i , as a function of the linear predictor η_i .

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 it similarly, i.e. we can use phrases such as “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 on both sides of Eq. 6.13, obtaining:

$$\frac{\pi_i}{1-\pi_i} = \frac{P(Y_i = 1)}{P(Y_i = 0)} = \exp(\beta_0) \exp(\beta_1 x_{i1}) \exp(\beta_2 x_{i2}) \cdots \exp(\beta_k x_{ik}). \quad (6.14)$$

Through this formulation, we can clearly see 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 know explore the practical usage of logistic regression and its different interpretations through an example.

Example 9 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, which in this case takes the form

$$P(Y_i = 1) = \frac{\exp(\eta_i)}{1 + \exp(\eta_i)},$$

with

$$\eta_i = \beta_0 + \beta_1 \text{age}_i + \beta_2 \text{gender}_i,$$

where the variable age is measured in years, and the variable “gender” takes 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.

	Estimate	Std. error	p-value
$\hat{\beta}_0$ (intercept)	-5.53	1.67	<0.01
$\hat{\beta}_1$ (age)	0.06	0.02	<0.01
$\hat{\beta}_2$ (gender)	1.45	0.85	0.09

Table 6.1 Coefficients, standard errors and p-values for the regression model in Example 9.

Plugging the estimated coefficients into the model equation yields:

$$\eta_i = \log\left(\frac{\pi_i}{1 - \pi_i}\right) = -5.53 + 0.06\text{age}_i + 1.45\text{gender}_i.$$

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:

- $\hat{\beta}_0$: 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 here.
- $\hat{\beta}_1$: For a given gender, an age increase of one 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 one 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 $e^{0.06} = 1.06$.
- $\hat{\beta}_2$: 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 that of a female patient of the same age by a (multiplicative) factor of $e^{1.45} = 4.26$.

The predicted probability of heart disease by age and gender are depicted in Figure 6.3. From the plot, we can clearly see that the risk of heart disease increases for older people and that males have a higher risk than females.

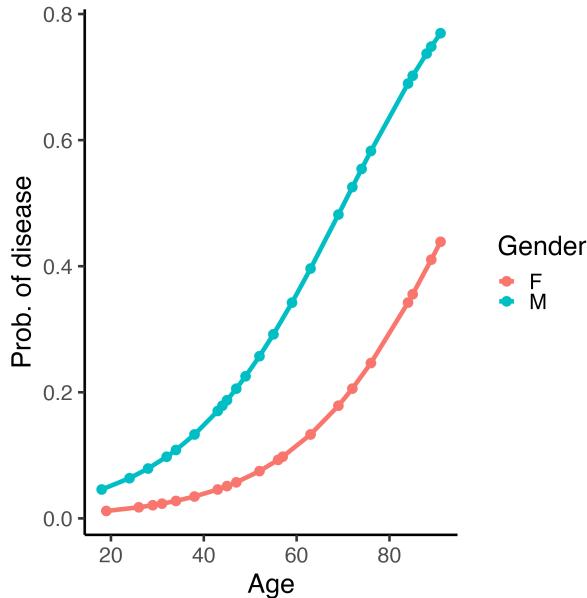


Fig. 6.3 Predicted probability of heart disease by age and gender, estimated via the logistic regression model.

6.4 Generalized Linear Models

In the previous section, we showed how the linear modelling framework can be extended to handle binary response variables by focusing on modelling the conditional expectation of the response variable, i.e. $E(Y|x)$, and transforming the output of the linear model in a suitable manner. We will now show that this generalization is not only possible in the binary case, but also for a wide variety of different response variable types, such as e.g. 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 Section A.5 for more details).
2. The expected value of the response value, $E(Y_i|x)$, is connected to the linear predictor $\eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik}$ through a response function $h()$. In other words, we have:

$$E(Y|x) = \mu = h(\eta)$$

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 modelled. 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 normal distributions 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 10 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:

Sales	Score	Price	Promotion
19	4.1	27€	No
6	3.8	43€	No
29	4.6	85€	Yes
21	4.0	60€	No
...

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 Figure 6.4.

Given that we already 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 covariates 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:

$$P(Y_i = y|x) = \exp(\eta_i),$$

with

$$\eta_i = \beta_0 + \beta_1 \text{review_score}_i + \beta_2 \text{price}_i + \beta_3 \text{promotion}_i$$

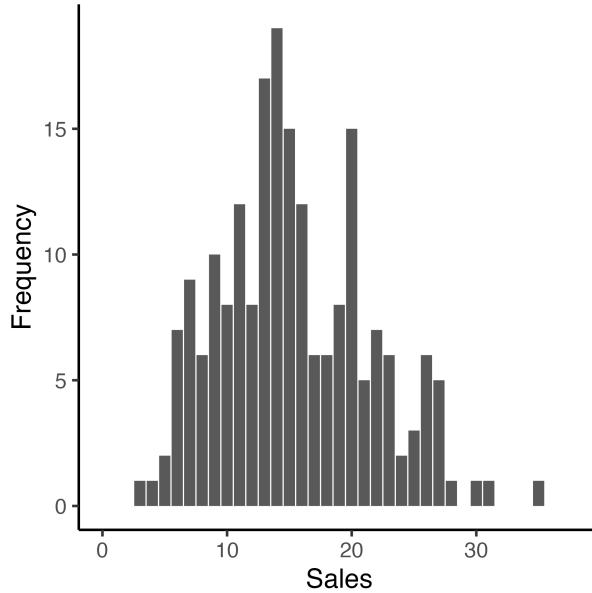


Fig. 6.4 Barplot of "Sales", the response variable in our regression model. The shape of the data is consistent with the choice of a Poisson distribution.

where the variable "review score" is measured in stars, the variable "price" in Euros, and the variable "promotion" takes 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 10, together with their related standard errors and p-values.

	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 6.2 Coefficients, standard errors and p-values for the regression model in Example 10.

Plugging the estimated coefficients into the model equation yields:

$$\eta_i = 0.642 + 0.473\text{review_score}_i + 0.001\text{price}_i + 0.166\text{promotion}_i.$$

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:

- $\hat{\beta}_0$: The intercept gives the log-expectation of heart disease 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.
- $\hat{\beta}_1$: For a given price and promotion, an increase of 1 star in review is associated with an increase of 0.473 in the logarithm of weekly sales. Equivalently, we can say that an increase of one year in age is associated to a multiplicative increase in sales by a factor of $e^{0.473} = 1.605$, that is a whopping 60% increase in sales.
- $\hat{\beta}_2$: 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.
- $\hat{\beta}_3$: 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 $e^{0.177} = 1.194$, i.e. an increase of 19% in weekly sales, *ceteris paribus*.

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 theatre 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
 Theatre: Expenses for theatre visits in CHF
 Theatre_prevYear: Expenses for theatre 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 theatre 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 theatre expenses?
- b. Eliminate all non-significant 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 ($^{\circ}\text{F}$) at take-off as well as the (non-)occurrence of thermal overstress of a certain unit. The data contains the following variables:

<code>flight</code>	Flight number
<code>temp</code>	Temperature in $^{\circ}\text{F}$
<code>td</code>	Thermal overstress (1 = Yes / 0 = No)

Our goal is to understand the relationship between temperature and thermal overstress.

Download the data set `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 $^{\circ}\text{F}$ 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 $^{\circ}\text{C}$? 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°F as on the date of the Challenger disaster.

- g. For which temperature does the probability of thermal overstress equal 0.5?

Chapter 7

Model Selection

The choice of the right model is apparently crucial. This was already discussed in Chapter 3.1 and we refer to Figure 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 just related to selecting a probability model, but also to specify 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

Over the previous chapters, we learned about different ways to model random variables by characterizing them via probability distributions. In Chapter 6, we introduced regression models that allowed 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 socio-economic 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, i.e. $x = (x_1, x_2, \dots, x_q)$ and we also allow the output variable to be multi-dimensional. Apparently, this is exactly the setting

of regression, which we covered in the previous chapter. But we want to treat the problem from a much broader perspective here. Our aim is 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. $F(y|x, \theta)$. The distribution depends on a set of parameters $\theta = (\theta_1, \theta_2, \dots, \theta_p)$. 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 \mathcal{F} , as motivated in Chapter 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 change the notation slightly. We define a model \mathcal{M} as a pair, composed of a probability model \mathcal{F} and a structural model \mathcal{H} , which defines how the parameters of the distribution depend on the input variables x . To be specific, a model is defined as

$$\mathcal{M} = (\mathcal{F}, \mathcal{H})$$

where the two model components are given through

$$\mathcal{F} = \{F(\cdot; \gamma), \gamma \in \Gamma\} \quad (7.1)$$

$$\mathcal{H} = \{h(\cdot; \theta) \text{ with } h : \mathcal{X} \rightarrow \Gamma \text{ and } \theta \in \Theta\}. \quad (7.2)$$

We here call \mathcal{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 $\gamma \in \Gamma$ we have $F(\cdot, \gamma)$ to be a fully specified distribution function. In particular, if the density exists then we can equivalently use

$$\mathcal{F} = \{f(\cdot; \gamma), \gamma \in \Gamma\} \quad (7.3)$$

with $f(\cdot, \gamma)$ as density to the distribution function $F(\cdot, \gamma)$. This is exactly the setup that we introduced in Chapter 3.1 and which we employed in the subsequent chapters so far.

Besides the probability model, we also have a **structural model** denoted by \mathcal{H} . In the area of machine learning, this is also called the hypothesis space. The model defines how the input variables x influence the parameters of the model. The function $h(\cdot, \theta)$ thereby depends on parameters θ . The structure of the function $h(\cdot, \theta)$ itself is assumed to be known but the parameters θ are unknown and need to be estimated. In particular, the value range of $h(\cdot; \theta)$, 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

$$\{h(x, \theta) \text{ for } x \in \mathcal{X} \text{ and } \theta \in \Theta\} \subseteq \Gamma \quad (7.4)$$

The constraint (7.4) guarantees, that for any value of x and any value of θ we obtain $F(., h(x, \theta)) \in \mathcal{F}$, which means we obtain a proper distribution.

The notation above can be motivated by looking at the logistic regression discussed in Chapter 6.3. Here, the probability model \mathcal{F} is a binomial distribution with $\pi \in [0, 1]$ as the only parameter. The structural model \mathcal{H} is built from the logistic function $h(.)$ given in (6.11) and the linear structure for the linear predictor η given in (6.10). Defining with $x = (1, x_1, x_2, \dots, x_q)^\top$ and $\theta = (\beta_0, \beta_1, \beta_2, \dots, \beta_q)^\top$ 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 above broad definition encompasses several different types of tools used to model dependencies, both in statistics and machine learning. While the latter will be exemplified in more depth in the last chapter, one of the oldest statistical methods, which is still one of the most used ones for real-world applications, namely regression, was discussed in the previous chapter. In this chapter, we approach the problem of choosing an appropriate model \mathcal{M} from a general perspective. In principle, this encompasses both, selecting a probability model \mathcal{F} as well as a structural model \mathcal{H} . We will introduce the Akaike information criterion (AIC), which, in fact, allows the comparison of different models and, hence, the selection of a better one. We also motivate the idea of why model selection can be built by dividing the data into one part, which is used for model fitting, and one part used for model selection, commonly known as cross-validation. In most practical applications, the focus is thereby to select the "best" structural model \mathcal{H} , while the probability model \mathcal{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 \mathcal{H} and probability component \mathcal{F} we want to simplify the notation for better readability. We therefore sometimes write the model class \mathcal{M} as

$$\mathcal{M} = \{F(\cdot; \theta); \theta \in \Theta\} \tag{7.5}$$

which formally stands

$$\mathcal{M} = \{F(\cdot; h(\cdot; \theta)); F \in \mathcal{F} \text{ and } h(\cdot, \cdot) \in \mathcal{H}\}$$

This notation accommodates that models comprise of structural and probability components, without explicitly listing these.

7.2 Training and Test Data

Let us dive now deeper into 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 $G(\cdot|x)$ with $x \in \mathcal{X}$. For simplicity, we look at two possible models which we denote as $\mathcal{M}_1 = (\mathcal{F}_1, \mathcal{H}_1)$ and $\mathcal{M}_2 = (\mathcal{F}_2, \mathcal{H}_2)$. As said above, in many situations these models will be nested, meaning that one model is a subset of the other. This means that the probability model is the same, i.e. $\mathcal{F}_1 = \mathcal{F}_2$ and the structural models are nested in that $\mathcal{H}_2 \subset \mathcal{H}_1$, or vice versa. Assuming \mathcal{H}_1 to be the larger model this can formally be written as

$$\mathcal{H}_1 = \{h(\cdot; \theta) \text{ with } \theta \in \Theta_1\} \text{ and } \mathcal{H}_2 = \{h(\cdot; \theta) \text{ with } \theta \in \Theta_2\} \text{ with } \Theta_2 \subset \Theta_1. \quad (7.6)$$

We call (7.6) a nested model structure. Alternatively, we may also think of non-nested models. If $\mathcal{F}_1 = \mathcal{F}_2$, this can be the case if

$$\mathcal{H}_1 = \{h(\cdot; \theta), \theta \in \Theta_1\} \text{ and } \mathcal{H}_2 = \{h(\cdot; \theta), \theta \in \Theta_2\} \text{ with } \Theta_2 \not\subset \Theta_1 \text{ and } \Theta_1 \not\subset \Theta_2. \quad (7.7)$$

In this case, the probability model is the same but the structural models differ and they are not subsets of another, i.e. $\mathcal{H}_1 \not\subset \mathcal{H}_2$ and $\mathcal{H}_2 \not\subset \mathcal{H}_1$. But we may even consider models, where the probability models differ, that is $\mathcal{F}_1 \neq \mathcal{F}_2$. In this general case we write $F_1(\cdot; \theta_1)$ for elements of \mathcal{F}_1 and accordingly $F_2(\cdot; \theta_2)$ for elements of \mathcal{F}_2 . Note that for nested models the distributional models are the same and in this case, we omit the subscript and write $F(\cdot; \theta_1)$ and $F(\cdot; \theta_2)$, respectively. The possible different setups for model selection including nested and non-nested scenarios are sketched in 7.1.

We refer to our results from Section 2.1, where we introduced the Kullback-Leibler divergence to define the best model in a set of models \mathcal{F} . We extend this now by allowing for the comparison of models. In fact, the model with the smaller distance to the true data-generating process should be preferred, where the Kullback-Leibler divergence is taken as a distance measure. 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

$$KL(G(\cdot); F_2(\cdot; \hat{\theta}_2)) = \int \log \frac{g(y)}{f_2(y; \hat{\theta}_2)} dG(y).$$

We should prefer model 1 if its divergence is smaller, that is we prefer model \mathcal{M}_1 if

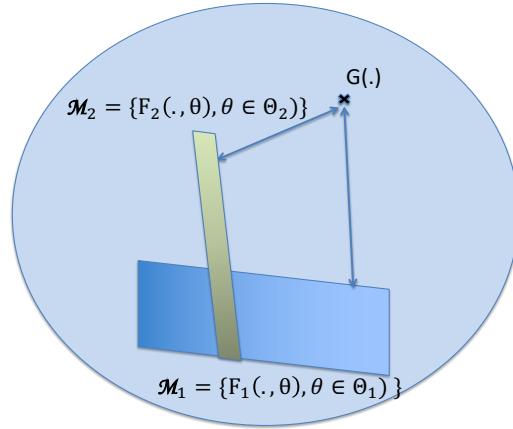
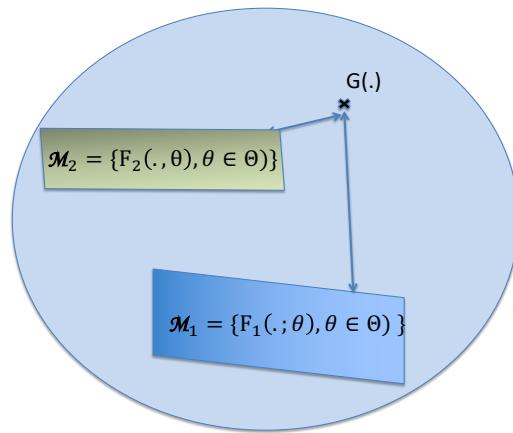
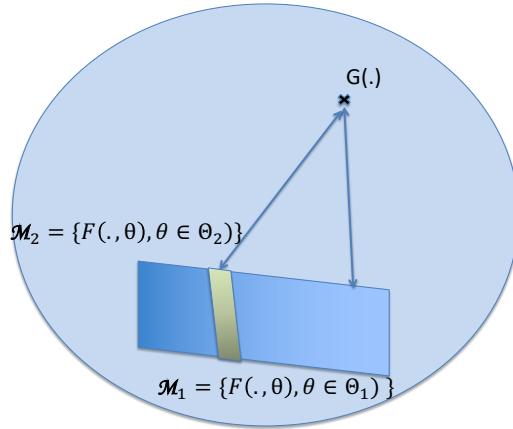


Fig. 7.1 Different settings for model selection. The top row shows nested models, the middle row shows disjoint models, bottom row visualizes non-nested and non-disjoint models

$$\begin{aligned} KL(G(\cdot); F_1(\cdot; \hat{\theta}_1)) - KL(G(\cdot); F_2(\cdot; \hat{\theta}_2)) < 0 &\Leftrightarrow \int \log \frac{f_1(y; \hat{\theta}_1)}{f_2(y; \hat{\theta}_2)} dG(y) > 0. \\ &\Leftrightarrow \text{prefer model } \mathcal{M}_1 \end{aligned} \quad (7.8)$$

Note that we can not 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, \dots, y_n have been drawn from $G(\cdot)$. This would lead to the quantity

$$\frac{1}{n} \sum_{i=1}^n \log \frac{f_1(y_i; \hat{\theta})}{f_2(y_i; \hat{\theta})}. \quad (7.9)$$

However, this quantity does not converge to the integral in (7.8) since we violate the fundamental rule which states "*Never make use of the same data twice*". Note that the integral in (7.8) refers to a new variable Y and in fact we can write (7.8) as the selection rule:

$$\text{If } E \left(\log \frac{f_1(Y; \hat{\theta}_1)}{f_2(Y; \hat{\theta}_2)} \right) > 0 \text{ we prefer } \mathcal{M}_1. \quad (7.10)$$

Now it becomes obvious that the expectation is with respect to Y , some new variable, but $\hat{\theta}_1$ as well as $\hat{\theta}_2$ is based on the data y_1, \dots, y_n . In contrast, in (7.9) we make use of the data twice, first in the sum, where we sum over y_1, \dots, y_n , as well as in the estimate, which is calculated from the same data y_1, \dots, y_n . This is a conceptual flaw that in fact will inevitably lead to overfitting. This is easily seen for nested models. If $\mathcal{F}_1 = \mathcal{F}_2$ and $\mathcal{H}_2 \subset \mathcal{H}_1$ 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 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 (7.9) which will be done in two ways. We start by making the summands explicitly independent of the estimates. This can be done by using $n - 1$ observations for fitting and the remaining observation for estimating the integral. Going through all observations this would lead to

$$\frac{1}{n} \sum_{i=1}^n \log \frac{f_1(y_i; \hat{\theta}_{1,-i})}{f_2(y_i; \hat{\theta}_{2,-i})} > 0 \text{ we prefer } \mathcal{M}_1, \quad (7.11)$$

where $\hat{\theta}_{k,-i}$ is the maximum likelihood estimate in model k calculated from all observations except of the i -the one. Hence we take the data $(y_1, \dots, y_{i-1}, y_{i+1}, \dots, y_n)$ to calculate $\hat{\theta}_{k,-i}$, where $k = 1$ or $k = 2$ and we use the left out observation y_i to estimate the expectation in (7.10). This approach is called **leave-one-out cross-validation**, as sketched below.

1. For $i = 1, \dots, n$ 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 (7.11). If this is larger than 0 prefer model \mathcal{M}_1 otherwise select model \mathcal{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.

1. We divide n into n_{train} and n_{test} , such that $n = n_{\text{train}} + n_{\text{test}}$. Typical decompositions are in the order of 4 to 1.
2. From the database we draw a sample without replacement of size n_{test} leading to the test data. We denote with $\mathcal{N}_{\text{test}} \subset \{1, \dots, n\}$ the index set of selected data points.
3. We train the models on $\mathcal{N}_{\text{train}}$ where $\mathcal{N}_{\text{train}} = \{1, \dots, n\} \setminus \mathcal{N}_{\text{test}}$ and obtain estimates $\hat{\theta}_1$ and $\hat{\theta}_2$.
4. We calculate

$$\sum_{i \in \mathcal{N}_{\text{test}}} \log \frac{f_1(y_i; \hat{\theta}_1)}{f_2(y_i; \hat{\theta}_2)}.$$

If this is > 0 we prefer model \mathcal{M}_1 otherwise we select model \mathcal{M}_2 .

It appears that such a procedure is somewhat data-hungry since we omit all observations in $\mathcal{N}_{\text{test}}$ for fitting the model. This means that if the database is small and further 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. Apparently, any combination is possible as shown in the bottom plot of Figure 7.2. In this case, one selects sets of observations for testing and fits the data with 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.

1. We divide the data into k groups and denote the resulting index sets with $\mathcal{N}_1, \dots, \mathcal{N}_k$ such that
 - (i) $\mathcal{N}_j \cap \mathcal{N}_l = \emptyset$
 - (ii) $|\mathcal{N}_j| \approx |\mathcal{N}_l|$
 - (iii) $\bigcup_{l=1}^k \mathcal{N}_l = \{1, \dots, n\}$
2. We fit the two models on $\mathcal{N} \setminus \mathcal{N}_j$ separately for all $j = 1, \dots, k$ leading to estimates $\hat{\theta}_{1,-\mathcal{N}_j}$ and $\hat{\theta}_{2,-\mathcal{N}_j}$
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 $k = n$ and this is equivalent to leave-out-one cross-validation.

7.3 AIC and BIC

The above procedures require to refit 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 (7.9) and suggest a correction which has been proposed about 50 years ago already by Akaike (1973). Before going into technical details we remind the reader to take a second look at Figure 3.2. We observe that the Kullback-Leibler divergence $KL(G(\cdot), F(\cdot; \hat{\theta}))$ is larger than for the optimal parameter, that is $KL(G(\cdot), F(\cdot; \hat{\theta})) > KL(G(\cdot), F(\cdot; \theta_0))$. The subsequently introduced Akaike Information Criterion aims to find a correction for that. The exact proof is a little bit clumsy to follow and will therefore be given at the end of this section.

Definition 7.1 The **Akaike Information Criterion** (AIC) for model class \mathcal{M} is defined as

$$AIC(\mathcal{M}) = -2 \sum_{i=1}^n \log f(y_i; \hat{\theta}) + 2p \quad (7.12)$$

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 (7.5). The AIC can be seen as an estimate

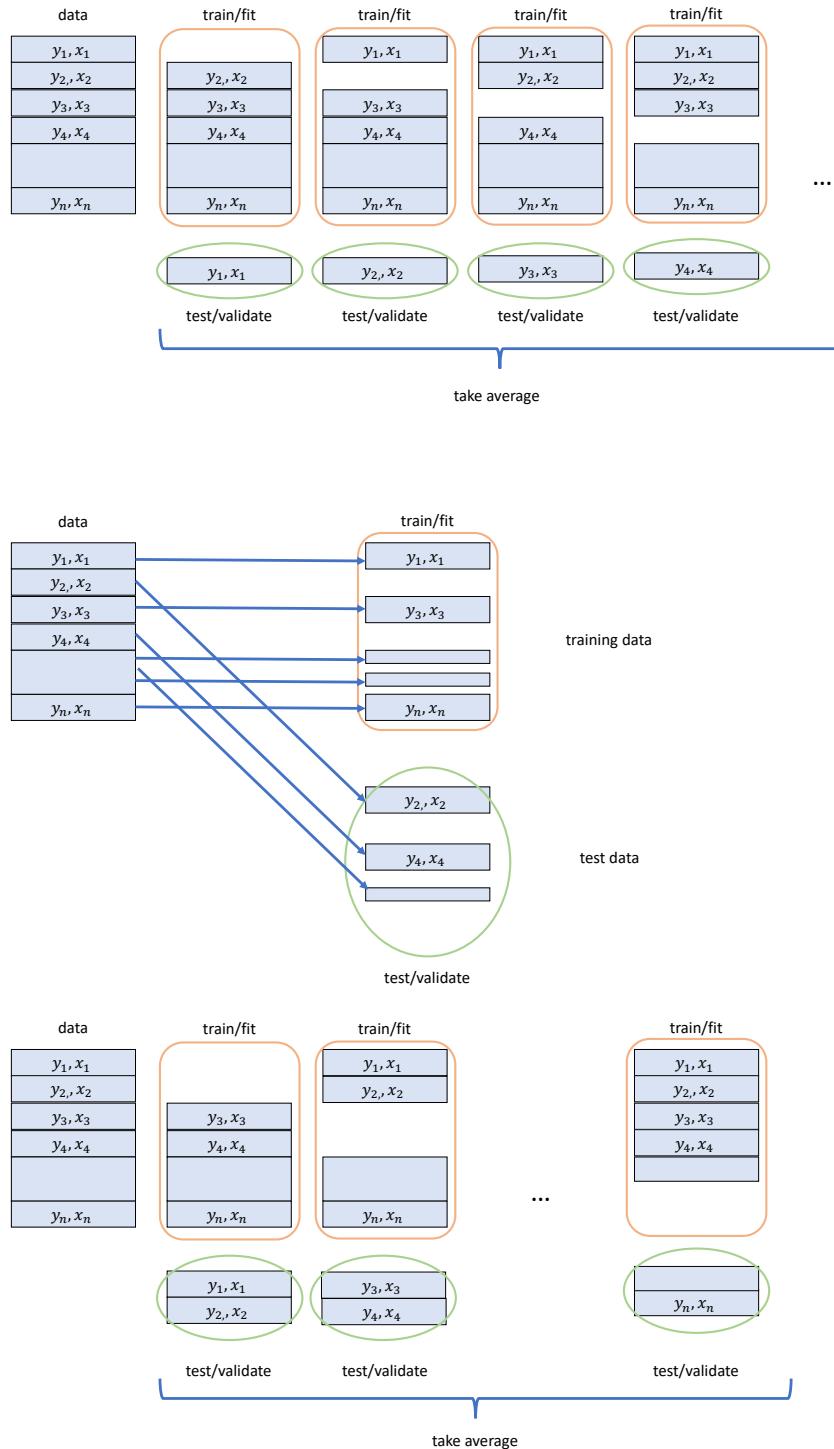


Fig. 7.2 Different versions of splitting data. The top row shows leave-one-out cross-validation. The middle row gives a split to training and test data. The bottom plot shows k -fold cross-validation.

of the quantity

$$2E\{KL(G(\cdot), F(\cdot; \hat{\theta}))\} - 2 \int \log(g(y)) dG(y), \quad (7.13)$$

where the expectation is carried out over the data Y_1, \dots, Y_n .

The factor 2 in (7.12) and (7.13) is arbitrary and in principle not necessary, but used for conventional reasons. Note that with the result (7.13) we can employ the AIC for model selection. In particular, this means that $AIC(\mathcal{M}_1) - AIC(\mathcal{M}_2)$ is an estimate for the following quantity.

$$\begin{aligned} & 2E\{KL(G(\cdot), F_1(\cdot; \hat{\theta}_1))\} - 2E\{KL(G(\cdot), F_2(\cdot; \hat{\theta}_2))\} \\ &= -2E \left\{ \int \log \left(\frac{f_1(y; \hat{\theta}_1)}{f_2(y; \hat{\theta}_2)} \right) dG(y) \right\} \end{aligned}$$

which suggests that if the term is negative then model \mathcal{M}_1 should be preferred and if it is positive then \mathcal{M}_2 is preferable.

The main advantage of the calculation of the AIC is that it does not require refitting the model, which is necessary for 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 if there are constraints on the parameter. 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 (7.12). We will pick up this point again in a later chapter.

Let us write the AIC in the case of normally distributed data. Assume $Y_i \sim N(\mu, \sigma^2)$, 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

$$l(\theta; \sigma^2) = \sum_{i=1}^n \left(\frac{1}{2} \log(\sigma^2) - \frac{(y_i - h(x_i; \theta))^2}{2\sigma^2} \right) \quad (7.14)$$

Differentiating (7.14) with respect to θ and σ^2 yields the maximum likelihood estimates and we find that the variance is estimated through

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (y_i - h(x_i; \hat{\theta}))^2. \quad (7.15)$$

Further calculations show, that the log likelihood at its maximum (ignoring terms not dependent on the parameters) takes the form

$$l(\hat{\theta}; \hat{\sigma}^2) = -\frac{n}{2} \log \left\{ \sum_{i=1}^n (y_i - h(x_i; \hat{\theta}))^2 \right\}. \quad (7.16)$$

Consequently, the AIC results to

$$AIC(\mathcal{M}) = n \log \left\{ \sum_{i=1}^n (y_i - h(x_i; \hat{\theta}))^2 \right\} + 2p, \quad (7.17)$$

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 (7.17) measures the goodness of fit, that is the empirical prediction error. How good 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 Figure 7.3, upper plot.

A companion of the AIC is the BIC, short for **Bayesian Information Criterion**. This is defined as

$$BIC(\mathcal{M}) = -2 \sum_{i=1}^n \log f(y_i; \hat{\theta}) + \log(n)p. \quad (7.18)$$

Hence, we replace the "2" in the latter term in (7.12) 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

$$BIC(\mathcal{M}) = n \log \left\{ \sum_{i=1}^n (y_i - h(x_L; \hat{\theta}))^2 \right\} + \log(n)p. \quad (7.19)$$

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 Figure 7.3.

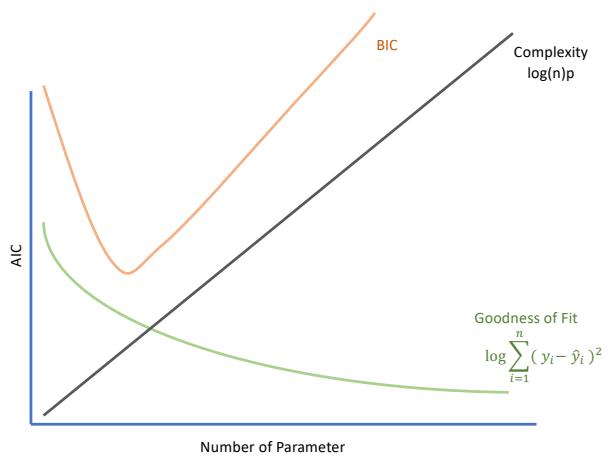
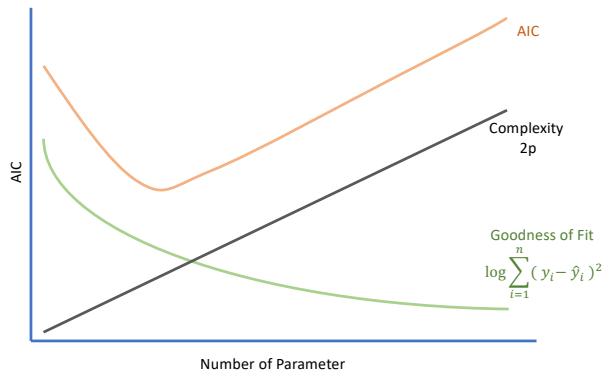


Fig. 7.3 AIC and BIC behavior by balancing goodness of fit with the complexity of a model

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

$$\begin{aligned} & E_{Y_1, \dots, Y_n} \{ KL(G(\cdot), F(\cdot; \hat{\theta}(Y_1, \dots, Y_n))) \} \\ = & \text{const} - E_{Y_1, \dots, Y_n} \left\{ \int \log f(y; \hat{\theta}(Y_1, \dots, Y_n)) dG(y) \right\}, \end{aligned} \quad (7.20)$$

where we made the dependence of our estimate $\hat{\theta}$ on the data y_1, \dots, y_n explicit. The constant term can be ignored in the following. Note that we have two integrals in the second term in (7.20): the inner integral over y in the Kullback-Leibler divergence and the outer integral resulting from the expectation with respect to Y_1, \dots, Y_n . To be explicit, we can write the expectation in the second component of (7.20) as

$$\int \left[\int \log(f(y; \hat{\theta}(y_1, \dots, y_n))) dG(y) \right] dG(y_1, \dots, y_n) \quad (7.21)$$

where $G(y_1, \dots, y_n) = \prod_i G(y_i)$. 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 $\hat{\theta}(y_1, \dots, y_n)$ as $\hat{\theta}$, but it should be kept in mind that $\hat{\theta}$ depends on y_1, \dots, y_n and not on y . This gives the approximation of (7.21) as

$$\begin{aligned} & \int \int \left\{ \log f(y; \theta_0) + \frac{\partial \log f(y; \theta_0)}{\partial \theta^\top} (\hat{\theta} - \theta_0) - \frac{1}{2} (\hat{\theta} - \theta_0)^\top J(\theta_0) (\hat{\theta} - \theta_0) \right\} dG(y) \prod_{i=1}^n dG(y_i) \\ \approx & \int \log(f(y; \theta_0)) dG(y) - \frac{1}{2} \int (\hat{\theta} - \theta_0)^\top I(\theta_0) (\hat{\theta} - \theta_0) \prod_{i=1}^n dG(y_i). \end{aligned} \quad (7.22)$$

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 (7.22). Making use of the asymptotic normality of the maximum likelihood estimate, this can be asymptotically approximated by $\text{tr}(I^{-1}(\theta_0)V(\theta_0))$ where $V(\theta_0)$ is the variance of the score function, see (4.26). Note that neither the integral in the first component in (7.22) 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, \dots, y_n . We do know that the data y_1, \dots, 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 (7.21), but leave $\hat{\theta}$ fixed for now. That is, we replace the expectation over $G(y)$ with the arithmetic mean using y_1, \dots, y_n . To be specific, we calculate

$$\frac{1}{n} \sum_{i=1}^n \log f(y_i; \hat{\theta}). \quad (7.23)$$

This is clearly just the likelihood function at its maximum divided by the sample size. We again expand (7.23) around θ_0 and obtain an approximation of (7.23) with

$$\frac{1}{n} \sum_{i=1}^n \left\{ \log f(y; \theta_0) + s_i(\theta_0; y_i)^\top (\hat{\theta} - \theta_0) - \frac{1}{2} (\hat{\theta} - \theta_0)^\top J_i(\theta_0; y_i) (\hat{\theta} - \theta_0) \right\}. \quad (7.24)$$

Note that the first component in (7.24) can approximate the first component in (7.22). This means

$$\frac{1}{n} \sum_{i=1}^n \log f(y_i; \theta_0) \rightarrow \int \log f(y; \theta_0) dG(y) \quad (7.25)$$

for increasing n , which can in fact be proven to be a consistent estimate. Moreover, because with increasing sample size n

$$\frac{1}{n} \sum_i J_i(\theta_0; y) \rightarrow I(\theta_0), \quad (7.26)$$

we can argue that the integrand of the third component in (7.24) 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, \dots, Y_n and using the asymptotic distribution of the maximum likelihood estimate we get

$$E_{Y_1, \dots, Y_n} \left(\frac{1}{2} (\hat{\theta} - \theta_0)^\top I^{-1}(\theta_0) (\hat{\theta} - \theta_0) \right) = \text{tr}(I^{-1}(\theta_0) V(\theta_0)).$$

What remains is the second component in (7.24). If we subtract this and take (7.25), (7.26) and the approximation (7.22), we get

$$\frac{1}{n} \sum_{i=1}^n \log f(y_i; \hat{\theta}) - \frac{1}{n} \sum_{i=1}^n s_i^\top(\theta_0; y_i) (\hat{\theta} - \theta_0) \rightarrow \int \int \log f(y; \hat{\theta}) dG(y) \prod_{i=1}^n dG(y_i).$$

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 (4.16) gives

$$\frac{1}{n} \sum_i s_i^\top(\theta_0; y_i) (\hat{\theta} - \theta_0) \approx \frac{1}{n} \sum_i s_i^\top(\theta_0; y_i) I^{-1}(\theta_0) \sum_j s_j(\theta_0; y_j).$$

Bearing in mind that $E(s_i(\theta_0; y_i)) = 0$, if we take the expectation with respect to Y_1, \dots, Y_n , we get

$$E_{Y_1, \dots, Y_n} \left(\frac{1}{n} \sum_i s_i(\theta_0; y_i)^\top (\hat{\theta} - \theta_0) \right) \approx \frac{1}{n} \text{tr}(I^{-1}(\theta_0) V(\theta_0)). \quad (7.27)$$

As both $I(\theta_0)$ and $V(\theta_0)$ depend on the unknown distribution $G(\cdot)$, we can not calculate their values. But if $G(\cdot) \in \mathcal{F}$ we have $V(\theta_0) = I(\theta_0)$ so that

$$\frac{1}{n} \text{tr}(I^{-1}(\theta_0) V(\theta_0)) = \frac{1}{n} p,$$

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). \square

Proof Derivation of the BIC: Very similar to the AIC is the Bayesian Information Criterion (BIC). Assume we have a set of models M_1, \dots, M_K , each of which corresponds to a distributional model, such that

$$M_k \Leftrightarrow Y_i \sim f_k(y; \theta_k) \quad i.i.d.$$

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 M_k , this is given by

$$P(M_k|y) = \frac{f(y|M_k)P(M_k)}{\int f(y)P(M_k)dM_k} = \frac{\int f_k(y; \theta_k)f_{\theta_k}(\theta_k)d\theta_k}{\int f_k(y; \theta_k)f_{\theta_k}(\theta_k)d\theta_k} P(M_k), \quad (7.28)$$

where $P(M_k)$ is the prior belief in model M_k . The denominator $f(y)$ is calculated by summing over all models

$$f(y) = \sum_k \int f_k(y; \theta_k)f_{\theta_k}(\theta_k)d\theta_k.$$

Although (7.28) 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

$$l_k(\theta_k) = \sum_{i=1}^n \log f_k(y_i; \theta_k)$$

be the log-likelihood for model M_k and define with $\hat{\theta}_k$ the corresponding Maximum Likelihood estimate. With Laplace approximation, the integral component in (7.28) is given by

$$\int \exp(l_k(\theta_k))f_{\theta_k}(\theta_k)d\theta_k \approx \left(\frac{2\pi}{n}\right)^{\frac{p_k}{2}} \exp(l_k(\hat{\theta}_k))f_{\theta_k}(\hat{\theta}_k)\left|\frac{1}{n}I_k(\hat{\theta}_k)\right|^{-\frac{1}{2}}$$

where p_k is the dimension of θ_k . Taking the log of the right-hand side gives

$$\frac{p_k}{2}\log(2\pi) - \frac{p_k}{2}\log(n) + l_k(\hat{\theta}_k) - \frac{1}{2}\log\left|\frac{1}{n}I_k(\hat{\theta}_k)\right| + \log f_{\theta_k}(\hat{\theta}_k).$$

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)

$$BIC_k = -2l_k(\hat{\theta}_k) + \log(n)p_k.$$

7.4 AIC, BIC and Hypothesis Testing

We may consider model selection also as a testing problem, or vice versa. This means we test on hypotheses, that one or multiple components of the parameter vector θ can be ignored, i.e. $H_0 : \theta_j = 0$ for $j \in \mathcal{J} \subset \{1, \dots, p\}$. To connect the two ideas consider the setting where we assume two nested statistical models M_0 and M_1 with $M_0 \subset M_1$. In other words

$$M_k = \{f(y; \theta), \theta \in \Theta_k\} \text{ for } k = 0, 1$$

with $\Theta_0 \subset \Theta_1$. Assume that $p_0 = |\Theta_0|$ and $p_1 = |\Theta_1|$ are the dimensions of the corresponding parameter spaces and let $p = p_1 - p_0 > 0$. In Chapter 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 \mathcal{M}_0 and accordingly $\hat{\theta}_1$ is the corresponding estimate in \mathcal{M}_1 . It can be shown (see the end of the section), that if model \mathcal{M}_0 holds, then

$$2\{l(\hat{\theta}_1) - l(\hat{\theta}_0)\} \sim \chi_p^2. \quad (7.29)$$

Formulating the hypothesis $H_0 : \mathcal{M} = \mathcal{M}_0$ and the alternative $H_1 : \mathcal{M} = \mathcal{M}_1$, we obtain the decision rule

$$\text{"}H_1\text{"} \Leftrightarrow 2\{l(\hat{\theta}_1) - l(\hat{\theta}_0)\} > \chi_{p,1-\alpha}^2,$$

where $\chi_{p,1-\alpha}^2$ 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

$$\text{"}H_1\text{"} \Leftrightarrow 2\{l(\hat{\theta}_1) - l(\hat{\theta}_0)\} > 2p,$$

while the BIC corresponds to

$$\text{"}H_1\text{"} \Leftrightarrow 2\{l(\hat{\theta}_1) - l(\hat{\theta}_0)\} > \log(n)p.$$

The different critical values, i.e. the threshold values to the right in the decision rules above, are shown in the upper plot in Figure 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.

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 $P(\text{"}H_1\text{"}|H_0)$ using the three decision rules given above. This is shown in the bottom plot of Figure 7.4. We see that both, AIC as well as 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 towards 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. We refer to Claeskens and Hjort (2008) for a more detailed discussion in this direction.

Proof We derive the asymptotic distribution of the likelihood ratio (7.29). Let us for simplicity assume

$$\theta = (\underbrace{\theta_1, \dots, \theta_{p_0}}_{=: \theta^{(0)}}, \underbrace{\theta_{p_0+1}, \dots, \theta_{p_0+p}}_{=: \theta^{(1)}}),$$

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

$$l(\hat{\theta}_1) - l(\theta_0) \xrightarrow{a} \chi^2_{p_0+p}. \quad (7.30)$$

Following similar arguments as in Section 4.4 and using that under H_0 we get $\hat{\theta}_0^{(0)} \approx \hat{\theta}_1^{(0)}$, it follows

$$\begin{aligned} 2\{l(\hat{\theta}_1) - l(\hat{\theta}_0)\} &= 2\{l(\hat{\theta}_1) - l(\theta_0) - (l(\hat{\theta}_0) - l(\theta_0))\} \\ &\approx (\hat{\theta}_1 - \theta_0)^T I^{-1}(\theta_0)(\hat{\theta}_1 - \theta_0) - (\hat{\theta}_0 - \theta_0)^T I^{-1}(\theta_0)(\hat{\theta}_1 - \theta_0) \\ &\approx (\hat{\theta}_1^{(1)} - \theta_0^{(1)})^T [I^{-1}(\theta_0)]^{(1)(1)} (\hat{\theta}_1^{(1)} - \theta_0^{(1)}) \end{aligned}$$

where $[I^{-1}(\theta_0)]^{(1)(1)}$ 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 estimate, we obtain the result. \square

7.5 Data Size and Dimension of the Model

Let us define the dimension of a model \mathcal{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. But we should also look at small data sets, meaning that n is small and asymptotic arguments might not hold. We therefore look at the ratio

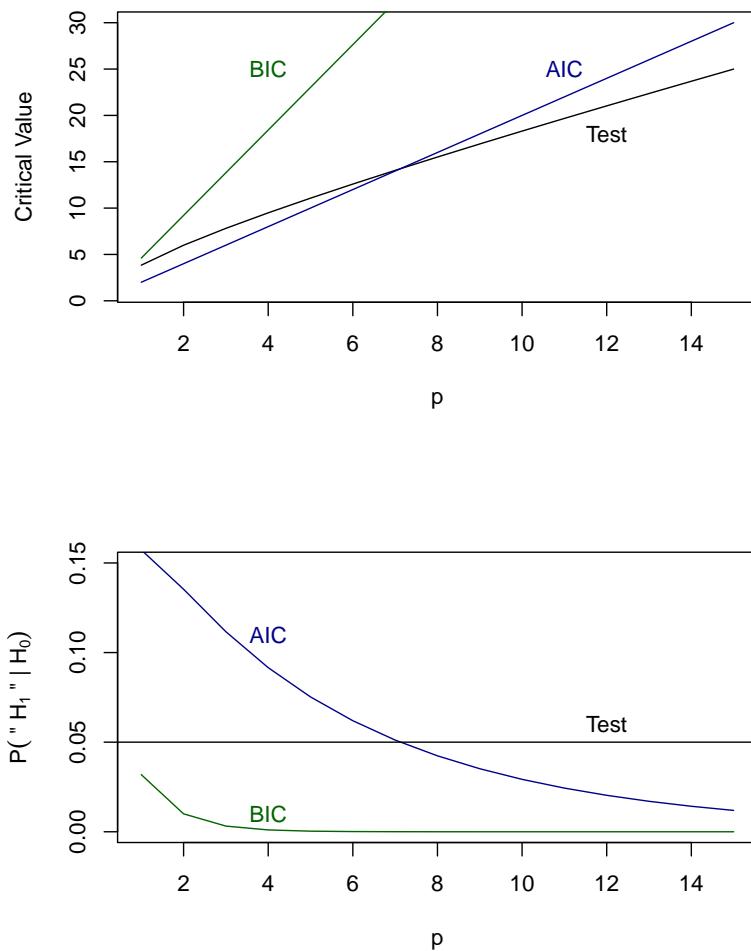


Fig. 7.4 Penalization values for AIC, BIC and hypothesis testing (upper plot) and resulting significance level (bottom plot)

$$\frac{|\Theta|}{n}. \quad (7.31)$$

We can think of three constellations for the above ratio, namely

$$\frac{|\Theta|}{n} \ll 1 \quad \text{or} \quad \frac{|\Theta|}{n} \lesssim 1 \quad \text{or} \quad \frac{|\Theta|}{n} > 1 \quad (7.32)$$

Model selection allows choosing the dimension of the model data-driven, that is $|\Theta|$ depends on n since the model is selected through some model selection criterion. As long as the ratio (7.31) is small, i.e. we are in the left-hand side situation in (7.32), there are multiple possibilities to do so and we will constrain the presentation here to the use of the AIC and BIC. Later we also want to discuss the two constellations in right-hand side cases in (7.32), i.e. when the ratio (7.31) takes values close or even larger than 1. The latter means that the model has more parameters than observations. This is beyond the classical statistical framework but applications in machine learning - labeled as deep learning – have shown that such models can have good, in fact, better prediction properties compared to models where the ratio (7.31) is smaller than 1. A discussion of this phenomenon requires some more technical details, in particular viewpoints from machine learning, which is why we postpone this to Chapter 15.3.2.

