Probability Theory and Statistics

With a view towards the natural sciences

Lecture notes

Niels Richard Hansen
Department of Mathematical Sciences
University of Copenhagen
November 2010

Preface

The present lecture notes have been developed over the last couple of years for a course aimed primarily at the students taking a Master's in bioinformatics at the University of Copenhagen. There is an increasing demand for a general introductory statistics course at the Master's level at the university, and the course has also become a compulsory course for the Master's in eScience. Both educations emphasize a computational and data oriented approach to science – in particular the natural sciences.

The aim of the notes is to combine the mathematical and theoretical underpinning of statistics and statistical data analysis with computational methodology and practical applications. Hopefully the notes pave the way for an understanding of the foundation of data analysis with a focus on the probabilistic model and the methodology that we can develop from this point of view. In a single course there is no hope that we can present all models and all relevant methods that the students will need in the future, and for this reason we develop general ideas so that new models and methods can be more easily approached by students after the course. We can, on the other hand, not develop the theory without a number of good examples to illustrate its use. Due to the history of the course most examples in the notes are biological of nature but span a range of different areas from molecular biology and biological sequence analysis over molecular evolution and genetics to toxicology and various assay procedures.

Students who take the course are expected to become users of statistical methodology in a subject matter field and potentially also developers of models and methodology in such a field. It is therefore intentional that we focus on the fundamental principles and develop these principles that by nature are mathematical. Advanced mathematics is, however, kept out of the main text. Instead a number of math boxes can be found in the notes. Relevant, but mathematically more sophisticated, issues are treated in these math boxes. The main text does not depend on results developed in

the math boxes, but the interested and capable reader may find them illuminating. The formal mathematical prerequisites for reading the notes is a standard calculus course in addition to a few useful mathematical facts collected in an appendix. The reader who is not so accustomed to the symbolic language of mathematics may, however, find the material challenging to begin with.

To fully benefit from the notes it is also necessary to obtain and install the statistical computing environment R. It is evident that almost all applications of statistics today require the use of computers for computations and very often also simulations. The program R is a free, full fledge programming language and should be regarded as such. Previous experience with programming is thus beneficial but not necessary. R is a language developed for statistical data analysis and it comes with a huge number of packages, which makes it a convenient framework for handling most standard statistical analyses, for implementing novel statistical procedures, for doing simulation studies, and last but not least it does a fairly good job at producing high quality graphics.

We all have to crawl before we can walk – let alone run. We begin the notes with the simplest models but develop a sustainable theory that can embrace the more advanced ones too.

Last, but not least, I owe a special thank to Jessica Kasza for detailed comments on an earlier version of the notes and for correcting a number of grammatical mistakes.

<div align="right">
November 2010
Niels Richard Hansen
</div>

Contents

1

Introduction

1.1   Notion of probabilities

Flipping coins and throwing dice are two commonly occurring examples in an introductory course on probability theory and statistics. They represent archetypical experiments where the outcome is uncertain – no matter how many times we roll the dice we are unable to predict the outcome of the next roll. We use probabilities to describe the uncertainty; a fair, classical dice has probability 1/6 for each side to turn up. *Elementary* probability computations can to some extent be handled based on intuition, common sense and high school mathematics. In the popular dice game Yahtzee the probability of getting a Yahtzee (five of a kind) in a single throw is for instance

$$\frac{6}{6^5} = \frac{1}{6^4} = 0.0007716.$$

The argument for this and many similar computations is based on the *pseudo theorem* that the probability for any event equals

$$\frac{\text{number of favourable outcomes}}{\text{number of possible outcomes}}.$$

Getting a Yahtzee consists of the six favorable outcomes with all five dice facing the same side upwards. We call the formula above a pseudo theorem because, as we will show in Section 2.4, it is only the correct way of assigning probabilities to events under a very special assumption about the probabilities describing our experiment. The special assumption is that all outcomes are equally probable – something we tend to believe if we don't know any better, or can see no way that one outcome should be more likely than others.

However, without some training most people will either get it wrong or have to give up if they try computing the probability of anything except the most elementary

events – even when the pseudo theorem applies. There exist numerous tricky probability questions where intuition somehow breaks down and wrong conclusions can be drawn if one is not extremely careful. A good challenge could be to compute the probability of getting a Yahtzee in three throws with the usual rules and provided that we always hold as many equal dice as possible.