Here we want to look at the case of small samples, where n is so small that asymptotic arguments do not necessarily hold. We will demonstrate how small sample corrections can circumvent problems here and propose a small sample correction for the AIC.

For now, we start exploring how the dimension of Θ can be chosen data-driven. We here particularly look at both, AIC and BIC and employ them for model selection. But the same ideas can be carried out through cross-validation, for the price of increased numerical effort for refitting the model. We will not go into detail in that direction and solely make use of AIC and BIC. We have used the latter already for comparing two models. We extend this approach now to a cascade of K nested models $\mathcal{M}_1 \subset \mathcal{M}_2 \subset \dots \mathcal{M}_K$. Exemplary and for demonstration purposes we motivate the idea using a simple linear regression model. Let x_1, \dots, x_K be a set of covariates leading to the overall linear model

$$Y_i = \underbrace{\theta_0 + x_{i1}\theta_1 + x_{i2}\theta_2 + \dots + x_{iK}\theta_K}_{=: h(x; \theta)} + \epsilon_i. \quad (7.33)$$

We aim to find the “relevant” input variables for this model by excluding “irrelevant” covariates. To do so we require some notation. Let \mathcal{M}_K be the model (7.33) with all covariates included. Excluding a single covariate can be notated by the K index sets $\mathcal{I}_{(K-1)(k)} = \{1, \dots, K\} \setminus \{k\}$ and we define with $\mathcal{M}_{(K-1)(k)}$ the regression model with covariates $k \in \mathcal{I}_{(K-1)(k)}$ included. In formula this means

$$\mathcal{M}_{(K-1)(k)} \Leftrightarrow Y_i = \theta_0 + \sum_{j \in \mathcal{I}_{(K-1)(k)}} x_{ij}\theta_j + \epsilon_i. \quad (7.34)$$

Excluding two covariates leads to $\binom{K}{2}$ different index sets which we index with $\mathcal{I}_{(K-2)(k)}$ where $k = 1, \dots, \binom{K}{2}$. Accordingly we can define the index sets $\mathcal{I}_{(k)(l)}$

where $k = 1, \dots, K$ and $l = 1, \dots, \binom{K}{k}$ and the corresponding model sets is denoted as $\mathcal{M}_{(k)(l)}$. 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

$$\hat{\mathcal{M}} = \arg \min\{AIC(\mathcal{M}_{(k)(l)})\}. \quad (7.35)$$

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 $\mathcal{I}_0 = \emptyset$ and set $k = 0$.

2. Let $\mathcal{M}_k = \{1, \dots, K\} \setminus \mathcal{I}_k$ and calculate the AIC for the $|\mathcal{M}_k|$ different models

$$Y_i = \mu + \sum_{k \in \mathcal{I}_k} x_{ik} \theta_k + x_{ij} \theta_j + \epsilon_i$$

for $j \in \mathcal{M}_k$ and choose the one with the minimal AIC value. Let the corresponding index of the covariate be l and define $\mathcal{I}_{k+1} = \mathcal{I}_k \cup \{l\}$. The corresponding AIC value is defined as AIC_{k+1} .

3. If $AIC_{k+1} < AIC_k$ set $k = k + 1$ and jump back to step 2. Otherwise, select the model with \mathcal{M}_k as the best model.

We demonstrate the procedure with the following simulation setup. We simulate from

$$Y_i = \sum_{k=1}^K x_{ik} \theta_k + \epsilon, \quad (7.36)$$

where $\epsilon \sim N(0, 1)$. 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 $N(0, 1)$ distribution. In Figure 7.5 – 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 existent values but also additional spurious ones. The true coefficients are shown in the bottom left plot in Figure 7.5, where we see the selected coefficients as bullet points and the unselected

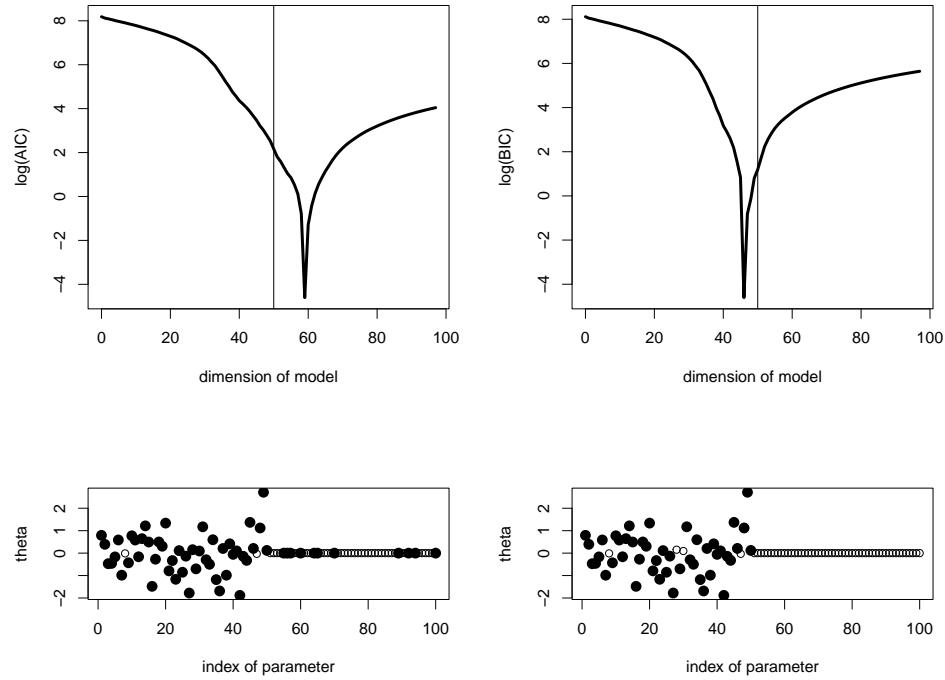


Fig. 7.5 Performance of AIC (left-hand side) and BIC (right-hand side) in the simulation example. The upper row shows the AIC and BIC values, respectively. The bottom shows the true coefficients with the selected one indicated with points

as circles. We experience a well-known weakness of the AIC which has the tendency to choose too complex models.

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 Figure 7.5. 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 Figure 7.5. 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 and hence on the ratio (7.31). 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 Figure 7.6 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

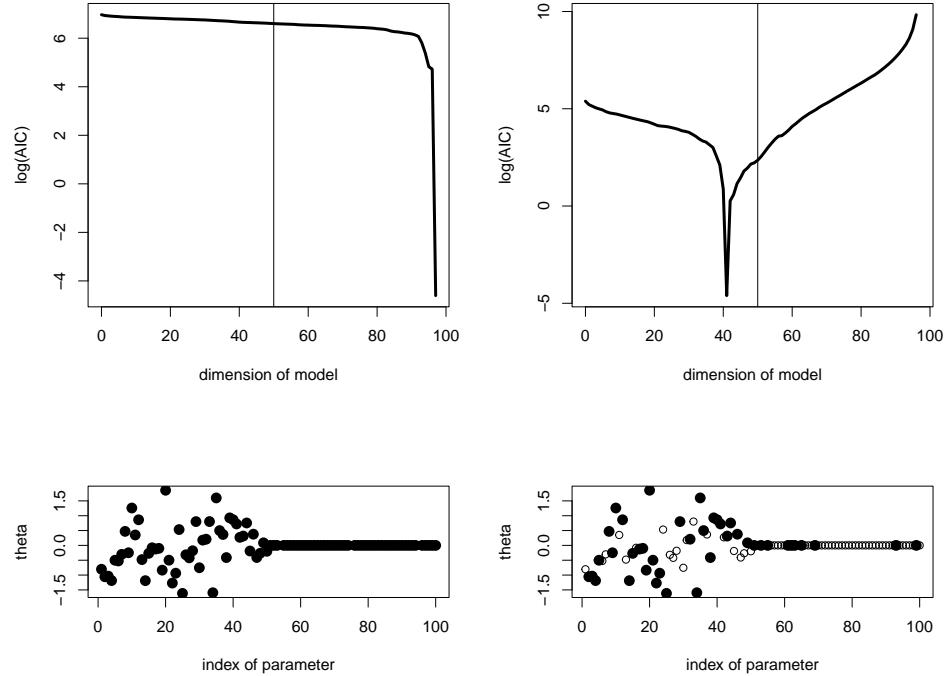


Fig. 7.6 Performance of AIC (left-hand side) and corrected AIC (right-hand side) in the simulation example for small n . Upper row shows the AIC and corrected AIC values, respectively. The bottom shows the true coefficients with the selected one indicated with points

is defined as follows:

$$AIC_c = -2l(\hat{\theta}) + 2p \left(\frac{n}{n-p-1} \right). \quad (7.37)$$

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 (7.37) gets large. We demonstrate the effect of the correction using the above simulation. The results are shown in Figure 7.6 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.

7.6 Exercises

Exercise 1

- Briefly define the Akaike and the Bayesian Information Criteria. What is their function? What is the main (empirical) difference between the two?
- 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?
- 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?
- Suppose we decide to select our model based on AIC instead of BIC. What are the AIC values for the two models?
- Do the two criteria agree? Which of the two criteria imposes a heavier penalty on the number of parameters?

Exercise 2

The data set `teengamb` in the R-package `faraway` contains information on the gambling behaviour of teenagers in Great Britain, and contains the following variables:

variable	description
<code>gamble</code>	money spent on gambling in pounds per year
<code>status</code>	socio-economic status (based on the job of the parents)
<code>income</code>	income (in pounds per week)
<code>verbal</code>	test-score based on the number of properly defined words (max. 12)
<code>sex</code>	sex (0 = male, 1 = female)

- You can fit a linear model in R using the function `lm`. Use the AIC in order to pick between *Model 1*

$$\text{gamble} = \gamma_0 + \gamma_1 \text{status} + \gamma_2 \text{income} + u$$

and *Model 2*

$$\text{gamble} = \psi_0 + \psi_1 \text{verbal} + \psi_2 \text{sex} + \epsilon.$$

Which model would you prefer?

- 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 4-fold cross validation that splits the data into 4 equally sized blocks (you can assume that the data can be evenly divided by 4), iteratively fits 4 models using three of the four blocks and determines the MSEP on the leftover fourth block, respectively. Then, calculate the average of the 4 MSEPs computed in each iteration and return this value.

Chapter 8

Simulation Based Inference

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 get in the foreground. Imagine Mendel with his inheritance experiments described in Example 6 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 it 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 so natural to generate random numbers with the many available software packages, that 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 better to call the generated random numbers "pseudo-random numbers", even though we do not want to be that explicit here. 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 20th 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

$$x_n := (ax_{n-1}) \bmod m \quad (8.1)$$

where "mod" is the modulo operator, that is we divide ax_{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 Figure 8.1 in the top plot. Calculating the resulting empirical distribution function leads to the middle plot shown in Figure 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 Figure 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 independent draws, even though apparently they are functionally dependent.

From now on we drop the term "pseudo" and just speak about generated random variables U_n on $[0, 1]$. We will 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 $g : R \rightarrow R^+$. Consider for instance $g()$ to be some function that is proportional to some density function $f()$, say. 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

$$\int_{-\infty}^{\infty} g(y) dy \quad (8.2)$$

assuming that the function is integrable. If we transform y to some value lying in the interval $[0, 1]$ we obtain $u = h(y)$, where $y = h^{-1}(u)$ and $u \in [0, 1]$. A convenient choice is to work with the logit function (also called the sigmoid function), which takes the form

$$u = u(y) = \frac{\exp(y)}{1 + \exp(y)} \Leftrightarrow y = y(u) = \log\left(\frac{u}{1-u}\right)$$

Then using standard substitution rules integral (8.2) can be rewritten to

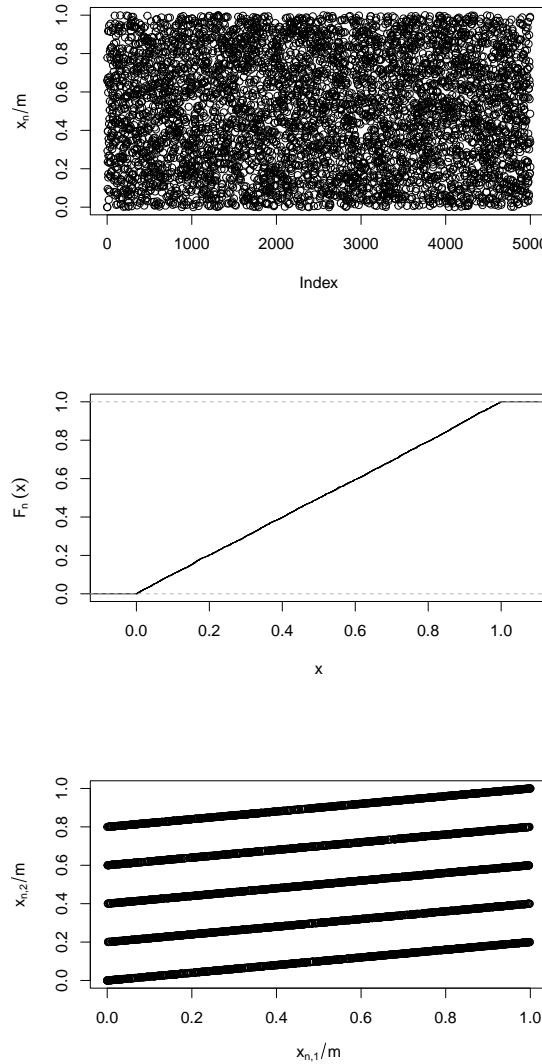


Fig. 8.1 Pseudo random numbers generated through (8.1) with starting value $x_0 = 1$ (top plot) and the resulting empirical distribution function (middle plot). The bottom plot shows two sequences of random numbers with different starting values.

$$\int_{-\infty}^{\infty} g(y) dy = \int_0^1 g(y(u)) \frac{dy(u)}{du} du = \int_0^1 g(y(u)) \underbrace{\frac{1}{u(1-u)}}_{=: \tilde{g}(u)} du \quad (8.3)$$

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

$$E(\tilde{g}(U)) = \int_0^1 \tilde{g}(u) du. \quad (8.4)$$

Drawing now independent uniformly distributed random variables U_i with $i = 1, \dots, N$ we obtain based on the law of large numbers (see A.6)

$$\bar{\tilde{g}} := \frac{1}{N} \sum_{i=1}^N \tilde{g}(U_i) \rightarrow E(\tilde{g}(U)) \quad (8.5)$$

as N increases. Hence, we can generate a sample U_i 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

$$E(\tilde{g}(U)) \in [\bar{\tilde{g}} - 1.96 \frac{\hat{\sigma}}{\sqrt{N}}, \bar{\tilde{g}} + 1.96 \frac{\hat{\sigma}}{\sqrt{N}}] \quad (8.6)$$

where

$$\hat{\sigma}^2 = \frac{1}{N-1} \sum_{i=1}^N (\tilde{g}(u_i) - \bar{\tilde{g}})^2. \quad (8.7)$$

It is interesting that statistical concepts can be used for numerical calculations and we can also utilize the derived results to quantify uncertainty. We want to demonstrate this by calculating the integral

$$\int_{-\infty}^{\infty} \exp\left\{\frac{1}{2}(y-\mu)^2\right\} \quad (8.8)$$

where $\mu \in R$. 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 interval and draw $N = 1000$ uniformly distributed random variables as proposed above. In Figure 8.2 we show the resulting values $\tilde{g}(U_i)$ for $\mu = 0$ on the top while the bottom plot shows $\tilde{g}(U_i)$ if $\mu = 4$. The resulting confidence intervals for the value of the integral are based on the 1000 samples [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

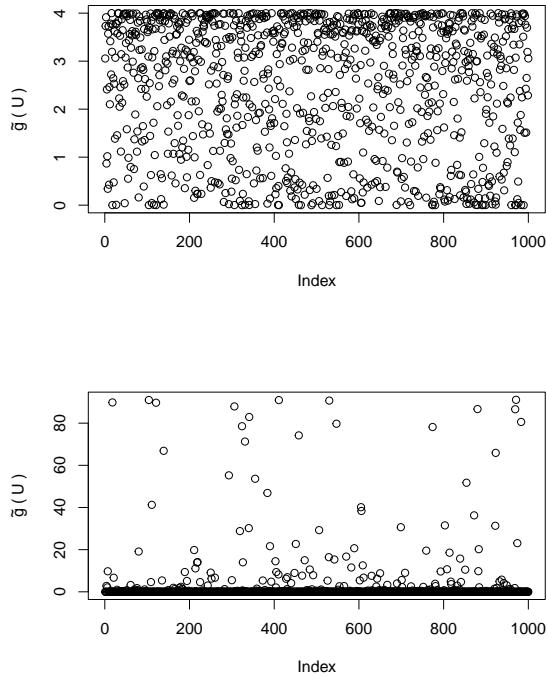


Fig. 8.2 Random numbers $\tilde{g}(U_i)$ for $\mu = 0$ (top) and $\mu = 4$ (bottom)

quality of the approximation depends on the variance which is much larger for the latter case, as can also be seen in Figure 8.2. We will look deeper into this issue later but emphasize for now that integration can be carried out by simulation and classical statistical inference allows us to assess the quality of the approximation.

8.2 Simulating from a Distribution Function

We have derived above how uniformly distributed random variables can be simulated. The next question we want to tackle is how to produce generally distributed random variables. We start with the setting where the distribution function $F(y)$ of a continuous random variable Y is known and invertible. We can then draw a random sample by utilizing the inverse F^{-1} of the distribution function.

Property 8.1 Let $Y = F^{-1}(U)$ where U is uniformly distributed on $[0, 1]$. Then Y is distributed with distribution function $F(\cdot)$ □

The proof of the statement is rather simple.

Proof Note that for a uniformly distributed random variable on $[0, 1]$ one has $P(U \leq u) = u$ for $u \in [0, 1]$. Hence

$$\begin{aligned} P(Y \leq y) &= P(F^{-1}(U) \leq y) \\ &= P(F(F^{-1}(U)) \leq F(y)) \\ &= P(U \leq F(y)) = F(y). \end{aligned}$$

This concludes the proof. □

The simulation idea is sketched in Figure 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 Figure 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 property 8.1.

Let us now go a step ahead and assume that the distribution function of 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 can not apply property 8.1. In this case, we can employ so-called **rejection sampling** which works as follows. Assume that we can easily simulate from a distribution $F^*(\cdot)$ and let $f^*(\cdot)$ denote the corresponding density. In other words, we know $F^*(\cdot)$ and its inverse and the corresponding density $f^*(\cdot)$. Assume further that there exists a finite constant a such that $f(y) < af^*(y)$ for all $y \in R$. Then the rejection sampling procedure is as follows.

1. Draw Y^* from $F^*(\cdot)$ using Property 8.1 .
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. We can show that the accepted values Y^* that

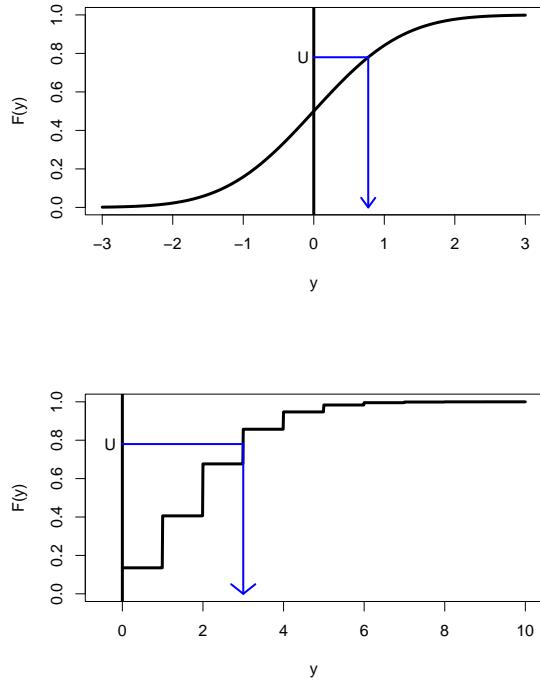


Fig. 8.3 Sketch for drawing random variables with known distribution

pass step 3 above are in fact drawn from the distribution $F(\cdot)$ and not $F^*(\cdot)$. This is shown in the following proof.

Proof Note that values, which pass step 3 have the distribution function

$$P\left(Y^* \leq y | U \leq \frac{f(Y^*)}{af^*(Y^*)}\right). \quad (8.9)$$

We need to prove that this equals $F(y)$. Following Bayes rule we obtain

$$P\left(Y^* \leq y | U \leq \frac{f(Y^*)}{af^*(Y^*)}\right) = \frac{P\left(Y^* \leq y, U \leq \frac{f(Y^*)}{af^*(Y^*)}\right)}{P\left(U \leq \frac{f(Y^*)}{af^*(Y^*)}\right)}. \quad (8.10)$$

Let us first look at the numerator. This can be rewritten as

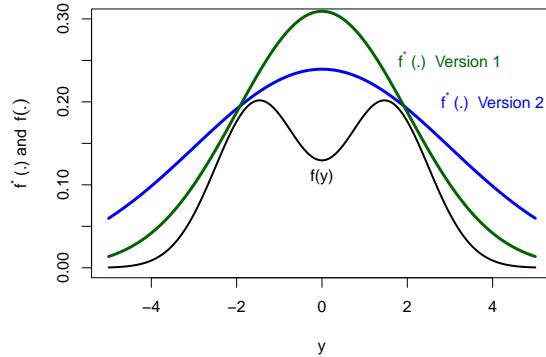


Fig. 8.4 Sketch for rejection sampling with two different proposal distribution

$$\begin{aligned}
 P\left(Y^* \leq y, U \leq \frac{f(Y^*)}{af^*(Y^*)}\right) &= \int_{-\infty}^y P\left(U \leq \frac{f(y^*)}{af^*(y^*)}\right) f^*(y^*) dy^* \\
 &= \int_{-\infty}^y \frac{f(y^*)}{af^*(y^*)} f^*(y^*) dy^* \\
 &= \frac{1}{a} F(y)
 \end{aligned}$$

The denominator in (8.10) is the same as the numerator with y set to infinity. Since $F(\infty) = 1$ the result follows immediately. \square

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 (8.10) 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 Figure 8.4, where we show two possible proposal distributions $F^*(\cdot)$ 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 $1/a$ can thereby also be interpreted considering the 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 mean value $1/a$.

An alternative to rejection sampling is **importance sampling**. Assume we are interested in calculating the expected value $E(h(Y))$ for some known function $h(\cdot)$ for the random variable Y , which is distributed according to $F(\cdot)$. Like above we can not directly sample from $F(\cdot)$. Instead, we again can simulate from $F^*(\cdot)$ and rewrite the quantity of interest as follows.

$$\begin{aligned}
E_F(h(Y)) &= \int_{-\infty}^{\infty} h(y) f(y) dy \\
&= \int_{-\infty}^{\infty} h(y) \frac{f(y)}{f^*(y)} f^*(y) dy \\
&= E_{F^*}\left(h(Y) \frac{f(Y)}{f^*(Y)}\right),
\end{aligned}$$

where the index at the expectation refers to the distribution the expectation is taken over. Using now *i.i.d* samples $Y_i^*, i = 1, \dots, N$ drawn from $F^*(\cdot)$ we obtain

$$\frac{1}{N} \sum_{i=1}^N h(y_i^*) \frac{f(y_i^*)}{f^*(y_i^*)} \rightarrow E\left(h(Y)\right).$$

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 weak accuracy, if the distribution $F^*(\cdot)$ from which we sample is not a good proxy for the distribution we want to sample from.

8.3 Simulating Proportional to a Density (MCMC)

The following simulation technique is a pillar-stone in Bayesian statistics, as we will see in the next chapter. So far have shown how to derive simulations if the distribution function is known, or if only the density $f(\cdot)$ is known but not the distribution function. We want to extend the simulation task now to settings where the density is only known up to an unknown proportionality factor. Hence, instead of the true density $f(y)$ one only knows $\tilde{f}(y)$ which is proportional to the density, i.e. $\tilde{f}(y) \propto f(y)$. Putting it differently, we have

$$f(y) = \frac{1}{a} \tilde{f}(y) \Leftrightarrow \tilde{f}(y) = af(y) \quad (8.11)$$

where constant a is unknown. The procedure goes back to Metropolis et al. (1953) and Hastings (1970) and is heavily used in Bayesian statistics, as further discussed in Chapter 9. We present the procedure here in its general form. The idea is to construct a Markov chain $Y^{*(1)}, Y^{*(2)}, Y^{*(3)}, \dots, Y^{*(T)}, \dots$ as sketched in Figure 8.5. Given the current value of the chain, denoted by $Y^{*(t)}$, we draw the next value from the distribution function $Q(y|y^{*(t)})$. This distribution function is also called transition probability. The Markov chain is said to have a stationary distribution $F(y)$ if for increasing t the probability of $Y^{*(t)}$ converges to $F(y)$. This requires a

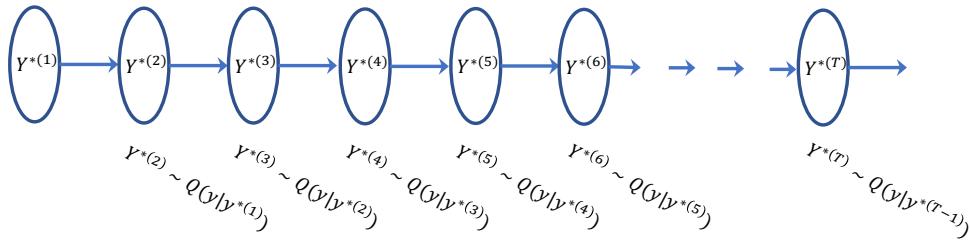


Fig. 8.5 Markov chain with transition probabilities $Q(y^{*(t)} | y^{*(t-1)})$.

number of technical details which lie beyond the scope of the material in this book. We, therefore, 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 general framework is usually referred to as **Markov Chain Monte Carlo**, or in short **MCMC**.

The procedure presented here is called Metropolis-Hastings algorithm. The goal is to derive a transition probability $Q(y|y^{(t)})$ such that the Markov chain does have a stationary distribution that is in fact is $F(y)$, where $F(y)$ is the distribution function to density $f(y)$, which, as said above, is only known up to a proportional function $\tilde{f}(y)$. 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 the proposal distribution

$$Y^* \sim H(y|y^{*(t)}) \quad (8.12)$$

where $H(\cdot|y^{*(t)})$ is a known distribution from which we can easily simulate. Note that $H(\cdot|y^{*(t)})$ can depend on $y^{*(t)}$ and let $h(y|y^{*(t)})$ denote the corresponding density.

3. Define the acceptance probability

$$\alpha(y^*, y^{*(t)}) = \min\left(1, \frac{\tilde{f}(y^*)}{\tilde{f}(y^{*(t)})} \frac{h(y^{*(t)}|y^*)}{h(y^*|y^{*(t)})}\right) \quad (8.13)$$

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 takes the required number of steps.

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 (8.11). The normalization constant however cancels out when calculating the acceptance probability in (8.13). 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 (8.13) 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 variants of the algorithm. The rejection step 5 above can for instance be reformulated to defining $y^{*(t+1)} = y^{*(t)}$, meaning that the Markov chain does not move.

We want to demonstrate the performance of the algorithm with a simple example. In Figure 8.6 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 centred at $\mu = y^{*(t)}$. This is a symmetric distribution so that the ratio $h(y^{*(t)}|y^*)/h(y^*|y^{*(t)}) \equiv 1$ in (8.13). 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 Figure 8.7 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 Figure 8.7. 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 (8.13). 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 20 percent. Clearly, this shows that the procedure works best if the acceptance probability is not too high, with a rule of thumb stating it should be in the order of 20 to 30 percent.

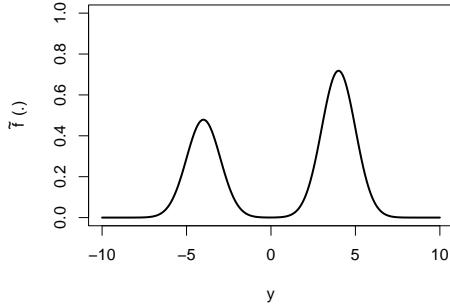


Fig. 8.6 Function $f^*(y)$ which is proportional to a density

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 Figure 8.7 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 Chapter 12.1) from the 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 $Y_i^* = Y^{*(d_0+i*d)}$, 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 $F(y)$.

8.4 Simulating from a Multivariate Normal Distribution

The above sections dealt with univariate random variables. But how can we simulate multivariate random variables with a prespecified dependence structure? We only give some general ideas here and it should be clear that multivariate dependencies can be complex, and so would simulation processes.

Most central is certainly the multivariate normal distribution which we want to treat first. The aim is to generate simulations from

$$Y \sim N(\mu, \Sigma)$$

with $Y = (Y_1, \dots, Y_p)^\top$ and Σ as prespecified 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 $Z^* = (Z_1^*, \dots, Z_p^*)^\top$. Next, we decompose Σ through singular value decomposition to

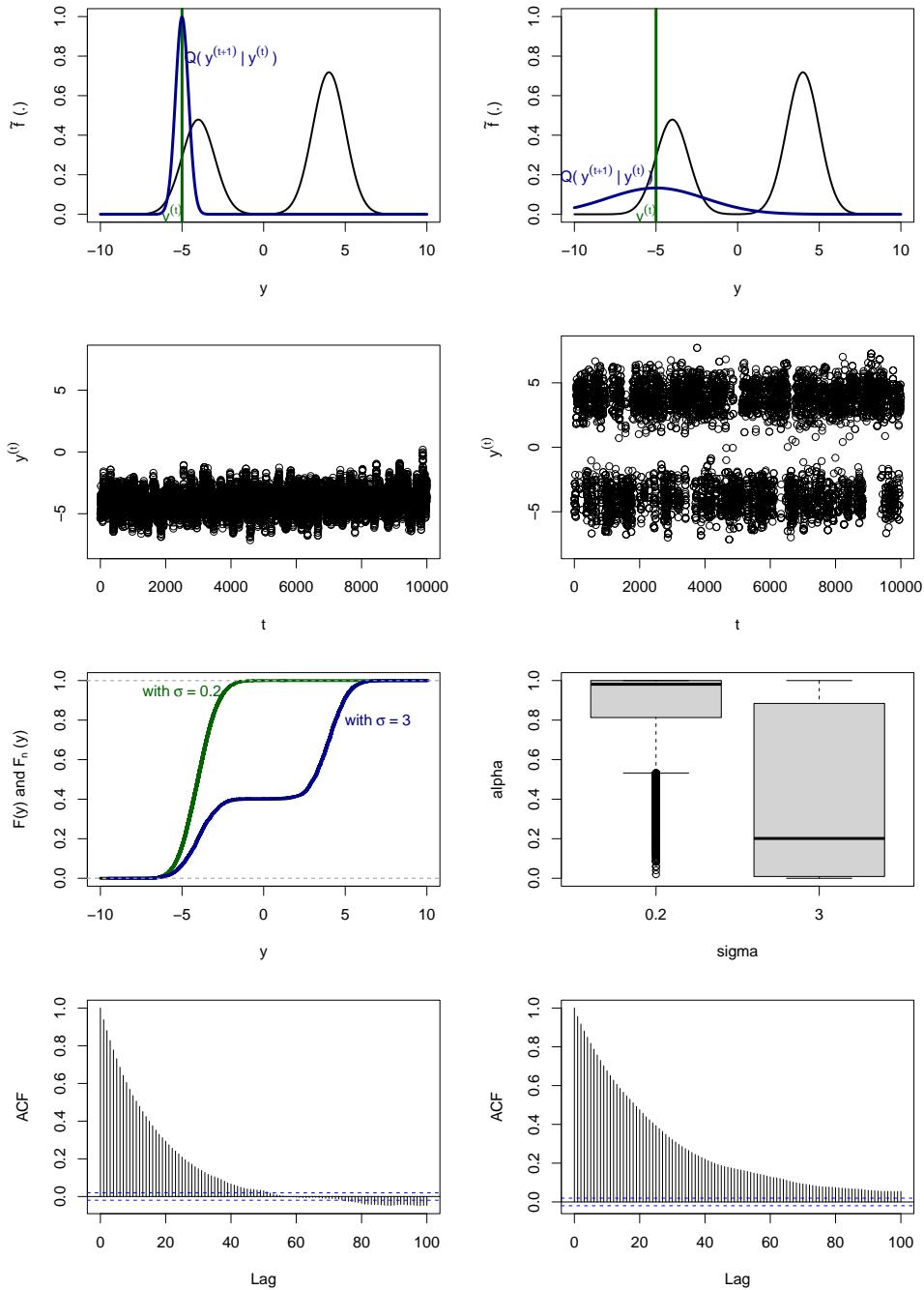


Fig. 8.7 Top row: Function $f^*(y)$ and two normal proposal distributions. Left-hand side with $\sigma = 0.2$, right-hand side with $\sigma = 3$.

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 (8.13) for the two settings. Markov chain of first (top) and second (bottom) order.

Bottom row: Autocorrelation $\text{Cor}(Y^{(t)}, Y^{(t-d)})$ for different lags d .

$$\Sigma = L \Lambda L^\top, \quad (8.14)$$

where L are (orthonormal) eigenvectors of Σ and $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_p)$ 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 $j = 1, \dots, p$. With this property and (8.14) we can define

$$\Sigma^{\frac{1}{2}} = L \Lambda^{\frac{1}{2}} L^\top, \quad (8.15)$$

where $\Lambda^{1/2} = \text{diag}(\sqrt{\lambda_1}, \dots, \sqrt{\lambda_p})$. If we now define

$$Y^* = \Sigma^{\frac{1}{2}} Z^* + \mu, \quad (8.16)$$

we obtain, since $Z^* \sim N(0, I_p)$ with I_p as p dimensional identity matrix that,

$$\begin{aligned} \text{Var}(Y^*) &= \Sigma^{\frac{1}{2}} I \Sigma^{\frac{1}{2}} \\ &= L \Lambda^{\frac{1}{2}} \underbrace{L^\top L}_{=I_p} \Lambda^{\frac{1}{2}} L^\top \\ &= L \Lambda^{\frac{1}{2}} \Lambda^{\frac{1}{2}} L^\top \\ &= L \Lambda L^\top \\ &= \Sigma. \end{aligned}$$

Consequently $Y^* \sim N(\mu, \Sigma)$, as desired.

A different approach to simulate multivariate random variables is to factorize the distribution. Note that we can rewrite any multivariate distribution through

$$\begin{aligned} f_Y(y_1, \dots, y_p) &= f_{Y_1}(y_1) \\ &\times f_{Y_2|Y_1}(y_2|y_1) \\ &\times f_{Y_3|Y_1, Y_2}(y_3|y_1, y_2) \\ &\cdots f_{Y_p|Y_1, \dots, Y_{p-1}}(y_p|y_1, \dots, y_{p-1}). \end{aligned} \quad (8.17)$$

For a multivariate normal distribution, we can derive the following properties. Assume that Y is decomposed into two subvectors Y_A and Y_B , so that $Y = (Y_A^\top, Y_B^\top)^\top$. With

$$Y = \begin{pmatrix} Y_A \\ Y_B \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_A \\ \mu_B \end{pmatrix}, \begin{pmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{pmatrix}\right) \quad (8.18)$$

we denote the subvectors and submatrices of μ and Σ , respectively. Then the following two properties hold

$$Y_A \sim N(\mu_A, \Sigma_{AA}) \quad (8.19)$$

$$Y_B | (Y_A = y_A) \sim N(\mu_B + \Sigma_{BA} \Sigma_{AA}^{-1} (y_A - \mu_A), \Sigma_{BB} - \Sigma_{BA} \Sigma_{AA}^{-1} \Sigma_{AB}). \quad (8.20)$$

With these results, we can now easily derive the separate components in (8.17). For instance for $p = 2$ we first simulate

$$Y_1^* \sim N(\mu_1, \Sigma_{11})$$

and with simulated value y_1^* use (8.20) and draw Y_2^* from

$$Y_2^* | y_1^* \sim N(\mu_2 + \Sigma_{21} \Sigma_{11}^{-1} (y_1^* - \mu_1), \Sigma_{22} - \Sigma_{21}^2 / \Sigma_{11}),$$

which in turn provides a random variable drawn from

$$Y^* = \begin{pmatrix} Y_1^* \\ Y_2^* \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\right). \quad (8.21)$$

8.5 Gibbs Sampling and Markov Fields

The previous section was constrained to multivariate normal distributions. We want to extend the view now towards simulations from general multivariate distributions. Assume the multivariate random vector $Y = (Y_1, \dots, Y_p)^\top$ which we want to simulate from

$$Y^* \sim F(\cdot) = F_{Y_1, \dots, Y_p}(\cdot). \quad (8.22)$$

The index notation used (8.22) indicates the set of random variables covered by the distribution function. The notation will be helpful, as will be seen below. The multivariate distribution function in (8.22) might not be known, but we assume the conditional distributions to be known for each component of Y given the remaining variables. That is we know

$$F_{Y_j|Y_{-j}}(\cdot|y_{-j}) \quad (8.23)$$

where index $-j$ refers to all components of the vector except the j -th one, that is $Y_{-j} = (Y_1, \dots, Y_{j-1}, Y_{j+1}, \dots, Y_p)$. The idea is to construct a Markov chain of simulations $Y^{*(t)}$, comparable to MCMC proposed above. In each 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

$$Y_j^* | (Y_{-j} = y_{-j}) \sim F_{Y_j|Y_{-j}}(\cdot|y_{-j})$$

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 $j = 1, \dots, p$ simulate (and hence overwrite) the j -th component of Y^* through

$$Y_j^* \sim F_{Y_j|Y_{-j}}(\cdot|y_{-j}^*) \quad (8.24)$$

4. Set $Y^{*(t+1)} = Y^*$ and proceed with step 2 until the sufficient number of iterations have been carried out.

The algorithm is sketched in Figure 8.8 for two variables. We demonstrate the performance with a simple example. We consider a bivariate normally distributed random variable with conditional distributions

$$Y_1|y_2 \sim N(0.75y_2, 1) \text{ and } Y_2|y_1 \sim N(0.75y_1, 1). \quad (8.25)$$

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 Figure 8.9. We see speedy convergence of the algorithm, even if the starting point is far away from the centre of the distribution.

The simulated data come from a joint normal distribution and in this case, we can even calculate the parameters exactly. We write

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}\right). \quad (8.26)$$

Using (8.19) and (8.20) and the simulation setup (8.25) to obtain

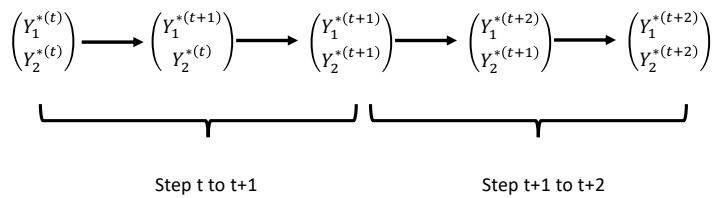


Fig. 8.8 Sketch of Gibbs sampling

$$\sigma_{12}/\sigma_{22} = 0.75 \text{ and } \sigma_{11} - \sigma_{12}^2/\sigma_{22} = 1.$$

Due to the symmetry of the simulation setup (8.25) we have $\sigma_{11} = \sigma_{22} =: \sigma^2$ and apparently we also generally have $\sigma_{12} = \sigma_{21}$. This allows 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 to 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 a case where a joint distribution exists and modify the previous example by replacing the simulation steps to

$$Y_1|y_2 \sim N(0.75y_2, 1) \text{ and } Y_2|y_1 \sim N(0.75y_1, \exp(y_1)/(1 + \exp(y_1))). \quad (8.27)$$

Hence, we induce variance heteroscedasticity for the second component. The resulting simulated data (with starting point $(0,0)$) are shown in Figure 8.10. In this case, the joint distribution has no simple structure.

The last example and the simulation steps (8.24) of the Gibbs algorithm show another interesting and important property. We do not need the full distribution for simulations, instead, we can only use the conditional distributions for each single component of the random vector. In other words, if we only have for each com-

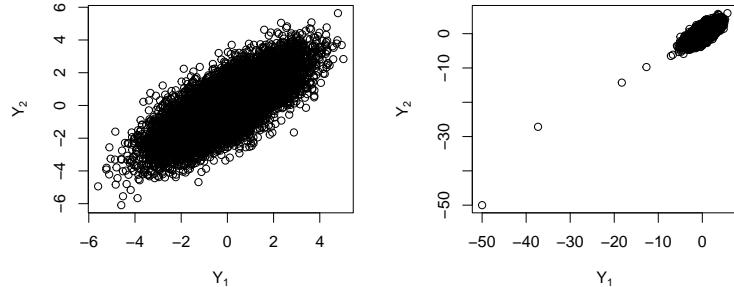


Fig. 8.9 Example of Gibbs sampling with two different starting points

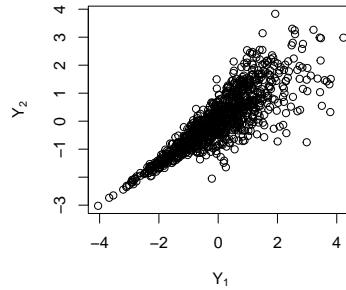


Fig. 8.10 Example of Gibbs sampling more complex conditional distributions

ponent Y_j the conditional distribution given all other variables Y_{-j} , we can make use of Gibbs sampling to obtain random draws from the unconditional distribution. Indeed, it can be shown that if we iteratively simulate from (8.23), the generated data are drawn from the unconditional distribution (8.22), if the Markov chain has a stationary distribution. This property is sketched in a proof at the end of this section.

The Gibbs sampling idea can easily be combined with the different sampling strategies that we discussed in Section 8.1. Note that we assumed in (8.24) that we know the conditional distribution functions for Y_j given Y_{-j} . If we instead only know the condition density $f_{Y_j|Y_{-j}}(\cdot|y_{-j})$, we can replace the simulation step (8.24) in the algorithm above by rejection sampling. If we know the conditional densities $f_{Y_j|Y_{-j}}(\cdot|y_{-j})$ 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 Chapter 9.

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, consider $Y = (Y_1, \dots, Y_p)$. We might want to assume some neighbourhood 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 Figure 8.11. In the top plot, we show a Markov structure of first order. This means that Y_j is dependent on the neighbours 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 and it allows for the factorization

$$f_{Y_1, \dots, Y_p}(y_1, \dots, y_p) = f_{Y_1}(y_1) \prod_{j=2}^p f_{Y_j|Y_{j-1}}(y_j|y_{j-1}).$$

The bottom plot in Figure 8.11 extends the setting towards a Markov chain of second order. In this case, one can use the factorization

$$f_{Y_1, \dots, Y_p}(y_1, \dots, y_p) = f_{Y_1, Y_2}(y_1, y_2) \prod_{j=3}^p f_{Y_j|Y_{j-1}, Y_{j-2}}(y_j|y_{j-1}, y_{j-2}).$$

Let 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 , that are the directly dependent quantities as shown in Figure 8.11. To simulate data from such a setup, one replaces the simulation step (8.24) with the simplified version

$$Y_j^*|Y_{-j} \sim F_{Y_j|Y_{\mathcal{N}_j}}(\cdot|y_{\mathcal{N}_j})$$

where we now only condition on the direct "neighbours".

The framework can be extended to Markov fields, which are sketched in Figure 8.12, top plot. Such an approach is suitable if the data carry some "spatial" structure. We can now write Y in matrix form $Y = (Y_{ij}, i = 1, \dots, N_1, j = 1, \dots, N_2)$ and induce a neighbour 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 $|j - l| > 1$ given the direct form neighbours, as visualized in Figure 8.12 in the bottom plot. Sampling can now be carried out with Gibbs sampling by replacing the simulation step in the algorithm above through

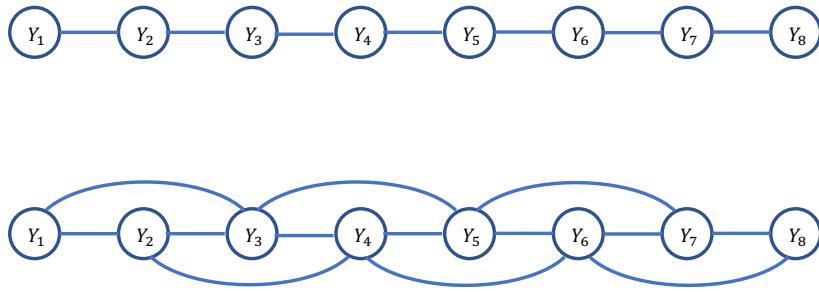


Fig. 8.11 Markov chain of first (top) and second (bottom) order

$$Y_j^* | y_{\mathcal{N}_j}^* \sim F_{Y_j|y_{\mathcal{N}_j}}(\cdot | Y_{\mathcal{N}_j} = y_{\mathcal{N}_j}^*) \quad (8.28)$$

Iterating over all indexed variables concludes one step in the Gibbs sampler.

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 however 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 Figure 8.13 in the top left plot. We evaluate the function at 25 points, sketched together with the isolines of the function in Figure 8.13 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 Figure 8.12. We assume that Y_{ij} given the neighbours follows a normal distribution

$$Y_{ij}^* | y_{\mathcal{N}_{ij}}^* \sim N(\bar{y}_{\mathcal{N}_{ij}}, \sigma_t^2), \quad (8.29)$$

where $\bar{y}_{\mathcal{N}_{ij}}$ 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 (8.29). To start, we just simulate

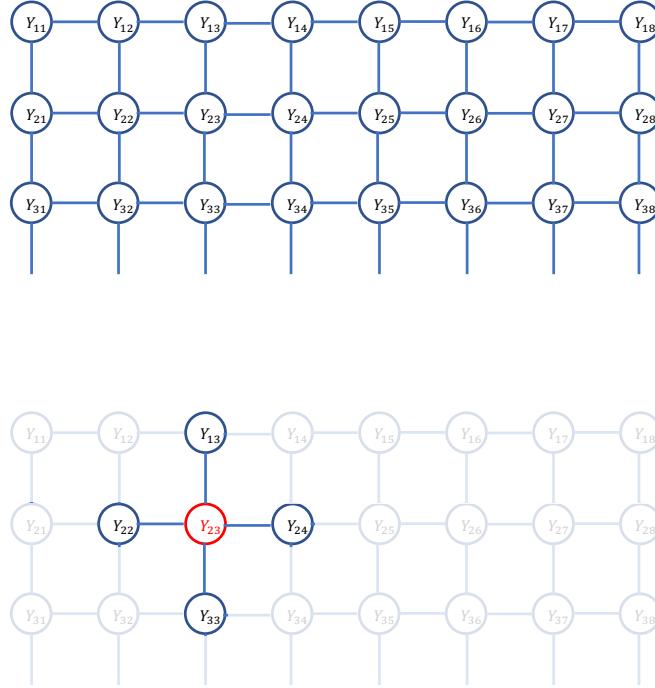


Fig. 8.12 Markov field (top) and dependence structure when conditioning on the neighbours.

from a normal distribution $Y_{ij}^{*(0)} \sim N(0.5, 0.5^2)$, where again, the 25 pixels from the function remain unchanged. This gives an image as shown in the middle left plot of Figure 8.13. Though this looks like random noise, bear in mind that 25 pixels are in fact fixed. We then start the Gibbs sampling using (8.29), except for the 25 evaluated points. We run the algorithm for 600 loops, where the variance in (8.29) is thereby shrinking in t with order $\sigma_t = 1/\log(t)$. We plot the simulated values $Y_{ij}^{*(600)}$ 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.

We can now also analyse 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 Figure 8.13. 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.

Proof We want to sketch that a single Gibbs step has the unconditional distribution as 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 note with $f_2(\cdot)$ the marginal density for Y_2 and $f_{1|2}(\cdot|y_2)$ 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 afterwards 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, that marginal distribution is the stationary distribution in the Gibbs algorithm.

Note that

$$\begin{aligned} P(Y_1^* \leq y_1^*) &= \int P(Y_1^* \leq y_1^* | y_1) f_1(y_1) dy_1 \\ &= \int \int F_{1|2}(y_1^* | y_2^*) f_{2|1}(y_2^* | y_1) f_1(y_1) dy_1 dy_2^* \\ &= \int F_{1|2}(y_1^* | y_2^*) \left(\int f_{2|1}(y_2^* | y_1) f_1(y_1) dy_1 \right) dy_2^* \\ &= \int F_{1|2}(y_1^* | y_2^*) f_2(y_2^*) dy_2^* \\ &= F_1(y_1^*) \end{aligned}$$

As can be seen, the reasoning is rather straightforward. \square

8.6 Uncertain Input Parameters

Usually, each simulation requires the specification of input (or tuning) parameters, which we can call θ . These parameters are often not completely known and there is uncertainty about properly specifying the parameter for the simulation. We may quantify this uncertainty with respect to some probability distribution $f_\theta(\theta)$ for $\theta \in \Theta$. Given θ we draw Y^* from

$$Y|\theta \sim F(\cdot; \theta) \quad (8.30)$$

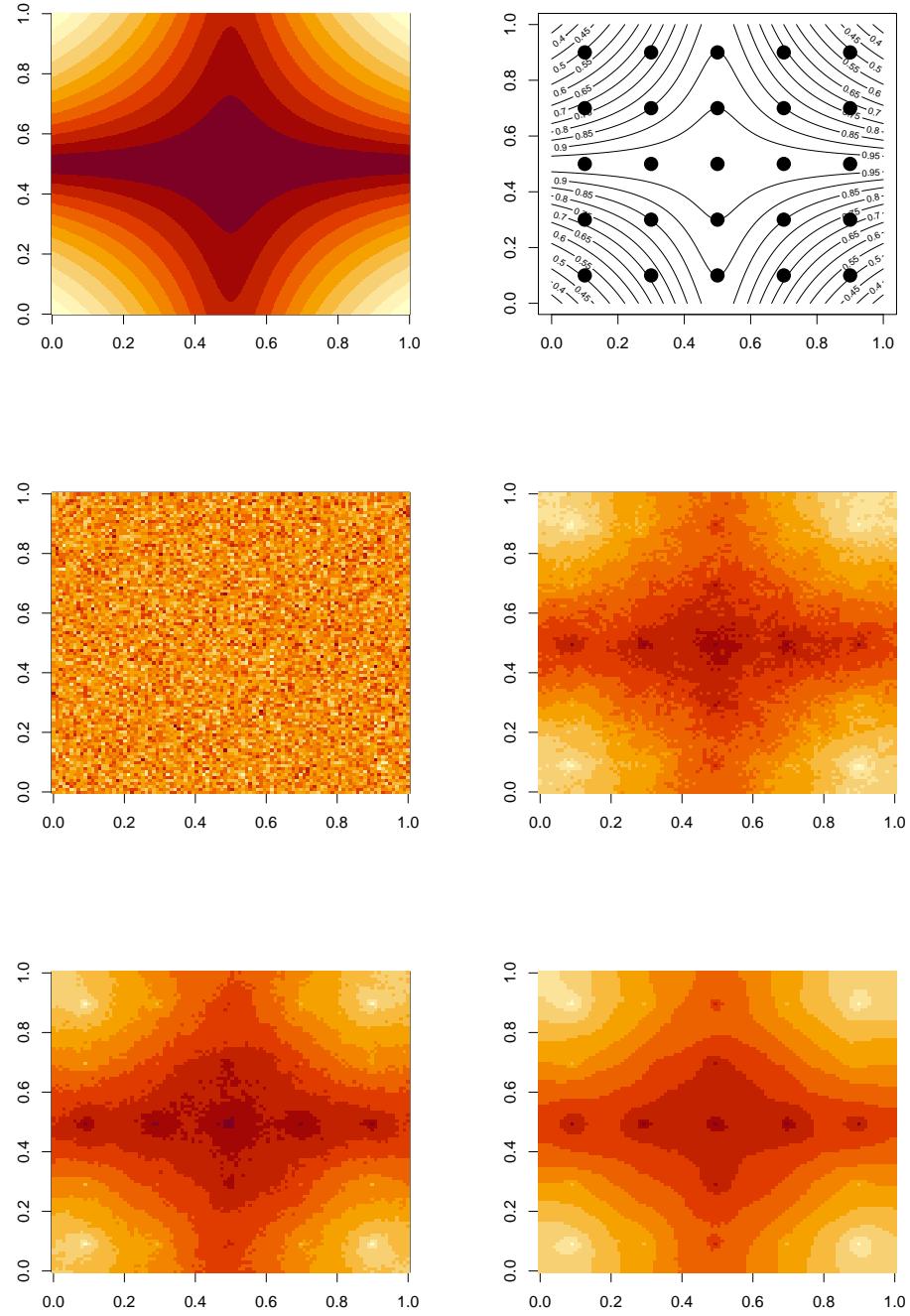


Fig. 8.13 Top row: Function $h(y_1, y_2)$ and the 25 data points at which the function is evaluated. Middle row: Simulated Gaussian Markov Random Field. Starting value to the left, after 1000 iterations to the right. Bottom row: Maximum (left) and mean (right) of the last 100 simulated values.

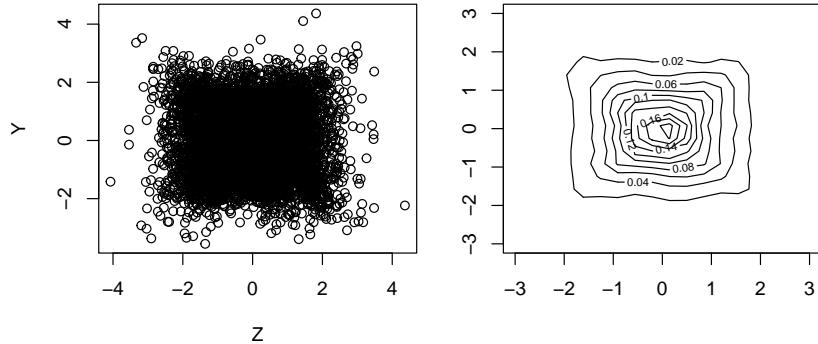


Fig. 8.14 Bivariate normal random variables with varying correlation

The uncertainty about the parameter leads to the distribution

$$Y \sim F()$$

where

$$F(y) = P(Y \leq y) = \int_{\theta \in \Theta} \int_{-\infty}^y f(y; \theta) f_\theta(\theta) d\theta dy$$

We can simulate from $F()$ by first simulating the parameter and then simulating from (8.30). To be specific:

1. Draw θ^* from $f_\theta(\cdot)$
2. Draw Y^* from $F(\cdot; \theta^*)$

Example 11 Assume we aim to draw a bivariate normal distribution but we are uncertain about the correlation. We assume the correlation ρ^* to be distributed be uniformly distributed 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 Figure 8.14. The left-hand side shows 10000 simulated values, and the right-hand side 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.

8.7 Bootstrapping

8.7.1 The Plug-In Principle

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 Glevenco-Cantelli theorem (see Section A.6). We assume a sample drawn from the general data generating process (2.3). Hence

$$Y_i \sim G() \text{ i.i.d. for } i = 1, \dots, n. \quad (8.31)$$

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 \mathcal{F} . Hence, we change the notation and rewrite the data generating process (8.31) to

$$Y_i \sim F() \text{ i.i.d. for } i = 1, \dots, n. \quad (8.32)$$

where $F(\cdot)$ is unknown. Let $F_n(y)$ denote the empirical distribution function defined as

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

Glevenco-Cantelli now states that for n increasing we obtain that $F_n(y)$ converges to $F(y)$ for all $y \in R$.

We have the intention to draw random numbers from (8.32). If we were able to do so, we could apparently 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 can not 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 (8.32) by its empirical version. That is we propose to draw (new) samples from

$$Y_i^* \sim F_n() \text{ i.i.d. for } i = 1, \dots, n. \quad (8.33)$$

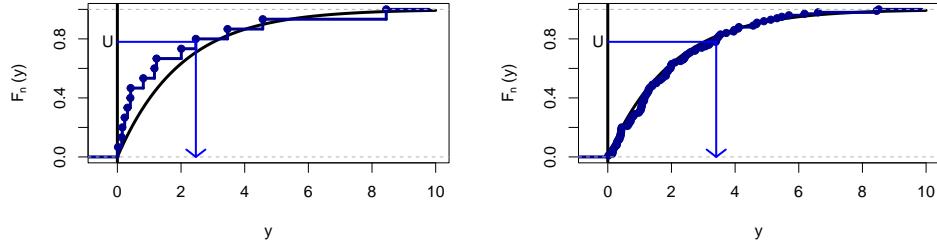


Fig. 8.15 Sampling from the empirical distribution function with sample size $n = 15$ (left) and $n = 100$ (right)

Note that this is possible, unlike drawing from $F(\cdot)$, since we know $F_n(\cdot)$ and can directly apply the ideas from the earlier parts of this chapter, see also Figure 8.3. In fact we can easily see that (8.33) is equal to

$$\text{Draw } Y_i^* \text{ from the data } (y_1, \dots, y_n) \text{ with replacement, for } i = 1, \dots, n. \quad (8.34)$$

This is known as **bootstrapping**, or more specifically as non-parametric bootstrapping. With Glevenko-Cantelli we see that the approach is justifiable since we just replace $F(\cdot)$ by its estimate $F_n(\cdot)$. We visualize the idea in Figure 8.15, where we show the empirical distribution function for exponentially distributed random variables for two different sample sizes.

We want to motivate the use of the bootstrap in more detail using the so-called plug-in principle. Let $y = (y_1, \dots, y_n)$ denote the data drawn from (8.32) 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 functional in the form

$$t(y) = T(F_n) \quad (8.35)$$

where T is a functional, that is a function applied to a function. For instance, the arithmetic mean can be written as

$$T(F_n) = \bar{y} = \sum_{i=1}^n \frac{y_i}{n} = \int_{-\infty}^{\infty} y dF_n(y)$$

utilizing the notation introduced in Chapter 1.2. If we replace the empirical distribution by its true distribution we obtain

$$T(F) = \int_{-\infty}^{\infty} y dF(y) = \mu.$$

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. Assume for instance that we want to estimate the variance of $\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$. Knowing $F(\cdot)$ this can be calculated through

$$\begin{aligned} T(F) = \text{Var}(\bar{Y}) &= \int \dots \int (\bar{y} - \mu)^2 dF(y_1) \dots F(y_n) \\ &= \int \dots \int \left(\frac{1}{n} \sum_{i=1}^n (y_i - \mu) \right)^2 dF(y_1) \dots F(y_n) \\ &= \frac{1}{n^2} \int \dots \int \sum_{i=1}^n (y_i - \mu)^2 dF(y_1) \dots F(y_n) \\ &\quad + \underbrace{\frac{1}{n^2} \int \dots \int \sum_{i=1}^n \sum_{j \neq i}^n (y_i - \mu)(y_j - \mu) dF(y_1) \dots F(y_n)}_{=0} \\ &= \frac{1}{n^2} \sum_{i=1}^n \int (y_i - \mu)^2 dF(y_i) = \frac{1}{n} \sigma^2 \end{aligned} \tag{8.36}$$

where $\sigma^2 = \text{Var}(Y_i), i = 1, \dots, n$. We can replace the above formula using the plug-in principle by replacing $F(\cdot)$ through $F_n(\cdot)$. First, we replace the unknown mean μ through the arithmetic mean \bar{y} , as derived above. Secondly, we replace the unknown distribution function through the empirical one, meaning we apply the plug-in principle which gives

$$T(\hat{F}_n) = \int \dots \int (\bar{y}^* - \bar{y})^2 dF_n(y_1^*) \dots F_n(y_n^*)$$

where $\bar{y}^* = \sum_{i=1}^n \bar{y}_i^*/n$, 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 (8.36) which leads us to

$$T(\hat{F}_n) = \frac{1}{n} \hat{\sigma}^2$$

where again using the plug-in principle

$$\hat{\sigma}^2 = \int (y^* - \bar{y})^2 dF_n(y) = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2.$$

While the plug-in principle is helpful, it is not always applicable. Most often we can not 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

$$Y_i^* \sim F_n(), \text{ i.i.d. for } i = 1, \dots, n. \quad (8.37)$$

This step can be repeated several times. The idea can be motivated with a simple example. We assume a sample (y_1, \dots, y_n) of size n from which we aim to estimate the variance. Figure 8.16 shows our data of size $n = 100$ in the top left plot. Calculating the empirical variance of these data leads $\hat{\sigma}^2 = 0.92$. We aim to quantify uncertainty for this variance estimate through bootstrapping. Do so we draw B bootstrap samples denoted by $(Y_1^{*(b)}, \dots, Y_n^{*(b)})$. Each sample is drawn from y_1, \dots, y_n with replacement which is equivalent to

$$Y_i^{*(b)} \sim F_n(), \text{ i.i.d. for } i = 1, \dots, n \text{ and } b = 1, \dots, B$$

Three such bootstrap samples are shown in Figure 8.16, top right plot. For each of the B bootstrap samples we calculate the resulting variance estimate, that is

$$\hat{\sigma}^{2(b)} = \sum_{i=1}^n (Y_i^{*(b)} - \bar{Y}^{*(b)})^2$$

where $\bar{Y}^{*(b)}$ 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 Figure 8.16. 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 easily carried out by taking the distribution of the bootstrapped values $\hat{\sigma}^{2(b)}$ which is shown in the bottom right plot of Figure 8.16. Using the $\alpha/2$ and the $1 - \alpha/2$ quantiles as boundaries provides a bootstrapped confidence interval for the estimate. In Figure 8.16 we show this for $\alpha = 0.1$ leading to the interval $[0.7, 1.15]$.

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 machine learning framework which will be discussed in Chapter 14.1. This can be seen as trained prediction $h(x, \hat{\theta})$ with $\hat{\theta}$ as trained (fitted) parameters. We may now want to quantify the uncertainty of the model, given that it is trained on the finite data (x_i, y_i) , where x_i are the input variables. If we consider both, x_i as well as y_i as random we can group them as pairs and bootstrap directly from these pairs. In other words, we draw B bootstrap samples

$$(x_i^{*(b)}, y_i^{*(b)}) \sim \{(x_i, y_i), i = 1, \dots, n\} \text{ with replacement} \quad (8.38)$$

for $b = 1, \dots, B$. From each bootstrap sample $(x_i^{*(b)}, y_i^{*(b)})$, $i = 1, \dots, n$ we retrain the model leading to $h(x, \hat{\theta}^{(b)})$ bootstrapped models. This allows now to quantify the uncertainty of the models based on the finite sample.

Example 12 Assume we have two classifiers and want to compare them. Consider for instance the case that two machine learner were trained for classifying images in "cats" and "dogs", to make it simple. The results of the two procedures applied to some test data of size $n = 1000$ are shown as confusion matrices in Table 8.1, which are structure-wise built like the error matrices from Table 5.3. Looking at the numbers it looks as if classifier 2 is slightly better. It classifies more cat as cats and more dogs as dogs. In fact, calculating the accuracy for these tables as defined in (5.25) on gets that classifier 1 gets 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 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. We want to tackle this question using bootstrapping.

The main idea to make the bootstrap work, in this case, is that if the hypothesis holds, then 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

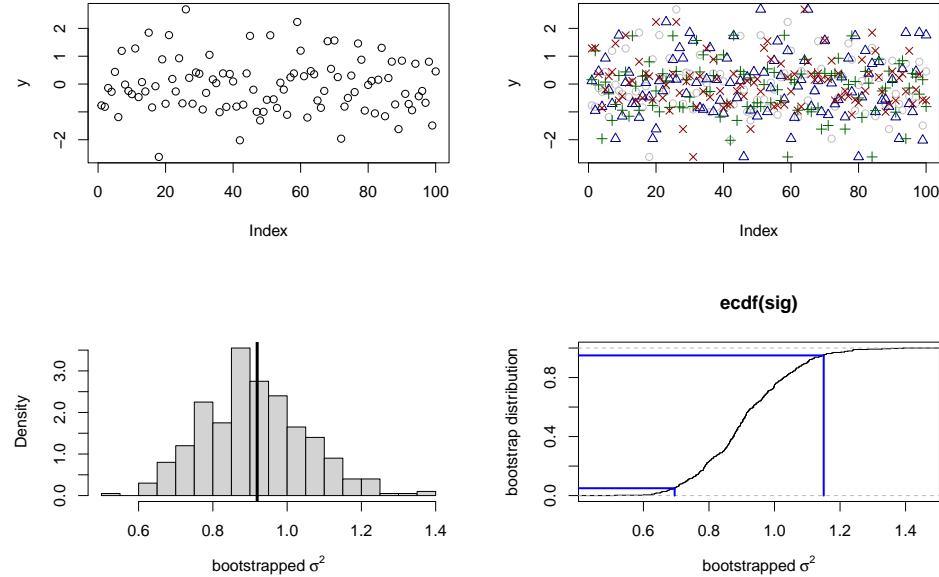


Fig. 8.16 Top row: Sample data of size $n = 100$ (left hand side) and original data (grey) with three bootstrap samples, coded in different colors and shapes of points (right-hand side)
 Bottom row: Variance of $B = 400$ bootstrap samples and empirical variance of the original data shown as vertical line (left-hand side) and resulting distribution function, which allows constructing confidence intervals (right-hand side)

Classifier 1			Classifier 2		
true class	decision		true class	decision	
	"cat"	"dog"		"cat"	"dog"
cat	235	57	cat	250	42
dog	82	626	dog	78	630
	317	683		328	672
	1000			1000	

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.

$$(z_i, c_{1,i}, c_{2,i}) \text{ for } i = 1, \dots, n \quad (8.39)$$

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

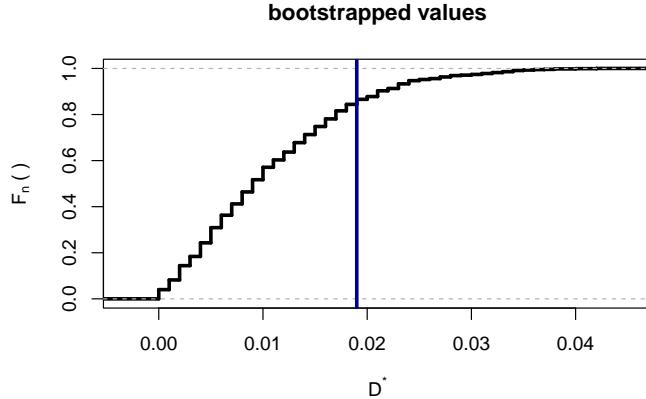


Fig. 8.17 Bootstrapped absolute differences in accuracy

$$(z_i^*, c_{1,i}^*, c_{2,i}^*) \quad (8.40)$$

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

$$(z_i^*, c_{(1+R_i^*)i}^*, c_{(2+R_i^*)i}^*) \quad (8.41)$$

where $R_i^* \sim B(1, 0.5)$ are binomial and i.i.d., which guarantees the random swap 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 $d = 0.091$. We show the resulting empirical distribution function of the bootstrapped values D^* in Figure 8.17 and indicate the observed value as a vertical line. We conclude that the differences found in Table 8.1 can purely be explained by random variation and in fact, the resulting p-value is on the order of 0.15. \triangleright

8.8 Permutation

To complete the chapter we need 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 elaborated 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

$$Y_i \sim F_Y(\cdot) \text{ i.i.d., } i = 1, \dots, n \text{ and } Z_j \sim F_Z(\cdot) \text{ i.i.d., } j = 1, \dots, m. \quad (8.42)$$

We want 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 Chapter 5 and use the Wilcoxon test statistics, the Kolmogorov-Smirnov statistics or any other suitable measure. To exemplify the idea we use

$$D = \sup_y |F_{n,Y}(y) - F_{m,Z}(y)| \quad (8.43)$$

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

$$x_i \sim F(), \text{ i.i.d. for } i = 1, \dots, (m+n). \quad (8.44)$$

Test statistics D in (8.45) 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 $y^* = (x_1^*, \dots, x_n^*)^\top$ and the latter m elements as z^* , i.e. $z^* = (x_{n+1}^*, \dots, x_{n+m}^*)^\top$. Using these permuted samples leads to

$$D^* = \sup_y |F_{n,Y^*}(y) - F_{m,Z^*}(y)| \quad (8.45)$$

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^* . Repeating this step B times provides B values $D^{*(b)}$, 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 13 In Figure 8.18 we show rents for apartments in Munich for surveys taken in two 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 it with the inflation factor. This gives the plot on the right-hand side. We can now calculate the Kolmogorov-Smirnov distance resulting

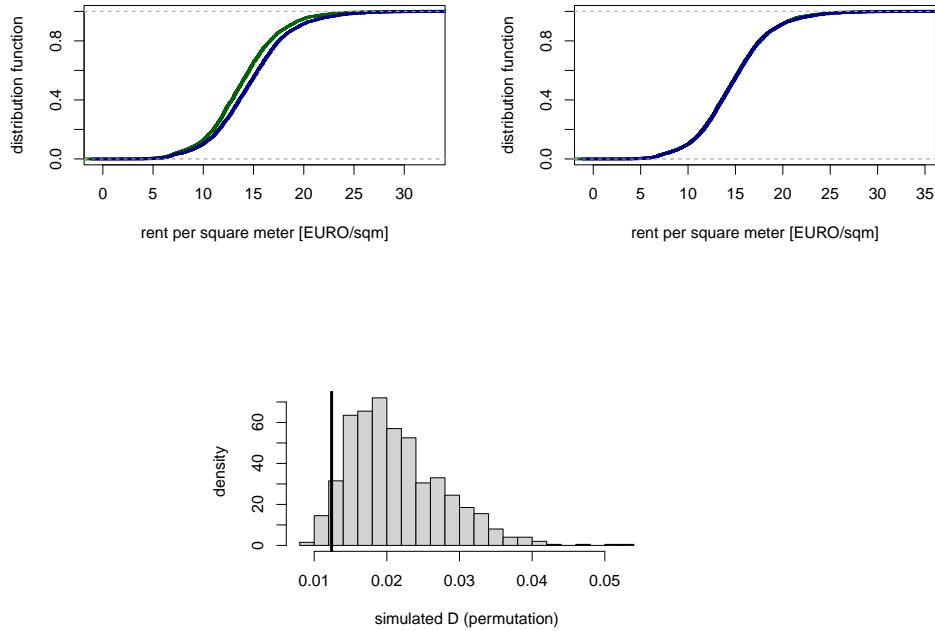


Fig. 8.18 Top row: Rent data (left) and corrected by inflation factor (right).
Bottom row: Simulated distances D^* through permutation and observed value D shown as vertical line.

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 Figure 8.18. 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. \triangleright

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.

		tomorrow		
		S	C	R
today				
	S	0.50	0.25	0.25
	C	0.20	0.40	0.40
	R	0.20	0.50	0.30

- Calculate the probability of the weather being sunny in two days, given that it is sunny today.
- 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?
- Visualise 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 $\hat{\theta} = 2.48$.

- We would like to quantify the uncertainty of $\hat{\theta}$. Implement a non-parametric bootstrap method to estimate $se(\hat{\theta})$.
- 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?
- Estimate the bias of $\hat{\theta}$ using non-parametric bootstrapping. Compute a bias-corrected estimate as well.
- 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.

Chapter 9

Bayesian Inference

The terminology Bayesian statistics is used if the uncertainty about parameters is formulated with distributions. This is a general setup which 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 Bayes 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 described in 1763 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 make the main principles of Bayes statistics clear. Following the notation of Chapter 7 we consider the model $\mathcal{M} = \{f(\cdot; \theta), \theta \in \Theta\}$, 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 Chapter 2 that uncertainty is well quantifiable with probability arguments.

Let us demonstrate the idea with a simple example. Assume a binomial model with

$$Y \sim B(n, \pi)$$

such that π is the parameter of interest. This leads to the likelihood

$$f(y; \pi) = \binom{n}{y} \pi^y (1 - \pi)^{n-y}. \quad (9.1)$$

We postulate our prior knowledge (or prior uncertainty) on π with a beta distribution, i.e.

$$\pi \sim Beta(\alpha, \beta).$$

The beta distribution is a continuous distribution on the interval $[0, 1]$ with density depending on the parameters $\alpha > 0$ and $\beta > 0$:

$$p(\pi; \alpha, \beta) = \frac{\pi^{\alpha-1} (1-\pi)^{\beta-1}}{B(\alpha, \beta)} \quad (9.2)$$

where $B(\alpha, \beta)$ is the so called Beta-function. The Beta-function serves as a normalization constant. It formally results to

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} = \int_0^1 u^{\alpha-1} (1-u)^{\beta-1} du,$$

with $\Gamma(\cdot)$ as 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 to

$$p(\pi|y) = \frac{f(y;\pi)p(\pi)}{\int_0^1 f(y;\tilde{\pi})p(\tilde{\pi})d\tilde{\pi}} \propto \pi^{\alpha+y-1}(1-\pi)^{\beta+n-y-1} \quad (9.3)$$

with \propto referring to proportionality, that is omitting all multiplicative constants that do not depend on π . With (9.3) and (9.2) we readily find that

$$\pi|y \sim Beta(\alpha+y, \beta+n-y). \quad (9.4)$$

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.

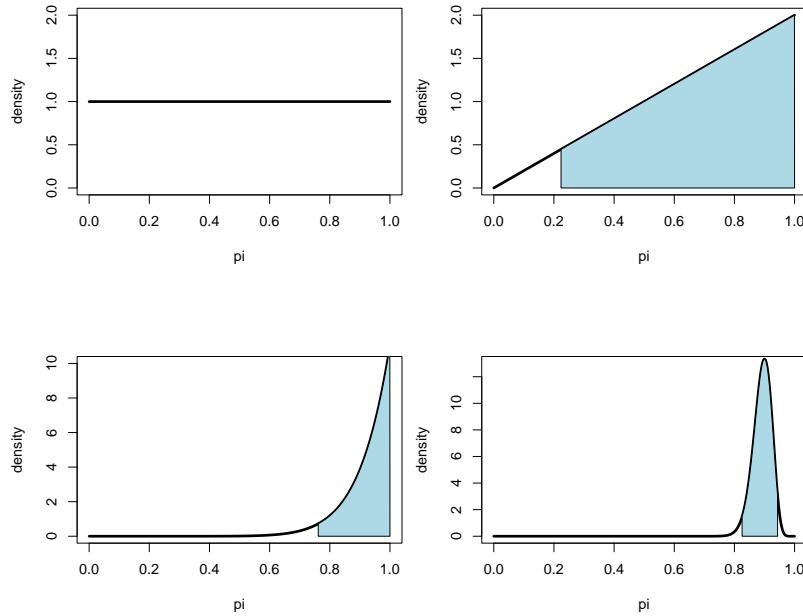


Fig. 9.1 Uniform prior beta distribution (top left). The posterior distribution for $n = 1$ and $y = 1$ (top right). Posterior for $n = 10$ and $y = 10$ (bottom left). Posterior for $n = 100$ and $y = 90$ (bottom right). Blue-shaded areas give 95 percent probability mass.

We demonstrate the idea in Figure 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 Figure 9.1. We also visualize 95 probability mass,

which gives an interval previously defined as credibility interval in Definition 3.11. If $n = 10$ and $y = 10$, that is in all 10 trials we observe success, 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 $\bar{y} = 1$ and consequently the central limit theorem fails to hold. This highlights a quite central advantage of Bayes 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.

Let us summarize the Bayesian principles before we go into more technical details.

1. We assume a distribution model $\mathcal{M} = \{f(\cdot; \theta), \theta \in \Theta\}$ for the data.
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.
4. The posterior distribution allows drawing inference, e.g. by calculating intervals with 95 percent probability mass.

9.2 Prior and Posterior

Let us get more formal now. With $\mathcal{M} = \{f(\cdot; \theta); \theta \in \Theta\}$ we define our probability model for the data y_1, \dots, y_n , which are assumed to be *i.i.d.* For θ we formulate our knowledge (or better missing knowledge) as prior distribution

$$\theta \sim p(\cdot; \gamma), \quad (9.5)$$

where parameter $\omega \in \Omega$ is called hyperparameter with Ω as hyperparameter space. We define the prior structure as the model class

$$\mathcal{P} = \{p(\cdot; \omega), \omega \in \Omega\}$$

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 \mathcal{M} we obtain the posterior distribution

$$p_{\theta|y}(\theta|y_1, \dots, y_n; \gamma) = \frac{\prod_{i=1}^n f(y_i; \theta) p(\theta; \omega)}{f(y; \omega)} \quad (9.6)$$

where

$$f(y; \omega) = \int_{\theta \in \Theta} \prod_{i=1}^n f(y_i; \theta) p(\theta; \omega) d\theta. \quad (9.7)$$

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 \mathcal{F} and \mathcal{P} are chosen such that the posterior is again in \mathcal{P} . We call this conjugate prior.

Definition 9.1 For a given probability model \mathcal{F} , we call \mathcal{P} the set of conjugate prior distributions, if it exists and if the posterior $p_{\theta|y}(\cdot) \in \mathcal{P}$. 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 the previous section, 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 \mathcal{F} and \mathcal{P} in Table 9.1.

Parameter	\mathcal{F}	\mathcal{P}
π	Binomial distribution	Beta distribution
λ	Poisson distribution	Gamma distribution
μ	Normal distribution	Normal distribution
λ	Exponential distribution	Gamma distribution

Table 9.1 Examples for conjugate distributions

Example 14 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. Note that if

$$p(\lambda) \propto \lambda^{\alpha-1} \exp(-\lambda\beta),$$

then $\lambda \sim \text{Gamma}(\alpha, \beta)$. This in turn shows for the posterior

$$p_{\lambda|y}(\lambda|y) \propto \left\{ \prod_{i=1}^n \left(\lambda^{y_i} \exp(-\lambda) \right) \right\} \lambda^{\alpha-1} \exp(-\lambda\beta),$$

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

$$\lambda|y \sim \text{Gamma}\left(\alpha + \sum_{i=1}^n y_i, \beta + n\right).$$

Hence, prior and posterior distribution come from the same model class. \triangleright

Example 15 Assume $Y_i \sim N(\mu, \sigma^2)$ and let σ^2 be known for simplicity. We assume the prior $\mu \sim N(\delta, \tau^2)$. 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

$$\begin{aligned} \mu|y &\propto \left\{ \prod_{i=1}^n \exp\left(-\frac{1}{2} \frac{(y_i - \mu)^2}{\sigma^2}\right) \right\} \exp\left\{-\frac{1}{2\tau^2} \mu^2 + \frac{\delta}{\tau^2} \mu\right\} \\ &\propto \exp\left\{-\frac{1}{2} \sum_{i=1}^n \frac{(y_i - \mu)^2}{\sigma^2} - \frac{1}{2\tau^2} \mu^2 + \frac{\delta}{\tau^2} \mu\right\} \\ &\propto \exp\left\{-\frac{1}{2} \left[\frac{n}{\sigma^2} + \frac{1}{\tau^2} \right] \mu^2 + \left[\sum_{i=1}^n \frac{y_i}{\sigma^2} + \frac{\delta}{\tau^2} \right] \mu\right\}. \end{aligned}$$

Consequently, by collecting the terms and using the above proportionality statements we get

$$\mu|y \sim N\left(\left(\frac{n}{\sigma^2} + \frac{1}{\tau^2}\right)^{-1} \left(\sum_{i=1}^n \frac{y_i}{\sigma^2} + \frac{\delta}{\tau^2}\right), \left(\frac{n}{\sigma^2} + \frac{1}{\tau^2}\right)^{-1}\right)$$

Note that if we set $\delta = 0$, that is we assume that the mean value μ is 'a priori' centred around 0, and setting $\tau^2 \rightarrow \infty$, that is we assume a prior such that possible values of μ range from very small to very large values, then we obtain approximately

$$\mu|y \sim N\left(\bar{y}, \frac{\sigma^2}{n}\right).$$

This result looks familiar when we exchange μ and \bar{y} and do not take a Bayesian view but let μ be a parameter. Note also that for $\tau^2 \rightarrow \infty$ the prior distribution gets degenerated, meaning that the integral $\int p(\mu; \delta=0, \tau^2 \rightarrow \infty) d\mu \rightarrow \infty$. We call this an *improper prior*. Still, even with an improper prior, the posterior distribution is well-defined. \triangleright

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 $n \rightarrow \infty$ and assume that the parameter space Θ is finite-dimensional.

Property 9.1 Bernstein - von Mises Theorem. For increasing sample size n and appropriately chosen prior we find

$$\theta - \hat{\theta} \xrightarrow{a} N(0, I^{-1}(\hat{\theta}))$$

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 occurs 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 Chapter 8.3.

Proof Looking at (9.6) we find

$$\begin{aligned} p_{\theta|y}(\theta|y) &\propto \exp\left\{\sum_{i=1}^n \log f(y_i; \theta) + \log p(\theta; \gamma)\right\} \\ &\propto \exp\left\{l_p(\theta)\right\} \end{aligned} \tag{9.8}$$

where

$$l_p(\theta) = \sum_{i=1}^n \log f(y_i; \theta) + \log p(\theta; \gamma) \tag{9.9}$$

$$= l(\theta) + \log p(\theta; \gamma). \tag{9.10}$$

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

$$\theta|y \stackrel{a}{\sim} N\left(\hat{\theta}_p, I_p^{-1}(\hat{\theta}_p)\right).$$

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. \square

9.3 Hyperparameters and Empirical Bayes

We have introduced prior distributions that depend on additional parameters ω , labelled 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

$$p(\theta; \omega) \propto \text{const}. \quad (9.11)$$

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 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

$$p_{\theta|y}(\theta|y) \propto \prod_{i=1}^n f(y_i; \theta) \cdot \text{const}. \quad (9.12)$$

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

$$\tilde{f}(y; \omega) := \int_{\theta \in \Theta} f(y; \theta) p(\theta; \omega) d\theta. \quad (9.13)$$

In other words, instead of working with the distribution $f(y; \theta)$ we work with the probability model $\tilde{\mathcal{F}} = \{\tilde{f}(\cdot; \gamma); \gamma \in \Gamma\}$. Choosing a value for hyperparameter ω can be done by utilizing the results from Chapter 4 and select ω to maximize the resulting likelihood

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

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 shown to be quite useful and we will revisit the idea in Chapter 13. **CHECK REFERENCE**

9.4 Inference based on MCMC

We now get to the heart of the Bayesian machinery. Looking at formula (9.6) we find

$$p_{\theta|y}(\theta|y_1, \dots, y_n) \propto \underbrace{\prod_{i=1}^n f(y_i; \theta)}_{L(\theta)} p(\theta; \omega). \quad (9.14)$$

Both components on the right-hand side in (9.14) are numerically available. The first is the likelihood, 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 an unknown constant. This is exactly the setup that we discussed in Chapter 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 Metropolis-Hastings 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

$$\hat{\theta} = \frac{1}{T-d} \sum_{t=d+1}^T \theta^{*(t)}$$

where d determines the burning in phase.

The algorithm is then as follows.

1. Let $\theta = (\theta_1, \dots, \theta_p)$ 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, that is

$$\theta^* \sim q(\cdot | \theta^{*(t)})$$

3. Calculate the acceptance probability through

$$\alpha(\theta^*, \theta^{*(t)}) = \min \left\{ 1, \frac{L(\theta^*) p(\theta^*; \omega)}{L(\theta^{*(t)}) p(\theta^{*(t)}; \omega)} \frac{q(\theta^{*(t)} | \theta^*)}{q(\theta^* | \theta^{*(t)})} \right\}$$

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. If $p > 2$ it is therefore advisable to combine the MCMC with Gibbs sampling as described in Chapter 8.5. This means, that the simulation steps 2 to 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

$$\begin{aligned} p_{\theta_j | y, \theta_{-j}}(\theta_j | y_1, \dots, y_n, \theta_{-j}) &\propto p_{\theta | y}(\theta | y_1, \dots, y_n) \\ &\propto \underbrace{\prod_{i=1}^n f(y_i; \theta)}_{L(\theta)} p(\theta; \omega), \end{aligned} \tag{9.15}$$

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 $\theta = (\theta_1, \dots, \theta_p)$ 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 $j = 1, \dots, p$ 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

$$\theta_j^{**} \sim q(\cdot | \theta_j^*)$$

- c. Calculate the acceptance probability through

$$\alpha(\theta^{**}, \theta^*) = \min \left\{ 1, \frac{L(\theta^{**})p(\theta^{**}; \omega)}{L(\theta^*)p(\theta^*; \omega)} \frac{q(\theta^* | \theta^{**})}{q(\theta^{**} | \theta^*)} \right\}$$

- 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 16 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 Figure 9.2 (top left). We can model the data by using the Poisson distribution but inflating the zero probability. With $Y \in \{0, 1, \dots\}$ being the count this leads to the model

$$P(Y = k; \pi, \lambda) = \begin{cases} \pi + (1 - \pi) \frac{\lambda^0}{0!} e^{-\lambda}, & \text{for } k = 0 \\ (1 - \pi) \frac{\lambda^k}{k!} e^{-\lambda}, & \text{for } k > 0 \end{cases}, \quad (9.16)$$

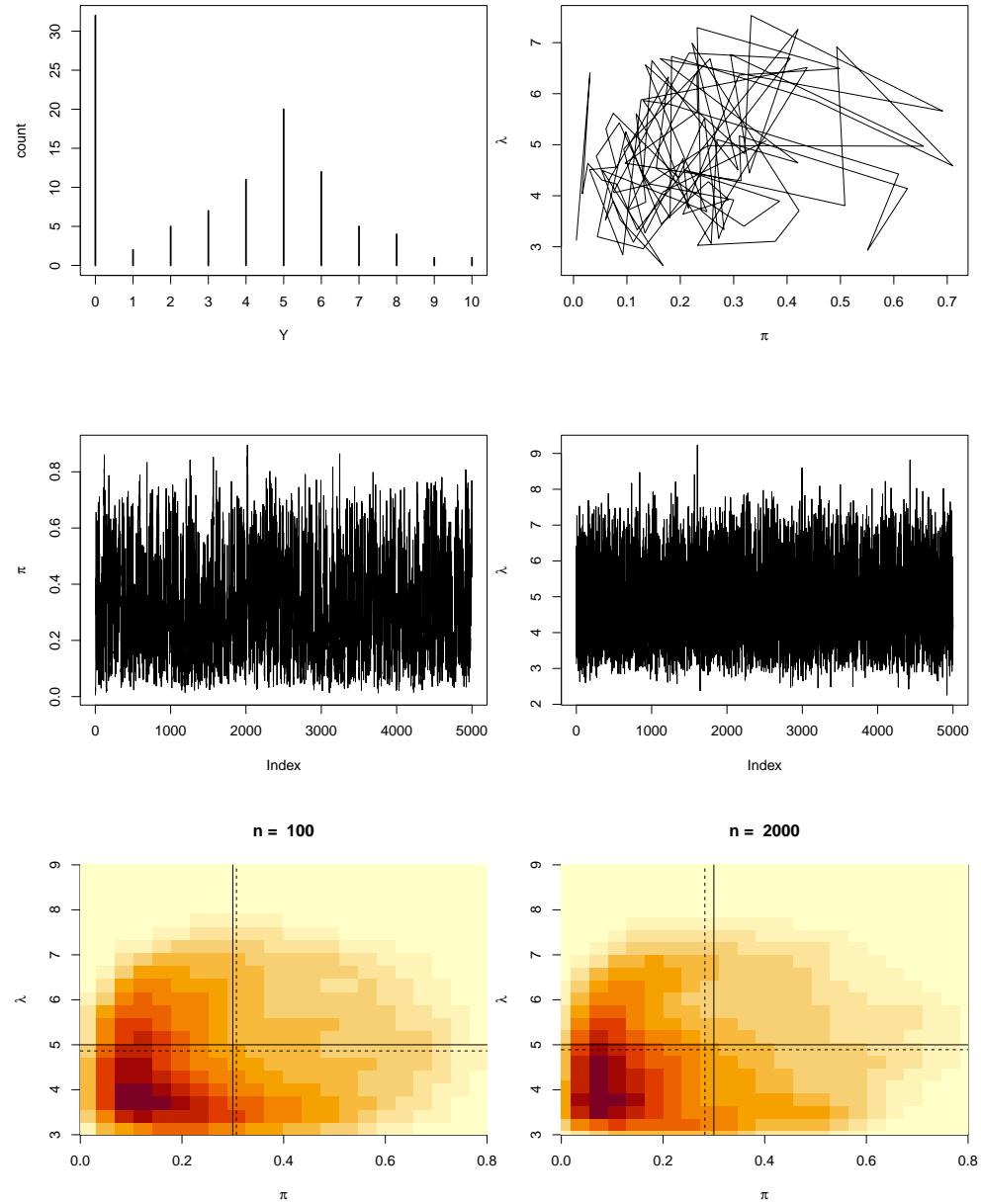


Fig. 9.2 Posterior distribution through MCMC for zero-inflated count data. Data (top left) and the first 300 steps of the MCMC chain (top right). MCMC steps for π (middle left) and λ (middle right). Joint posterior distribution for (π, λ) for $n = 100$ and for $n = 2000$

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

$$\pi \sim \text{Uniform}[0, 1] = \text{Beta}(1, 1) \quad (9.17)$$

$$\lambda \sim p(\lambda; \omega) = \omega e^{-\lambda\omega} \quad (9.18)$$

with ω as hyperparameter. For ω fixed and data y_1, \dots, y_n , the posterior is then proportional to

$$\lambda, \pi | y_1, \dots, y_n \propto \prod_{i=1}^n P(Y = y_i; \lambda, \pi) \omega e^{-\lambda\omega}, \quad (9.19)$$

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 parameter. To do so we propose a value for $\eta = \text{logit}(\pi) = \log(\pi/(1 - \pi))$ instead of π . Note that the logit transformation yields the prior for η through

$$\eta \sim p_\eta(\eta) = \frac{\partial \pi}{\partial \eta} \underbrace{p_\pi(\pi(\eta))}_{=\text{const}} = \pi(\eta)(1 - \pi(\eta)) \quad (9.20)$$

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

$$\eta^* | \eta^{*(t)} \sim N(\eta^{*(t)}, \sigma_\eta^2),$$

when we set $\sigma_\eta = 0.5$.

Similarly, we proceed for λ . Instead of modeling λ we take the log and model $\delta = \log(\lambda)$. This leads to the prior

$$\delta \sim p_\delta(\delta; \omega) = \frac{\partial \lambda}{\partial \delta} p_\lambda(\lambda(\delta); \omega) = \lambda \omega e^{-\lambda\omega} \quad (9.21)$$

with $\lambda = \exp(\delta)$. For the proposal we choose $\delta^* | \delta^{*(t)} \sim N(\delta^{*(t)}, \sigma_\delta^2)$ 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

$$\frac{\prod_{i=1}^n P(y_i; \lambda^*, \pi^*)}{\prod_{i=1}^n P(y_i; \lambda^{*(t)}, \pi^{*(t)})} = \exp \left\{ \sum_{i=1}^n \left[\log(P(y_i; \lambda^*, \pi^*)) - \log(P(y_i; \lambda^{*(t)}, \pi^{*(t)})) \right] \right\},$$

where $\lambda^* = \exp(\delta^*)$ and $\pi^* = \exp(\eta^*)/(1 + \exp(\eta^*))$. For large n it is advisable to calculate the first term as suggested.