Figure 1.1: The relative frequency of times that the dice sequence ⚅ ⚄ ⚅ comes out before the sequence ⚅ ⚀ ⚄ as a function of the number of times the dice game has been played.

The Yahtzee problem can in principle be solved by counting – simply write down all combinations and count the number of favorable and possible combinations. Then the pseudo theorem applies. It is a futile task but in principle a possibility.

In many cases it is, however, impossible to rely on counting – even in principle. As an example we consider a simple dice game with two participants: First I choose a sequence of three dice throws, ⚀ ⚄ ⚅, say, and then you choose ⚄ ⚀ ⚄, say. We throw the dice until one of the two sequences comes out, and I win if ⚀ ⚄ ⚅ comes out first and otherwise you win. If the outcome is

⚄ ⚅ ⚄ ⚅ ⚄ ⚀ ⚄ ⚄ ⚄ ⚄ ⚀ ⚄ ⚅

then I win. It is natural to ask with what probability you will win this game. In addition, it is clearly a quite boring game, since we have to throw a lot of dice and simply wait for one of the two sequences to occur. Another question could therefore be to ask how boring the game is? Can we for instance compute the probability for

having to throw the dice more than 100, or perhaps 500, times before any of the two sequences shows up? The problem that we encounter here is first of all that the pseudo theorem does not apply simply because there is an infinite number of favorable as well as possible outcomes. The event that you win consists of the outcomes being all finite sequences of throws ending with ⬚ ⬚ ⬚ *without* ⬚ ⬚ ⬚ occurring somewhere as three subsequent throws. Moreover, these outcomes are certainly not equally probable. By developing the theory of probabilities we obtain a framework for solving problems like this and doing many other even more subtle computations. And if we cannot *compute* the solution we might be able to obtain an answer to our questions using *computer simulations*. Moreover, the notes introduce probability theory as the foundation for doing statistics. The probability theory will provide a framework, where it becomes possible to clearly formulate our statistical questions and to clearly express the assumptions upon which the answers rest.
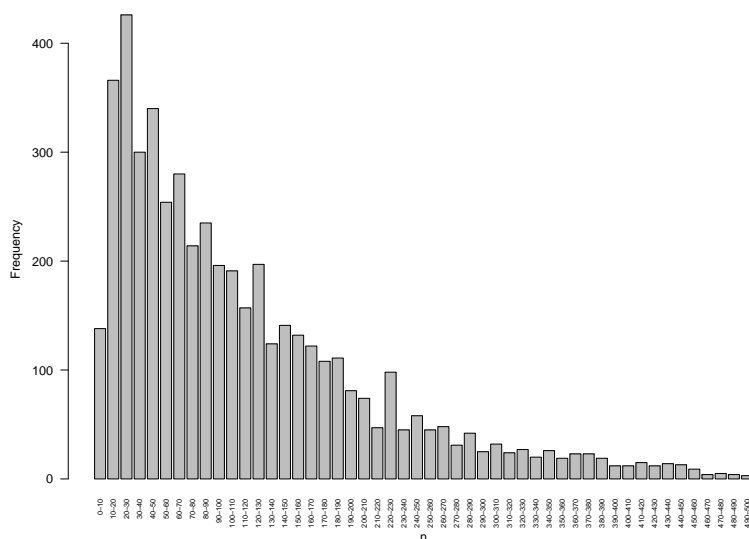
Figure 1.2: Playing the dice game 5000 times, this graph shows how the games are distributed according to the number of times, n, we had to throw the dice before one of the sequences ⬚ ⬚ ⬚ or ⬚ ⬚ ⬚ occurred.