We look at the performance of the MCMC steps in Figure 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 to estimate complex distributional models. \triangleright

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.

$$\hat{E}(\theta|y) = \frac{1}{T-k} \sum_{t=k+1}^T \theta^{*(t)}$$

where k is defines the length of the burn-in phase. We may also calculate the posterior variance to assess uncertainty, i.e.

$$\widehat{\text{Var}}(\theta|y) = \frac{1}{T-k} \sum_{t=k+1}^T \left\{ \theta^{*(t)} - \hat{E}(\theta|y) \right\}^2.$$

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 which lead to data comparable to the observed ones. Secondly, 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 here advisable to explicitly take the dependence of the likelihood on y into account. Note that the posterior results through

$$p(\theta|y) \propto L(\theta; y)p_\theta(\theta).$$

In some constellations, the likelihood itself can be clumsy and difficult to calculate. We consider such examples in Chapter 13, where we present ideas of latent variables. (**CHECK REFERENCES**) We will also look at cases where the likelihood has a clumsy normalization constant. 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 to 3 until the required number of simulated values $\theta^{*1}, \dots, \theta^{*N}$ 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 $p_\theta(\theta|y)$. To see this, note that for simulations that passed step 3 we denote the sampling probability as \tilde{p}_θ . Then

$$\begin{aligned}\theta^* &\sim \tilde{p}_\theta(\theta^*) \propto p_\theta(\theta^*) f(y; \theta^*) \\ &\propto p_\theta(\theta^* | y).\end{aligned}$$

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 16 on zero-inflated count data.

Example 17 Considering Figure 9.2 we have plotted zero-inflated count data in the top left plot. Using model (9.16) together with the priors (9.20) and (9.19) we can simulate π^*, λ^* and subsequently Y^* . We define the distance between the data y , seen in Figure 9.2 and the simulated values through

$$d(Y^*, y) = \sum_{k=0}^{10} (\bar{x}_k^* - x_k)^2$$

where

$$\bar{x}_k = \frac{1}{n} \sum_{i=1}^n 1\{y_i = k\} \text{ and } \bar{x}_k^* = \frac{1}{n} \sum_{i=1}^n 1\{Y_i^* = k\}.$$

Hence we compare the relative frequencies of the observed and simulated data up to an outcome 10. In Figure 15.3 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. \triangleright

9.6 Variational Bayes Reasoning

In the previous section, we focused on situations where the likelihood is not explicitly available. Hence, approximation refers to the distribution $f(y; \theta)$. We now look at a different approach in that we aim to approximate the posterior distribution. The most simple setup for this is the Bernstein-von Mises Theorem 9.1 or its penalized version, which we rewrite again. For large samples, i.e. for n sufficiently large, we have

$$\theta|y \stackrel{d}{\sim} N(\hat{\theta}_p, I_p^{-1}(\hat{\theta}_p)), \quad (9.22)$$

where $\hat{\theta}_p$ is the maximum of the penalized log likelihood

$$l_p(\theta) = \sum_{i=1}^n \log f(y_i; \theta) + \log p(\theta)$$

where the prior distribution 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 $p(\theta|y)$ through some approximation $q(\theta)$, where $q(\cdot)$ has a more simple structure than the original $p(\theta|y)$. For example, we might restrict $q(\theta)$ to be a normal distribution and hence aim to find the "closest" normal distribution to $p(\theta)$. Apparently, this resembles (9.22). We may also simplify the distribution further and assume independence. That is for $\theta = (\theta_1, \dots, \theta_p)$ we assume $q(\theta) = q_1(\theta_1)q_2(\theta_2) \cdots q_p(\theta_p)$. In general, let

$$q(\theta; \gamma) \in Q = \{q(\cdot; \gamma); \gamma \in \Gamma\}$$

be some set of easy-to-handle distributions for which we aim to find the element $q(\cdot; \hat{\gamma})$ which has the smallest distance to the true posterior distribution $p(\theta|y)$. As distance, we use the Kullback-Leibler divergence as in Chapter 3 but now we replace the order of the densities. Hence, we define

$$q(\cdot; \hat{\gamma}) = \arg \min_{q \in Q} \left\{ \text{KL}(q(\cdot; \gamma), p(\cdot|y)) \right\}$$

with

$$\text{KL}(q(\cdot; \gamma), p(\cdot|y)) = \int \log \frac{q(\theta; \gamma)}{p(\theta|y)} q(\theta; \gamma) d\theta.$$

Note that the integral is with respect to the density $q(\theta; \gamma)$ which will be utilized later. With $p(\theta|y) = f(y; \theta)p(\theta)/f(y)$ we can rewrite the above to

$$\text{KL}(q(\cdot; \gamma), p(\cdot|y)) = - \int \log \frac{f(y; \theta)p(\theta)}{q(\theta; \gamma)} q(\theta; \gamma) d\theta + \underbrace{\int \log f(y) q(\theta; \gamma) d\theta}_{=\log f(y)} \quad (9.23)$$

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

$$\text{LB}(q(\cdot; \gamma)) = \int \log \frac{f(y; \theta)p(\theta)}{q(\theta; \gamma)} q(\theta; \gamma) d\theta \quad (9.24)$$

with respect to γ .

As mentioned before, the class of distributions Q is chosen such that calculations are feasible so that the lower bound $\text{LB}(\cdot)$ can either be calculated analytically or obtained through simulations. This also applies to the derivative, which we aim to set to zero. We get

$$\begin{aligned} \frac{\partial \text{LB}(q(\cdot; \gamma))}{\partial \gamma} &= \int \log f(y; \theta)p(\theta) \frac{\partial q(\theta; \gamma)}{\partial \gamma} d\theta - \int \frac{\partial \log q(\theta; \gamma)}{\partial \gamma} q(\theta; \gamma) d\theta \\ &\quad - \int \log q(\theta; \gamma) \frac{\partial q(\theta; \gamma)}{\partial \gamma} d\theta. \end{aligned}$$

The second component above is zero for Fisher regular distributions, see (4.5). The remaining two components can be rewritten to

$$\frac{\partial \text{LB}(q(\cdot; \gamma))}{\partial \gamma} = E_{q(\cdot; \gamma)} \left\{ \frac{\partial \log q(\theta; \gamma)}{\partial \gamma} \left[\log(f(y; \theta)p(\theta)) - \log(q(\theta; \gamma)) \right] \right\}$$

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

$$\theta^{*(b)} \sim q(\cdot; \gamma^{(t)}) \text{ for } b = 1, \dots, B$$

and calculate the estimate of the derivative through

$$\frac{\partial \widehat{\text{LB}}(q(\cdot; \gamma^{(t)}))}{\partial \gamma} = \frac{1}{B} \sum_{b=1}^B \left\{ \frac{\partial \log q(\theta^{*(b)}; \gamma^{(t)})}{\partial \gamma} \left[\log(f(y; \theta^{*(b)}) p(\theta^{*(b)})) - \log(q(\theta^{*(b)}; \gamma^{(t)})) \right] \right\} \quad (9.25)$$

Following standard numerical optimization techniques one updates $\gamma^{(t)}$ through

$$\gamma^{(t+1)} = \gamma^{(t)} + r^{(t)} \frac{\partial \widehat{\text{LB}}(q(\cdot; \gamma^{(t)}))}{\partial \gamma} \quad (9.26)$$

with $r^{(t)}$ some (possibly vector valued) learning rate. We refer exemplarily to Kingma and Welling (2014), 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 18 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.

$$\lambda \sim p(\lambda; \omega) = \omega e^{-\lambda \omega}$$

with $\omega = 0.01$. Like above we look at transformed parameters, i.e. $\eta = \text{logit}(\pi) = \log(\pi/(1-\pi))$ and $\delta = \log(\lambda)$ 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.

$$\begin{aligned} q(\eta, \delta; \gamma) &= \underbrace{N(\eta; \mu_\eta, \sigma_\eta^2)}_{=: q_1(\eta; \mu_\eta, \sigma_\eta^2)} * \underbrace{N(\delta; \mu_\delta, \sigma_\delta^2)}_{=: q_2(\delta; \mu_\delta, \sigma_\delta^2)} . \end{aligned}$$

The parameters are collected in $\gamma = (\mu_\eta, \log(\sigma_\eta^2), \mu_\delta, \log(\sigma_\delta^2))$. 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 corresponds 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

$$\begin{aligned}\eta^{*(b)} &\sim N(\mu_\eta^{(t)}, \sigma_\eta^{2(t)}) \\ \delta^{*(b)} &\sim N(\mu_\delta^{(t)}, \sigma_\delta^{2(t)})\end{aligned}$$

for $b = 1, \dots, B$ 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 $\hat{\gamma}$ and we can plot the resulting density $q(\cdot; \hat{\gamma})$ which we visualize in Figure 9.3 for λ and π . \triangleright

Variational Bayes methods got 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, we have no measure that quantifies how close the approximation $q(\cdot; \gamma)$ is to the true posterior distribution. We refer to Fox and Roberts (2012) for more details.

9.7 Bayes and Regularization

NOT IN FINAL FORM We are sometimes faced with the fact that the number of observations that we need to estimate is large compared to the available data. Considering the p dimensional parameter $\theta = (\theta_1, \theta_2, \dots, \theta_p)$, 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 small samples, in the example below of order $n = 1$ or $n = 2$. Apparently, proper data-driven statistical inference is hardly possible with such small samples. Hence the example here serves as a demonstration only, to visualize how the likelihood behaves if $n < p$ or $n \approx p$.

We look at the situation $Y_i \sim N(\mu, \sigma^2)$ with *i.i.d.* samples and we set $n = 1$. Hence, there are two parameters which need to be estimate, based on a simple observation. The log likelihood results to

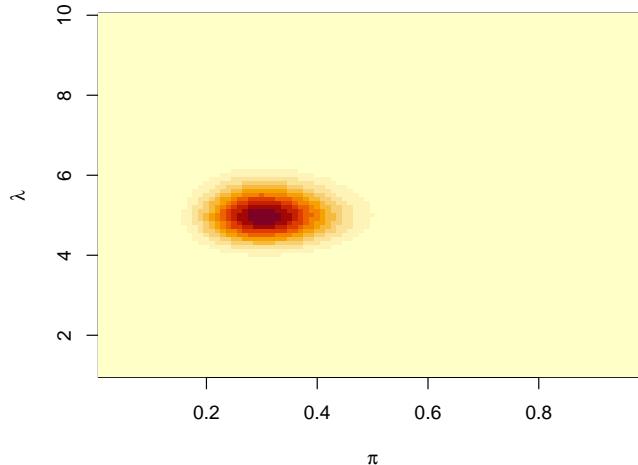


Fig. 9.3 Variational approximation of the posterior, based on the product of two normal distributions for $\text{logit}(\pi)$ and $\log(\lambda)$.

$$l(\mu, \sigma^2) = \frac{1}{2} \log(\sigma^2) - \frac{1}{2} \frac{(y_1 - \mu)^2}{\sigma^2}$$

We plot the likelihood (not the log likelihood) for a single observation in Figure 9.4, 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 $\hat{\mu} = y_1$ and $\hat{\sigma}^2 = 0$. This is not a useful result, since we obtain a degenerated distribution with variance 0.

We therefore assume weakly informative prior distributions for both, the mean value μ as well as the variance σ^2 , as shown in the middle plot of Figure 9.4. For the mean value μ we assume a normal prior with large variance and for σ^2 we use a inverse gamma distribution with a large range of values. 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 plot. While there is 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 ([REFERENCES](#)). Formally we define the penalized log likelihood as

$$l_p(\theta, \omega) = l(\theta) + \log(p(\theta; \omega)) \quad (9.27)$$

where ω are some hyper parameters. The maximum penalized likelihood, corresponding to the posterior mode, is achieved by setting

$$0 = \frac{\partial l_p(\theta, \omega)}{\partial \theta} = \frac{\partial l(\theta)}{\partial \theta} + \frac{\partial \log(p(\theta; \omega))}{\partial \theta}. \quad (9.28)$$

The remaining question is the specification of the hyper parameters ω . In principle, these should be set such that the prior is uninformative. Sometime it is also possible to estimate the parameters, but that requires some more simplified distribution model.

To Be Continued

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 pre-planning 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:

$$\begin{aligned} P(D_1) &= 0.4 \\ P(D_2) &= 0.1 \\ P(D_3) &= 0.3 \\ P(D_4) &= 0.2 \end{aligned}$$

- 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?

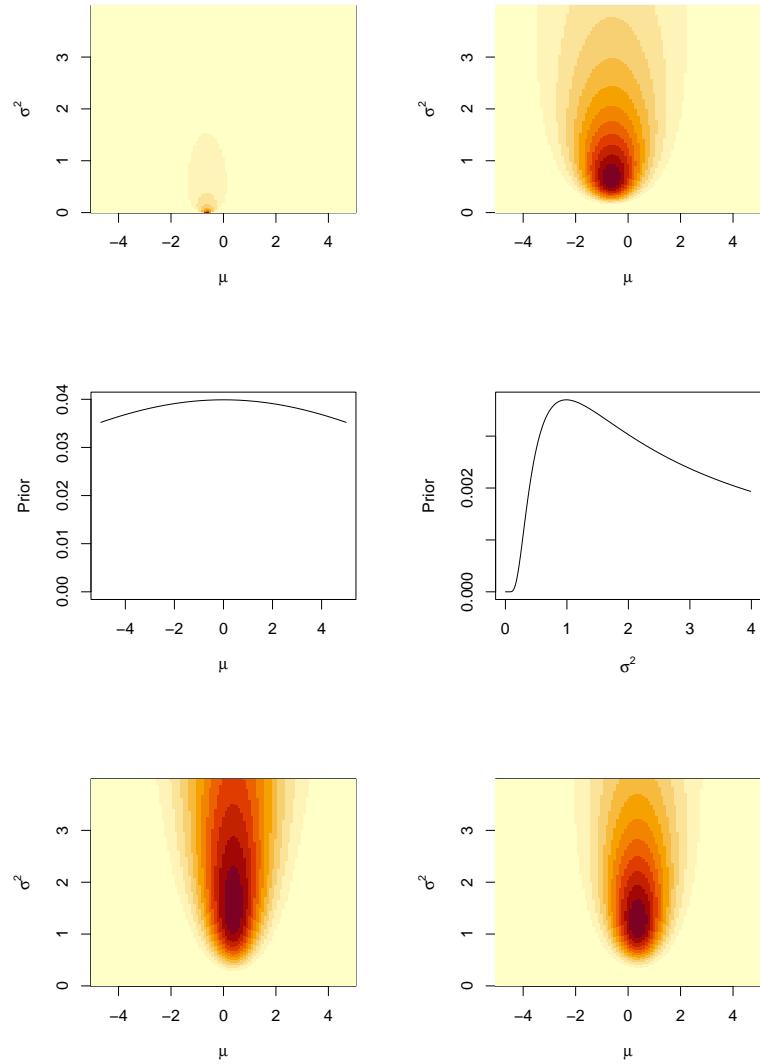


Fig. 9.4 Likelihood function for $n = 1$ (left) and $n = 2$ (right) for normally distributed data (top row). Prior structure for μ and σ^2 (middle row) and resulting posterior distribution for $n = 1$ (left) and $n = 2$ (right) shown in bottom row.

- b. In the final E2 survey, 71 out of 200 respondents were interested. Does the best department change?

Exercise 2

Let X_1, \dots, X_n be an i.i.d. sample from a geometric distribution with parameter π , i.e. $X_i \sim Geom(\pi), i = 1, \dots, n$.

- a. We assume a beta distribution as prior for the unknown parameter π , i.e. it holds $X_i|\pi \sim Geom(\pi)$ and $\pi \sim Beta(\alpha, \beta)$. 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 $Beta(\alpha, \beta)$ 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 $Beta(\alpha, \beta)$ 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 $\{Y_t^*, t \in \mathbb{N}_0\}$ with stationary distribution $\pi = (\pi_H, \pi_R, \pi_W)$. We assume a symmetric proposal density. **Here is some definition missing!**

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:

$$p_{ij} = P(Y_{t+1}^* = j | Y_t^* = i) = P(Y_1^* = j | Y_0^* = i) = \begin{cases} q_{ij} \alpha_{ij} & i \neq j, \\ 1 - \sum_{k \in S \setminus \{i\}} q_{ik} \alpha_{ik} & i = j, \end{cases} \quad i, j \in S,$$

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 .

Chapter 10

Inference in Extreme Data

In today's world, we observe more and more extreme events. Be it extreme droughts, extreme rainfalls or extreme water levels of rivers. This chapter presents the main ideas on how to deal with data that occur in the tails of a distribution. This also includes rare events, which happen to occur very seldom. In both cases, statistical inference and the quantification of uncertainty can not rely on asymptotic statements like the central limit theorem. However, other tools have been developed and extreme events have their own limit theorem, which is presented in this chapter.

10.1 Rare Events

Rare events occur seldom and/or with a low probability. These can be extreme events, like high water levels leading to flooding or extreme heat waves, or events that occur extremely seldom, like accidents in nuclear power plants or plane crashes. We will discuss extreme events in the next subsection. In this section, we focus on the case of rare occurrences of events.

We want to motivate the idea with a simple example. Assume we are interested in the proportion π of students that are drug addicted. We draw a sample on n students and question, whether the student is drug addicted ($Y = 1$) or not ($Y = 0$). We assume for simplicity that students give the correct answers. Otherwise, we would need to rely on randomized response experiments, which we want to ignore here (see van den Hout and van der Heijden, 2002). In our sample of $n = 10$ students we have not a single student who answered with yes, that is $y_i = 0$ for all $i = 1, \dots, n$. Apparently, the methods derived above in Chapter 4 break down. The log-likelihood function, assuming a binomial distribution, takes the form

$$l(\pi) = 10\log(1 - \pi), \quad (10.1)$$

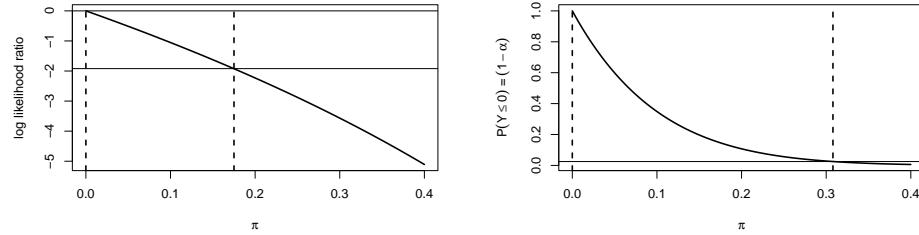


Fig. 10.1 Confidence intervals based on (void) asymptotic Chi-squared approximation (left) and exact calculation (right-hand side) for binomial data with $n = 10$ and $y = 0$

which is visualized in Figure 10.1. The maximum is achieved at $\hat{\pi} = 0$. Note that the first order derivative at the maximum of likelihood is unequal to zero, which contradicts the derivations from Chapter 4. In fact, following blindly the asymptotic results of Chapter 4 would lead to useless results, since with $\hat{\pi} = 0$ the resulting variance of the estimate is zero as well, which in turn would mean that $\pi = 0$. Certainly, this is not the case. As a first idea we could look at the log likelihood as plotted on the right-hand side in Figure 10.1. If we use the results derived for the likelihood ratio, we would state with (see equation (4.19)) that

$$2[l(\hat{\pi}) - l(\pi)] \sim \chi^2_1.$$

Hence, any test on $H_0 : \pi = \pi_0$ would be rejected if

$$l(\hat{\pi}) - l(\pi) > \frac{\chi^2_{1,1-\alpha}}{2}$$

with $\chi^2_{1,1-\alpha}$ as $1 - \alpha$ quantile of a Chi-squared distribution. Taking the equivalence between testing and confidence intervals and considering the shape of the likelihood above (10.1) would lead to the confidence

$$\left[0, \exp\left(\frac{\chi^2_{1,1-\alpha}}{2n}\right) \right],$$

which we visualize as a vertical line in Figure 10.1 on the left-hand side. It needs to be stressed that the interval does not have any mathematical justification and in fact, is too small. In other words, the asymptotic arguments derived in Chapter 4 do not hold since $\partial l(\hat{\theta})/\partial \theta \neq 0$.

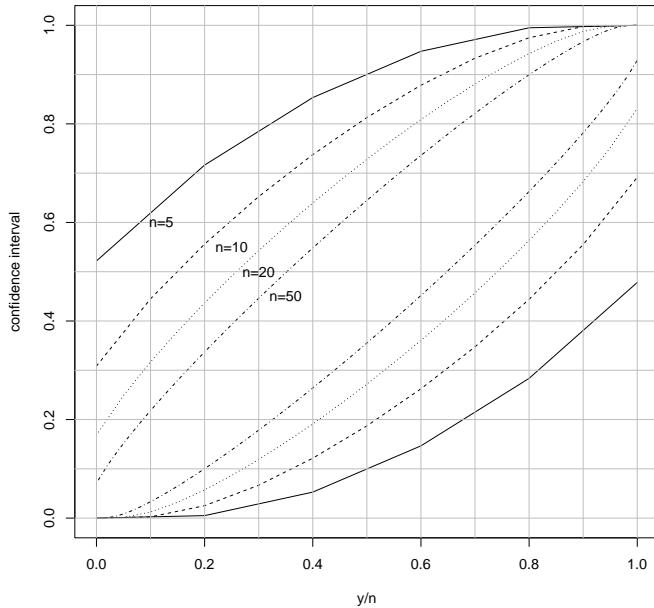


Fig. 10.2 Exact confidence intervals based on Clopper-Pearson approach

However, the shape of the Binomial distribution allows us to calculate exact confidence intervals and hence to draw valid statistical inference even for rare events. The idea goes back to Clopper and Pearson (1934). Let us assume that for n trials we observe $y = 0$. We can now calculate the probability that we observe $Y = 0$ for a binomial distribution with given n but varying π . We plot this curve for $n = 10$ and take $\alpha/2$ as threshold value in Figure 10.1 on the right-hand side. Hence, we look at the value π_r such that

$$P(Y \leq 0, \pi_r) = \frac{\alpha}{2}.$$

This gives the right-hand side end of the confidence interval (upper value). The lower bound is in this case zero. The general construction principle is as follows. 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 interval is called *exact confidence interval*, which is shown for different values of n and different outcomes of y/n in Figure 10.2. We also refer to Somerville and Brown (2013) and Thulin (2014) for further details and a discussion of alternative approaches.

The idea is also applicable to the Poisson distribution. Consider here the case that we have n independent Poisson distributed random variables

$$Y_i \sim \text{Poisson}(\lambda) \text{ i.i.d.}$$

but we have observed rare events. This means that λ is (extremely) small and in the most extreme case we didn't observe an event at all, that is $y_i = 0$ for $i = 1, \dots, n$. Though the data give 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 are, like above, interested in deriving a confidence interval for λ without relying on the asymptotic arguments given in Section 4. Before doing so let us first look at the likelihood function. We obtain

$$L(\lambda) = \prod_{i=1}^n \frac{\lambda^{y_i}}{y_i!} e^{-\lambda} \propto \lambda^{\sum_{i=1}^n y_i} e^{-n\lambda}.$$

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 go back to Garwood (1936). 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 Figure 10.3. Note that one needs to divide the values by n in the case of *i.i.d.* samples with intensity λ .

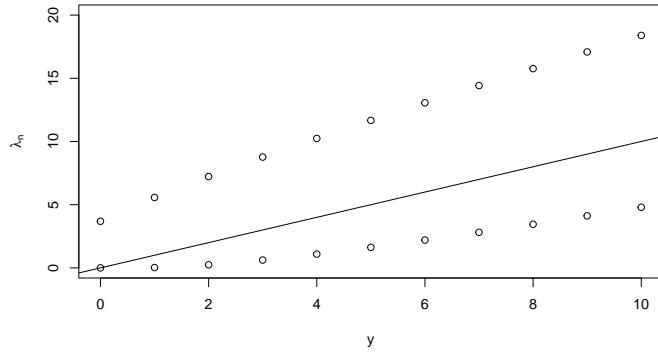


Fig. 10.3 Exact confidence intervals for Poisson distribution

10.2 Extreme Data

In the previous section, we looked at rare events. We extend this now to extreme events, which apparently are rare as well. Examples include extreme water levels in a river or extreme temperatures. To start, we look at an example.

Example 19 In Figure 10.4 we show in the upper plot the daily maximum temperature for the last years recorded on Helgoland, a German island in the North Sea. We clearly see the annual variation. We look at two quantities. First, we calculate the yearly maximum of the daily maximum temperatures which is shown in the middle plot. An overall upward trend is quite apparent, but we do not aim to discuss this at this point. Secondly, we also want to look at extreme minima, for which we calculate the lowest daily maximum temperature per year. This is plotted in the bottom plot. We also include the zero line in this plot as the freezing point. Points above the zero line indicate years, which did not have a single day with continuous frost. In this section, we aim to find a proper distribution model for these data. \triangleright

Let us define formally what we have looked at in the previous example. We are interested in the maximum (or minimum) of data. We, therefore, assume the classical setup

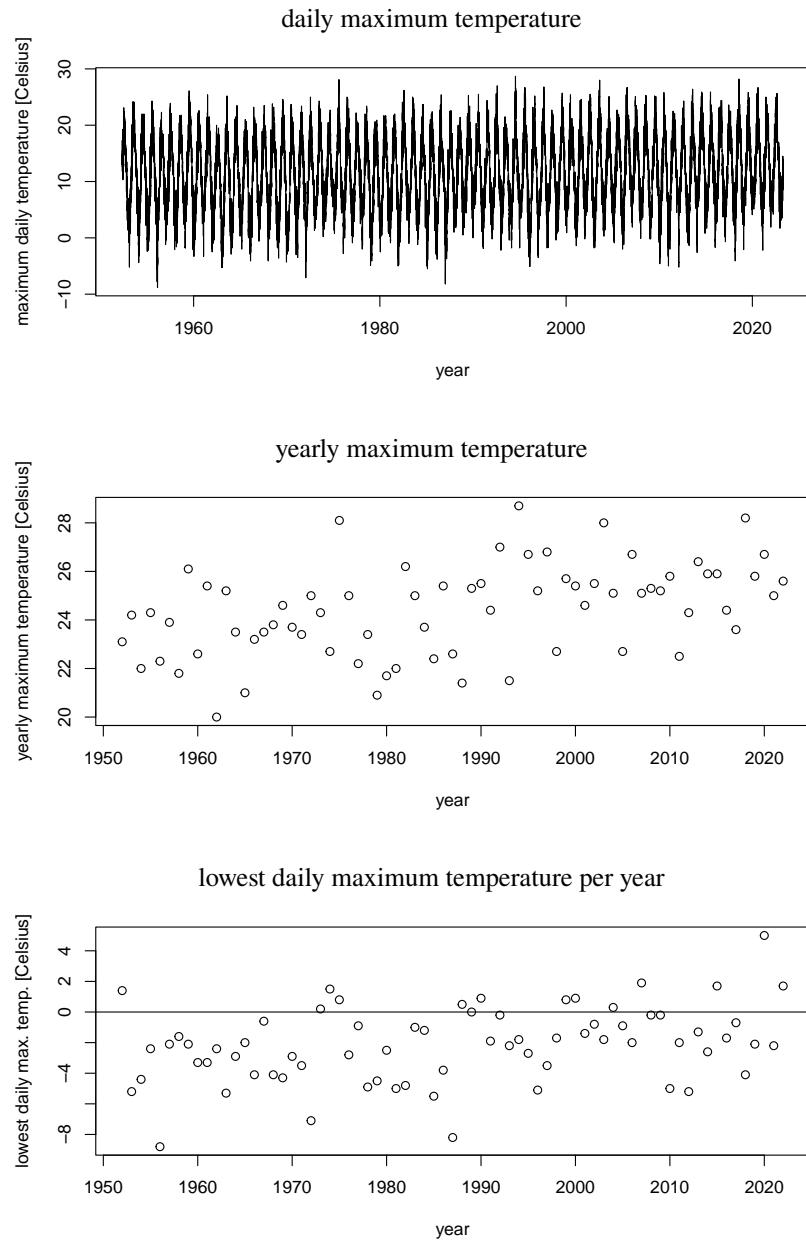


Fig. 10.4 Temperature on the island Helgoland in the North Sea

$$Y_i \sim F(\cdot) \text{ i.i.d., } i = 1, \dots, n$$

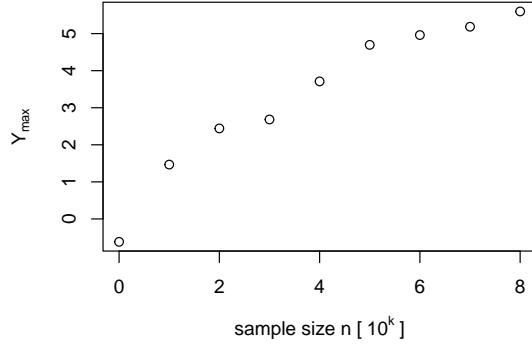


Fig. 10.5 Maximum of n standard normally distributed random variables with $n = 10^k$

and our interest is in the distribution of

$$Y_{\max,n} = \max\{Y_1, \dots, Y_n\}.$$

It is not difficult to derive the distribution function of Y_{\max} , since based on the *i.i.d.* assumption we get

$$\begin{aligned} F_{Y_{\max,n}}(y) &= P(Y_{\max,n} \leq y) \\ &= P(Y_1 \leq y, \dots, Y_n \leq y) \\ &= P(Y_1 \leq y) \cdots P(Y_n \leq y) \\ &= F^n(y). \end{aligned}$$

If we take the limit $n \rightarrow \infty$ we do not obtain a useful limit distribution. If the support of $F(\cdot)$ is bounded, that is, there is an overall maximum value c , such that

$$c = \sup\{y : F(y) < 1\},$$

it follows that

$$\lim_{n \rightarrow \infty} F_{Y_{\max,n}}(y) = \begin{cases} 0 & \text{for } y < c \\ 1 & \text{for } y \geq c \end{cases}$$

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 so that $F_{Y_{\max,n}}(y)$ becomes degenerated. We visualize this in Figure 10.5 where we plot the maximum of $n = 10^k$ simulated $N(0, 1)$ variables. With k increasing, the maximum increases and goes to ∞ with increasing n . Hence we need to normalize the data in some reasonable way. We therefore assume that there are sequences a_n and b_n , which can depend on the sample size, such that

$$\frac{Y_{\max,n} - b_n}{a_n}$$

converges to some non-degenerated distribution with $n \rightarrow \infty$. Hence,

$$\begin{aligned}\lim_{n \rightarrow \infty} P\left(\frac{Y_{\max,n} - b_n}{a_n} \leq y\right) &= \lim_{n \rightarrow \infty} P(Y_{\max,n} \leq a_n y + b_n) \\ &= \lim_{n \rightarrow \infty} F^n(a_n y + b_n) = H(y)\end{aligned}$$

where $y \in \mathbb{R}$ and H is some non-degenerated distribution. The latter is called **extreme value distribution**.

Property 10.1 Let $\{Y_i\}_{i=1}^\infty$ be i.i.d. random variables and $Y_{\max,n} := \max(Y_1, \dots, Y_n)$.

If there are sequences of real numbers $\{a_n\}_{n=1}^\infty, \{b_n\}_{n=1}^\infty$ with

$$\frac{M_n - b_n}{a_n} \xrightarrow{d} Z \text{ as } n \rightarrow \infty,$$

then Z follows one of the following three distributions:

1. Extreme value distribution or Gumbel distribution:

$$H_1(z) = \exp(-\exp(-z))$$

2. (Inverted) Weibull distribution:

$$H_2(z) = \begin{cases} \exp(-(-z)^{1/k}) & \text{for } z \leq 0 \\ 1 & \text{for } z > 0 \end{cases}$$

3. Fréchet Pareto distribution:

$$H_3(z) = \begin{cases} 0 & \text{for } z < 0 \\ \exp(-z^{-1/k}) & \text{for } z \geq 0 \end{cases}$$

We can consider Property 10.1 as a limit theorem for the tails in the distribution. In other words, if the focus is not on the centre 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. 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 USA. Once there, he studied extreme values and can today be seen as the founding father of extreme value theory.

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

$$H(z) = \begin{cases} \exp(-(1+\gamma z)^{-1/\gamma}) & \text{for } \gamma \neq 0 \\ \exp(-\exp(-z)) & \text{for } \gamma = 0 \end{cases} \quad (10.2)$$

with $z = (x - \mu)/\sigma$.

Looking at (10.2) we can interpret the parameters, similarly to other distributions. Parameter μ gives 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 usability if the focus is on analyzing extreme events. The field is well elaborated and we motivated the main ideas here, only. We refer to Coles et al. (2001) or Haan and Ferreira (2006) for further details.

We conclude by demonstrating the ideas with the example from above.

Example 20 CHANGE NUMBERING In example 19 we looked at the maximum yearly temperature. We now take the maximum yearly values shown in the middle plot in Figure 10.4 and estimate the parameters of the GEV distribution using standard maximum likelihood. This leads to parameter estimates

$$\hat{\mu} = 23.76 (0.25)$$

$$\hat{\sigma} = 1.88 (0.17)$$

$$\hat{\gamma} = -0.30 (0.08)$$

with standard errors given in brackets. Since γ is estimated negatively we obtain a Fréchet-Pareto distribution. We plot the resulting distribution in Figure 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 Figure 10.6. The corresponding estimates (and standard errors) are

$$\hat{\mu} = -0.25 (0.05)$$

$$\hat{\sigma} = 2.50 (0.21)$$

$$\hat{\gamma} = -3.06 (0.31).$$

We can now return to the question of whether the data are identically distributed, that is whether the maximum temperature in the early years has 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 Figure 10.6. A clear shift towards 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 degrees (Celsius). \triangleright

	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 10.1 Parameter estimates for the two time periods

Proof For a sketch of the proof of Property 10.1 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

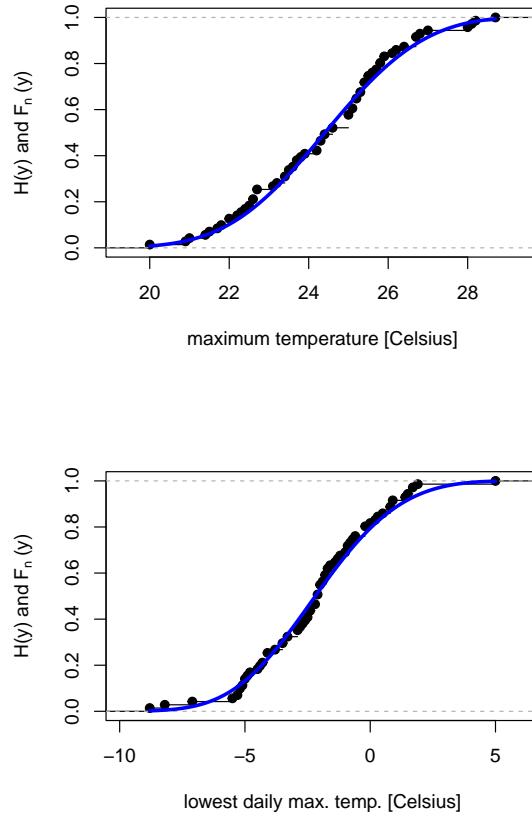


Fig. 10.6 Distribution of yearly maximum temperature and fitted extreme value distribution (upper plot). Distribution of lowest daily maximum temperature and fitted extreme value distribution (bottom plot)

$$\begin{matrix}
 Y_{11} & Y_{12} & \dots & Y_{1K} \\
 Y_{21} & Y_{22} & \dots & Y_{2K} \\
 \vdots & \vdots & \ddots & \vdots \\
 Y_{T1} & Y_{T2} & \dots & Y_{TK}
 \end{matrix} \tag{10.3}$$

where $n = T * K$. Consider for instance Y_{t1}, \dots, Y_{tk} as the daily temperature in year t , like in the example above. We can then define

$$Y_{t,\max,K} = \max\{Y_{t1}, \dots, Y_{tk}\}$$

as the row maxima, which in the example is the annual maximum temperature. The overall maximum is then

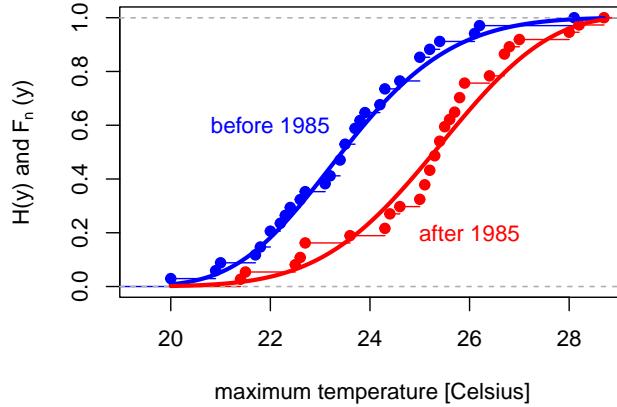


Fig. 10.7 Distribution of yearly maximum temperature and fitted extreme value distribution for years before and after 1985

$$Y_{\max,n} = \max \{ Y_{t,\max,K} \}.$$

This is also called taking the "block maximum", where the blocks here correspond to years. Looking at Figure 10.4 we have plotted $Y_{t,\max,K}$ in the middle plot.

With this notation we have

$$Y_{\max,n} = \max_{1 \leq t \leq T} \max_{1 \leq k \leq K} (Y_{tk}) = \max_{1 \leq t \leq T} \{ Y_{t,\max,K} \}. \quad (10.4)$$

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 $\frac{Y_{\max,TK} - b_{LK}}{a_{LK}}$ also converges to Z , for $T \rightarrow \infty$ and $K \rightarrow \infty$. 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 $K \rightarrow \infty$ and all $1 \leq t \leq T$.

Assuming independence we can assume $Y_{t,\max,K} \sim F_{Y_{\max,K}}$, i.e. all row maxima have the same distribution. Using (10.4) we get

$$F_{Y_{\max,TK}}(y) = F_{Y_{\max,K}}^T(y)$$

where $F_{Y_{\max,TK}}(\cdot)$ 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

$$F_Z(b_T + a_T z) = [F_Z(z)]^T. \quad (10.5)$$

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. \square

10.3 Exercises

Exercise 1

- a.
- b.
- c.

Chapter 11

Multivariate Data

We will look at multivariate data in this chapter. We assume that y has multiple components which interact with each other. There is no causal order in that the first component influences the second, the first two influence the third, and so on. Instead, we consider the variables to be on "equal footing", which means that they influence each other mutually. We want to provide methods that allow us to quantify associations, in particular beyond simple bivariate correlations.

11.1 Models for contingency tables

We now consider multivariate data and let $Y = (Y_1, \dots, Y_p)$ be a p -dimensional random vector. We begin with discrete valued quantities and assume that $Y_j \in \{1, \dots, K_j\}$ for $j = 1, \dots, p$. Hence, Y can be seen to take values in the p -dimensional cube with side length K_1, K_2, \dots, K_p . We visualize a 4 by 2 by 3-dimensional cube in Figure 11.1. The outcome $Y = (k_1, k_2, k_3)$ means that Y lies in the element (k_1, k_2, k_3) .

With multiple multivariate observations, we run into notational ambiguities. We have an index for the component of Y as well as an index for the observation number. This would lead to the use of double subscripts, which will make the formulae look rather complicated. In particular, we will also need multiple subscripts to capture higher-order mutual dependence among the multivariate data, which would require multiple subscript indices. We therefore want to pursue a change of notation in this chapter, which hopefully is not confusing but contributes to the readability of the material. To be specific, we will notate the observation index as superscript put in brackets. Hence, $Y^{(i)} = (Y_1^{(i)}, \dots, Y_p^{(i)})$ is the p -dimensional random vector of the i observation. The data are then recorded as multivariate independent observations denoted by $Y^{(1)}, \dots, Y^{(n)}$. With this notational change, we can write the cell counts as

$$n_{k_1 k_2 \dots k_p} = \sum_{i=1}^n 1\{y^{(i)} = (k_1, k_2, \dots, k_p)\}$$

We also need counts for the margins of the table which we denote by replacing the index with a dot, that is for instance

$$n_{k_1 k_2 \dots k_{p-1}, \cdot} = \sum_{i=1}^n 1\{y_1^{(i)} = k_1\} 1\{y_2^{(i)} = k_2\} \dots 1\{y_{p-1}^{(i)} = k_{p-1}\}$$

or

$$n_{k_1, \dots, \cdot} = \sum_{i=1}^n 1\{y_1^{(i)} = k_1\}.$$

Accordingly

$$n = n_{\cdot, \dots, \cdot} = \sum_{i=1}^n 1.$$

This corresponds to a multivariate p -dimensional contingency table with an example given in the following.

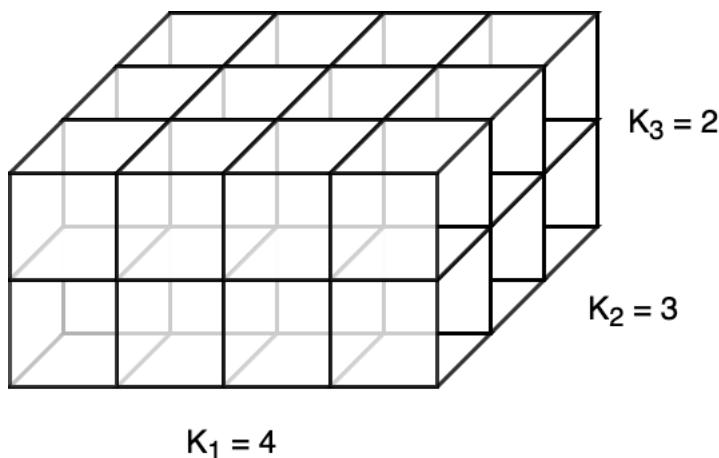


Fig. 11.1 Three dimensional contingency table represented as 3D object

Example 21 We consider the following example as a demonstration. The data come from a survey of ($n = 3185$) rented households in the city of Munich and we look at the 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 : centre (yes or no)
- Y_3 : new contract (yes or no)
- Y_4 : district (one 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 apparently does also contain structural zeros, since some districts are central while others are non-central. It is more common 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 is 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.

floor space	center	
	no	yes
< 50	228	227
50 - 100	1455	803
> 100	207	205

Table 11.1 Munich rental guide survey, tabulated with respect to floor space and center

new contract	district						
	Allach	Altstadt	Au-Haidhausen	Aubing	Berg am Laim	Bogenhausen	...
no	17	38	131	18	50	111	...
yes	12	20	75	36	33	73	...

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 $Y = (Y_1, \dots, Y_p)$ be a discrete-valued random variable. The distribution of Y follows a **log-linear model** if

$$P(Y = y; \theta) \propto \exp\left\{\theta_0 + \sum_j s_j(y)\theta_j\right\} \quad (11.1)$$

where $s_j(\cdot)$ are some linearly independent statistics calculated from $y = (y_1, \dots, y_p)$ and $\theta = (\theta_0, \theta_1, \dots)$ is the parameter vector.

This is a very general setup which we need to specify further to make the idea clear. We do this first for $p = 2$, which is a squared contingency table. In this case, it is helpful to use a double index for the statistics, which we want to motivate subsequently. As statistics we use

$$\begin{aligned} s_{i_1 \cdot}(y) &= 1\{y_1 = i_1\} \text{ for } 1 \leq i_1 < K_1 \\ s_{\cdot i_2}(y) &= 1\{y_1 = i_2\} \text{ for } 1 \leq i_2 < K_2 \\ s_{i_1 i_2}(y) &= 1\{y_1 = i_1, y_2 = i_2\} \text{ for } 1 \leq i_1 < K_1 \text{ and } 1 \leq i_2 < K_2. \end{aligned}$$

In other words $s_{i_1 \cdot}(\cdot)$ counts the observations for which $y_1 = i_1$ while $s_{i_1 i_2}(\cdot)$ is the count of outcomes with $y_1 = i_1$ and $y_2 = i_2$. Note that the above indices do not run over all possible outcomes and, in fact, we postulate $i_1 < K_1$ and $i_2 < K_2$. We need these constraints to make the statistics linearly independent, and hence the model identifiable. This is necessary since we have for instance $1\{y_1 = K_1\} = 1 - \sum_{i_1=1}^{K_1-1} s_{i_1 \cdot}(y)$.

Assume now we have independent observations $Y^{(1)}, \dots, Y^{(n)}$. The general form of the log-linear model (11.1) then becomes for observations $Y^{(1)}, \dots, Y^{(n)}$

$$\begin{aligned} P(Y^{(1)} = y^{(1)}, \dots, Y^{(n)} = y^{(n)}) &\propto \prod_{i=1}^n \exp\left\{\theta_0 + \sum_{i_1=1}^{K_1-1} s_{i_1 \cdot}(y^{(i)})\theta_{i_1 \cdot} + \sum_{i_2=1}^{K_2-1} s_{\cdot i_2}(y^{(i)})\theta_{\cdot i_2} + \sum_{i_1=1}^{K_1-1} \sum_{i_2=1}^{K_2-1} s_{i_1 i_2}(y^{(i)})\theta_{i_1 i_2}\right\} \\ &= \exp\left\{\theta_0 + \sum_{i_1=1}^{K_1-1} n_{i_1 \cdot} \theta_{i_1 \cdot} + \sum_{i_2=1}^{K_2-1} n_{\cdot i_2} \theta_{\cdot i_2} + \sum_{i_1=1}^{K_1-1} \sum_{i_2=1}^{K_2-1} n_{i_1 i_2} \theta_{i_1 i_2}\right\}. \end{aligned}$$

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 $n_{i_1 \cdot} = \sum_{i_2=1}^{K_2} n_{i_1 i_2}$.

To simplify the formula we set $K_1 = 2$ and $K_2 = 2$ to obtain a two-by-two table. In this case, we obtain the probability (likelihood) for the data

$$P(Y^{(1)} = y^{(1)}, \dots, Y^{(n)} = y^{(n)}) \propto \exp\{\theta_0 + n_{1 \cdot} \theta_{1 \cdot} + n_{\cdot 1} \theta_{\cdot 1} + n_{11} \theta_{11}\}$$

If we now set all parameters with double index to zero, which in the two-by-two table is just one parameter, we obtain that the distribution factorizes to a component dependent only on the outcome of Y_1 and a component dependent only on the outcome of Y_2 . Consequently, in this case, we have that Y_1 and Y_2 are independent.

The log-linear model does not define a proper distribution since the normalization constant is missing. We can easily add this by employing 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 $N_{i_1 i_2} = \sum_{i=1}^n \mathbb{1}\{Y_1^{(i)} = i_1, Y_2^{(i)} = i_2\}$. We then assume

$$N_{i_1 i_2} \sim \text{Poisson}(\lambda_{i_1 i_2})$$

where

$$\lambda_{i_1 i_2} = \exp\{\theta_0 + \theta_{i_1 \cdot} + \theta_{\cdot i_2} + \theta_{i_1 i_2}\}$$

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 likelihood for the observed counts $n_{i_1 i_2}$

$$\begin{aligned} P(N_{i_1 i_2} = n_{i_1 i_2}, 1 \leq i_1 \leq K_1, 1 \leq i_2 \leq K_2) \\ &= \prod_{i_1=1}^{K_1} \prod_{i_2=1}^{K_2} \frac{\lambda_{i_1 i_2}^{n_{i_1 i_2}}}{n_{i_1 i_2}!} \exp\{-\lambda_{i_1 i_2}\} \\ &\propto \prod_{i_1=1}^{K_1} \prod_{i_2=1}^{K_2} \exp\{n_{i_1 i_2} \log \lambda_{i_1 i_2}\} \\ &= \exp\left\{n_{\cdot \cdot} \theta_0 + \sum_{i_1=1}^{K_1-1} n_{i_1 \cdot} \theta_{i_1 \cdot} + \sum_{i_2=1}^{K_2-1} n_{\cdot i_2} \theta_{\cdot i_2} + \sum_{i_1=1}^{K_1-1} \sum_{i_2=1}^{K_2-1} n_{i_1 i_2} \theta_{i_1 i_2}\right\}, \end{aligned}$$

where $n_{\cdot \cdot}$ is the number of observations. This clearly 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 a multinomial model results, that is

$$P(N_{i_1 i_2} = n_{i_1 i_2}, i_1 = 1, \dots, K_1, i_2 = 1, \dots, K_2 | n_{\cdot \cdot} = n) = \binom{n_{\cdot \cdot}}{n_{11} \dots n_{K_1 K_2}} \prod_{i_1=1}^{K_1} \prod_{i_2=1}^{K_2} \pi_{i_1 i_2}^{n_{i_1 i_2}}$$

with $\pi_{i_1 i_2} = \lambda_{i_1 i_2} / \sum_{i_1=1}^{K_1} \sum_{i_2=1}^{K_2} \lambda_{i_1 i_2}$. Hence, the information on the overall sample size is not informative with respect to the log-linear structure.

Log-linear models get 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 as well as 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 ... by K_p dimensional contingency table. This is, returning to Figure 11.1, a p -dimensional cube with side length $K_1, K_2, K_3, \dots, K_p$. We define with N_{i_1, \dots, i_p} the random cell counts and with n_{i_1, \dots, i_p} the observed ones and we utilize the bullet point notation in that a dot as subscript indicates summation over the index. This leads to the following clumsy formula

$$\begin{aligned} P(N_{i_1, \dots, i_p} = n_{i_1, \dots, i_p}, i_1 = 1, \dots, K_1; i_2 = 1, \dots, K_2, \dots) \\ \propto \exp \left\{ \theta_0 + \sum_{i_1=1}^{K_1-1} n_{i_1, \dots, \cdot} \theta_{i_1, \dots, \cdot} + \sum_{i_2=1}^{K_2-1} n_{\cdot, i_2, \dots, \cdot} \theta_{\cdot, i_2, \dots, \cdot} + \dots + \sum_{i_p=1}^{K_p-1} n_{\cdot, \dots, i_p} \theta_{\cdot, \dots, i_p} \right. \\ + \sum_{i_1=1}^{K_1-1} \sum_{i_2=1}^{K_2-1} n_{i_1 i_2, \dots, \cdot} \theta_{i_1 i_2, \dots, \cdot} + \dots + \sum_{i_{p-1}=1}^{K_{p-1}-1} \sum_{i_p=1}^{K_p-1} n_{\cdot, \dots, i_{p-1} i_p} \theta_{\cdot, \dots, i_{p-1} i_p} + \dots \\ \left. + \sum_{i_1=1}^{K_1-1} \sum_{i_2=1}^{K_2-1} \dots \sum_{i_p=1}^{K_p-1} n_{i_1 i_2 \dots i_p} \theta_{i_1 i_2 \dots i_p} \right\}. \end{aligned}$$

Though the formula looks messy, it has a very clear structure. The components involved depend only on one variable, on all two-wise pairs, on all triples and so on up to all p variables. We call the terms with more than one variable involved interaction terms. Hence $\theta_{i_1 i_2, \dots, \cdot}$ is the pairwise interaction of Y_1 and Y_2 for outcome i_1 and i_2 and consequently $\theta_{i_1 i_2 i_3, \dots, \cdot}$ is a three-wise interaction. We see that interaction, which corresponds to mutual dependence, can be pretty cumbersome for discrete value multivariate data and hence the aim is to find a more parsimonious structure of the model. Before doing so let us formalize the complex model again. To simplify notation we just write 1 for variable Y_1 , 2 for variable Y_2 and so on. We also simplify the view by taking $p = 3$ for the moment and looking at the binary case only, i.e. $K_1 = K_2 = K_3 = 2$. This gives a 2 by 2 by 2 table with 8 cells. The full model, also called the saturated model, has 8 parameters, namely $\theta_0, \theta_{1..}, \theta_{.1..}, \theta_{11..}, \theta_{1..1}, \theta_{.11..}$ and $\theta_{111..}$. If we condition on the sample size, then θ_0 is given through the condition. We can denote the model as

$$1 * 2 * 3,$$

which means that the three way interaction between variables Y_1, Y_2 and Y_3 exists. If we impose that $\theta_{111..} = 0$, that is we exclude the three-way interaction, then the highest

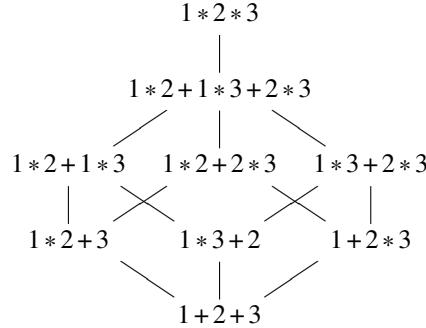
order interactions are between Y_1 and Y_2 , between Y_1 and Y_3 as well as between Y_2 and Y_3 . Formally, we can write this as

$$1 * 2 + 1 * 3 + 2 * 3.$$

If we further restrict $\theta_{11} = 0$, 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. This leads to some model hierarchy which is visualized as follows



The statistical challenge is now to find a good model given the data in the contingency table. Formally this can be done by testing whether parameters are equal to zero or by model selection using, e.g. the Akaike Information Criterion which we discussed in Chapter 7. This apparently extends to higher-order contingency tables with $p > 3$. It should also be clear that the model hierarchy visualized for $p = 3$ above gets even more complex for $p > 3$, which is why we want to simplify the structure even further. In fact, some of the models above allow for conditional independence interpretation which we want to explore further.

We say that Y_1 and Y_2 are conditionally independent given Y_3 if and only if

$$P(Y_1 = i_1, Y_2 = i_2 | Y_3 = i_3) = P(Y_1 = i_1 | Y_3 = i_3)P(Y_2 = i_2 | Y_3 = i_3)$$

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

$$Y_1 \perp\!\!\!\perp Y_2 | Y_3 \text{ or } 1 \perp\!\!\!\perp 2 | 3. \quad (11.2)$$

Applying Bayes rule shows

$$P(Y_{i_1} = i_1, Y_{i_2} = i_2 | Y_{i_3} = i_3) \propto P(Y_{i_1} = i_1, Y_{i_2} = i_2, Y_{i_3} = i_3)$$

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 $Y_1 \perp\!\!\!\perp Y_2 | Y_3$ we have

$$P(Y_{i_1} = i_1, Y_{i_2} = i_2 | Y_{i_3} = i_3) \propto h_{13}(i_1, i_3)h_{23}(i_2, i_3)$$

for some functions $h_{13}(\cdot)$ and $h_{23}(\cdot)$ for all possible outcomes i_1, i_2 and i_3 . This even holds vice versa.

Considering the structure of the log-linear model we see that this holds if all parameters that involve components i_1 and i_2 together are equal to zero. Putting it differently, we have

$$Y_1 \perp\!\!\!\perp Y_2 | Y_3 \Leftrightarrow \theta_{i_1 i_2 \cdot} = 0 \text{ and } \theta_{i_1 i_2 i_3} = 0 \text{ for all } i_1, i_2 \text{ and } i_3.$$

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 . We can also interpret the conditional independence as follows:

$$P(Y_1 = i_1 | Y_2 = i_2, Y_3 = i_3) = P(Y_1 = i_1 | Y_3 = i_3).$$

In other words, given the outcome of Y_3 , any additional information on the outcome of Y_2 does not change the distribution of Y_1 .

The above hierarchy of the model structure can be simplified to models with conditional independence structure. This is discussed in more depth in the next section. Estimation 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 19 (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:

floor space * center + center * district + floor space * new contract

The model simplifies quite drastically and we need to elaborate on the interpretation of the results. \triangleright

11.2 Graphical Models

We motivated the idea of conditional independence in the previous section and we get more formal now. Let $Y = (Y_1, \dots, Y_p)$ be a p -dimensional random variable with $f_{123\dots p}(y_1, \dots, y_p)$ as corresponding density or probability function. We define with A,B and C the following subsets

$$\begin{aligned} A \cap B &= A \cap C = B \cap C = \emptyset \\ A \cup B \cup C &\subset \{1, \dots, p\} \\ A \neq \emptyset \text{ and } B \neq \emptyset \end{aligned}$$

We say that A is conditionally independent of B given C, denoted as

$$A \perp\!\!\!\perp B | C$$

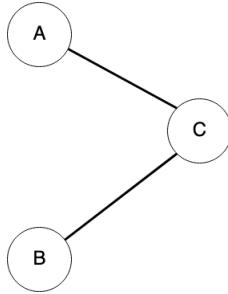
if and only if

$$f_{A \cup B | C}(y_A, y_B | y_C) = f_{A|C}(y_A | y_C) f_{B|C}(y_B | y_C). \quad (11.3)$$

This holds if and only if the density decomposes to

$$f_{A \cup B \cup C}(y_A, y_B, y_C) = h_{A \cup C}(y_A, y_C) h_{B \cup C}(y_B, y_C).$$

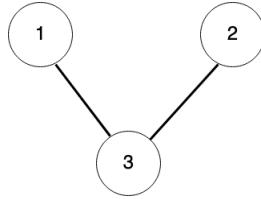
We can visualize the conditional independence with a graph, which explains the terminology **graphical model**. To be concrete we visualize that A and B are conditionally independent of C as follows:



The missing edge between A and B translates to conditional independence given C. This graph in fact is for sets of variables and we want to extend this to single variables. To do so, define a Graph G as pair $G = (E, V)$, where $E = \{1, \dots, p\}$ is the set of nodes and $V \subset E \times E$ is a set of edges. The edges are of less interest, but the missing edges are relevant concerning conditional independence. We link these two concepts through

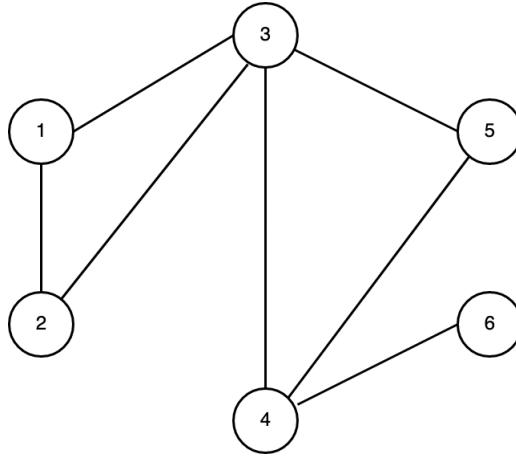
$$i \perp\!\!\!\perp j | E \setminus \{i, j\} \Leftrightarrow (i, j) \notin V \text{ and } (j, i) \notin V.$$

Exemplary for $p = 3$, the following graph can be interpreted as $1 \perp\!\!\!\perp 2 | 3$:



Let us now get to more dimensions. We give in Figure 11.2 a graphical model for six variables. There are 8 edges missing in the graph and these translate to the independencies listed below:

$$\begin{aligned} & 1 \perp\!\!\!\perp 4 | \{2, 3, 5, 6\} \\ & 1 \perp\!\!\!\perp 5 | \{2, 3, 4, 6\} \\ & 1 \perp\!\!\!\perp 6 | \{2, 3, 4, 5\} \\ & 2 \perp\!\!\!\perp 4 | \{1, 3, 5, 6\} \\ & 2 \perp\!\!\!\perp 5 | \{1, 3, 4, 6\} \\ & 2 \perp\!\!\!\perp 6 | \{1, 3, 4, 5\} \\ & 3 \perp\!\!\!\perp 6 | \{1, 2, 4, 5\} \\ & 5 \perp\!\!\!\perp 6 | \{1, 2, 3, 4\} \end{aligned}$$

**Fig. 11.2** Graphical Model for 6 variables

The statement is also called *pairwise conditional independence*. Looking at the graph in Figure 11.2 we can also make other statements, which in fact can be shown to hold equivalently in a conditional independence framework. Note that all paths from the set of nodes $\{4, 5, 6\}$ to the set of nodes $\{1, 2\}$ go via 3. This translates to

$$\{1, 2\} \perp\!\!\!\perp \{4, 5, 6\} | 3.$$

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 above eight conditional independence statements, we have similar decomposition properties which indeed confirm the above decomposition. This is called *global conditional independence*. We refer to Lauritzen (1996) or Maathuis et al. (2018) for details. Generally, we can say that for variable sets A, B and C it holds that $A \perp\!\!\!\perp B | C$ if every path from $i \in A$ to $j \in B$ goes over elements in set C for all $i \in A$ and all $j \in B$. Hence in Figure 11.2 we also have $\{3, 5\} \perp\!\!\!\perp 6 | 4$. There is a further conditional independence property. We can also condition on the direct neighbours in the graph and achieve independence from the remaining nodes. For instance, variable 2 has neighbors 1 and 3 and we can state

$$2 \perp\!\!\!\perp \{4, 5, 6\} | \{1, 3\}.$$

Or in general, let $N(i)$ be the set of neighbors of i , that is

$$\mathcal{N}(i) = \{j : (i, j) \in V \text{ or } (j, i) \in V\}.$$

Then

$$i \perp\!\!\!\perp E \setminus \{i \cup \mathcal{N}(i)\} \mid \mathcal{N}(i).$$

This is called *local conditional independence*. The three independence formulations are in fact equivalent.

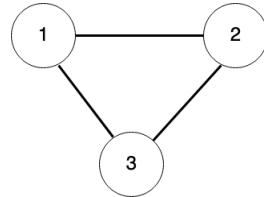
Property 11.1 Given a graphical model, the three conditional independence interpretations *pairwise conditional independence*, *local conditional independence* and *global conditional independence* are equivalent. \square

The proof can for instance be found in Whittaker (2009).

For given *i.i.d.* data $Y^{(i)} = (Y_1^{(i)}, \dots, Y_p^{(i)})$ we can now aim to select a model which has as many conditional independencies as possible based on some model selection criteria. This again induces some model hierarchy which we show exemplarily for three variables in Figure 11.3. Apparently for more than three variables the set of possible models gets large.

Example 20 CONTINUED FROM ABOVE 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 Figure 11.4. We can interpret this exemplary as follows. Given that an apartment is in the centre, 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. \triangleright

Graphical Models are a flexible tool and we utilized the idea already in Chapter 8 when we introduced the simulation of multivariate random variables. The graph quite easily allows 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:



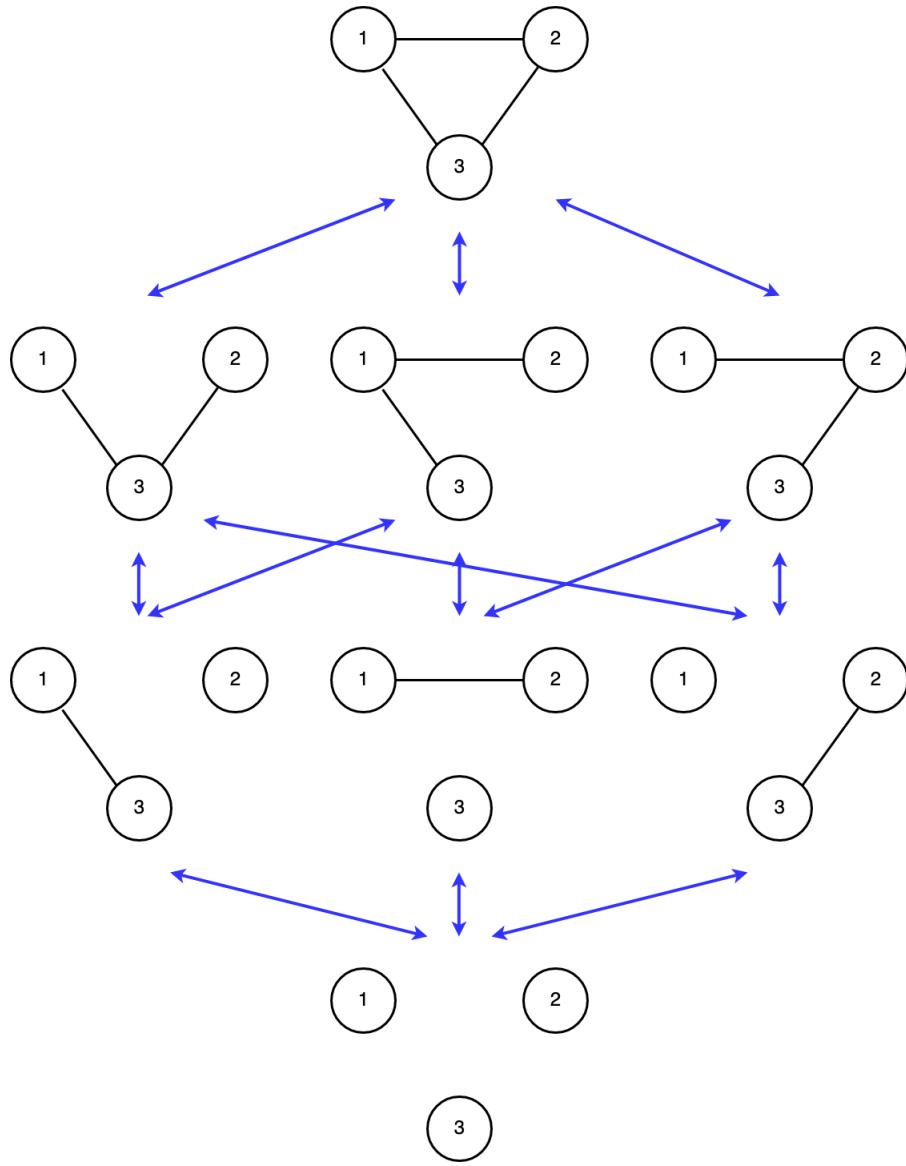


Fig. 11.3 Model hierarchy of graphical models for three variables

This follows since conditional in 3 we obtain no independence of Y_1 and Y_2 . At first, this looks somewhat surprising. Although Y_1 and Y_2 are independent, that is any information on Y_1 does not provide information on Y_2 , the picture changes once we observe Y_3 . However, the following example motivates that information on Y_3 can lead to dependence of Y_1 and Y_2 . Assume a technical system with two

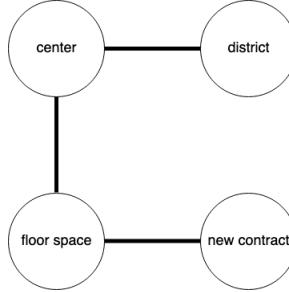


Fig. 11.4 Graphical Model rental survey example

components. We define with Y_1 the operating status for component 1, with $Y_1 = 1$ if the component is running and $Y_1 = 0$ otherwise. Accordingly, Y_2 expresses the information for component 2. The two variables are considered to be Binomial and independent of each other, that is the technical operation of the one system is not dependent in any case on the other. However, the full system only runs if both components are working and fails if one of the two is not working. This can also be modelled with a binary variable.

$$\begin{aligned}
 Y_1 &= \begin{cases} 1 & = \text{technical failure in component 1} \\ 0 & = \text{normal operation} \end{cases} \\
 Y_2 &= \begin{cases} 1 & = \text{technical failure in component 2} \\ 0 & = \text{normal operation} \end{cases} \\
 Y_3 &= \max(Y_1, Y_2) = \begin{cases} 1 & \text{system break down} \\ 0 & \text{system working} \end{cases}
 \end{aligned}$$

We model the components as

$$\begin{aligned}
 Y_1 &\sim B(1, \pi_1) \\
 Y_2 &\sim B(1, \pi_2) \\
 Y_3 | Y_1, Y_2 &\sim B(1, \pi_3(y_1, y_2))
 \end{aligned}$$

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 let us assume that the technical system breaks down, that is $Y_3 = 1$ and we question whether component 1 or component 2 caused the problem. If we check component 1 and obtain that component 1 is running, that is $Y_1 = 0$, this leads to an increased probability that the system failure is due to a failure of component 2. In fact, ignoring variable Y_3 and applying the Bayes rule we have

$$P(Y_2 = 1 | Y_1 = 1) = \frac{P(Y_1 = 1, Y_2 = 1)}{P(Y_1 = 1)} = \frac{\pi_1 \pi_2}{\pi_1} = \pi_2.$$

Hence we have independence of Y_1 and Y_2 . But if we condition on Y_3 we get

$$\begin{aligned} P(Y_1 = 1, Y_2 = 1 | Y_3 = 1) &= \frac{P(Y_1 = 1, Y_2 = 1, Y_3 = 1)}{P(Y_3 = 1)} \\ &= \frac{P(Y_1 = 1, Y_2 = 1)}{P(\max(Y_1, Y_2) = 1)} \\ &= \frac{\pi_1 \pi_2}{1 - P(Y_1 = 0, Y_2 = 0)} \\ &= \frac{\pi_1 \pi_2}{1 - (1 - \pi_1)(1 - \pi_2)} \end{aligned}$$

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

$$\begin{aligned} P(Y_2 = 1 | Y_3 = 1, Y_1 = 0) &= \frac{P(Y_2 = 1, Y_1 = 0, Y_3 = 1)}{P(Y_1 = 0, Y_3 = 1)} \\ &= \frac{P(Y_1 = 0, Y_2 = 1)}{P(Y_1 = 0, Y_2 = 1)} \\ &= 1. \end{aligned}$$

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 towards **directed acyclic graphs** or in short **DAG**. We sketch the resulting DAG for the above discussed example in Figure 11.5. Nodes 1 and 2 are not connected but both lead to node 3.

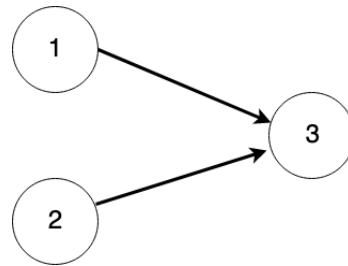


Fig. 11.5 Directed Acyclic Graph (DAG) for three variables

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 $\text{Pa}(i)$ the parents. For instance in Figure 11.5 we have $\text{Pa}(1) = \emptyset$, $\text{Pa}(2) = \emptyset$ and $\text{Pa}(3) = \{1, 2\}$. Moreover, the nodes in the DAG are ordered so that $\text{Pa}(1) = \emptyset$ holds in general and node n has no following node, $n \notin \text{Pa}(j)$ for $j = 1, \dots, p - 1$. Moreover in general there are no loops or cycles, that is

$$\text{Pa}(i) \subset \{1, \dots, i - 1\}.$$

The conditional independence statements are then

$$i \perp\!\!\!\perp \{1, \dots, i - 1\} \setminus \text{Pa}(i) \mid \text{Pa}(i)$$

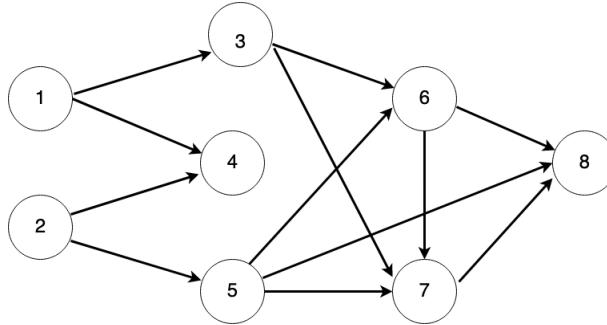
for all $i = 1, \dots, p$. Hence, if we condition on the parents of a variable in the DAG then all variables beyond the direct parents which lie "before" the variable in the DAG do not provide any information about Y_i . moreover, we can also state that two variables are generally (i.e. unconditionally) independent if they have no common ancestors and are not connected by an active 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 Chapter 8 and will return to this in the time series examples in Chapter 12. To fully understand the ideas of a DAG let us look at a more complex example. Consider the DAG in Figure 11.6. 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 and we refer exemplarily to Cowell et al. (2007). In general, for a multivariate distribution on variables Y_1, \dots, Y_p we can factorize the distribution to

$$f_{Y_1 \dots Y_p}(y_1, \dots, y_p) = \prod_{j=1}^p f_{Y_j | \text{Pa}(j)}(y_j | y_{\text{Pa}(j)}).$$

Hence, for the DAG in Figure 11.6 we get (in short notation)

**Fig. 11.6** DAG for eight variables

$$f_{12\dots 8} = f_1 f_2 f_3|_1 f_4|_{12} f_5|_2 f_6|_{35} f_7|_{356} f_8|_{567}.$$

An interesting statement results for the conditional independence structure which results from a DAG. Note that looking at the DAG in Figure 11.5 we saw with the example that

$$Y_1 \perp\!\!\!\perp Y_2 | Y_3,$$

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 a graphical model from the DAG by

1. putting an undirected edge between all nodes in $\text{Pa}(i)$ for $i = 1, \dots, p$ (that is we are "marrying" the parents).
2. replace all directed edges by undirected ones.

We demonstrate this in Figure 11.7, where we include the "moralized" edges as dashed lines. This graph now describes all conditional independence statements as a graphical model. Apparently, we loose the information of two independencies which disappeared due to looking at all p variables. DAGs are a central tool in causality theory, which we do not want to touch here. We instead refer to Peters et al. (2017) and Pearl (2009).

11.3 Multivariate Normal Distribution

The most commonly used multivariate distribution is certainly the multivariate normal distribution. We worked with the distribution already in the simulation section in Chapter 8.4. We want to show now, that the multivariate normal distribution again

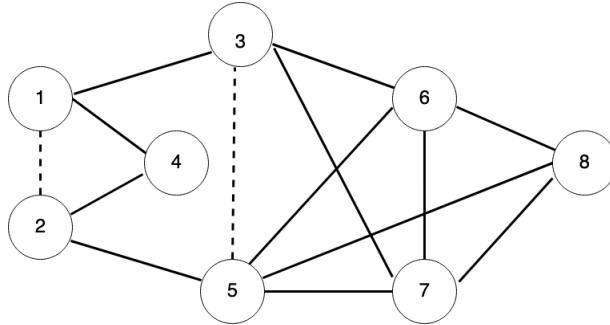


Fig. 11.7 Moralized graph from Figure 11.6. Induced edges are dashed.

has a log-linear form and with the results from above we can quite easily relate a multivariate normal distribution to graphical models.

Definition 11.2 Let $Y = (Y_1, \dots, Y_p)$ be a continuous random vector. Y follows a multivariate normal distribution with mean $\mu = (\mu_1, \dots, \mu_p)^\top$ and $p \times p$ dimensional variance matrix Σ if

$$f(y; \mu, \Sigma) = \frac{1}{(2\pi)^{\frac{p}{2}}} |\Sigma|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} (y - \mu)^\top \Sigma^{-1} (y - \mu) \right\}. \quad (11.4)$$

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 (11.4) we obtain the maximum likelihood estimates

$$\begin{aligned} \hat{\mu} &= \frac{1}{n} \sum_{i=1}^n y^{(i)} = \left(\frac{1}{n} \sum_{i=1}^n y_1^{(i)}, \dots, \frac{1}{n} \sum_{i=1}^n y_q^{(i)} \right)^\top \\ \hat{\Sigma} &= \frac{1}{n} \sum_{i=1}^n (y^{(i)} - \hat{\mu})(y^{(i)} - \hat{\mu})^\top. \end{aligned}$$

Let $\Omega = \Sigma^{-1}$ be the inverse variance matrix, commonly called the concentration matrix and denote with Ω_{jl} the j, l -th element of Ω . This allows to rewrite (11.4) to

$$f(y; \mu, \Sigma) \propto \exp \left\{ -\frac{1}{2} \sum_{j=1}^p \Omega_{jj} y_j^2 + \frac{1}{2} \sum_{j=1}^p \sum_{l>j}^p \Omega_{jl} y_j y_l \right\} \quad (11.5)$$

This shows immediately, that if $\Omega_{jl} = 0$ then the density decomposes as follows. Let $C = \{1, \dots, p\} \setminus \{j, l\}$ be the index set without coefficients j and l . Then

$$f(y; \mu, \Sigma) \propto h_{j \cup C}(y_j, y_C) h_{+l \cup C}(y_l, y_C)$$

which, based on (11.3) implies that Y_j is conditionally independent of Y_l given the remaining variables. Hence, we have the equivalence

$$\Omega_{jl} = 0 \Leftrightarrow Y_j \perp\!\!\!\perp Y_l | Y_C. \quad (11.6)$$

The concentration matrix is informative as it carries conditional independencies.

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 = Number of movies watched

B = Number of books read

V = Hours spent playing videogames

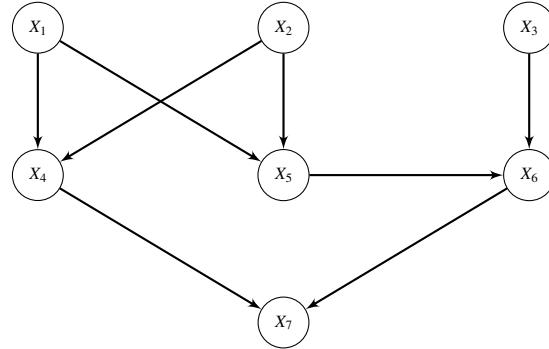
with $M \leq 7$, $B \leq 2$, and $V \leq 20$. We are interested in studying the behaviour of the students with respect to these three aspects.

- Can you sketch a contingency table for this data? If so, do it, if not, explain why.
- 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.
- 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)?

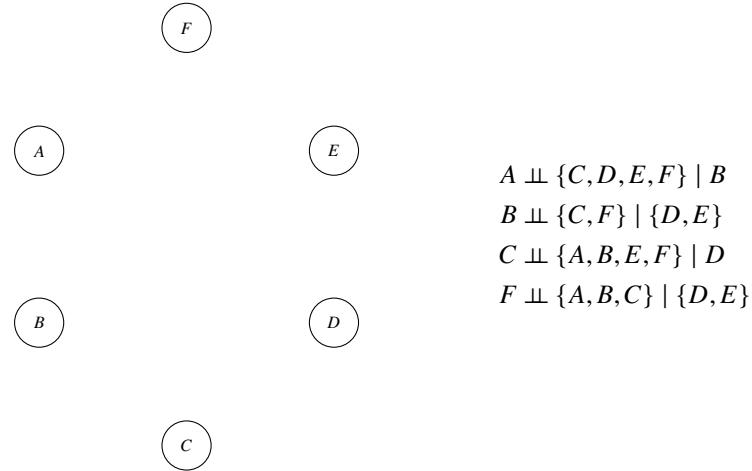
- d. Given the assumptions made in point c), are the variables M and V independent? Are they conditionally independent? Sketch a dependency graph of the three variables.

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 , $i = 1, \dots, n$, state the variables that it is independent of. Additionally, factorise 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, then remove edges one by one.

**Exercise 3**

Let $\mathbf{x} \in \mathbb{R}^n$ be a random vector with $E(\mathbf{x}) = \mathbf{0}$ and $V = \text{Cov}(\mathbf{x})$. Assuming that \mathbf{x} follows a multivariate normal distribution and that $\text{Cov}(\mathbf{x})$ is a diagonal matrix, show that x_1, \dots, x_n are independent.

Chapter 12

Non i.i.d. Data

The underlying lens on data analysis fostered in this book is to assume that the data are generated via a probabilistic distribution that we approximate through a statistical model. In this context, the i.i.d. assumption states that each observation is an independent observation of some data-generating process. From the practical perspective, this assumption naturally leads to a likelihood factorized into separate low-dimensional components, which is easier to take care of during estimation discussed in Chapter 3. At the same time, this assumption forms the foundation of asymptotic theory from Chapter 4. Once this assumption, however, does not hold anymore, all these advantages vanish. In Chapter , we studied first where some independence assertions are not permissible. As a bypass, the framework of Graphical Models was introduced to cope with such issues in a reasonably abstract way. The current chapter aims to tackle a related problem from the angle of specific data types. In the following chapters, we cover data with additional structure in the form of temporal order (Chapter 12.1), nesting structure (Chapter 12.2), and location (Chapter 12.3). Consecutively, we provide an introduction to data types that are not in the realm of Euclidean space anymore, i.e., where distances between outcomes of the random variable can not be calculated anymore through the typical Euclidean distance. For Chapter 12.4, we introduce issues and ways to cope with dependence in the context of networked data. Finally, Chapter 12.5 will tackle functional observations, where observations are not any more scalar values but entire functions.

12.1 Time Series

We assume in the following that the data are collected over time, which we denote as time series. We write

$$Y_1, Y_2, Y_3, \dots, Y_T$$

and assume discreteness and equidistance of the timepoints $t = 1, 2, \dots, T$. We assume further that there is some correlation over time in that $\text{Cov}(Y_t, Y_{t'}) \neq 0$ for $1 \leq t, t' \leq T$. We define with

$$\mu_Y(t) = E(Y_t)$$

the mean function and with

$$\gamma_Y(t, t') = \text{Cov}(Y_t, Y_{t'})$$

the covariance function. To exemplify the idea, we look at the following examples. Figure 12.1 shows the course of the annual global average temperature for the years 1850 to 2019. The data show the difference in the average over the considered years. We see a clear trend in that $\mu_y(t)$ has been increasing since about 1975. We include an estimate $\hat{\mu}_y(t)$ in the plot, which allows to look at the remaining fluctuation $y_t - \hat{\mu}_y(t)$, shown in the bottom of Figure 12.1. Details on how to estimate $\mu_y(t)$ are sketched later in the chapter. As a second example we look at the German stock market index DAX for the years 2010 to 2022. This is shown in Figure 12.2. We again see a slight upwards trend and in particular, a sudden fall in early 2020, which is when the COVID-19 pandemic hit the world and the stock market suffered. We can correct for the trend by looking at so-called log-returns, that is we define the time series

$$\tilde{Y}_t = \log\left(\frac{Y_t}{Y_{t-1}}\right).$$

We plot this in the bottom of Figure 12.2. In this plot we see additional features of the data, namely the variation of the DAX. There are two periods that stick out, namely 2012 and again 2020, where we observe a larger fluctuation of \tilde{Y}_t . As final example, we look at the daily temperature in the city of Augsburg for the years 2003-2022, which are shown in the upper plot of Figure 12.3. We see a clear annual variation with winters and summers clearly exposed. We will later show how to extract this annual pattern and look at the remaining fluctuation. In other words, we can extract the seasonal variation from the data and look at the resulting seasonal-corrected data. This is shown in the bottom plot of Figure 12.3.

To start our look at time series we first set the mean structure to zero, that is $\mu_Y(t) \equiv 0$. We further assume some (weak) stationarity in that the covariance function does only depend on the time lag $|t - t'|$ but not on the concrete value of t , that is $\gamma_Y(t, t')$ can be written as

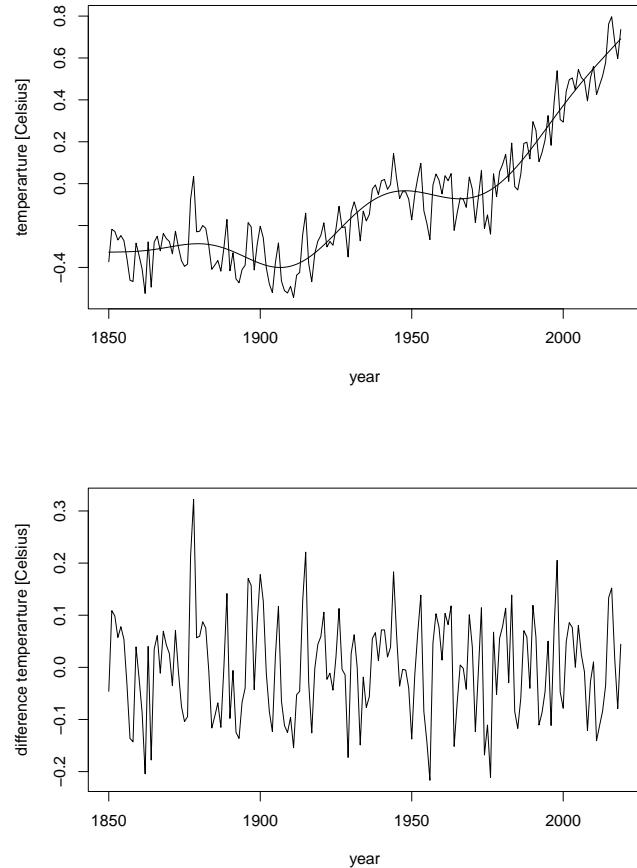


Fig. 12.1 Variation of global annual temperature around the mean value (top plot) and the detrended variation (bottom plot)

$$\gamma_Y(t, t') = \gamma_Y(|t - t'|)$$

In this case, we call

$$\gamma_Y(h) = \text{Cov}(Y_t, Y_{t+h}) = \text{Cov}(Y_{t+h}, Y_t)$$

the **autocovariance function** with $h = 0, 1, \dots, T$. This leads to the variance matrix

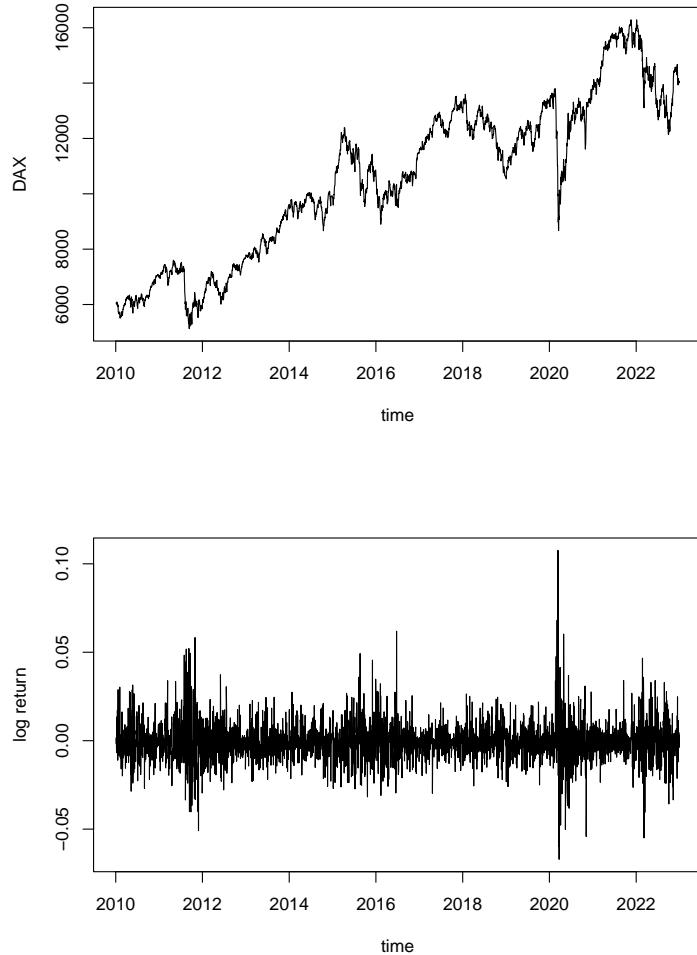


Fig. 12.2 German stock market index (DAX) (top plot) and log returns (bottom plot)

$$\Gamma = \begin{bmatrix} \gamma(0) & \gamma(1) & \gamma(2) & \dots & \gamma(T) \\ \gamma(1) & \gamma(0) & \gamma(1) & \dots & \gamma(T-1) \\ \gamma(2) & \gamma(1) & \gamma(0) & \dots & \gamma(T-2) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \gamma(T) & \gamma(T-1) & \gamma(T-2) & \dots & \gamma(0) \end{bmatrix}.$$

Since $\gamma_Y(0) = \text{Var}(Y_t)$ we obtain with

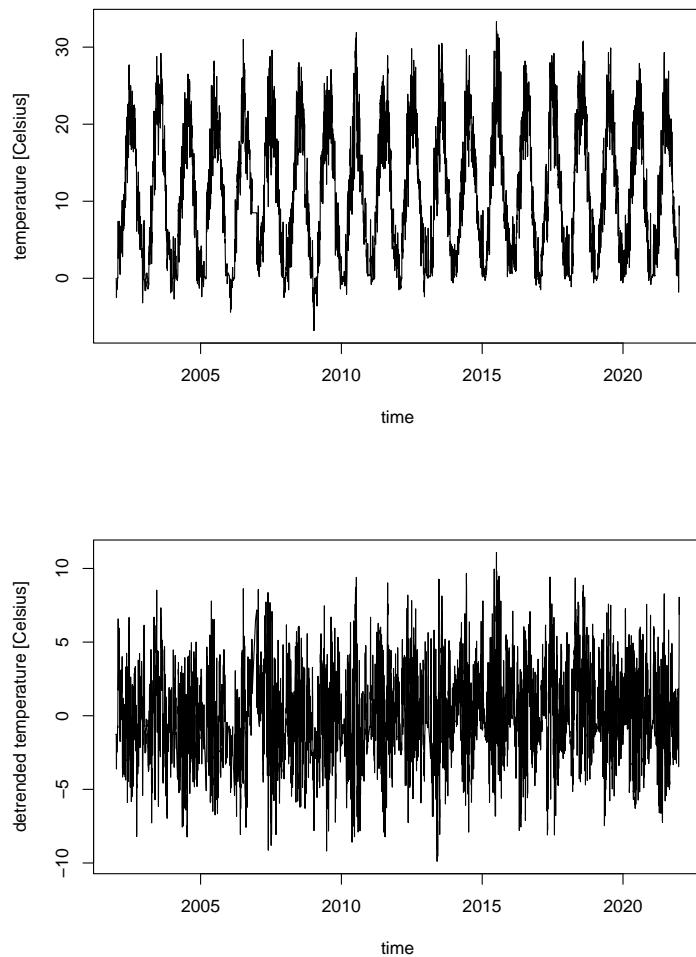


Fig. 12.3 Annual daily temperature in the city of Augsburg (top plot) and deseasoned data (bottom plot)

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)}, \quad (12.1)$$

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

$$\gamma_Y(h) = \begin{cases} \sigma^2 & \text{for } h = 0 \\ 0 & \text{for } h \neq 0 \end{cases}.$$

We call a time series $U_t \sim N(0, \sigma^2)$ with $\gamma_U(h) = 0$ for $h \neq 0$ white noise.

It will be of particular interest to estimate the autocorrelation/autocovariance function from the data y_1, \dots, y_T . Assuming a constant mean $E(X_t) = \mu$ we can estimate μ through the arithmetic mean $\bar{y} = \sum_{t=1}^T y_t / T$. The autocovariance function is then estimated through

$$\hat{\gamma}_Y(h) = \frac{1}{T} \sum_{t=1}^{T-h} (y_t - \bar{y})(y_{t+h} - \bar{y})$$

for $0 \leq h < T$. Note that the sum is over $T - h$ elements but we divide it by T and not by $T - h$. This leads to a positive definite matrix estimate

$$\hat{\Gamma} = \begin{bmatrix} \hat{\gamma}(0) & \hat{\gamma}(1) & \hat{\gamma}(2) & \dots & \hat{\gamma}(T) \\ \hat{\gamma}(1) & \hat{\gamma}(0) & \hat{\gamma}(1) & \dots & \hat{\gamma}(T-1) \\ \hat{\gamma}(2) & \hat{\gamma}(1) & \hat{\gamma}(0) & \dots & \hat{\gamma}(T-2) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \hat{\gamma}(T) & \hat{\gamma}(T-1) & \hat{\gamma}(T-2) & \dots & \hat{\gamma}(0) \end{bmatrix}.$$

To see this let $x_t = y_t - \bar{y}$. Then $\hat{\Gamma}$ can be written as

$$\hat{\Gamma} = \frac{1}{T} DD^\top,$$

where D is the $T \times (2T - 1)$ dimensional matrix with entries

$$D = \begin{bmatrix} 0 & \dots & 0 & x_1 & \dots & x_{n-1} & x_n \\ \vdots & \ddots & x_1 & x_2 & \dots & x_n & 0 \\ 0 & \ddots & \ddots & \vdots & \ddots & \ddots & \vdots \\ x_1 & x_2 & \dots & x_n & 0 & \dots & 0 \end{bmatrix}.$$

This readily shows positive (semi) definiteness through

$$a^\top \hat{\Gamma} a = \frac{1}{T} [aD][aD]^\top \geq 0$$

for all $a \in \mathbb{R}^T$. From $\hat{\Gamma}$ we immediately obtain an estimate of R through

$$\hat{R} = \frac{\hat{\Gamma}}{\hat{\gamma}(0)}$$

with $\hat{\gamma}(0)$ as diagonal element of $\hat{\Gamma}$. We will utilize \hat{R} later. In fact the columns of \hat{R} , or to be more specific the first column is also known as the sample autocorrelation function defined as

$$\hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)}.$$

The estimate gives information about the serial correlation, which is shown in the following examples. In Figure 12.4 we show the resulting estimates $\hat{\gamma}(h)$ for the three examples from above. The upper plot gives the estimated autocorrelation for the trend-corrected annual temperature, the middle plot is for the log-returns of the DAX and the bottom plot shows the autocorrelation for the seasonal-corrected temperature in Augsburg. While for the first two examples we have a quickly decreasing autocorrelation, which is in principle zero for time lags of order two and above, this is not the case for the daily lags of the temperature in Augsburg.