Enough about dice games! After all, these notes are about probability theory and statistics with applications to the natural sciences. Therefore we will try to take examples and motivations from real biological, physical and chemical problems, but it can also be rewording intellectually to focus on simple problems like those from a dice game to really understand the fundamental issues. Moreover, some of the questions that we encounter, especially in biological sequence analysis, are similar in nature to those we asked above. If we don't know any better – and in many cases we don't – we may regard a sequence of DNA as simply being random. If we don't

have a clue about whether the sequence encodes anything useful, it may just as well be random evolutionary leftovers. Whether this is the case for (fractions of) the intergenic regions, say, of eukaryote genomes is still a good question. But what do we actually mean by random? A typical interpretation of a sequence being random is to regard the DNA sequence as the outcome of throwing a four sided dice, with sides A, C, G, and T, a tremendously large number of times.

One purpose of regarding DNA sequences as being random is to have a *background model*. If we have developed a method for detecting novel protein coding genes, say, in DNA sequences, a background model for the DNA that is not protein coding is useful. Otherwise we cannot tell how likely it is that our method finds *false* protein coding genes, i.e. that the method claims that a segment of DNA is protein coding even though it is not. Thus we need to compute the probability that our method claims that a random DNA sequence – with random having the meaning above – is a protein coding gene. If this is unlikely, we believe that our method indeed finds truly protein coding genes.

A simple gene finder can be constructed as follows: After the start codon, ATG, a number of nucleotides occur before one of the stop codons, TAA, TAG, TGA is reached for the first time. Our protein coding gene finder then claims that if more than 99 (33 codons) nucleotides occur before any of the stop codons is reached then we have a gene. So what is the chance of getting more than 99 nucleotides before reaching a stop coding for the first time? The similarity between determining how boring our little dice game is should be clear. The sequence of nucleotides occurring between a start and a stop codon is called an open reading frame, and what we are interested in is thus how the lengths of open reading frames are distributed in random DNA sequences.

The probability theory provides the tools for computing probabilities. If we know the *probability measure* – the assignment of probabilities to events – the rest is in principle just mathematical deduction. Sometimes this is what we need, but in many cases we are actually interested in questions that go beyond what we can obtain by pure deduction. Situations where we need the interplay between data – experimental data or computational data – and the probability theory, are where statistics comes into the picture. Hopefully the reader will realize upon reading these notes that the four sided dice model of DNA is rarely useful quantitatively, but may provide some qualitative insight into more practical problems we really want to solve.

1.2 Statistics and statistical models

Where probability theory is a deductive and mathematical science, statistics – and in particular applied statistics – is an inductive science, where we want to obtain knowledge and understanding about general relations and patterns from data. In other words, if we have observed certain events as outcomes of an experiment how

do we find the mechanisms that generated those outcomes, and how do we understand those mechanisms? What brings the topics of probability theory and statistics together is the use of probability theory to model the generating mechanisms, and the corresponding mathematical framework that comes along provides us with methods for doing inductive inference and for understanding and interpreting the results. Probability measures have the ability to capture unpredictable or uncontrollable variation (randomness) together with systematic variability, and are therefore ideal objects to model almost any kind of experimental data – with "experiments" to be understood in the widest possible sense. Theoretical statistics offers a range of *statistical models*, which for the present section can be thought of families of probability measures, and methods for transferring data into probability measures. The transfer of data into a probability measure is a cornerstone in statistics, which is often called *statistical inference*. The primary aim of the notes is to give a solid understanding of the concept of a statistical model based on probability theory and how the models can be used to carry out the inductive step of doing statistical inference.

To get a good understanding of what a statistical model is, one needs to see some examples. We provide here a number of examples that will be taken up and reconsidered throughout the notes. In this section we describe some of the background and the experiments behind the data. In the rest of the notes these examples will be elaborated on to give a fully fledged development of the corresponding statistical models and inference methods as the proper theoretical tools are developed.

Example 1.2.1 (Neuronal interspike times)**.** Neuron cells in the brain are very well studied and it is known that neurons transmit electrochemical signals. Measurements of a cells membrane potential show how the membrane potential can activate voltage-gated ion channels in the cell membrane and trigger an electical signal known as a spike.

At the most basic level it is of interest to understand the interspike times, that is, the times between spikes, for a single neuron in a steady state situation. The interspike times behave in an intrinsically stochastic manner meaning that if we want to describe the typical interspike times we have to rely on a probabilistic description.