12.1.1 ARMA Models

A central model class in time series analysis are Auto Regressive Moving Average (ARMA) models, which extend to Auto Regressive Integrated Moving Average (ARIMA) processes by taking the differences of the original data. The model class was originally proposed by Whittle (1951) but made generally accessible by Box and Jenkins (1970), see also Box et al. (2015). A short description on the history of time series analysis is also found in Tsay (2000). Let us start with the Auto-Regressive

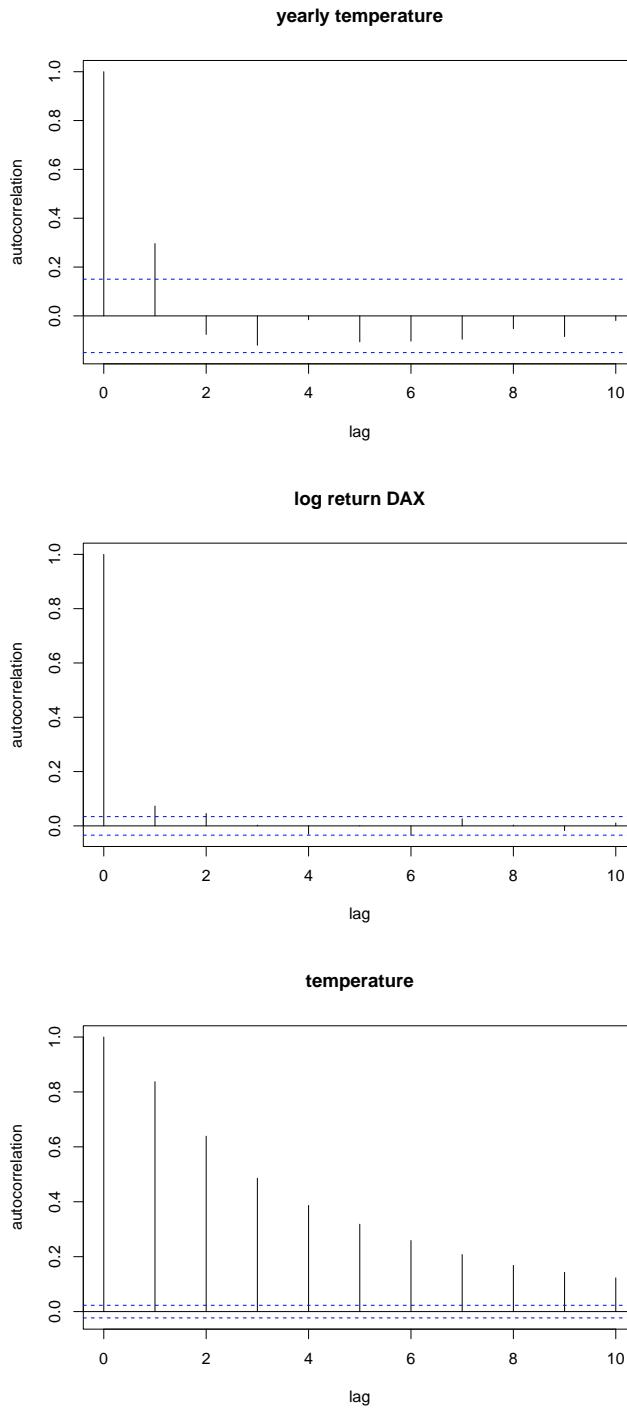


Fig. 12.4 Estimated autocorrelation function for the three examples in Figures 12.1, 12.2 and 12.3, from top to bottom.

(AR) model. We assume that Y_t is determined by the previous p values plus some white, independent noise added to it.

Definition 12.1 An Auto Regressive 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 with a so called backshift operator B . This is defined as follows. We write

$$\begin{aligned} BY_t &= Y_{t-1} \\ B^2 Y_t &= BY_{t-1} = Y_{t-2} \end{aligned}$$

or generally

$$B^p Y_t = Y_{t-p}.$$

This also holds for negative values of p , e.g., $B^{-1} Y_t = Y_{t+1}$, which makes sense for theoretical purposes, as we will see later.

With the backshift operator, we can rewrite the AR(p) process in a concise form to

$$\left(1 - \sum_{j=1}^p \phi_j B^j\right) Y_t = Z_t.$$

The easiest example is the AR(1) process, which in fact is quite often found in real data. In this case, we have

$$Y_t = \phi Y_{t-1} + Z_t,$$

where we dropped the subscript 1 at ϕ to avoid unnecessary notation. Note that ϕ plays the role of the autocovariance, since

$$\text{Cov}(Y_t, Y_{t-1}) = \phi \text{Cov}(Y_{t-1}, Y_{t-1}).$$

Assuming stationarity we obtain for the autocorrelation defined in (12.1) $\rho(1) = \phi$. It is not difficult to obtain in general $\rho(h) = \phi^h$. Apparently, we need $|\phi| < 1$, so that the correlation matrix Γ results as

$$\Gamma = \begin{bmatrix} 1 & \phi^1 & \phi^2 & \dots & \phi^\top \\ \phi^1 & 1 & \phi^1 & \dots & \phi^{T-1} \\ \phi^2 & \phi^1 & 1 & \dots & \phi^{T-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \phi^\top & \phi^{T-1} & \phi^{T-2} & \dots & 1 \end{bmatrix}.$$

The role of ϕ also becomes clear by converting the AR(1) process into an infinite series, later defined as Moving Average. 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 backwards, we obtain

$$Y_t = \sum_{j=0}^{\infty} \phi^j Z_{t-j}.$$

This in turn allows to derive the variance of Y_t through

$$\text{Var}(Y_t) = \sum_{j=0}^{\infty} (\phi^2)^j \sigma^2 = \frac{\sigma^2}{1 - \phi^2},$$

where the limit follows through the geometric series. Again we see the requirement that $|\phi| < 1$. In Figure 12.5 we show three examples of AR(1) processes. The top plot shows a series with $\phi = 0.9$, the middle plot with $\phi = 0$, which corresponds to independence, and the bottom plot 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 regularly observed series, e.g., daily temperature data have such features after correcting for seasonality. For zero correlation we do not see these phase patterns and for negative correlation we see the opposite, namely strongly alternating behavior. Finally, let us look at AR(1) processes with $|\phi| > 1$. The process in this case is said to explode, since its variance increases. This is shown in Figure 12.6, where we show the behavior of an AR(1) process with $\phi = 1.02$. We will come back to this in a more mathematical form later in the chapter.

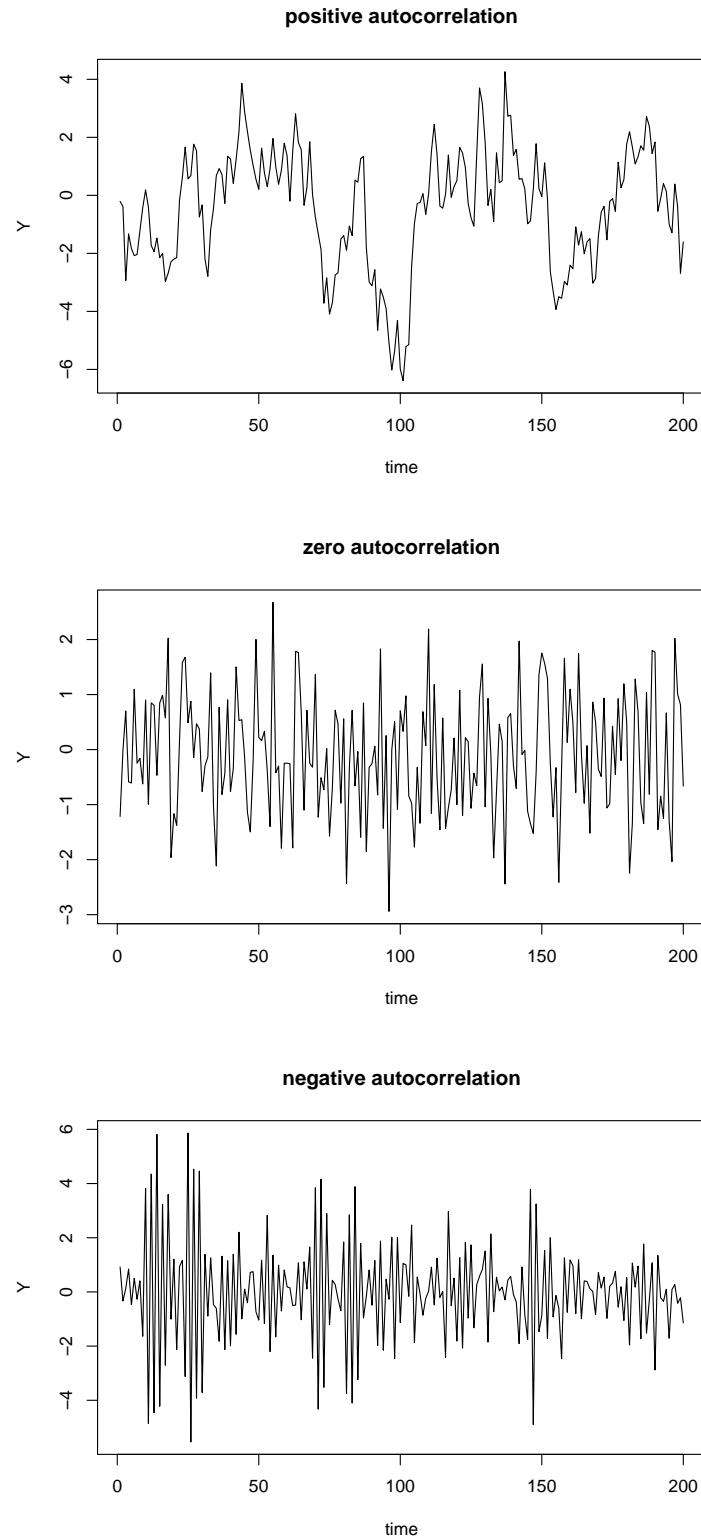


Fig. 12.5 Simulated AR(1) process with positive (top), zero (middle) and negative (bottom) autocorrelation

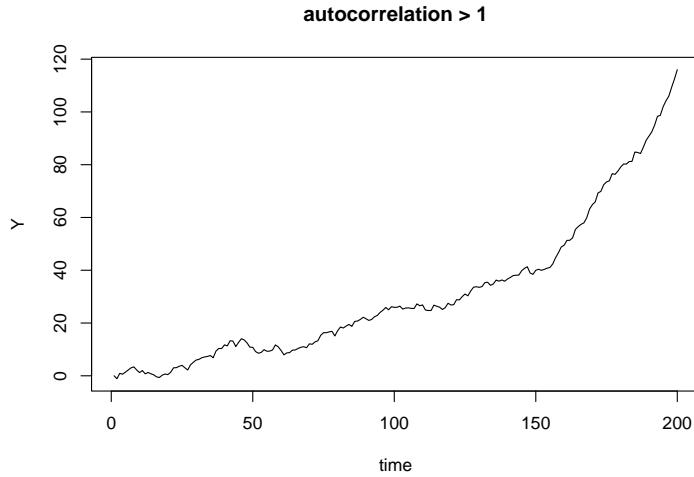


Fig. 12.6 Simulated AR(1) process with autocorrelation $\phi > 1$

Let us now move to the Moving Average.

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.

Like above we can make use of the backshift operator and write

$$Y_t = \sum_{i=0}^q \theta_i B^i Z_t$$

with $\theta_0 = 1$. We consider the MA(1) process, that is

$$Y_t = Z_t + \theta Z_{t-1},$$

where we dropped the subscript 1 for notational simplicity. For the autocovariance we obtain

$$\begin{aligned}
\gamma(0) &= \text{Cov}(Y_t, Y_t) \\
&= \text{Cov}(Z_t + \theta Z_{t-1}, Z_t + \theta Z_{t-1}) \\
&= \sigma^2 (1 + \theta^2) \\
\gamma(1) &= \text{Cov}(Y_t, Y_{t-1}) \\
&= \text{Cov}(Z_t + \theta Z_{t-1}, Z_{t-1} + \theta Z_{t-2}) \\
&= \theta \sigma^2 \\
\gamma(h) &= 0 \text{ for } h > 1
\end{aligned}$$

and accordingly

$$\rho(h) = \begin{cases} 1, & \text{for } h = 0 \\ \frac{\theta}{1+\theta^2}, & \text{for } h = 1 \\ 0, & \text{for } h > 1 \end{cases}.$$

In Figure 12.7 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 Figure 12.5 shows that the processes have completely different behavior.

We can now combine the two processes leading to ARMA(p, q).

Definition 12.3 The Auto Regressive Moving Average process of order p and q, in short ARMA(p, q) is defined as

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q}.$$

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 $\theta_0 = 1$. Let us get a little more formal and 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 to write an ARMA(p, q) process as

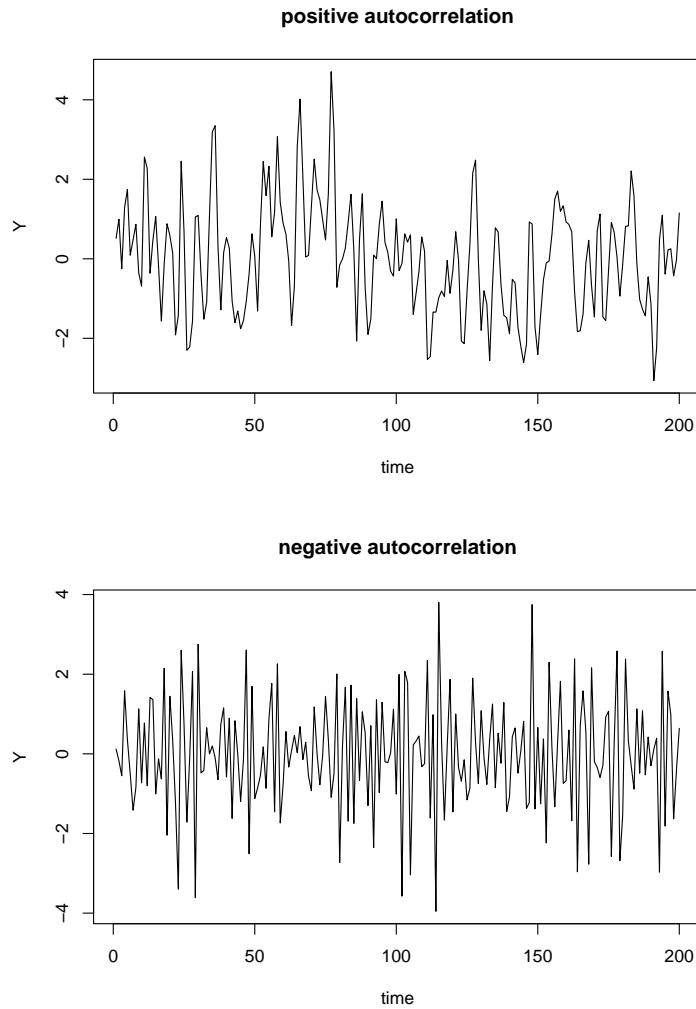


Fig. 12.7 Simulated MA(1) process with positive (top), and negative (bottom) correlation

$$\Phi(B)Y_t = \Theta(B)Z_t \quad (12.2)$$

This looks elegant and allows us to investigate some properties of ARMA(p, q) processes. First, it is important to note that the representation is not unique. Look at the white noise process

$$Y_t = Z_t,$$

with $Z_t \sim N(0, \sigma^2)$ i.i.d. We can write

$$Y_t - \phi Y_{t-1} = Z_t - \theta Z_{t-1},$$

which is the same white noise process but now written as ARMA(1, 1) process with

$$\Phi(B) = 1 - \phi B \text{ and } \Theta(B) = 1 + \theta B,$$

with $\theta = -\phi$. This will become relevant when we come to estimation, since significant coefficients for the AR, as well as for the MA part, do not imply the validity of an ARMA process.

In this respect it can be helpful to rewrite the process using the operator notation from above. In fact, we aim to find the "inverse" of $\Phi(B)$, which we denote as $\Phi^{-1}(B)$. This means, that

$$\Phi^{-1}(B)\Phi(B) = 1,$$

with 1 denoting the identity operator, that is $1Y_t = Y_t$. 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 as

$$\Phi^{-1}(B) = \sum_{j=0}^{\infty} \phi^j B^j = 1 + \phi^1 B^1 + \phi^2 B^2 + \dots$$

To see this note that we obtain a telescope sum

$$\begin{aligned} \Phi^{-1}(B)\Phi(B) &= (1 + \phi^1 B^1 + \phi^2 B^2 + \dots)(1 - \phi B) \\ &= 1 - \phi B + \phi B - \phi^2 B^2 + \phi^2 B^2 + \dots \\ &= 1, \end{aligned}$$

that is elements are added and subtracted. Since $|\phi| < 1$, the summands converge to zero. Such operations are also available for higher order operators, which are, however more difficult to construct (see McCullagh (2018)) and conditional on convergence of the operator. We look at the latter later. For now we assume that an inverse exists, which allows to write the ARMA(p, q) process in (12.2) as

$$Y_t = \Phi^{-1}(B)\Theta(B)Z_t,$$

which for the ARMA(1, 1) process results to

$$\begin{aligned} Y_t &= (1 + \phi^1 B^1 + \phi^2 B^2 + \dots)(1 - \theta B)Z_t \\ &= Z_t + (\phi + \theta)Z_{t-1} + \phi(\phi + \theta)Z_{t-2} + \dots \\ &= \sum_{i=0}^{\infty} \psi_i B^i Z_{t-i}, \end{aligned}$$

with $\psi_0 = 1$ and $\psi_i = \phi^{i-1}(\phi + \theta)$ for $i \geq 1$. Hence, any AR components can be transferred to infinite MA components. This holds generally, even for exploding processes. But for non-exploding processes we can construct the time series from the past, which we call causal.

Definition 12.4 An ARMA(p, q) process is **causal** if the process can be rewritten to

$$Y_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j} = \Psi(B)Z_t, \quad (12.3)$$

with $\Psi(B) = \sum_{j=0}^{\infty} \psi_j B^j$ and $\sum_{j=0}^{\infty} |\psi_j| < \infty$ where $\psi_0 = 1$.

Let us look at this in more depth for an AR(1) process. Apparently, if $|\phi| < 1$ we immediately get that the process is causal with $\psi_i = \phi^i$ and with a converging geometric series $\sum_{i=0}^{\infty} |\phi|^i$. But what happens if $|\phi| > 1$? We write

$$\Phi(B) = (1 - \phi B) = \left(1 - (\phi^{-1} B^{-1})^{-1}\right)$$

and expand the series around ϕ^{-1} . The "inverse" of $\Phi(B)$ results through

$$\Phi^{-1}(B) = - \sum_{j=1}^{\infty} \phi^{-j} B^{-j},$$

which converges since $|\phi^{-1}| < 1$. This can be seen, again by telescoping summation leading to

$$\begin{aligned}
\Phi^{-1}(B)\Phi(B) &= \left(-\sum_{j=0}^{\infty} \phi^{-j} B^{-j}\right) \left(1 - (\phi^{-1} B^{-1})^{-1}\right) \\
&= \left(-\phi^{-1} B^{-1} - \phi^{-2} B^{-2} - \dots\right) \left(1 - (\phi^{-1} B^{-1})^{-1}\right) \\
&= -\phi^{-1} B^{-1} + 1 - \phi^{-2} B^{-2} + \phi^{-1} B^{-1} + \dots \\
&= 1
\end{aligned}$$

This in fact shows, that the AR(1) process with $|\phi| > 1$ can be written as

$$\begin{aligned}
Y_t &= \Phi^{-1}(B)Z_t \\
&= -\sum_{j=0}^{\infty} \phi^{-j} B^{-j} Z_t \\
&= \sum_{j=0}^{\infty} \phi^{-j} Z_{t+j}.
\end{aligned}$$

In other words, the AR(1) process for $|\phi| > 1$ is determined by future events and hence it is not causal. This justifies again that we require $|\phi| < 1$ for any reasonable data application. We can extend this towards general ARMA(p, q) processes by reformulating the operator $\Phi(B)$ to a polynomial, written as $\Phi(z)$ with

$$\Phi(z) = \sum_{j=0}^{\infty} \phi_j z^j$$

with $\phi_0 = 1$. It can be shown that if $\phi(z) \neq 0$ for $|z| < 1$, then the process is causal and vice versa. The coefficients ψ_j in (12.3) are found by solving

$$\Psi(z) = \sum_{j=0}^{\infty} \psi_j z^j = \frac{\Theta(z)}{\Phi(z)}$$

for $|z| \leq 1$. We refer to Shumway et al. (2017).

Finally, we extend the ARMA process again towards an ARIMA process, standing for Auto Regressive Integrated Moving Average process. The "integrated" part results through 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 Figure 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.

12.1.2 Autocorrelation and Partial Autocorrelation

We already presented the autocorrelation function $\gamma(h)$ and its estimation above. We want to extend this here towards the partial autocorrelation, which in fact provides valued 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 $\text{Corr}(Y_t, Y_{t-h}) = \phi^h$. 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 seen in Chapter 8. In fact, the order of the AR process determines the order of the required history, as sketched in Figure 12.5. 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}, \dots, Y_{t-h} . In other words,

$$\hat{Y}_{t,h} = \tilde{\phi}_{1,h} Y_{t-1} + \cdots + \tilde{\phi}_{h,h} Y_{t-h},$$

where $\tilde{\phi}_h = (\tilde{\phi}_{1,h}, \dots, \tilde{\phi}_{h,h})$ minimizes the squared prediction error,

$$\tilde{\phi}_h = \arg \min E \left\{ \left(Y_t - \sum_{j=1}^h \phi_{h,j} Y_{t-h} \right)^2 \right\}.$$

Note that $\tilde{\phi}_{p,j} = \phi_j$ and for $h > p$ we have $\tilde{\phi}_{h,j} = 0$ for $j > p$. This in turn means $\hat{Y}_{t,h} = \hat{Y}_{t,p}$ for $h \geq p$. In other words, if h equals the order of the AR process, then the coefficients equal the coefficients of the process and for h larger than the order of the autocorrelation one has achieved the best linear approximation and all further coefficients are zero. The partial autocorrelation function of a stationary process is then defined as

$$\rho^{(p)}(h) = \text{Corr}(Y_t - \hat{Y}_{t,h-1}, Y_{t-h} - \hat{Y}_{t-h,h-1}),$$

where $\hat{Y}_{t,0} \equiv 0$. This is the correlation of Y_t with Y_{t-h} if we condition on the intermediate steps $Y_{t-1}, \dots, Y_{t-h+1}$. Note that for $h > p$ we have

$$\begin{aligned} \rho^{(p)}(h) &= \text{Corr}(Y_t - \hat{Y}_{t,h-1}, Y_{t-h} - \hat{Y}_{t-h,h-1}) \\ &= \text{Cov}(Y_t - \hat{Y}_{t,p}, Y_{t-h} - \hat{Y}_{t-h,p}) \\ &= \text{Cov}(Z_t, Z_{t-h}) \\ &= 0 \end{aligned}$$

Hence, the partial autocorrelation function allows nicely to get insight into the order of the process. We demonstrate this by looking at an AR(3) process as visualized in Figure 12.8. 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 we in contrast take the partial correlation function $\rho^{(p)}(h)$, given in the bottom plot, we see that the first three elements are non-zero but all remaining components beyond the order of 3, vanish.

We briefly sketch how the above quantities can be estimated. Note that we already discussed the estimation of the autocovariance matrix Γ above and we can build upon that. For an AR(p) process we obtain the so called Yale-Walker equation. Note that for $h = 1, \dots, p$

$$\begin{aligned} \gamma(h) &= \text{Corr}(Y_t, Y_{t-h}) \\ &= \text{Cov}(\phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \dots + \phi_p Y_{t-p}, Y_{t-h}) \\ &= \phi_1 \text{Cov}(Y_{t-1}, Y_{t-h}) + \dots + \phi_p \text{Cov}(Y_{t-p}, Y_{t-h}) \\ &= \phi_1 \gamma(|h-1|) + \phi_2 \gamma(|h-2|) + \dots + \phi_p \gamma(|h-p|) \end{aligned} \tag{12.4}$$

Moreover we have

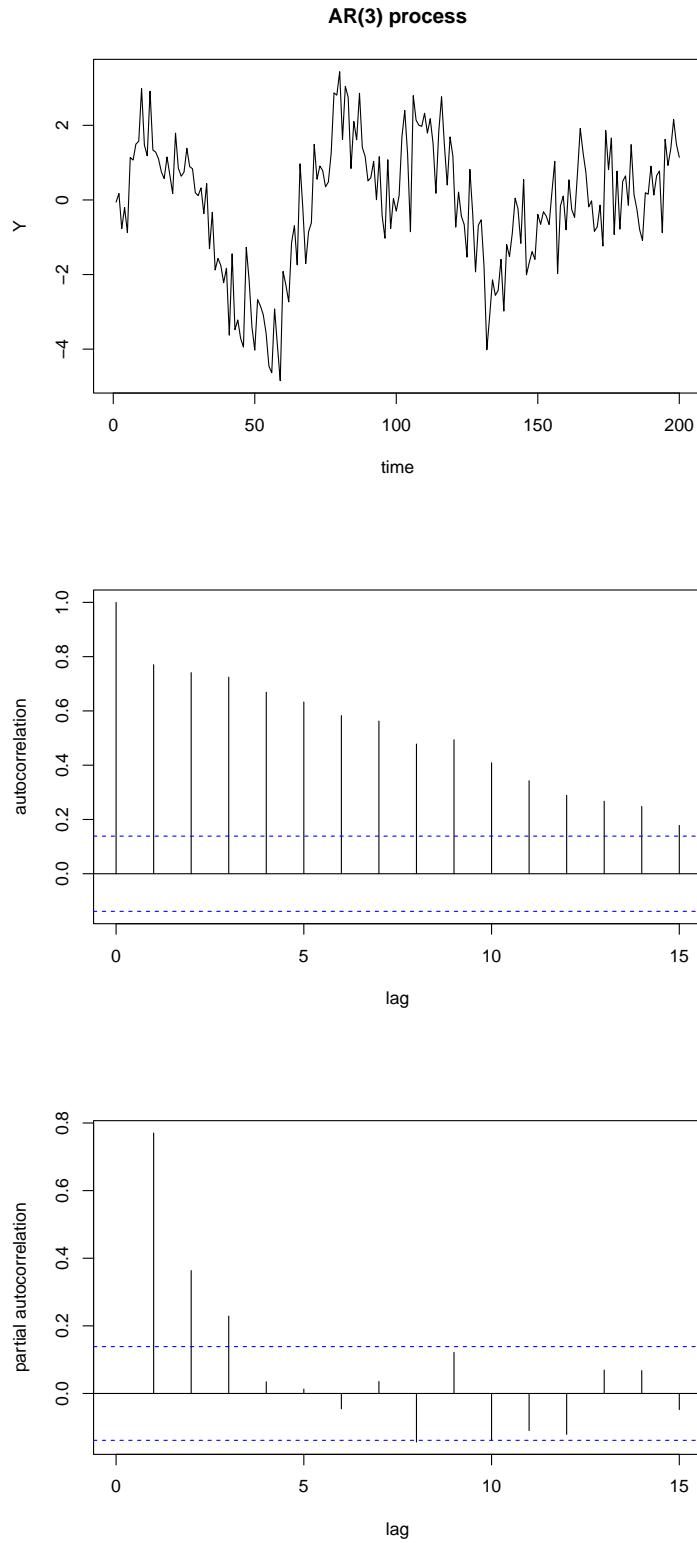


Fig. 12.8 Simulated AR(3) process (top), estimated autocorrelation function (middle) and estimated partial autocorrelation function (bottom)

$$\begin{aligned}
\sigma^2 &= \text{Var}(Z_t) \\
&= \text{Cov}(Z_t, Z_t) \\
&= \text{Cov}(Z_t, Y_t) \\
&= \text{Cov}(Y_t - \phi_1 Y_{t-1} - \phi_2 Y_{t-2} - \cdots - \phi_p Y_{t-p}, Y_t) \\
&= \gamma(0) - \phi_1 \gamma(1) - \phi_2 \gamma(2) - \cdots - \phi_p \gamma(p)
\end{aligned} \tag{12.5}$$

The above equations can be rewritten in matrix notation. Let Γ be the $p \times p$ covariance matrix with entries $\Gamma_{ij} = \gamma(|i - j|)$ and define with γ the p -dimensional vector $(\gamma(1), \dots, \gamma(p))^\top$. Then (12.4) can be written as

$$\Gamma\phi = \gamma$$

with $\phi = (\phi_1, \dots, \phi_p)^\top$. The second equation (12.5) results as

$$\sigma^2 = \gamma(0) - \phi^\top \gamma.$$

With the above estimate for Γ we readily have an estimate for γ , which is just a subvector from the matrix estimate $\hat{\Gamma}$. This leads to estimates $\hat{\phi}$ and $\hat{\sigma}^2$ through

$$\hat{\phi} = \hat{\Gamma}^{-1} \hat{\gamma} \text{ and } \hat{\sigma}^2 = \hat{\gamma}(0) - \hat{\gamma}^\top \hat{\Gamma}^{-1} \hat{\gamma}.$$

The estimates have reasonable asymptotic properties for an AR(p) process in that $T \rightarrow \infty$

$$\begin{aligned}
\sqrt{T}(\hat{\phi} - \phi) &\xrightarrow{a} N(0, \sigma^2 \Gamma^{-1}) \\
\hat{\sigma}^2 &\rightarrow \sigma^2
\end{aligned}$$

and for $h > p$

$$\sqrt{T}\hat{\phi} \xrightarrow{a} N(0, 1)$$

We may also apply maximum likelihood ideas, but we do not want to get further into details here. Instead we refer to literature, which focuses on time series analysis, see e.g.(??).

12.1.3 Forecasting

The central intention 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 a sufficiently long time window T has been observed. We may now question, how Y_{t+1} can be predicted with y_t, \dots, y_{t-p} being observed. This can be reformulated to constructing a prediction function $m(y_t, \dots, y_{t-p})$, such that the mean prediction error

$$E\left\{(Y_{t+1} - m(y_t, \dots, y_{t-p}))^2 | y_t, \dots, y_{t-p}\right\}$$

is small. The model $m(\cdot)$ can be of linear structure, that is

$$m(y_t, \dots, y_{t-p}) = \beta_1 y_t + \beta_2 y_{t-1} + \dots + \beta_{p-1} y_{t-p},$$

which lead to Auto Regressive models discussed above. But apparently we may also use non-linear models. We refer to De Gooijer (2017) for a detailed discussion.

12.2 Multilevel Data and Repeated Measurements

Notation (?): repeated observations $Y_{jt}, t = 1, \dots, n_j, j = 1, \dots, J$

12.3 Spatial Data

Maybe add subsections?

The famous geographer Waldo Rudolph Tobler coined the phrase, “*Everything is related to everything, but near things are more related than distant things*”, now referred to as Tobler’s first law of geography. This is mirrored in statistical analysis of spatial data, as we need to take into account that data are non-*i.i.d.* and ”near observations” are more correlated than ”distant ones”. Spatial data can have different structures and formats and modeling them requires a wide range of different tools. We provide a short overview here and refer to literature that covers one or another aspect in more depth. The distinction between the different data types is understood best with examples. In Figure 12.9 we plot data collected in a survey of households in Munich. Households were randomly selected (based on a random sample of the registry office) and the points show the location of the households in Munich.



Fig. 12.9 Geolocation of sampled apartments 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 due to 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 $i = 1, \dots, n$. We can consider the location itself as relevant information. For instance, we might question, is the intensity of rented apartments higher in some areas of Munich than in other areas? Presumably, in the suburbs, we find more family houses and hence fewer rented apartments. The opposite seems plausible for the city center. We consider the location s_i in this case as relevant information and aim to model the distribution that generates s_i . The easiest is to plot a two-dimensional density. In this case, we assume

$$s_i \sim f(\cdot), \quad i = 1, \dots, n$$

where $f(\cdot)$ is a two-dimensional density over the city area of Munich. Alternatively, one may consider s_i to be drawn from a two-dimensional Poisson process. We do not get into this field here but refer exemplary to Baddeley et al. (2007).

We focus on data settings where we have additional information observed about the locations s of all statistical units. In the rent example, this entails the rent Y_i of the i -th drawn apartment located in s_i . In this setting, Tobler's law states that the rent of two apartments is more correlated with the closer the apartments are (at least if we take other driving factors like the facilities of the apartment into account). Let Y_i denote the rent of apartment i , then Y_i and Y_j are correlated, where the dependence decreases with the distance between the apartments, that is, with $\|s_i - s_j\|$. This is typically modeled with a so-called isotropic covariance matrix, which means

$$\text{Cov}(Y_i, Y_j) = C(\|s_i - s_j\|). \quad (12.6)$$

In spatial statistics, it is more common to work with the variogram, which is defined through

$$\gamma(\|s_i - s_j\|) := \text{Var}(Y_i - Y_j) = \text{Var}(Y_i) + \text{Var}(Y_j) - 2\text{Cov}(Y_i, Y_j)$$

which with (12.6) results to

$$\gamma(\|s_i - s_j\|) = \text{Var}(Y_i) + \text{Var}(Y_j) - 2C(\|s_i - s_j\|).$$

Assuming that the variance does not depend on the location itself we see that

$$C(\|s_i - s_j\|) = -\gamma(\|s_i - s_j\|) + \sigma^2$$

where $\sigma^2 = \text{Var}(Y_i) = \text{Var}(Y_j)$. We plot the resulting fitted variogram in Figure 12.10, which shows that there is spatial correlation up to a distance of about 2 kilometers. Beyond that the variogram gets flat on the level of two times the variance. We include a smoothly fitted curve in the plot to account for the estimation variability of the estimates. The point estimates of the variogram are thereby derived as follows. We discretize the distances to $0 < d_1 < d_2 < d_3 \dots$. For a given distance d_k we then estimate

$$\hat{\gamma}(d_k) = \frac{1}{|N(d_k)|} \sum_{i,j \in N(d_k)} (Y_i - Y_j)^2$$

where

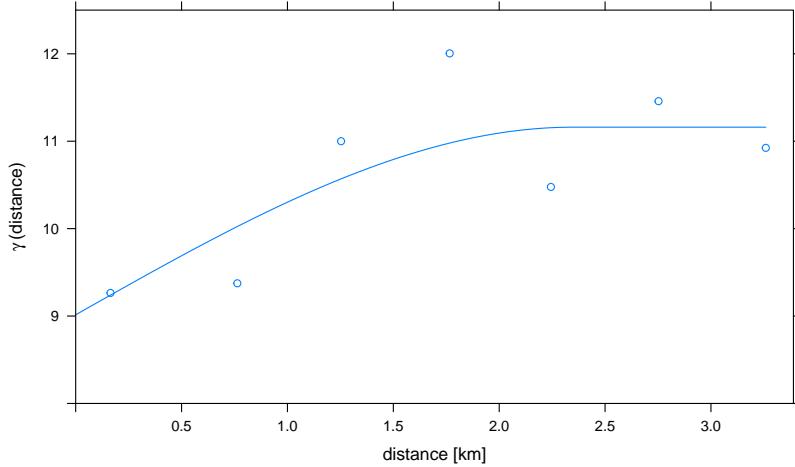


Fig. 12.10 Variogram for rents of apartments in Munich

$$N(d_k) = \left\{ (i, j) : \left| \|s_i - s_j\| - d_k \right| < \Delta \right\}.$$

Hence, $N(d_k)$ 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 latter needs to be specified in advance.

The definition of distance itself is relevant and at this point, it is worth emphasizing that distance in principle can be any kind of distance. Apparently, in most cases it will be the Euclidean distance, that is for the distance of geolocations s_i and s_j we take

$$d = \|s_i - s_j\|_2,$$

where $\|\cdot\|_2$ refers to scalar product of the coordinates. However, the definition of the variogram also applies to different distance measures, for instance, if the data "live" on a grid, like streets. In this case, the distance between two points might be calculated through the shortest driving distance between two points. We sketch this in Figure 12.11 where we look at the Euclidean distance on the left and the (street) network distance on the right. Conceptually this does not make a difference and the only thing required is to calculate the pairwise distances $d_{ij} = \|s_i - s_j\|$, where $\|\cdot\|$ here refers to any distance measure. We need though, that distance is symmetric, i.e. there are no one-way streets simply speaking.

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

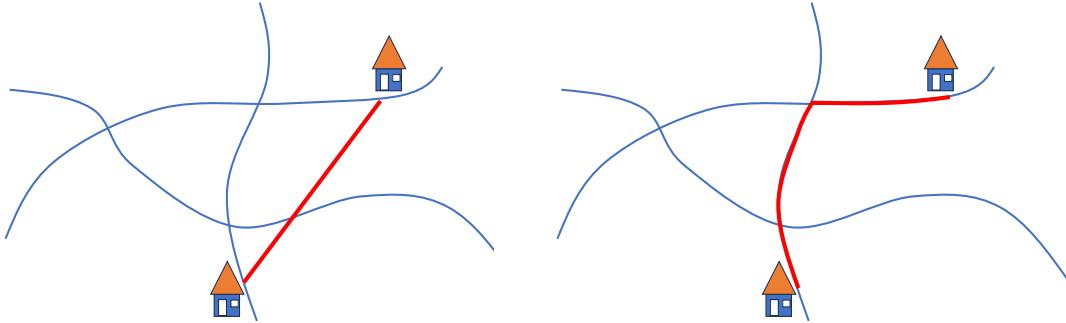


Fig. 12.11 Distance between two locations, Eukledian (left) and based of shortest route (right)

idea behind this is, that we decompose observations Y into a spatially correlated component and independent noise, very much in line with 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

$$Y_j = h(s_j) + \epsilon_j,$$

where $h(\cdot)$ is a smooth component that is estimated from the data.

We can consider $h(s)$ a spatial process with covariance structure $C(\cdot)$ dependent on the distance between two observations. The function $h(s)$ can be obtained through 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, \dots, s_n leading to data $Y(s_1), \dots, Y(s_n)$, or in short y_1, \dots, y_n . We assume that $Y(s)$ has covariance matrix V which is build from the covariance structure $C(|s_i - s_j|)$, that is $V_{ij} = \text{Cov}(Y(s_i), Y(s_j)) = C(|s_i - s_j|)$. We aim now to estimate the value of the process at location s_0 , given the observed values $(y(s_1), \dots, y(s_n))$ or, again, in short $y = (y_1, \dots, y_n)$. Utilizing properties of the multivariate normal distribution we obtain the predictor

$$\hat{Y}(s_0) = wV^{-1}y$$

where

$$\begin{aligned} w &= \text{Cov}(Y(s_0), (Y(s_1), \dots, Y(s_n))) \\ &= (C(|s_0 - s_1|), C(|s_0 - s_2|), \dots, C(|s_0 - s_n|)). \end{aligned}$$

The justification behind the formula is that $(Y(s_0), Y(s_1), \dots, Y(s_n))^\top$ can be considered to be normally distributed with

$$\begin{pmatrix} Y(s_0) \\ Y(s_1) \\ \vdots \\ Y(s_n) \end{pmatrix} \sim N\left(0, \begin{pmatrix} \sigma^2 & w^\top V \\ w & V \end{pmatrix}\right),$$

where $\sigma^2 = C(0) = \text{Var}(Y(s_0))$. Hence,

$$Y(s_0) | y \sim N(wV^{-1}y, \sigma^2 - wV^{-1}w).$$

We can easily extend the model by assuming that the mean value of $Y(s)$ is not zero but depends on some covariates $x(s)$ in a linear form, i.e. $E(Y(s)|x(s)) = x(s)\beta$. Defining $\tilde{Y}(s) = Y(s) - x(s)$ leads to the results above. We visualize the resulting predictions in Figure 12.12. We see that rents in the city center are generally more expensive, which is not surprising.

An alternative approach results by spatial regression. In this case we assume

$$Y(s) = h(s) + \epsilon$$

but now $h(s)$ is considered as smooth function which can be estimated through, for instance, spline smoothing (Wood, 2017). This is in fact using the ideas of Chapter 6 so that we do not repeat the ideas here.

Often, the geolocation of data is not given exactly but the location is categorized to administrative districts like city quarters or counties. In this case one can either approximate the location with the centroid of the district, or one can use the resulting lattice to incorporate the spatial structure. We demonstrate this for the city of Munich. In Figure 12.13, we sketch the lattice structure resulting from the districts and their neighborhood structure. The nodes are the districts, and an edge exists if the districts share a joint 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 $s_i \in \{1, \dots, K\}$ denotes now the district in which apartment i is located. Then Tobler's law states that apartments in one district 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

$$Y(s_i) = u(s_i) + \epsilon,$$

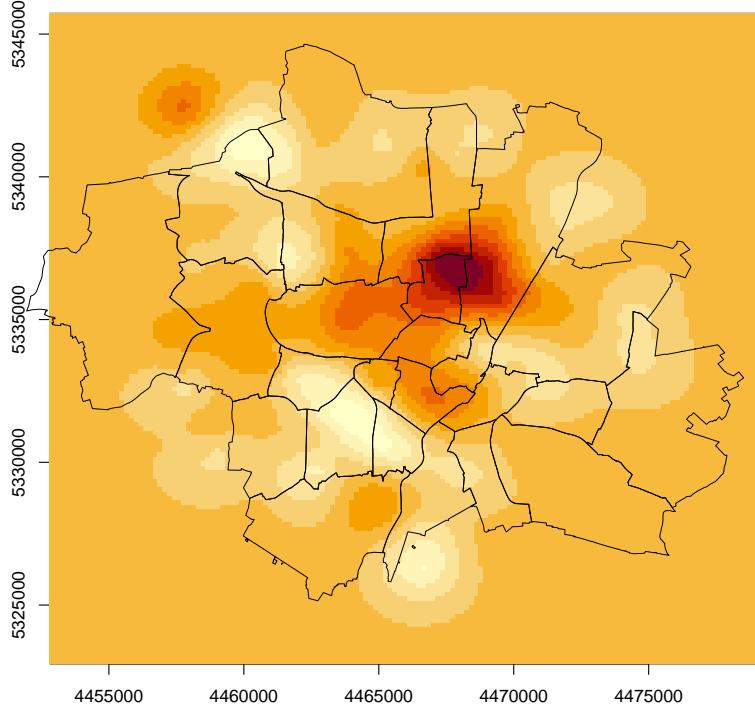


Fig. 12.12 Kriging of rents. Shown is the predicted average rent.

where $u(s)$ is the district effect of district s_i . Given that $s_i \in \{1, \dots, K\}$ we can write the district effects as $u(1), \dots, u(K)$ or in short u_1, \dots, u_K and we assume now a spatial structure for u_1, \dots, u_K . In fact, the neighborhood structure can be used to induce a graphical model for u_1, \dots, u_K in that we postulate

$$u_k \perp\!\!\!\perp u_l | (u_1, \dots, u_k) \setminus \{u_k, u_l\}$$

Shouldn't we, generally, abide by the rule with rvs X and observations x ? if and only if u_k and u_l are not neighbours. The spatial structure is now reduced to a spatial structure on the district level. We exemplify this using the rental data, which is shown in Figure 12.14.

It can be seen that spatially correlated data require special attention and, hence, sophisticated statistical treatment. This goes far beyond the scope of this section. We refer to the following literature for further reading. One of the most central books

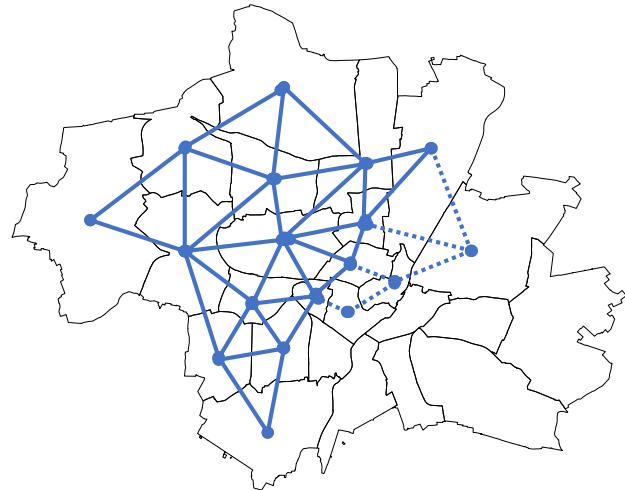


Fig. 12.13 Sketched lattice and neighborhood structure of districts in Munich

in this field is undoubtedly the monograph by Cressie (2015), which first appeared more than 30 years 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 this field, spatial econometrics developed, 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).

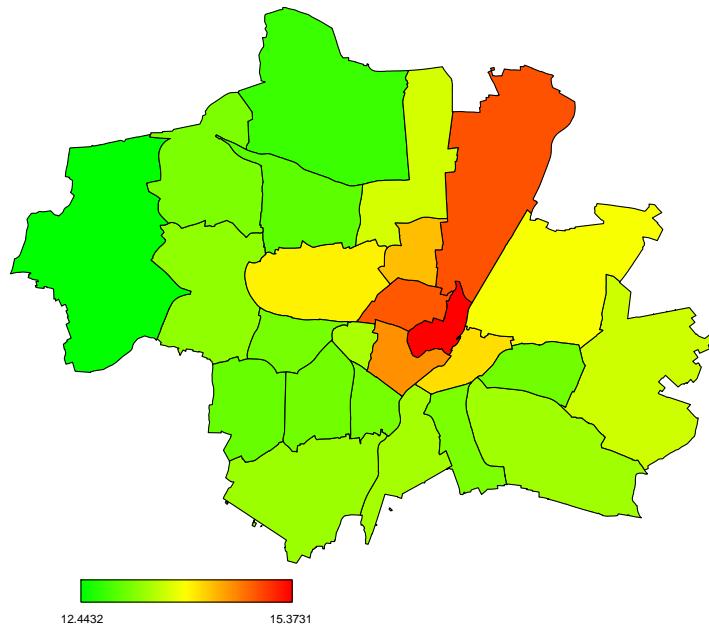


Fig. 12.14 Average rent per district based on spatial lattice data

12.4 Network Data

Quotes like “networks are ubiquitous” (see, e.g., Fienberg, 2012 and Goldenberg et al., 2010) are themselves “ubiquitous” and demonstrate the traction network data analysis has gained over the past decades. Here, networks represent relational structures between entities that are generally not conditionally independent of one another. Contrary to this recent trend, the bulk of statistical research from the last century is limited to the study of attributes of sampled population members under conditional dependence assumptions. 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 clearly affects smoking habits as well.

More abstractly, the question is: How is individual behavior affected by the behavior of other actors in the same network and the network itself? A term coined in the Social Sciences for these two mechanisms are “*selection*” and “*influence*”. Whereas “*selection*” concerns processes in which behavior affects the relations between actors in the network, the “*selection*” mechanism implies that relations between actors affects their behavior. Going further, it is even reasonable that one behavior can affect another behavior, just like social ties depend on one another. A “holistic” treatment of this issue probably calls to mind a multivariate model that jointly studies all social connections and behaviors in some population of interest. Since this treatment would, however, go well beyond the scope of this book, we simplify the situation by dividing our study into a model for the behavior and a model for the network. For the first part, the techniques for multivariate data introduced in Chapter can be employed, while novel methods are needed for the study of the network, which will be the topic of this chapter.

For this endeavor, we are interested in studying two types of factors that can have an effect on the probability of a tie between two population members: first, how one tie affects other ties (*endogenous* factors), and second, how ties are affected by things happening outside of the network (*exogenous* factors). Going back to the example above about smoking and relationships, an endogenous factor would be the tendency to be friends with friends of your friends, while the tendency to have friends that share one’s smoking habits would be an exogenous factor.

Historically, Moreno (1934) was among the first to challenge the assumption of conditional independence when modeling network links. In his study of runaway behavior among students in a girls’ school (depicted in Figure 12.15), he introduced the concept of a “sociogram”. This descriptive method revealed that the girls’ embedding in a social space led to clustered outbreaks, which led to the hypothesis that the girls’ position within a network affects their actions. Suppose we were to apply standard statistical models, such as logistic regression, to the binary indicator of whether a specific girl ran away. Moreno’s theory challenges the conditional independence assumption inherent in the classic regression model, casting doubt on its straightforward application. This observation underscores a core concept of the social network perspective: attributes of individuals in a study do not exist autonomously but rather emerge from structural or relational processes (Wasserman and Faust, 1994). As a result, the central characteristic of networks lies in the interdependence of relations among the studied subjects, making understanding this dependence a crucial goal of network analysis.

Concurrent with this development, logging data generated from interactions, communications, and transactions within networks has led to using networks in commercial contexts and unprecedented sizes. **Make it clear what logging data entails and how it connects to large networks?** Here, the primary focus is not on understanding the dependence but on identifying some latent structure capturing the dependence structure. This latent structure can, e.g., be a decomposition of the network into cohesive subgroups, where we expect many interactions between actors within the same group and fewer interactions between actors of different groups or actor-specific latent positions where the distance between actors in this latent space

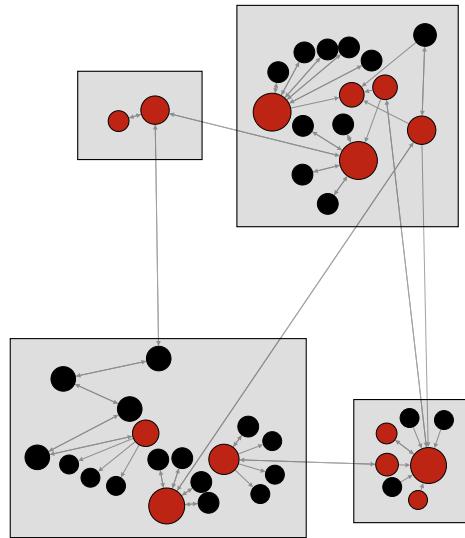


Fig. 12.15 Network indicating the feelings of attraction between pupils at Hudson School for Girls. The color of nodes is red if the respective girl was among the runaways and black otherwise. Additionally, the rectangles represent the cottages in which the girls lived. The size of the nodes is given by the sum of in- and out-degree (**Will be explained later**). This figure is inspired by Figure 1 in Borgatti et al. (2009) and reproduced in Fritz (2022).

governs the probability of them interacting. Community detection implies the former type of latent structure and involves identifying cohesive groups or clusters within a network.

Community detection techniques can identify cohesive subgroups within the studied actors based on shared characteristics, behaviors, or preferences. In this context, cohesive subgroups are subsets of actors that behave in a similar way. In most cases, this entails that more edges are expected within the subgroups than between the subgroups. These segments can then enable tailored marketing strategies, product offerings, and customer experiences to specific community needs. Additionally, community detection can aid in detecting anomalous behavior. Financial institutions analyze transaction networks to detect fraudulent activities, such as clusters of accounts involved in suspicious transactions, which trigger alerts for further investigation. Community detection within organizations sheds light on employee collaboration patterns, allowing businesses to identify cohesive groups, foster cross-functional teamwork, improve knowledge sharing, and enhance productivity. **Maybe add some citations.**

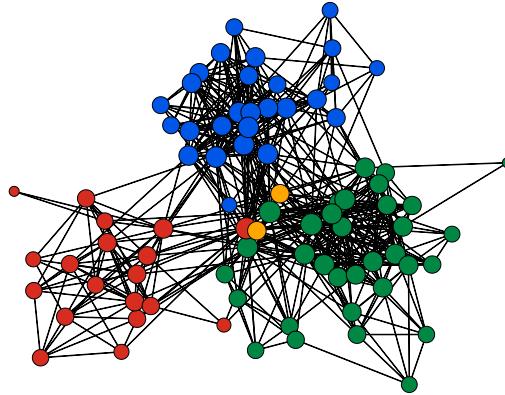


Fig. 12.16 Network of personal friendships of a faculty of a UK university. The color of the nodes represents the affiliation of the academics.

Given this context, this chapter will first provide an abstract representation of networks as mathematical objects called *graphs* in Section 12.4.1. Equipped with this knowledge, Section ?? delves deeper into ways to analyze data descriptively. As the number of actors within a network grows, these descriptive analyses become difficult to interpret. Therefore, we introduce two ways to model a network with the help of a probabilistic model over the set of all possible networks in Section 12.4.2. This stochasticity of graphs differentiates the approaches introduced in this chapter from the deterministic graphs used in Section 11.2 that encode independence relations between random variables. Similar to how we motivated the study of networks in the previous paragraphs, approaches based on implying and testing for a particular dependence structure based on Exponential Random Graph Models are sketched in Section 12.4.2.1, whereas Section 12.4.2.2 exhibits tools to identify subgroups of actors in networks with community detection tools based on Stochastic Block Models.

12.4.1 Representation and Description

Networks encode edges between actors. Since the analysis of network data is of interest in a wide array of disciplines ranging from Statistics, Social Sciences to Physics, the field is increasingly multidisciplinary with equally tailored jargon to each science. To keep the presentation precise and clear, we stick with the terminology of edges and actors. Alternative names one might find in the literature for edges are connections, relations, or ties, and for actors, either nodes, agents, or vertices.

For a running example, we study the friendship network of a university in the UK (Nepusz et al., 2008). The network encompasses 81 actors divided into three groups based on their school affiliation. An edge in this network indicates a personal relationship between two respective actors and is derived from answers in a questionnaire. Arguably, the most straightforward way to visualize such data is to plot the nodes on a two-dimensional plane and indicate connections between nodes by straight lines drawn between respective nodes. The color and size of the nodes in a plot can represent additional categorical or continuous information on the level of each actor. The above-mentioned field of sociometry later evolved into graph drawing and mainly revolves around algorithms that find the most informative ways to place these nodes in two-dimensional space so that interesting structures are identifiable. Since it is beyond the scope of this chapter to detail principles and examples of these algorithms, we refer to Brandes and Sedlmair (2019) for a technical introduction to the field. Still, Figure 12.16 serves as an excellent example of how applying the most common algorithm in that field, namely the Fruchterman-Reingold algorithm (Fruchterman and Reingold, 1991), uncovers many connections between actors with the same affiliation and fewer connections if this is not the case.

While

- Start with the classic representation as graphs
- Connect this part with the Graphical Models part
- How can binary networks be viewed as tabular data?
- Arguably, knowing how to represent network data is of questionable help when network data is observed and needs to be analyzed
- Use drawing algorithms
- Define other descriptive properties such as the degree structure, triangular closure, centrality

12.4.2 Models

12.4.2.1 Dependence Modeling

- Let's take one step back and think about the general purpose of data analysis involving network data.
- What purposes are common?
 1. Link prediction (mention this with references to papers, but I would not add a subsection on it, first, since one can do this with)
 2. Community detection
 3. Dependence modeling
- Spend 1.5 pages detailing how each can be pursued

- End with a connector to functional data (maybe some biological FMRI data application, where FDA is used to derive a network and SNA to understand it?)

12.4.2.2 Community Detection

12.5 Functional Data

Functional data analysis (FDA) uses curves, images, or functions as observations and is closely related to time series analysis and longitudinal analysis for repeated measurements. In the following, we will use curves as an 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 . FDA aims to achieve goals similar to statistical scalar or multivariate data analyses. In contrast to repeated measurements and longitudinal data which often involve only a few repeated observations $Y_{jt}, t = 1, \dots, n_j$ over time on the same subject j , functional data typically involves a large (or, theoretically, an infinite) number of measurements $Y_j(t), t \in \mathcal{T}$ per observation unit j observed on a continuous domain \mathcal{T} . More formally speaking, $Y \in L^2(\mathcal{T})$ is assumed to be the realization of a zero-mean square integrable stochastic process. In practice, as only a finite number of observed values of each function can be stored, functions Y are observed on grid points $t_{i,r} \in \mathcal{T}$, yielding the observations $y_i(t_{i,r}), r = 1, \dots, R_i, i = 1, \dots, n$. 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., $t_{j+1} - t_j = t_{k+1} - t_1 \forall j, k \in \{1, \dots, R - 1\}$, the data can be represented and stored in an $R \times n$ -matrix.

Methods for analyzing longitudinal and functional data traditionally also differ, with functional data treating observed sequences per measurement unit as one observation. This is, in turn, more closely related to the viewpoint of time-series analysis, which typically considers a single observation of a stochastic process on an equidistant time grid from a single observation unit. In other words, if multiple replications of a time series are available, this could be a good indicator to apply FDA.

12.5.1 Analysis and Properties of Functional Data

12.5.1.1 Descriptive Analysis

Functional data can be analyzed in different ways. The most common approach is to plot all functions as curves over their respective observed time points using a line plot (see Figure 12.17 for an example). In contrast to descriptive statistics for scalar data,

other visualization techniques become notably more intricate for functional data. As there is no natural order or obvious definition of outliers for multiple functions, a concept of centrality is required. Possible ways to represent data and identify outliers are, e.g., given in Hyndman and Shang (2010), defining outliers based on magnitudes or shapes, or utilizing functional principal component analysis (described in more detail below).

12.5.1.2 Smoothing Approaches

A possible pre-processing step before further analyzing functional data is to apply pre-smoothing of the curves (Ramsay and Silverman, 2005) to remove observation noise using nonparametric regression or smoothing techniques (cf. Figure 12.17, right plot). Whereas this allows obtaining truly functional observations, it is challenging to account for the uncertainty in estimating the smoothed versions of observations Y_i 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 and 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 and Carroll, 2006). However, instead of subsequently working with the pre-smoothed functions, it 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. Another common method to incorporate some form of denoising as preprocessing of functional data is to (first) use a functional principal component analysis (see, e.g., Goldsmith et al., 2011).

12.5.1.3 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-Loëve expansion, stating that any stochastic process Y can be represented as an infinite series of uncorrelated random variables. 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

$$Y(t) = \mu(t) + \sum_{m=1}^{\infty} \xi_m \phi_m(t), \quad t \in \mathcal{T}. \quad (12.7)$$

In (12.7), μ represents the mean function of the stochastic process Y , the ϕ_m 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

$$\text{Cov}(Y(s), Y(t)) = \sum_{m=1}^{\infty} e_m \phi_m(s) \phi_m(t), \quad (12.8)$$

with eigenvalues e_m , by convention sorted as $e_1 \geq e_2 \geq \dots \geq 0$. As for the 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 Equation 12.7 can be truncated to provide an approximation to Y with a finite number of basis functions. This allows for a sample of independent replications of Y to be represented by a set of finite functions, common to all observations, and individual scores that capture function-specific deviations from the mean.

Next 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 12.18 visualizes the functional principal components for the before-shown curves from Figure 12.17.

12.5.1.4 Variations in Amplitude and Phase

Many of the previously introduced methods assume that different observed functions $y(t)$ and $y'(t)$ can be meaningfully compared for the same time point t and hence differ mainly in amplitude, i.e., the variation along the y -axis. Functions, however, can also differ in behavior along the time domain. For example, natural phenomena might result in a similar functional measurement but delayed in time. This so-called phase variation is an important aspect of FDA and may confound the analysis if not accounted for. If phase variation is present and functions are not *aligned*, this hamper also statements made for specific points in time. Hence, depending on the context, we may either align functions before analysis, which is known as registration or warping, or, e.g., analyze the amplitude and phase variations separately.

The process of aligning the set of functions $y_i, i = 1, \dots, n$ to a general mean μ can be described as discovering appropriate *warping functions* ϱ_i , which create a monotonic mapping of \mathcal{T} onto itself, ensuring optimal alignment of $y_i(\varrho_i(t))$ to $\mu(t)$ for each $t \in \mathcal{T}$. Consequently, ϱ_i encapsulates the varying 'velocities' (phase) across different functions x_i within the range of \mathcal{T} (Ramsay and Silverman, 2005). Traditional methods for registration typically involve identifying specific landmarks (such as peaks, troughs, zero-crossings, or similar derivative values) across functions and interpolating ϱ_i linearly between these landmarks.

More advanced techniques account for the entire function and often seek to minimize the \mathcal{L}^2 distance between a function y_j and an aligned function y_i over ϱ . This is expressed as:

$$\inf_{\varrho} \|y_i \circ \varrho - y_j\|_{\mathcal{L}^2}^2 = \inf_{\varrho} \int_{\mathcal{T}} (y_i(\varrho(t)) - y_j(t))^2 dt. \quad (12.9)$$

Despite its comprehensiveness, this method has issues including a lack of symmetry (warping y_i to y_j is not the same as warping y_j to y_i) and the notorious *pinching problem* (Marron et al., 2015). The latter refers to the potential for the distance in Equation (12.9) to become small or zero, even when y_i and y_j are not warped counterparts, due to a warping function γ selectively contracting or expanding areas of \mathcal{T} based on the similarity between y_i and y_j . To address these concerns, ? suggested employing the *elastic distance*, optimized over warping based on the Fisher-Rao metric. This metric conveniently reduces to the \mathcal{L}^2 distance in Equation (12.9) for optimally aligned square-root-velocity transformed curves. The elastic distance avoids the pinching problem and can be used to align functions or to derive, for example, a mean estimate post-alignment by eliminating all phase information. For the weather example from before, we artificially create a misalignment by shifting one of the three curves by 14 days into the future and another curve by 14 days into the past. We then apply the elastic alignment. The result is given in Figure 12.19.

12.5.2 Functional Regression

In many situations that involve functional data, there is also an interest in using the functions as either covariates, called scalar-on-function regression (SOFR), as response with scalar covariates, which is referred to as function-on-scalar regression (FOSR), or both in a function-on-function regression (FOFR). We first discuss the different model types and then give a brief overview of how to estimate these models.

12.5.2.1 Model Types

Scalar-on-Function Regression One of the most often utilized models in scalar-on-function regression is the *simple functional linear model*, given by

$$Y_i = \alpha + \int_{\mathcal{T}} x_i(t)\beta(t)dt + \epsilon_i, \quad i = 1, \dots, n, \quad (12.10)$$

where Y_i is now a scalar random variable, $x_i(t)$ a given realization of $X_i(t) \in L^2(\mathcal{T})$, β is the *coefficient function* describing the “importance” of the t th time point of $x_i(t)$ on the response. The error term $\epsilon_i \stackrel{iid}{\sim} \mathcal{N}(0, \sigma^2)$ usually follows the same assumptions as in the standard linear model. While the normality assumption is not always necessary, it is particularly relevant when utilizing likelihood-based techniques to estimate unknown model parameters.

Analogous to the standard LM and GLM, the model (12.10) can be extended to incorporate multiple functional predictors, generalized to non-normal responses, or combined with scalar predictors such as scalar linear effects, random effects, or

non-linear additive model terms.

Function-on-Scalar Regression When the response $Y(t), t \in \mathcal{T}$, is functional and all say p covariates are scalar, a simple functional model is the *linear function-on-scalar model*, represented as:

$$Y_i(t) = \alpha(t) + \sum_{j=1}^p x_{ij}\beta_j(t) + \epsilon_i(t), \quad i = 1, \dots, n, \quad t \in \mathcal{T}. \quad (12.11)$$

In this representation, $\alpha(t)$ is the functional intercept, which can be interpreted as the functional average of Y . $\beta_j(t), j = 1, \dots, p$ are (varying-)coefficient functions defining the effect of the covariates on the i th outcome value of Y , while $\epsilon_i(t)$ is the error function derived from a random process with zero mean and covariance function $\Sigma(s, t), s, t \in \mathcal{T}$, e.g., a Gaussian process $\epsilon(t) \sim \mathcal{GP}(0, \Sigma(s, t))$. Again, more complex models can be constructed (see, e.g., Scheipl et al., 2015) and, in particular, generalized to other distribution assumptions and additive model terms (see, e.g., Scheipl et al., 2016).

Function-on-Function Regression If both the response and predictor(s) are functional and expected to be observed over the identical domain \mathcal{T} , both previous approaches (SOFR, FOSR) can be combined by adapting (12.11) for functional covariates:

$$Y_i(t) = \alpha(t) + \sum_{j=1}^p x_{ij}(t)\beta_j(t) + \epsilon_i(t), \quad i = 1, \dots, n, \quad t \in \mathcal{T}. \quad (12.12)$$

The model (12.12) is called the *linear concurrent model* as it provides a concurrent effect for covariates on the response for every time point $t \in \mathcal{T}$. Depending on the relationship between covariates and response, various relaxations of Equation 12.12 exist. As for the SOFR model, we can integrate the functional predictors over the whole time domain if we assume that all time points of the covariates have an influence on every single time point of the response:

$$Y_i(t) = \alpha(t) + \sum_{j=1}^p \int_{\mathcal{S}_j} x_{ij}(s)\beta_j(s, t)ds + \epsilon_i(t), \quad i = 1, \dots, n, \quad t \in \mathcal{T}. \quad (12.13)$$

Here, \mathcal{S}_j symbolizes the domain of the j th functional predictor. Especially when response and covariates share the same time domain \mathcal{T} , it might be important to allow for broader integration limits than in (12.13). One plausible effect in practice is the so-called *historical effect*, which replaces $\int_{\mathcal{S}_j}$ in (12.13) by $\int 0^\top$, but more

general integration limits that depend on t can be constructed (see Brockhaus et al., 2017; Rügamer et al., 2018).

12.5.2.2 Estimation and Uncertainty Quantification

Various procedures exist to estimate the different functional regression models. A widely used semi-parametric approach is basis expansions for the (smooth) coefficient function $\beta(t)$ or $\beta(s, t)$ using, e.g., B-spline functions or the basis obtained by applying FPCA. Combining this with a numerical integration scheme used as an approximation for the involved integrals in the model, the objective function can be written as a (penalized) linear (mixed) model and estimated with commonly used approaches such as OLS, PLS, or PIRLS (see, e.g. Greven and Scheipl, 2017, for details). A similar approach can also be used for models with functional outcomes, where the integration over the response domain is replaced by a weighted sum using numerical integration. If both response and covariates share the same basis, the estimation problem can also be directly transferred to the basis coefficient space by evaluating both sides of the model using the same basis and directly working in the coefficient space (Goldsmith et al., 2011). In addition to these semi-parametric approaches, various non-parametric approaches using, e.g., Kernel-based methods (see Ferraty and Vieu, 2006).

Uncertainty quantification, i.e., tests and confidence intervals for functional regression models can, in many, cases be related to inference procedures in linear or additive models. In SOFR models this, e.g., boils down to testing whether the whole functional predictor $\int x_i(t)\beta(t)dt = 0$, which corresponds to testing if all basis coefficients of a basis representation used to represent the non-linear function are equal to zero.

12.6 Exercises

Exercise 1

- a.
- b.
- c.

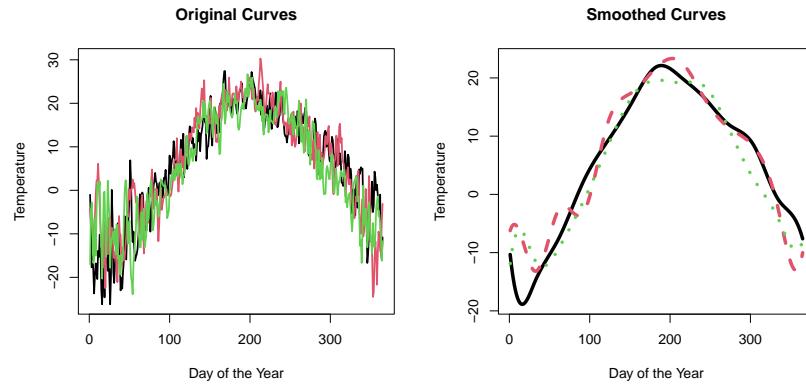


Fig. 12.17 Line plot of original (left) and smoothed functional data showing temperatures in Montreal with different lines corresponding to different years.

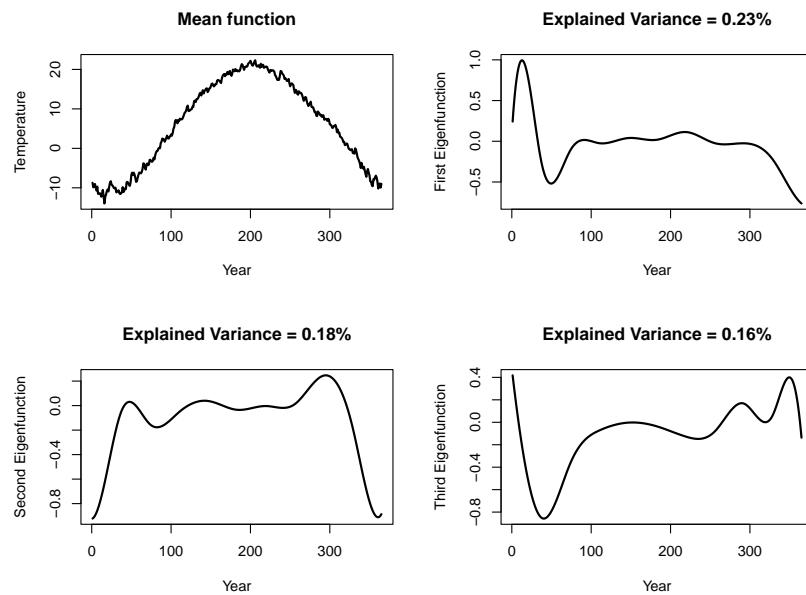


Fig. 12.18 Mean curve and first three functional principal components of the Montreal weather data.

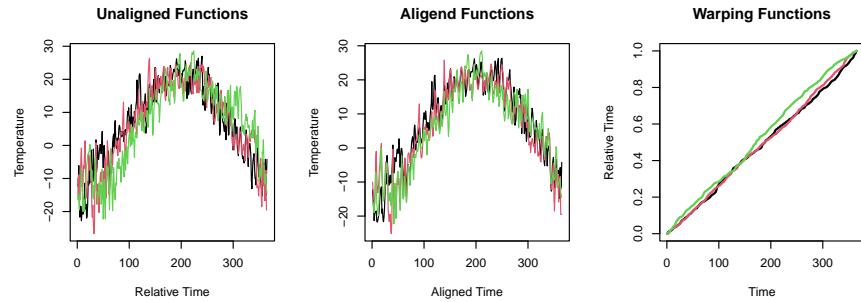


Fig. 12.19 Three curves shifted in time (left) and aligned to their mean (center) using the warping functions that transform the original time to the aligned time (right).

Chapter 13

Latent and Missing Data

- When do missing data occur?
- What are the naive solutions? - ζ Complete-case analysis
- Show example based on simulation where the CCA analysis fails
- Contrast the example with a model-based imputation and innuend to its introduction in the next sections

13.1 Types of Missingness

- Start by defining the classic types of missingness
- But focus should be put on applications where one or the other type is valid or invalid
- Note that this is eventually a untestable assumption from the mathematical perspective, therefore, substantive understanding of what means what and to what extent the assumption is fulfilled is a substantive matter

13.2 Mixture Models

- Bring up the example of throwing a different dice and forgetting which dice was used in which experiment
- Use this example to motivate imputation and, in turn, EM (but maybe not yet refer to it as EM but just model-based imputation)
- maybe include a diagram (see paper by Do and Batzoglou "What is the expectation maximization algorithm")
- This example should be split into a few parts scattered in 12.2

- Use the example to give raise mixture models and define them with some mathematical rigor (maybe start by saying that the mixture components do not have to be Binomial experiments but can be any distribution introduced so far)

13.3 Expectation Maximization

- Continue example from last section and draw a diagram to make thus 10% easy to follow

13.4 Complexity through Latent Structures

- Missing data is not the only application of EM algorithms
- Lead to latent variables that are part of the model but

13.5 Exercises

Exercise 1

- a.
- b.
- c.

Chapter 14

Relating Machine Learning and Statistics

This chapter transitions our focus from traditional statistics to the field of machine learning, emphasizing their strong connection and shared principles. It aims to clarify the relationship between statistical methods and machine learning, showing how they complement and strengthen each other. We start by outlining essential machine learning concepts and approaches, highlighting their foundations in statistical theory. The emphasis is on viewing machine learning algorithms as natural extensions of statistical models that are 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.

14.1 Machine Learning from a Statistics Perspective

In supervised machine learning, the goal is to approximate the data generating process $G(\cdot|x)$ in (2.1) using a machine learning model from the so-called **hypothesis space** (or model space) \mathcal{H} , introduced earlier as *structural model*. Instead of approximating or modeling G with a set of distribution functions, this hypothesis space is usually defined by a set of functions

$$\mathcal{H} = \{h(\cdot; \theta), \theta \in \Theta\}. \quad (14.1)$$

The (machine learning) *model* $h(\cdot; \theta) : \mathcal{X} \rightarrow \Gamma$, with $\Gamma \subseteq \mathbb{R}^v, v \in \mathbb{N}$ approximates the unknown mapping from the covariate or feature space \mathcal{X} to the a space of predictions Γ . The parameter (vector) θ defines a single instance (a machine learning model) of \mathcal{H} . For example, if $h(x; \theta) = x^\top \theta$, \mathcal{H} is the space of linear regression models with covariates (features) x and coefficients $\theta \in \Theta = \mathbb{R}^P$. In many of the classic machine learning applications, this hypothesis space is usually not paired with a distribution assumption or model \mathcal{F} , but 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 \mathcal{F} .

14.1.1 Empirical Risk Minimization

Learning in machine learning is the combination of a hypothesis space \mathcal{H} , a *risk* \mathcal{R} and an optimization routine. We first define a loss function

$$\ell : \mathcal{Y} \times \mathcal{H} \rightarrow \mathbb{R}; (y, h(x; \theta)) \mapsto \ell(y, h(x; \theta)) \quad (14.2)$$

which we want to minimize with respect to $h \in \mathcal{H}$. The loss is usually given by the application of interest, e.g., to find a model h such that the squared deviations of h and y are minimized. Acknowledging that the true data generating process $G(\cdot|x)$ 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*

$$\mathcal{R}(h) := \int \ell(y, h(x; \theta)) dG(y|x). \quad (14.3)$$

As (14.3) involves the unknown true data generating process, the theoretical risk is approximated by its empirical analog, the *empirical risk*, which for a given data set of i.i.d. observations $(x_i, y_i)_{i=1, \dots, n}$ is

$$\mathcal{R}_{emp}(h) := \sum_{i=1}^n \ell(y_i, h(x_i; \theta)). \quad (14.4)$$

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 approximating θ_0 via

$$\hat{\theta}_{ERM} = \arg \min_{\theta \in \Theta} \mathcal{R}_{emp}(h(\cdot; \theta)). \quad (14.5)$$

Putting (14.4) and (14.5) together, we can directly see the relationship between *estimating* the parameter θ using maximum likelihood estimation and *learning* the parameter θ using empirical risk minimization. If we choose ℓ to be the negative log-likelihood of a parametric distribution specified through $\gamma \in \Gamma$, $\hat{\theta}_{ERM} = \hat{\theta}_{ML}$. In this case, the model h (implicitly) describes the parametrized transformation from x to the distribution parameters γ of some distribution F . For example, in a regression

setting with \mathcal{H} being the class of linear models, one typically defines the loss to 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 with a Bernoulli distribution assumption, minimizing the mean squared error is equivalent to maximum likelihood estimation of the mean of a Gaussian distribution, or 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. Note that also many examples exist where there is no direct correspondence to a distribution assumption, e.g., the Hinge loss (Cortes and Vapnik, 1995). The rationale, in this case, is not to directly approximate the data generating process G but to optimize a pre-defined 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).

Classification 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, h is usually called a *classifier*. A further distinction is made if h returns a probability for every possible value in \mathcal{Y} , in which case h is called a probabilistic classifier or, if a score for every value in \mathcal{Y} is returned, h is called a scoring function or scoring classifier. A logistic regression, for example, is a probabilistic classifier, returning probabilities $\pi = \mathbb{P}(Y = 1|x)$ and $1 - \pi = \mathbb{P}(Y = 0|x)$ for $\mathcal{Y} = \{0, 1\}$. The additive predictor of logistic regression, in turn, is a scoring function producing a score (the logits) that can be normalized into probabilities.

14.2 Regularization, Penalization and Bayes

One of the most crucial parts in machine learning is their regularization. Being often highly overparametrized (cf. Section 15.3.2) and expressive, there is a great risk of overfitting the given dataset as described in more detail in the following subsection.

14.2.1 Regularized Risk Minimization

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 (14.4) to the so-called regularized risk

$$\mathcal{R}_{reg}(h) := \sum_{i=1}^n \ell(y_i, h(x_i; \theta)) + \lambda \cdot \text{pen}(\theta), \quad (14.6)$$

where $\text{pen} : \Theta \rightarrow \mathbb{R}^+$ 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. The explicit form of the penalty depends on the model class. Information criteria introduced in Section 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 -norm¹ counts the number of non-zero elements in θ . If we choose the empirical risk to be twice the negative log-likelihood and $\lambda = 2$, we can then consider the optimization of the AIC as a way of performing regularized risk minimization.

14.2.2 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 Section 7, we could ask if it is possible to directly find the AIC-optimal model, i.e.,

$$\hat{\theta} = \arg \min_{\theta \in \Theta} \mathcal{R}_{emp}(h(\cdot; \theta)) + \|\theta\|_0. \quad (14.7)$$

This problem — optimizing a model with the so-called l_0 -norm 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 non-zero entries, the l_1 -norm is defined as

$$\text{pen}(\theta) = \|\theta\|_1 = \sum_{j=1}^p |\theta_j|, \quad (14.8)$$

i.e., sums up the absolute values of all entries in θ . This penalty is also known as the lasso penalty (Tibshirani, 1996) and can be solved with 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 much more common in statistics and optimization as it requires specialized optimization routines. 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.

¹ Technically, it is not an actual norm as it violates the absolute homogeneity property.

Another common penalty in both machine learning and statistics is the L_2 - or ridge penalty:

$$\text{pen}(\theta) = \|\theta\|_2^2 = \sum_{j=1}^p \theta_j^2, \quad (14.9)$$

which, contrary to the name, is a squared l_2 -norm penalty. As can be seen in Figure XXX, the ridge penalty cannot enforce sparsity but instead regularizes the model by shrinking the coefficients toward zero. For linear models and the L_2 -loss, for example, the regularized risk minimizer $\hat{\theta}_{RRM}$ has the closed form solution

$$\hat{\theta}_{RRM} = (X^\top X + \lambda I_p)^{-1} X^\top y. \quad (14.10)$$

For models with no explicit analytical solution, a common strategy in machine learning is to perform gradient descent. The gradient of a regularized risk objective with ridge penalty (divided by 2 for convenience) is

$$\frac{\partial}{\partial \theta} \{\mathcal{R}_{reg}(h(\cdot; \theta))\} = \frac{\partial}{\partial \theta} \left\{ \sum_{i=1}^n \ell(y_i, h(x_i; \theta)) + \frac{\lambda}{2} \|\theta\|_2^2 \right\} = \frac{\partial}{\partial \theta} \{\mathcal{R}_{emp}(h(\cdot; \theta))\} + \lambda \theta. \quad (14.11)$$

The gradient descent updates the model via

$$\theta^{[\text{new}]} = \theta^{[\text{old}]} - \nu \frac{\partial}{\partial \theta} \{\mathcal{R}_{reg}(h(\cdot; \theta^{[\text{old}]})\}) \quad (14.12)$$

with learning rate $\nu \in (0, 1]$. Expanding (14.12) using (14.11), we see that the ridge penalty induces a *decay* of the old parameter value defined by λ :

$$\begin{aligned} \theta^{[\text{new}]} &= \theta^{[\text{old}]} - \nu \frac{\partial}{\partial \theta} \{\mathcal{R}_{reg}(h(\cdot; \theta^{[\text{old}]})\}) \\ &= \theta^{[\text{old}]} - \nu \left(\frac{\partial}{\partial \theta} \{\mathcal{R}_{emp}(h(\cdot; \theta^{[\text{old}]})\}) + \lambda \theta^{[\text{old}]} \right) \\ &= \theta^{[\text{old}]} (1 - \nu \lambda) - \nu \frac{\partial}{\partial \theta} \{\mathcal{R}_{emp}(h(\cdot; \theta^{[\text{old}]})\}). \end{aligned} \quad (14.13)$$

In machine and particularly deep learning, it is common to apply *weight decay* during optimization — in other words to impose a ridge penalty, but not explicitly through a regularized objective as in (14.9) but instead implicitly by altering the optimization routine as in (14.13).

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):

$$\lambda \approx \frac{1}{\nu T}. \quad (14.14)$$

In other words, 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.

14.2.3 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:

$$P(\theta|X, y) \propto P(y|X, \theta)P(\theta), \quad (14.15)$$

where $P(\theta|X, y)$ is the posterior distribution, $P(y|X, \theta)$ 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 :

$$\theta \sim \mathcal{N}(0, \tau^2 I). \quad (14.16)$$

This prior expresses 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

$$P(y|X, \theta) \propto \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - x_i^\top \theta)^2\right). \quad (14.17)$$

The posterior distribution is then given by

$$P(\theta|X, y) \propto \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - x_i^\top \theta)^2\right) \exp\left(-\frac{1}{2\tau^2} \|\theta\|_2^2\right). \quad (14.18)$$

Taking the negative log of the posterior results in the following objective function:

$$-\log P(\theta|X, y) \propto \sum_{i=1}^n (y_i - x_i^\top \theta)^2 + \frac{\sigma^2}{\tau^2} \|\theta\|_2^2. \quad (14.19)$$

Minimizing this negative log-posterior is hence equivalent to minimizing the regularized risk function with a ridge penalty:

$$\mathcal{R}_{reg}(h) = \sum_{i=1}^n (y_i - x_i^\top \theta)^2 + \lambda \|\theta\|_2^2, \quad (14.20)$$

where $\lambda = \frac{\sigma^2}{\tau^2}$.

The equivalence between minimizing 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.

14.3 Uncertainty in Machine Learning

While there is still a debate about whether there is a generally valid categorization of uncertainties in machine learning, it cannot be denied that there are (at least) three areas of uncertainty quantification (UQ) in machine learning that practitioners are interested in: quantifying the uncertainty of the model's prediction, quantifying the uncertainty in the model's parameters, and quantifying the uncertainty in the data generating process. We will cover these three topics in the following by describing the goal of these UQ approaches and providing some basic concepts on how UQ methods for these three types work in practice.

14.3.1 Prediction Intervals

Bootstrap, Resampling in general, CV, Conformal

14.3.2 Parameter Uncertainty**14.3.3 Distributional Learning****14.4 Exercises****Exercise 1**

- a.
- b.
- c.

Chapter 15

TAKE OUTS

15.1 Uncertainty and Prediction

The above discussion refers to finding a good algorithm, or call it estimate, based on the data y_1, \dots, y_n . It is important to reflect, though, that even with massive data aleatoric uncertainty may remain. This becomes important when we talk about prediction. To see this, we include again input variables in the data generating process as given in (2.1). This process is approximated through the statistical model $\mathcal{F} = \{F(\cdot; \theta = h(x)), \theta \in \Theta, x \in X\}$ where X refers to some set in which the input variables are observed and recorded and function $h(\cdot)$ allows to modify the parameter θ based on x . This additional function $h(\cdot)$ typically depends on additional parameters, which we will explain in more depth in Chapter 6. Let now $\hat{h}(\cdot)$ be the trained or estimated function which gives for a some value x_0 the parameter estimate $\hat{\theta} = \hat{h}(x_0)$. For value x_0 we might want to predict the outcome Y . We could for instance take the mean value as predicted value, i.e.

$$\hat{Y}|x = \int y dF(y|\hat{\theta}) \quad (15.1)$$

or for discrete valued outcomes we could use the mode

$$\hat{Y}|x = \arg \max_{y \in \mathcal{Y}} \{P(Y = y|\hat{\theta})\} \quad (15.2)$$

where, as introduced in Chapter (??), we have $P(Y = y; \cdot) = F(y; \cdot) - F(\hat{y}; \cdot)$ and \mathcal{Y} as the set of possible realizations of Y . Considering the image classification task from above we have \mathcal{Y} as set of animals, numerically coded with 1, 2, etc. and \hat{Y} as animal category with the highest probability. In the example shown in Figure 2.1 it becomes clear, that \hat{Y} for given x is not equal to Y , which means we can never make a 100% correct prediction if there is aleatoric uncertainty. We can also

call this uncertainty the inevitable prediction uncertainty and it shows that in every situation in which prediction is applied, we will observe wrong predictions simply due to the randomness of Y shown in (2.1). It is therefore important to realize that aleatoric uncertainty is the main driver of prediction uncertainty and remains even with infinite data bases. Prediction uncertainty is thereby often measured in squared losses, i.e.

$$E((Y - \hat{Y})^2) = \int (y - \hat{Y})^2 dG(y). \quad (15.3)$$

We will show later that this squared loss has some interesting properties. First, it corresponds to assuming normally distributed data and secondly, the formula above decomposes to what will be called (squared) bias-variance-trade-off. Though this is treated in a later chapter it is important at this point to realize that coping with uncertainty is not restricted to dealing with data but also with estimating new values, i.e. prediction.

15.2 Squared Loss

to which we will now come back. We look at prediction models where we aim to model the data generating process defined in (2.1) as aleatoric uncertainty, that is

$$Y \sim G(\cdot|x). \quad (15.4)$$

We aim to approximate the unknown process $G(\cdot|x)$ through some parametric model class

$$Y|x \sim F(\cdot|x; \theta) \in \mathcal{F}, \quad (15.5)$$

where θ depends on x through some function $h(\cdot)$ which itself might have parameters. To be specific we set $\theta = h(x; \gamma)$ with $\gamma \in \Gamma$ as some parameters. The specification of the model class thereby needs two decisions. First, for given input variables x we need to choose a suitable distribution model \mathcal{F} for the outcome Y . And secondly, we need to determine how the input variables x influence the model parameter θ , that is how function $h(x; \gamma)$ is chosen. The latter part distinguishes the different prediction models which will be more formally discuss in Chapter 6. We here focus on the choice of an appropriate probability model \mathcal{F} . This should incorporate a suitable quantification of the aleatoric uncertainty.

If y is binary taking values 0 and 1, there is no selection problem since the only distribution for binary data is the binomial distribution. Hence in this case one

models the probability $P(Y = 1|x; \theta)$. If Y takes more than two outcomes one may use a multinomial distribution or a Poisson distribution, where the latter is quite useful for count data. If Y is continuous the set of possible distributions is in principle unlimited, often however one works with a normal distribution. Assuming that the variance does not depend on x this leads to the log likelihood function

$$l(\theta) = -\frac{1}{2} \sum_{i=1}^N (y_i - \mu(x_i))^2 \quad (15.6)$$

where $\mu(x_i)$ is a model that relates input variables x to the mean value μ . Using the notation from above we can set $\mu(x_i) = h(x_i; \gamma)$. Maximizing (3.5) appears to be equivalent to minimizing the squared loss function

$$\text{loss}(\mu) = \sum_{i=1}^N (y_i - \mu(x_i))^2,$$

which is a commonly used loss function in machine learning models. The squared loss aims to minimize the squared error between an observation y_i and its prediction $\mu(x_i)$. In other words, using a squared loss is equivalent to assuming the response variable Y to be normally distributed.

15.3 overparameterized models

To illustrate the idea we turn our attention to a straightforward example. We take a small sample with just a single value $y_1 = 1$ and assume that $Y_1 \sim N(\theta, \sigma^2 = 1)$. The resultant log likelihood is displayed in the upper row of Figure 15.1. A notable observation is that the likelihood is flat and that the maximum is not exposed. This outcome is to be expected, stemming from the small sample size, or more precisely, the ratio of parameter count to observation count. In this instance, we are confronted with a single observation relative to a single parameter. While this may appear somewhat implausible, it is crucial to acknowledge that modern machine learning, particularly within the realm of deep learning, frequently involves training models endowed with a greater number of parameters than observations. Consequently, the traditional constraint of a low parameter-to-observation ratio, commonplace in statistics, is not applicable to today's machine learning endeavors. However, to address this situation, regularization techniques become imperative, and Bayesian methodologies can serve as a viable approach.

Looking at the lower plot of Figure 15.1, we present the log posterior under the assumption of a standard normal distribution as a prior for the parameter θ . With this shift, the optimization challenge becomes notably convex, thus facilitating

numerical solutions. We shall revisit this aspect in Chapter 14.2. To conclude for now, we deduce that assigning a specific distribution to the parameter θ yields two effects. First, it allows us to look at the posterior distribution, which is the classical Bayesian step and extensively discussed in this book. Second, when our primary concern is the posterior mode, a Bayesian model contributes to numerical stability, especially when dealing with high-dimensional models.

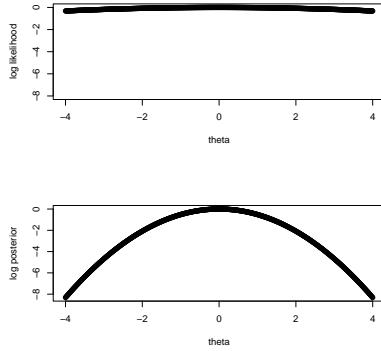


Fig. 15.1 Log likelihood function for a sample of size 1 (upper plot). Resulting log posterior density in case of a normal prior distribution

15.3.1 Overfitting and Generalization

[Deep Learning, Generalization]

15.3.2 Overparameterized Models and Implicit Regularization

New developments in the field of Deep Learning have gone new routes by heavily overparameterizing the model, see e.g. We discussed high dimensional models already in Chapter ?? and revisit this setup here again. Let $\mathcal{F} = \{F(\cdot; \theta), \theta \in \Theta\}$ be the model. Unlike above, we now assume that Θ is high dimensional, and in fact, we even allow the dimension of Θ to be larger than n . Let K be the dimension of the parameter and we explicitly assume $K > n$. We remain with the definition that the optimal parameter θ_0 results through minimizing the Kullback-Leibler divergence which gives the implicit definition

$$E\left(\frac{\partial f(y; \theta_0)}{\partial \theta}\right) = 0. \quad (15.7)$$

For maximum likelihood estimation, we replace the expectation in (15.7) through its empirical version leading to the estimating equation

$$\sum_{i=1}^n \frac{\partial f(y_i; \theta)}{\partial \theta} = 0. \quad (15.8)$$

Taking the second order derivative gives

$$J(\theta) = \sum_{i=1}^n \frac{\partial^2 f(y_i; \theta)}{\partial \theta \partial \theta^\top}, \quad (15.9)$$

which results as a sum with n summands, each being a $K \times K$ matrix which are not guaranteed to have full rank. In this case, the rank of $J(\theta)$ grows with n , but if $K > n$ it is not guaranteed that $J(\theta)$ has full rank. This, in turn, means that the optimization problem of finding the closest model is not convex, and in fact, it allows for multiple solutions. We want to demonstrate this with a simple regression model. As before let x_1, \dots, x_K be a high dimensional set of covariates leading to the model

$$Y_i = x_{i1}\theta_1 + x_{i2}\theta_2 + \dots + x_{iK}\theta_K + \epsilon_i, \quad (15.10)$$

where the residual is assumed to be normal and x_{ij} is the entry for the j -th covariate in the i -th observation. Following the results from Section ?? and making use of matrix notation we can rewrite the model to

$$Y = X\theta + \epsilon, \quad (15.11)$$

leading to the estimate

$$\hat{\theta} = (X^\top X)^{-1} X^\top Y. \quad (15.12)$$

Note that X has dimension $n \times K$, but if K is larger than n , as assumed here, the matrix $X^\top X$ is not invertible in (15.12). Hence, we need to modify the estimation and impose some additional structure on the parameters. We motivated this already in ???, but here we are more specific in the context of overparameterized models. We also refer to Hastie et al. (2022) for more technical details on this setup.