A more ambitious goal is to relate interspike times to external events such as visual stimuli and another goal is to relate the interspike times of several neurons. ⋄

Example 1.2.2 (Motif counting)**.** Special motifs or patterns in the genomic DNA-sequences play an important role in the development of a cell – especially in the way the regulation of gene expression takes place. A motif can be a small word, in the DNA-alphabet, or a collection of words as given by for instance a regular expression or by other means. One or more proteins involved in the expression regulation mechanisms are then capable of binding to the DNA-segment(s) that corresponds to the word or word collection (the binding site), and in doing so either enhance or suppress the expression of a protein coding gene, say.

The typical datasets we consider in relation to motifs are computational in nature. Of course the genomes are biological, but we employ one or several computational approaches that give us the location, say, of all the occurrences of the motif. It is rarely the case that the computational procedure will find only biologically meaningful binding sites, and we want a statistical model of the random occurrence of motifs in DNA-sequences. In particular we are interested in the distribution of how many computationally predicted binding sites we will find in a sequence of length n, how the distribution depends upon n and perhaps also on the nucleotide composition of the DNA-sequences. As for the open reading frames, if we are comfortable with the four sided dice model of DNA, the problem is a probability problem. But again we are not, and we would like instead to model directly the predicted binding site occurrences.

A slight variation of counting the number of occurrences is to just record whether a motif is present in a sequence or not, or whether it is present at a given location or window in the sequence. ◇

Example 1.2.3 (Forensic Statistics)**.** In forensic science one of the interesting problems from the point of view of molecular biology and statistics is the ability to identify the person who committed a crime based on DNA-samples found at the crime scene. One approach known as the short tandem repeat (STR) analysis is to consider certain specific tandem repeat positions in the genome and count how many times the pattern has been repeated. The technique is based on tandem repeats with non-varying flanking regions to identify the repeat but with the number of pattern repeats varying from person to person. These counts of repeat repetitions are useful as "genetic fingerprints" because the repeats are not expected to have any function and the mutation of repeat counts is therefore neutral and not under selective pressure. Moreover, the mutations occur (and have occurred) frequently enough so that there is a sufficient variation in a population for discriminative purposes. It would be of limited use if half the population, say, have a repeat count of 10 with the other half having a repeat count of 11.

Without going into too many technical details, the procedure for a DNA-sample from a crime scene is quite simple. First one amplifies the tandem repeat(s) using PCR with primers that match the flanking regions, and second, one extracts the sequence for each of the repeats of interest and simply count the number of repeats. Examples of STRs used include TH01, which has the pattern `AATG` and occurs in intron 1 of the human tyrosine hydroxylase gene, and TPOX, which has the same repeat pattern but is located in intron 10 of the human thyroid peroxidase gene.

One tandem repeat is not enough to uniquely characterize an individual, so several tandem repeats are used. A major question remains. Once we have counted the number of repeats for k, say, different repeat patterns we have a vector $(n_1, \ldots, n_k)$ of repeat counts. The Federal Bureau of Investigation (FBI) uses for instance a standard set of 13 specific STR regions. If a suspect happens to have an identical vector of repeat counts – the same genetic fingerprint – we need to ask ourselves

what the chance is that a "random" individual from the population has precisely this genetic fingerprint. This raises a number of statistical questions. First, what kind of random procedure is the most relevant – a suspect is hardly selected completely at random – and second, what population is going to be the reference population? And even if we can come up with a bulletproof solution to these questions, it is a huge task and certainly not a practical solution to go out and count the occurrences of the fingerprint $(n_1, \ldots, n_k)$ in the entire population. So we have to rely on smaller samples from the population to *estimate* the probability. This will necessarily involve model assumptions – assumptions on the probabilistic models that we will use.

One of the fundamental model assumptions is the classical *Hardy-Weinberg equilibrium* assumption, which is an assumption about *independence* of the repeat counts at the two different chromosomes in an individual. ◇

Example 1.2.4 (Sequence evolution). We will consider some of the most rudimentary models for evolution of biological sequences. Sequences evolve according to a complicated interaction of random mutations and selection, where the random mutations can be single nucleotide substitutions, deletions or insertions, or higher order events like inversions or crossovers. We will only consider the substitution process. Thus we consider two sequences, DNA or proteins, that are evolutionary related via a number of nucleotide or amino acid substitutions. We will regard each nucleotide position (perhaps each codon) or each amino acid position as unrelated to each other, meaning that the substitution processes at each position are independent. We are interested in a model of these substitution processes, or alternatively a model of the evolutionary related pairs of nucleotides or amino acids. We are especially interested in how the evolutionary distance – measured for instance in calendar time – enters into the models.