If $K > n$ we need to replace the inverse in (15.12). We can use the generalized inverse, or we can simply add a ridge penalty. That is we get the estimate through

$$\hat{\theta} = (X^\top X + \lambda I_K)^{-1} X^\top Y. \quad (15.13)$$

where λ is some small number, which can even be estimated data driven. The estimate can in fact be motivated through Bayesian arguments which we want to motivate. We assume the prior

$$\theta \sim N(0, \rho^2 I_K)$$

and given θ we get Y through

$$Y|\theta \sim N(X\theta, \sigma^2 I_n).$$

This yields the log posterior for θ (assuming σ^2 to be known)

$$\log p(\theta|y) = -\frac{1}{2\sigma^2} (y - X\theta)^\top (y - X\theta) - \frac{1}{2\rho} \theta^\top \theta + const.$$

Differentiation with respect to θ leads to

$$\frac{\partial \log p(\theta|y)}{\partial \theta} = -\frac{1}{\sigma^2} X^\top (y - X\theta) - \frac{1}{\rho} \theta$$

which is solved by

$$\hat{\theta} = (X^\top X + \frac{\sigma^2}{\rho^2} I_K)^{-1} X^\top y. \quad (15.14)$$

We can then calculate the Kullback-Leibler divergence $KL(G(\cdot), F(\cdot, \hat{\theta}))$ or approximate this with cross-validation. The performance of the $KL(G(\cdot), F(\cdot, \hat{\theta}))$ is exemplified with different dimensions of θ . To do so we take the above model setup by now we explicitly simulate $\theta_k \sim N(0, 1)$ and order the components such that $|\theta_k| \geq |\theta_{k+1}|$. We set $\rho^2 = 0.5$ and calculate the Kullback-Leibler divergence using the exact simulation model. We take $n = 100$ as above but now we set $K = 200$ so that we have the double number of parameters compared to the number of data points. We give the corresponding (log) Kullback-Leiber divergence in Figure 15.2 and provide a smooth approximation for better visual impression. It is important to recognize that the curve has two minima, one for $K < n$ and a second one for $K > n$. The latter occurs since the number of parameters used in the simulation exceeds the sample size and the penalty induced by (15.14) in fact allows to obtain estimates through a Bayesian formulation of the model. The open question is of course how to optimally choose the prior and hence the penalty parameter. We do not want to

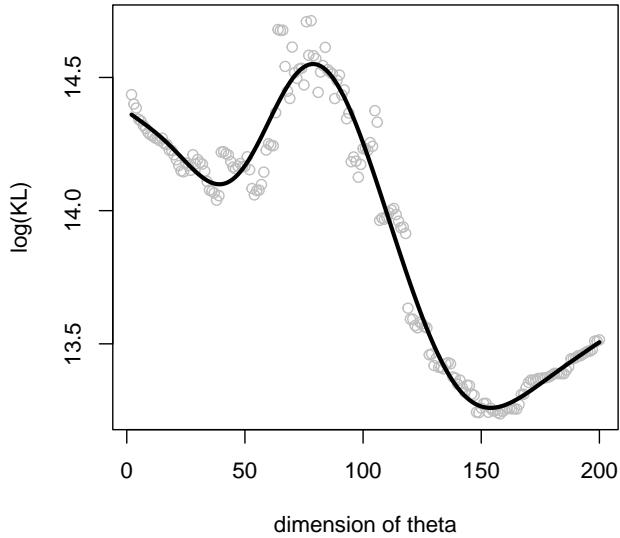


Fig. 15.2 Performance of Kullback-Leibler Divergence in an overparameterized model with ridge penalty estimation. Sample size in $n = 100$.

go further into details here but emphasize that overparameterized models combined with reasonable priors open the door for models and analyses beyond the classical statistical paradigm $K < n$.

15.3.3 ABC Extension

One may further develop the idea toward a regression or more generally a prediction setting. Let therefore $t(y) = (t_1(y), \dots, t_p(y))$ denote a p -dimensional vector of statistics, calculated from the data. Let further $t(Y^*)$ be the corresponding value for the simulated values. Apparently we have the causal chain

$$\theta^* \rightarrow t(Y^*),$$

which we now revert using a regression setup. To be specific, we assume

$$\theta^* = m(t(Y^*)) + \epsilon \quad (15.15)$$

for some residual ϵ . Hence, we revert the causal chain to

$$t(Y^*) \rightarrow \theta^*.$$

While this sounds very implausible, the approach does make sense. We refer to Beaumont et al. (2002), who proposed the idea based on regression models or Blum and François (2010), who used neural networks. Note that we can generate as many simulations as necessary to obtain a solid data base for fitting model (15.15). Assuming *i.i.d.* data we obtain for $i = 1, \dots, N$

$$\theta_{(i)}^* = m\left(t\left(Y_{(i)}^*\right)\right) + \epsilon_{(i)},$$

where the index (i) here refers to the i -th simulation setup. To estimate function $m(\cdot)$ we can use the full machinery. For simplicity we present the idea as linear regression, that is we replace $m(\cdot)$ by

$$m(t(Y^*)) = \beta_0 + t(y)\beta.$$

Example 21 We continue with the above example where we simulated $N = 10000$ $Y_{(j)}^*$ for $j = 1, \dots, 10000$. As statistics we use

$$X = (\bar{x}_0, \bar{x}_1, \bar{x}_2, \dots, \bar{x}_{10})$$

as before and define with $X_{(j)}$ the resulting values for the j -th simulation. We use for linear regression

$$\theta_{(j)}^* = \beta_0 + x_{(j)}\beta_x + \epsilon_{(j)}$$

and estimate coefficients β based on the simulations, leading to estimates $\hat{\beta}_0$ and $\hat{\beta}_x$. For the observed data we now predict

$$\hat{\theta} = \hat{\beta}_0 + x\hat{\beta}_x$$

with x as observed vector of statistics. Based on the regression model we can also build upon the normality of ϵ to construct a credibility region CR, such that

$$P(\theta \in \text{CR}(\hat{\theta})) \geq 1 - \alpha.$$

We here prefer to directly use the normality assumption and plot the resulting isolines. We visualize this in Figure 15.4. Note that the linear model was fitted for the unconstrained parameters $\eta = \text{logit}(\pi)$ and $\delta = \log(\lambda)$ while in Figure 15.4 we

rescale this to π and λ . We find that the approximate approach provides reasonable insight in this example. \triangleright

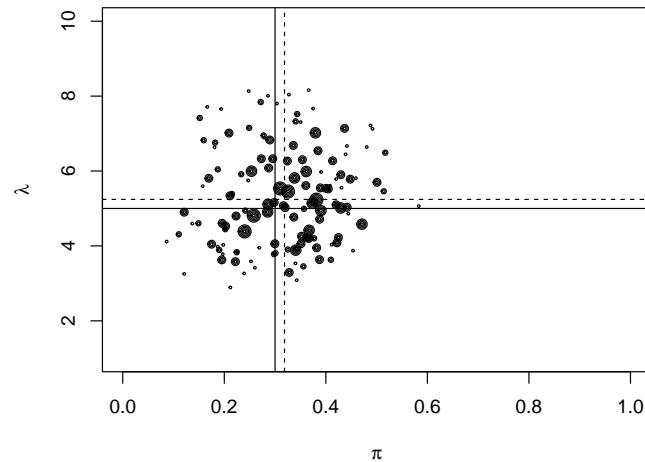


Fig. 15.3 Approximate Bayesian Computation. Simulated parameters π and λ , weighted by the reciprocal distance to the observed data

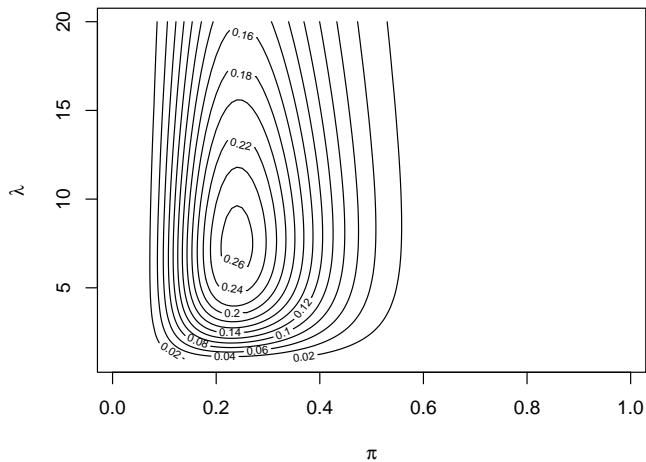


Fig. 15.4 Approximate Bayesian Computation. Regression based approach and predicted value using $N = 10000$ simulated values

Appendix A

Additional results, notions and technical details

Use the template *appendix.tex* together with the Springer document class SVMono (monograph-type books) or SVMult (edited books) to style appendix of your book.

A.1 Background: Measurement- and Probability Theory

Definition A.1 (Random Variable) If $(\Omega_1, \mathcal{F}_1, \mathbb{P})$ is a probability space and $(\Omega_2, \mathcal{F}_2)$ is a measurable space, then a \mathcal{F}_1 - \mathcal{F}_2 measurable mapping

$$X : \Omega_1 \rightarrow \Omega_2$$

is called a random variable (RV).

Definition A.2 (Image Measure) If $(\Omega_1, \mathcal{F}_1, \mathbb{P})$ is a measure space and $(\Omega_2, \mathcal{F}_2)$ is a measurable space and $f : \Omega_1 \rightarrow \Omega_2$ is measurable, then

$$\mu_f : \mathcal{F}_2 \rightarrow [0, \infty[$$

$$\mu_f(B) := \mu(f^{-1}(B))$$

is called the image measure μ_f of μ under f , and $(\Omega_2, \mathcal{F}_2, \mu_f)$ is a measure space.

Definition A.3 (Distribution of a Random Variable) $(\Omega_1, \mathfrak{F}_1, \mathbb{P})$ is a probability space and $(\Omega_2, \mathfrak{F}_2)$ is a measurable space, and $X : \Omega_1 \rightarrow \Omega_2$ is a random variable, then the image measure \mathbb{P}_X of \mathbb{P} under X is called the distribution of X .

In the following chapters, we will usually assume that we are dealing with measurable functions and can thus use Lebesgue and Riemann integral as interchangeable terms. More formally, we have the following:

Definition A.4 (Lebesgue Integral) Let λ be the Lebesgue measure, $t : \mathbb{R} \rightarrow \mathbb{R}$ be (λ) -measurable and t be Riemann-integrable on $[a, b] \subseteq \mathbb{R}$. Then t is Lebesgue-integrable and it holds that

$$\int_{[a,b]} t d\lambda = \int_a^b t(x) dx.$$

In particular, we will use the following relationship frequently:

Definition A.5 (This should be a theorem?) Let $X : \Omega \rightarrow \mathbb{R}^n$ be an n-dimensional real random variable and $t : \mathbb{R}^n \rightarrow \mathbb{R}$ a measurable function, such that $t \circ X : \Omega \rightarrow \mathbb{R}$ is integrable. Then it holds that

$$\int t d\mathbb{P}_X = \int t \circ X d\mathbb{P}.$$

Definition A.6 (Expected Value) Let $X : \Omega \rightarrow \mathbb{R}$ be a quasi-integrable real random variable, then

$$E(X) := \int X d\mathbb{P} = \int x d\mathbb{P}_X(x)$$

is the expected value of X .

In particular, due to Theorem A.5, it holds for a measurable function t and distribution \mathbb{P}_Y of random variable Y that

$$\int t(y) d\mathbb{P}_Y(y) = E(t(Y)). \quad (\text{A.1})$$

Definition A.7 Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $X_i : \Omega \rightarrow \mathbb{R}$, $i = 1, 2, \dots, n$.

Then the sequence $(X_n)_{n \in \mathbb{N}}$ converges in probability to $X : \Omega \rightarrow \mathbb{R}$, if

$$\mathbb{P}(|X_n - X| > \varepsilon) \rightarrow 0 \text{ as } n \rightarrow \infty \text{ for all } \varepsilon > 0;$$

We write $X_n \xrightarrow{\mathbb{P}} X$ and also speak of stochastic convergence.

Definition A.8 (Weak Law of Large Numbers) Let X_1, \dots, X_n be independent and identically distributed with $E(X_i) = \mu$ and $\text{Var}(X_i) = \sigma^2 < \infty$ for $i = 1, \dots, n$. Then it holds that

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

In other words, the sample mean \bar{X}_n converges in probability to the expected value μ .

Definition A.9 (Distribution Function) If $\mathbb{P} : \mathcal{B} \rightarrow [0, 1]$ is a probability measure on \mathbb{R} , then

$$F_{\mathbb{P}} : \mathbb{R} \rightarrow [0, 1], x \mapsto F_{\mathbb{P}}(x) := \mathbb{P}((-\infty, x])$$

is called the distribution function of \mathbb{P} .

Definition A.10 (Convergence in Distribution) Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $X_i : \Omega \rightarrow \mathbb{R}$ random variables with distribution functions F_i , $i = 1, 2, \dots, n$. Then the sequence $(X_n)_{n \in \mathbb{N}}$ converges in distribution to $X : \Omega \rightarrow \mathbb{R}$ with distribution function F , if

$$\lim_{n \rightarrow \infty} F_n(x) = F(x)$$

for all x where F is continuous. We write $X_n \xrightarrow{D} X$.

Fig. A.1 Please write your figure caption here

A.1.1 Subsection Heading

Instead of simply listing headings of different levels we recommend to let every heading be followed by at least a short passage of text. Furtheron please use the L^AT_EX automatism for all your cross-references and citations as has already been described in Sect. A.1.

For multiline equations we recommend to use the `eqnarray` environment.

$$\begin{aligned} \mathbf{a} \times \mathbf{b} &= \mathbf{c} \\ \mathbf{a} \times \mathbf{b} &= \mathbf{c} \end{aligned} \tag{A.2}$$

A.1.1.1 Subsubsection Heading

Instead of simply listing headings of different levels we recommend to let every heading be followed by at least a short passage of text. Furtheron please use the L^AT_EX automatism for all your cross-references and citations as has already been described in Sect. A.1.1.

Please note that the first line of text that follows a heading is not indented, whereas the first lines of all subsequent paragraphs are.

Table A.1 Please write your table caption here

Classes	Subclass	Length	Action Mechanism
Translation	mRNA ^a	22 (19–25)	Translation repression, mRNA cleavage
Translation	mRNA cleavage	21	mRNA cleavage
Translation	mRNA	21–22	mRNA cleavage
Translation	mRNA	24–26	Histone and DNA Modification

^a Table foot note (with superscript)

GK Hier der Rest von Kapitel 2

This book copes with uncertainty and how to grasp it and how to quantify it. Probability theory is a very helpful tool in this respect, which is where we start. We emphasize, however, that we do not aim to provide a gentle introduction in probability theory. Instead, that the following chapter is more meant as a collection of formulae and results that we will need later in the book. Readers, who are familiar with main concepts of probability can therefore glance over this chapter only, to

familiarize themselves with our notation. Readers without any prior knowledge in probability might also look at more fundamental introductions to the topic, see e.g. Dekking et al. (2005).

A.2 Notation and Background

A.3 Common Distributions

There are some distributions that appear so often in statistics that familiarity with their parameters and features would greatly benefit the reader. The distributions themselves depend on one or more parameters, which are conventionally denoted with Greek letters. We also follow the notational convention that the dependence of the distribution on the parameter is expressed with a semicolon, that is, the parameters of the distribution are listed after the semicolon. This convention is used throughout the rest of the book and is easily comprehended with the following standard distributions.

Definition A.11 Binomial distribution $B(n, p)$

- Application: Two outcomes: success — 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 $k = 0, \dots, n$
- Expectation: $E(Y) = np$
- Variance: $Var(Y) = np(1 - \pi)$

Definition A.12 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 $k = 0, 1, 2, \dots$
- Expectation: $E(Y) = \lambda$
- Variance: $Var(Y) = \lambda$

Definition A.13 Normal distribution $N(\mu, \sigma)$

- 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 A.14 Exponential distribution $Exp(\lambda)$

- 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 $Var(Y) = \frac{1}{\lambda^2}$

Definition A.15 t-distribution $t(n)$

- Application: Metric symmetrically distributed random variable with a considerable probability of extreme outcomes (“heavy tails”). Also used for statistical tests for the mean of normally distributed variables when the variance is unknown and estimated from the data and the degrees of freedom, n , is small.
- Density function: $f(y; d) = \frac{\Gamma(\frac{d+1}{2})}{\sqrt{d\pi}\Gamma(\frac{d}{2})} \left(1 + \frac{y^2}{d}\right)^{-\frac{d+1}{2}}$
where $\Gamma(\cdot)$ is the gamma function and $d = 1, 2, \dots$ 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 $d \geq 1$
- Variance: $Var(Y) = \frac{d}{d-2}$ for $d \geq 3$

Definition A.16 Chi-squared distribution $\chi^2(n)$

- Application: Squared quantities like sample variances, sum of n independent squared normal distributed random variables.

- Density function: $f(y; n) = \frac{y^{\frac{n}{2}-1} e^{-\frac{y}{2}}}{2^{\frac{n}{2}} \Gamma(\frac{n}{2})}$ where $n = 1, 2, 3, \dots$
- Expectation: $E(Y) = n$
- Variance: $Var(Y) = 2n$

A.4 Common Distributions

In the following section, we list frequently employed distributions, which will be used later in the book.

- Normal Distribution

- Chi-squared Distribution

Let Z_1, \dots, Z_n be n independently $N(0, 1)$ distributed random variables. Then

$$X^2 := \sum_{i=1}^n Z_i^2 \quad (\text{A.3})$$

follows a Chi-squared distribution with n degrees of freedom. We denote this as

$$X^2 \sim \chi_p.$$

The shape of a Chi-squared distribution for different degrees of freedom is shown in Figure ???

The density of the distribution takes the form ???

We can obtain a Chi-squared distribution from a multivariate normal distribution as follows. Assume that Y is vector-valued normally distributed, that is $Y \sim N(\mu, \Sigma)$ where μ is the vector of mean values and Σ is the variance matrix. Then

$$(Y - \mu)^\top \Sigma^{-1} (Y - \mu) \sim \chi_p \quad (\text{A.4})$$

with p as dimension of Y

- t -distribution

We can combine the normal distribution and the Chi-squared distribution to derive the so-called t -distribution. Let Z_1, \dots, Z_n be n independently $N(0, 1)$ distributed random variables. Then the ratio

$$\frac{Z_1}{\sqrt{\sum_{i=2}^n Z_i^2 / (n-1)}} \approx t_{(n-1)}, \quad (\text{A.5})$$

follows a t -distribution with $n - 1$ degrees of freedom

The exact density of the t -distribution is of less interest to us, but given here for completeness:

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 Figure ????

A.5 The Exponential Family

Several of the above distributions allow for a mathematical generalisation, that in turn allows for the development of theories and models that apply to the entire model class. One such class of distributions that is central to statistics is the exponential family of distributions. This class consists of many of the most common distributions in statistics, for example, the normal, the Poisson, the binomial distribution, and many more. Although these distributions are very different, as can be seen above, they share some structural properties.

Definition A.17 A class of distributions for a random variable Y is called an **exponential family**, if the density (or probability function) can be written in the form

$$f_Y(y; \theta) = \exp\{t^\top(y)\theta - \kappa(\theta)\}h(y), \quad (\text{A.6})$$

where $h(y) \geq 0$ and $t(y) = (t_1(y), \dots, t_p(y))^\top$ is a vector of known functions and $\theta = (\theta_1, \dots, \theta_p)^\top$ is a parameter vector.

The density (or probability function) given in (A.6) is very general, and thus looks a little complicated, but we will see in the subsequent examples that the different terms simplify quite substantially in most examples. Let us therefore consider the quantities in (A.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 **natural parameter**, and quantity $\kappa(\theta)$ simply serves as normalisation constant, such that the density integrates out to 1. Hence, $\kappa(\theta)$ is defined by the following:

$$1 = \int \exp\{t^\top(y)\theta\}h(y)dy \exp(-\kappa(\theta)) \\ \Leftrightarrow \kappa(\theta) = \log \int \exp\{t^\top(y)\theta\}h(y)dy.$$

As a consequence, by differentiating $\kappa(\theta)$ with respect to θ , we get

$$\frac{\partial \kappa(\theta)}{\partial \theta} = \frac{\int t^\top(y) \exp\{t^\top(y)\theta\}h(y)dy}{\int \exp\{t^\top(y)\theta\}h(y)dy} = \int t^\top(y)f_Y(y; \theta)dy = E(t^\top(Y)).$$

Finally, the quantity $h(y)$ depends on the random variable and 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 above introduced distributions can be written in the style of an exponential family. We thereby start with the normal distribution which can be rewritten as

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(y-\mu)^2}{\sigma^2}\right) \\ = \exp\left(-\frac{1}{2}\frac{y^2 - 2y\mu + \mu^2}{\sigma^2} - \frac{1}{2}\log(\sigma^2)\right) \frac{1}{\sqrt{2\pi}} \\ = \exp\left(\underbrace{\left(-\frac{y^2}{2}, y\right)}_{t^\top(y)} \underbrace{\left(\frac{1}{\sigma^2}\right)}_{\theta} - \underbrace{\frac{1}{2}\left(-\log\frac{1}{\sigma^2} + \frac{\mu^2}{\sigma^2}\right)}_{\kappa(\theta)}\right) \underbrace{\frac{1}{\sqrt{2\pi}}}_{h(y)}$$

where $\theta_1 = \frac{1}{\sigma^2}$ and $\theta_2 = \frac{\mu}{\sigma^2}$ such that

$$\kappa(\theta) = \frac{1}{2}\left(-\log(\theta_1) + \frac{\theta_2^2}{\theta_1}\right) = -\frac{1}{2}\left(\log(\theta_1) - \frac{\theta_2^2}{\theta_1}\right).$$

Note that:

$$\frac{\partial \kappa(\theta)}{\partial \theta_1} = -\frac{1}{2}\left(\frac{1}{\theta_1} + \frac{\theta_2^2}{\theta_1^2}\right) = -\frac{1}{2}(\sigma^2 + \mu^2) = E\left(-\frac{Y^2}{2}\right) = E(t_1(Y)) \\ \frac{\partial \kappa(\theta)}{\partial \theta_2} = \frac{\theta_2}{\theta_1} = \mu = E(Y) = E(t_2(Y))$$

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 $\theta = \log(\frac{\pi}{1 - \pi})$ is also known as log-odds ratio and $\kappa(\theta)$ results in

$$\kappa(\theta) = n \log(1 + \exp(\theta)).$$

As $\pi = \exp(\theta)/(1 + \exp(\theta))$ one gets

$$\frac{\partial \kappa(\theta)}{\partial \theta} = n \frac{\exp(\theta)}{1 + \exp(\theta)} = n\pi = E(Y) = E(t(Y)).$$

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.

A.6 Limiting Distributions and Inequalities

Glevenko-Cantelli

A.6.1 Chauchy-Scharz Inequality

Let X and Y be two random variables. Then

$$\begin{aligned}\text{Cov}(X, Y)^2 &= E((X - E(X))(Y - E(Y))^2 \\ &\leq E((X - E(X))^2)E((Y - E(Y))^2) = \text{Var}(X) * \text{Var}(Y)\end{aligned}\quad (\text{A.7})$$

A.6.2 Markov Inequality

There are a number of useful inequalities which we want to list. We also sketch the proof, if this is easy.

Property A.1 Markov Inequality. For a non-negative random variable Y we have

$$P(Y > a) \leq \frac{E(Y)}{a}.$$

The proof is quite simple, as sketched below.

Proof Note that $E(Y) > 0$. Then

$$\begin{aligned}E(Y) &= \int_0^\infty yf(y)dy \\ &= \int_0^a yf(y)dy + \int_a^\infty yf(y)dy \\ &\geq \int_a^\infty yf(y)dy \\ &\geq a \int_a^y f(y)dy = aP(Y \geq a)\end{aligned}$$

A.7 Fisher regular

We will work with distributions that have special properties which 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 20th century, see Box (1978).

Definition A.18 A distribution $f(y; \theta)$ is **Fisher-regular** if the following properties hold:

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

$$\int \frac{\partial}{\partial \theta} f(y; \theta) dy = \frac{\partial}{\partial \theta} \int f(y; \theta) dy.$$

We emphasize that Fisher-regularity is not a strong assumption and 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.

A.8 Exercises

Exercise 1

Random variables are a key concept in probability theory. Let us explore some of the basics.

- a. Give the definition of a probability space as well as the definition of a random variable.
- b. Random variables are usually identified with certain functions: Give the definition of a (cumulative) distribution function and a probability (density) function for a univariate real valued random variable. How are they connected to each other? Distinguish the discrete and continuous case.

- c. Often random variables are defined as functions of other random variables. This is sometimes called variable change. The following theorem provides a relationship between densities:

Let g be an invertible and differentiable function that transforms the random variable X to the random variable $Y = g(X)$, then:

$$f_Y(y) = \left| \frac{d}{dy} g^{-1}(y) \right| f_X(g^{-1}(y))$$

Given $X \sim N(\mu, \sigma^2)$ compute the density of $Y = e^X$, commonly referred to as log-normally distributed.

Exercise 2

An employee stays at the office a little bit longer than his contract requires for all 225 working days of the year. His additional working time on each day is described with an exponentially distributed random variable with an expected value of 5 minutes. This means its density is $f_Y(y) = \lambda e^{-\lambda y}$ with some λ (*hint: remember the properties of the exponential distribution*). Moreover, we assume that different days are independent.

- a. Use the central limit theorem to compute the approximate distribution of the employee's extra working time for a whole year.
- b. Compute the probability that the employee works more than 16 additional hours.

Exercise 3

Let's now play with the central limit theorem using R. The goal of this exercise is to perform simulations and plot the results to gain a graphical intuition of the theorem.

- a. Simulate a $Bin(100, 0.75)$ variable 1000 times and plot a simple histogram of the resulting vector. Overlay the corresponding expected density function given by the Central Limit Theorem.

Why does this work, even though we made a histogram of the raw simulated values of a binomial variable instead of a statistic derived from multiple samples?

Hint: Have a look at the R-commands `rbinom`, `curve`, `dnorm`.

- b. Write a function with parameters n , p , and r that plots a histogram of r standardised sample means Z_n (i.e. mean of 0 and variance of 1) of n draws from a Bernoulli distribution with parameter p . Again, overlay the expected distribution according to the central limit theorem (the standard normal density). You should fix the domain of the histogram (e.g. to $[-5, 5]$) so that comparisons are easier. Play around with the parameters n , p and r . What can you say about the speed and rate of convergence for different values of p ?

Hint: Have a look at the function `manipulate` from the `manipulate` package for interactively changing the parameters you want to adjust in the plot.

- c. Adjust your function from (b), so instead of simulating a binomial variable, it simulates r samples of size n from $\text{Exp}(\lambda)$. The parameters n and r remain as function parameters, p is replaced with λ . The function is supposed to compute the mean of each sample, standardize them and again plot the histogram against the standard normal. What can you say about the speed of convergence now?

References

- Agresti, A. (2012). *Categorical data analysis*, Volume 792. John Wiley & Sons.
- Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle. *Proceedings of the 2nd International Symposium on Information Theory*, 267–281.
- Aldrich, J. (1997). Ra fisher and the making of maximum likelihood 1912-1922. *Statistical science* 12(3), 162–176.
- Anselin, L. (1988). *Spatial econometrics: methods and models*, Volume 4. Springer Science & Business Media.
- Baddeley, A., I. Bárány, and R. Schneider (2007). Spatial point processes and their applications. *Stochastic Geometry: Lectures Given at the CIME Summer School Held in Martina Franca, Italy, September 13–18, 2004*, 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.
- Bayes, T. (1763). Lii. 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.
- Beaumont, M. A., W. Zhang, and D. J. Balding (2002). Approximate bayesian computation in population genetics. *Genetics* 162(4), 2025–2035.
- Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society* 57, 289–300.
- Berger, V. W. and Y. Zhou (2014). Kolmogorov–smirnov test: Overview. *Wiley statsref: Statistics reference online*.
- Berry, K. J., J. E. Johnston, and P. W. Mielke (2019). *A primer of permutation statistical methods*. Springer.
- Berry, K. J., J. E. Johnston, and P. W. Mielke Jr (2011). Permutation methods. *Wiley Interdisciplinary Reviews: Computational Statistics* 3(6), 527–542.
- Bivand, R. S., E. J. Pebesma, V. Gomez-Rubio, and E. J. Pebesma (2008). *Applied spatial data analysis with R*, Volume 747248717. Springer.

- Blei, D. M., A. Kucukelbir, and J. D. McAuliffe (2017). Variational inference: A review for statisticians. *Journal of the American statistical Association* 112(518), 859–877.
- Blum, M. G. and O. François (2010). Non-linear regression models for approximate bayesian computation. *Statistics and computing* 20, 63–73.
- Bochenek, B. and Z. Ustrnul (2022). Machine learning in weather prediction and climate analyses—applications and perspectives. *Atmosphere* 13(2), 180.
- Bonferroni (1936). *Teoria statistica delle classi e calcolo delle probabilità*. Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze.
- Borgatti, S. P., A. Mehra, D. J. Brass, and G. Labianca (2009, February). Network analysis in the social sciences. *Science* 323(5916), 892–895. Publisher: American Association for the Advancement of Science.
- Box, G. and G. Jenkins (1970). *Time Series Analysis: Forecasting and Control*. San Francisco, Holdan-Day.
- Box, G. E., G. M. Jenkins, G. C. Reinsel, and G. M. Ljung (2015). *Time series analysis: forecasting and control*. John Wiley & Sons.
- Box, J. F. (1978). *R.A. Fisher, the Life of a Scientist*. John Wiley & Sons Inc.
- Brandes, U. and M. Sedlmair (2019). *Network Visualization*, pp. 5–21. Cham: Springer International Publishing.
- Brockhaus, S., M. Melcher, F. Leisch, and S. Greven (2017). Boosting flexible functional regression models with a high number of functional historical effects. *Statistics and Computing* 27(4), 913–926.
- Claeskens, G. and N. L. Hjort (2008). Model selection and model averaging. *Cambridge Books*.
- Clopper, C. J. and E. S. Pearson (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* 26(4), 404–413.
- Coles, S., J. Bawa, L. Trenner, and P. Dorazio (2001). *An introduction to statistical modeling of extreme values*, Volume 208. Springer.
- Cortes, C. and V. Vapnik (1995). Support-vector networks. *Machine learning* 20(3), 273–297.
- Cowell, R. G., P. Dawid, S. L. Lauritzen, and D. J. Spiegelhalter (2007). *Probabilistic networks and expert systems: Exact computational methods for Bayesian networks*. Springer Science & Business Media.
- Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton University Press.
- Cressie, N. (2015). *Statistics for spatial data*. John Wiley & Sons.
- Dale, A. I. (1982). Bayes or laplace? an examination of the origin and early applications of bayes' theorem. *Archive for History of Exact Sciences*, 23–47.
- Davison, A. C. and D. V. Hinkley (1997). *Bootstrap Methods and their Application*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.
- De Gooijer, J. G. (2017). *Elements of nonlinear time series analysis and forecasting*, Volume 37. Springer.
- de Laplace, P. S. (1820). *Théorie analytique des probabilités*, Volume 7. Courcier.
- Dean, A., D. Voss, and D. Draguljić (2017). Response surface methodology. In *Design and analysis of experiments*, pp. 565–614. Springer.

- Dekking, F. M., C. Kraaikamp, H. P. Lopuhaa „, and L. E. Meester (2005). *A Modern Introduction to Probability and Statistics: Understanding why and how*. Springer Science & Business Media.
- DelSole, T. and M. K. Tippett (2021). Correcting the corrected aic. *Statistics & Probability Letters* 173, 109064.
- Devlin, K. (2010). The pascal-fermat correspondence: How mathematics is really done. *The Mathematics Teacher* 103(8), 579–582.
- Dikta, G. and M. Scheer (2021). *Bootstrap Methods*. Springer.
- Edwards, A. W. (1997). What did fisher mean by "inverse probability" in 1912–1922? *Statistical Science* 12(3), 177–184.
- Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics* 7(1), 1 – 26.
- Efron, B. (2010). *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Cambridge University Press, Cambridge, UK.
- Eicker, F. (1967). Limit theorems for regressions with unequal and dependent errors. In *Proc. Fifth Berkeley Sympos. Math. Statist. and Probability (Berkeley, Calif., 1965/66), Vol. I: Statistics*, Berkeley, Calif, pp. 59 – 82. Univ. California Press.
- Fahrmeir, L., T. Kneib, S. Lang, and B. Marx (2013). *Regression - Models, Methods and Applications*. Springer.
- Ferraty, F. and P. Vieu (2006). *Nonparametric Functional Data Analysis*. Springer Series in Statistics. New York: Springer.
- Fienberg, S. E. (2006). When did Bayesian inference become "Bayesian"? *Bayesian Analysis* 1(1), 1 – 40.
- 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. Publisher: Taylor & Francis.
- Fisher, R. (1922). On the interpretation of 2 from contingency tables, and the calculation of p. *ournal of the Royal Statistical Society*, 87 – 94.
- Fisher, R. (1955). Statistical methods and scientific induction. *Journal of the Royal Statistical Society: Series B (Methodological)* 17(1), 69–78.
- Fisher, R. (1958a). Cigarettes, cancer, and statistics. *The Centennial Review of Arts & Science* 2, 151–166. Publisher: 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, July). Lung Cancer and Cigarettes? *Nature* 182(4628), 108–108.
- Fox, C. W. and S. J. Roberts (2012). A tutorial on variational bayesian inference. *Artificial intelligence review* 38, 85–95.
- Fritz, C. (2022). *Statistical approaches to dynamic networks in society*. PhD Thesis, lmu.
- Fruchterman, T. M. J. and E. M. Reingold (1991, November). Graph drawing by force-directed placement. *Software: Practice and Experience* 21(11), 1129–1164. Publisher: John Wiley & Sons, Ltd.
- Fréchet, M. (1943). Sur l'extension de certaines évaluations statistiques au cas de petits échantillons. *Rev. Inst. Int. Statist* 11, 182 – 205.

- Gandini, S., E. Botteri, S. Iodice, M. Boniol, A. B. Lowenfels, P. Maisonneuve, and P. Boyle (2008, January). 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.
- Gelfand, A. E. (2000). Gibbs sampling. *Journal of the American statistical Association* 95(452), 1300–1304.
- Geman, S. and D. Geman (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI-6(6)*, 721–741.
- Goldenberg, A., A. X. Zheng, S. E. Fienberg, E. M. Airoldi, and Others (2010). A survey of statistical network models. *Foundations and Trends in Machine Learning* 2(2), 129–233. Publisher: Now Publishers, Inc.
- Goldsmith, J., J. Bobb, C. M. Crainiceanu, B. Caffo, and D. Reich (2011). Penalized functional regression. *Journal of computational and graphical statistics* 20(4), 830–851.
- Goldsmith, J., F. Scheipl, L. Huang, J. Wrobel, C. Di, J. Gellar, J. Harezlak, M. W. McLean, B. Swihart, L. Xiao, C. Crainiceanu, and P. T. Reiss (2020). *refund: Regression with Functional Data*. R package version 0.1-23.
- Goodfellow, I., Y. Bengio, and A. Courville (2016). *Deep Learning*. MIT Press. <http://www.deeplearningbook.org>.
- Goodfellow, I. J., J. Shlens, and C. Szegedy (2015). Explaining and harnessing adversarial examples. *stat 1050*, 20.
- Goodman, S. (2008). A dirty dozen: twelve p-value misconceptions. *Semin Hematol*, 235 – 240.
- Greven, S. and F. Scheipl (2017). A general framework for functional regression modelling. *Statistical Modelling* 17(1-2), 1–35.
- Grimmett, G. and D. Stirzaker (2001). *Probability and Random Processes, 3rd edition*. Oxford Uni Press.
- Gumbel, E. J. (1958). *Statistics of extremes*. Columbia university press.
- Haan, L. and A. Ferreira (2006). *Extreme value theory: an introduction*, Volume 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., A. Montanari, S. Rosset, and R. J. Tibshirani (2022). Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics* 50(2), 949 – 986.
- Hastings, W. K. (1970, 04). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 57(1), 97–109.
- Hoffman, M. D., D. M. Blei, C. Wang, and J. Paisley (2013). Stochastic variational inference. *Journal of Machine Learning Research*.
- 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.
- Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proc. Fifth Berkeley Sympos. Math. Statist. and Probability (Berkeley, Calif., 1965/66), Vol. I: Statistics*, Berkeley, Calif, pp. 221 – 233. Univ. California Press.
- Hüllermeier, E. and W. Waegeman (2021, March). Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning* 110(3), 457–506.
- Hurvich, C. M. and C.-L. Tsai (1989). Regression and time series model selection in small samples. *Biometrika* 76(2), 297–307.
- Hyndman, R. J. and H. L. Shang (2010). Rainbow plots, bagplots, and boxplots for functional data. *Journal of Computational and Graphical Statistics* 19(1), 29–45.
- Kauermann, G., H. Haupt, and N. Kaufmann (2012). A hitchhiker’s view on spatial statistics and spatial econometrics for lattice data. *Statistical Modelling* 12(5), 419–440.
- Kingma, D. P. and M. Welling (2014). Auto-encoding variational Bayes. in bengio and lecun (eds.). *2nd International Conference on Learning Representations*.
- Kiureghian, A. D. and O. Ditlevsen (2009, March). Aleatory or epistemic? Does it matter? *Structural Safety* 31(2), 105–112.
- Kolb, C., C. L. Müller, B. Bischl, and D. Rügamer (2023). Smoothing the edges: A general framework for smooth optimization in sparse regularization using hadamard overparametrization.
- 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., D. Tran, R. Ranganath, A. Gelman, and D. M. Blei (2017). Automatic differentiation variational inference. *Journal of machine learning research*.
- 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.
- Lauritzen, S. L. (1996). *Graphical models*, Volume 17. Clarendon Press.
- Lehmer, D. H. (1951). Mathematical methods in large-scale computing units. *Annu. Comput. Lab. Harvard Univ.* 26, 141–146.
- Li, X. and X.-L. Meng (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.
- Lindsay, B. G., J. Kettenring, and D. O. Siegmund (2004). A Report on the Future of Statistics. *Statistical Science* 19(3), 387 – 413.
- Maathuis, M., M. Drton, S. Lauritzen, and M. Wainwright (2018). *Handbook of graphical models*. CRC Press.
- Marin, J.-M., P. Pudlo, C. P. Robert, and R. J. Ryder (2012). Approximate bayesian computational methods. *Statistics and computing* 22(6), 1167–1180.

- Marron, J. S., J. O. Ramsay, L. M. Sangalli, and A. Srivastava (2015). Functional data analysis of amplitude and phase variation. *Statistical Science*, 468–484.
- 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., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics* 21(6), 1087–1092.
- Moreno, J. L. (1934). Who Shall Survive? A New Approach to the Problem of Human Interrelations. Place: Washington, DC Publisher: Nervous and Mental Disease Publishing Co.
- Morris, J. S. and R. J. Carroll (2006). Wavelet-based functional mixed models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 68(2), 179–199.
- Nelder, J. A. and R. W. Wedderburn (1972). Generalized linear models. *Journal of the Royal Statistical Society Series A: Statistics in Society* 135(3), 370–384.
- Nepusz, T., A. Petróczi, L. Négyessy, and F. Bazsó (2008). Fuzzy communities and the concept of bridgeness in complex networks. *Physical Review E* 77(1), 1–12.
- Neyman, J. and E. S. Pearson (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.
- Pearson, K. (1904). On the theory of contingency and its relation to association and normal correlation. *Mathematical Contributions to the Theory of Evolution XIII*.
- 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., D. Janzing, and B. Schölkopf (2017). *Elements of causal inference: foundations and learning algorithms*. The MIT Press.
- Ramsay, J. O. and B. W. Silverman (2005). *Functional Data Analysis*. Springer Series in Statistics. New York, NY: 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.
- Ripley, B. D. (2005). *Spatial statistics*. John Wiley & Sons.
- Robbins, H. and S. Monro (1951). A stochastic approximation method. *The Annals of Mathematical Statistics*, 400–407.
- Robert, C. and G. Casella (2004). *Monte Carlo Statistical Methods*. Springer.
- Ross, S. M. (2022). *Simulation*. academic press.
- Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, 1151–1172.
- Rügamer, D., S. Brockhaus, K. Gentsch, K. Scherer, and S. Greven (2018). Boosting factor-specific functional historical models for the detection of synchronization in

- bioelectrical signals. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 67(3), 621–642.
- Scheipl, F., J. Gertheiss, and S. Greven (2016). Generalized functional additive mixed models. *Electronic Journal of Statistics* 10, 1455–1492.
- Scheipl, F., A.-M. Staicu, and S. Greven (2015). Functional additive mixed models. *Journal of Computational and Graphical Statistics* 24(2), 477–501.
- Shumway, R. H., D. S. Stoffer, and D. S. Stoffer (2017). *Time series analysis and its applications, 4th edition*, Volume 4. Springer.
- Somerville, M. C. and R. S. Brown (2013). Exact likelihood ratio and score confidence intervals for the binomial proportion. *Pharmaceutical statistics* 12(3), 120–128.
- Stigler, S. M. (1975). Studies in the history of probability and statistics. xxxiv: Napoleonic statistics: The work of laplace. *Biometrika* 62(2), 503–517.
- Stigler, S. M. (2018). Richard price, the first bayesian. *Statistical Science* 33(1), 117–125.
- 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.
- Suess, E. A. and B. E. Trumbo (2010). *Introduction to probability simulation and Gibbs sampling with R*, Volume 307. Springer.
- Tager, I. B., S. T. Weiss, B. Rosner, and F. E. Speizer (1979). Effect of parental cigarette smoking on the pulmonary function of children. *American Journal of Epidemiology* 110(1), 15–26.
- Tavaré, S., D. J. Balding, R. C. Griffiths, and P. Donnelly (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, 01). 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.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 58(1), 267–288.
- Tibshirani, R. J. and B. Efron (1993). An introduction to the bootstrap. *Monographs on statistics and applied probability* 57, 1–436.
- 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*, Volume 34. Cambridge University Press.
- van de Schoot, R., S. Depaoli, R. King, B. Kramer, K. Märkens, M. G. Tadesse, M. Vannucci, A. Gelman, D. Veen, J. Willemsen, et al. (2021). Bayesian statistics and modelling. *Nature Reviews Methods Primers* 1(1), 1.
- van den Hout, A. and P. G. van der Heijden (2002). Randomized response, statistical disclosure control and misclassification: a review. *International Statistical Review* 70(2), 269–288.

- 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.
- Wasserman, L., A. Ramdas, and S. Balakrishnan (2020). Universal inference. *Proceedings of the National Academy of Sciences* 117(29), 16880–16890.
- Wasserman, S. and K. Faust (1994). *Social network analysis : Methods and applications*. Cambridge: Cambridge University Press.
- White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* 48, 817 – 838.
- Whittaker, J. (2009). *Graphical models in applied multivariate statistics*. Wiley Publishing.
- Whittle, P. (1951). *Hypothesis testing in time series analysis*, Volume 4. Almqvist & Wiksell boktr.
- Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin* 1(6), 80–83.
- Wood, S. N. (2017). *Generalized additive models: an introduction with R*. CRC press.
- Yao, F., H.-G. Müller, and J.-L. Wang (2005). Functional data analysis for sparse longitudinal data. *Journal of the American statistical association* 100(470), 577–590.
- Ziliak, S. T. (2008). Retrospectives: Guinnessometrics: the economic foundation of “student’s” t. *Journal of Economic Perspectives* 22(4), 199–216.

Index

- k*-fold cross-validation, 139
accuracy, 111
Akaike Information Criterion (AIC), 140
Aleatoric Uncertainty, 10
autocorrelation function, 259
autocovariance function, 257
Bayesian Information Criterion, 147
Bayesian Information Criterion (BIC), 143
Bernstein - von Mises Theorem, 199
bias, 30
bias-variance-tradeoff, 30
Bonferroni adjustment, 105
bootstrapping, 182
chi-squared statistics, 90
concentration matrix, 250
confidence interval, 36
consistency, 28
credibility interval, 43
cross-validation, 139
data generating process, 13
directed acyclic graphs (DAG), 247
empirical Bayes, 200, 201
Epistemic Uncertainty, 10
estimator, 16
exact confidence interval, 222
extreme value distribution, 226
F score, 112
false discovery rate, 111
family wise error rate, 104
Fisher information, 55
Fisher-regular, 327
forward selection, 152
Gaussian Markov Random Field, 176
generalized extreme value, 227
generalized extreme value distribution (GEV), 227
Gibbs sampling, 172
graphical model, 241
highest density interval, 44
hyperparameter, 196
Hypothesis Test, 85
i.i.d, 14
importance sampling, 164
Kolmogorov-Smirnov Test, 94
Kullback-Leibler Divergence, 19
likelihood function, 24
likelihood principle, 27, 70
likelihood ratio, 63

Likelihood-Ratio, 64
log-linear model, 235

Markov Chain Monte Carlo (MCMC),
166

Markov Inequality, 327

maximum likelihood, 55

maximum likelihood estimate, 25

mean squared error, 30

Neyman-factorisation, 33

Neyman-Pearson test, 100

null hypothesis, 76

p-value, 77

pairwise conditional independence, 243

pivotal statistic, 37

plug-in principle, 183

posterior distribution, 196

power of test, 96

prevalence, 111

prior distribution, 196

rejection sampling, 162

Score Function, 54

score function, 52

sensitivity, 111

specificity, 111

stationarity, 256

statistical model, 15

sufficient, 33

training and test data, 139

Wald test, 88

Wilcoxon-test, 93

Yale-Walker, 273