The models will from a biological point of view be very simplistic and far from realistic as models of real sequence evolution processes. However, they form the starting point for more serious models, if one, for instance, wants to enter the area of phylogenetics, and they are well suited to illustrate and train the fundamental concepts of a statistical models and the methods of statistical inference.

Obtaining data is also a little tricky, since we can rarely go out and read of evolutionary related sequences where we know the relation "letter by letter" – such relations are on the contrary established computationally using alignment programs. However, in some special cases, one can actually observe real evolution as a number of substitutions in the genome. This is for instance the case for rapidly evolving RNA-viruses.

Such a dataset was obtained for the H strain of the hepatitis C virus (HCV) (Ogata et al., Proc. Natl. Acad. Sci., 1991 (88), 3392-3396). A patient, called patient H, was infected by HCV in 1977 and remained infected at least until 1990 – for a period of 13 years. In 1990 a research group sequenced three segments of the HCV genome obtained from plasma collected in 1977 as well as in 1990. The three segments, denoted segment A, B and C, were all directly alignable without the need to introduce

Position	42	275	348	447	556	557	594	652	735	888	891	973	979	1008	1011	1020	1050	1059	1083	1149	1191	1195	1224	1266
H77	G	C	C	A	G	C	C	C	T	C	T	G	G	C	G	C	T	T	C	T	T	T	T	A
H90	A	T	T	G	A	T	T	T	C	T	C	A	A	T	A	T	A	C	T	C	C	A	C	G

Table 1.1: The segment position and nucleotides for 24 mutations on segment A of the hepatitis C virus.

insertions or deletions. The lengths of the three segments are 2610 (A), 1284 (B) and 1029 (C) respectively.

segment A

	H90 A	H90 C	H90 G	H90 T
H77 A		1	11	1
H77 C	4		1	20
H77 G	13	3		1
H77 T	3	19	1	

segment B

	H90 A	H90 C	H90 G	H90 T
H77 A		0	5	0
H77 C	1		0	8
H77 G	1	1		1
H77 T	2	6	0	

segment C

	H90 A	H90 C	H90 G	H90 T
H77 A		1	2	0
H77 C	1		2	5
H77 G	4	0		0
H77 T	1	3	1	

Table 1.2: Tabulation of all mutations in the three segments A, B and C of the hepatitis C virus genome from the 1977 H strain to the 1990 H strain.

In Table 1.2.4 we see the position for the first 24 mutations as read from the 5'-end of segment A out of the total of 78 mutations on segment A. In Table 1.2.4 we have tabulated all the mutations in the three segments. ◇

Example 1.2.5 (Toxicology). "Alle Dinge sind Gift und nichts ist ohne Gift; allein die Dosis macht, dass ein Ding kein Gift ist." Theophrastus Phillipus Auroleus Bombastus von Hohenheim (1493-1541). All compounds are toxic in the right dose – the question is just what the dose is.

The *dose-response* experiment is a classic. In a controlled experiment, a collection of organisms is subdivided into different groups and each group is given a specific dose of a compound under investigation. We focus here on whether the concentration is lethal, and thus subsequently we count the number of dead organisms in each group. How can we relate the dose to the probability of death? What is the smallest concentration where there is more than 2% chance of dying? What is the concentration where 50% of the organisms die – the so-called LD50 value?

We want a model for each of the groups that captures the probability of dying in that particular group. In addition, we want to relate this probability across the groups to the concentration – or dose – of the compound.

Figure 1.3 shows the data from an experiment with flies that are given different doses of the insecticide dimethoat. The figure also shows a curve that is inferred from the given dataset, which can be used to answer the questions phrased above. A major topic of the present set of notes is to develop the theory and methodology for how such a curve is inferred and how we assess the uncertainty involved in reporting e.g. the inferred value of LD50.
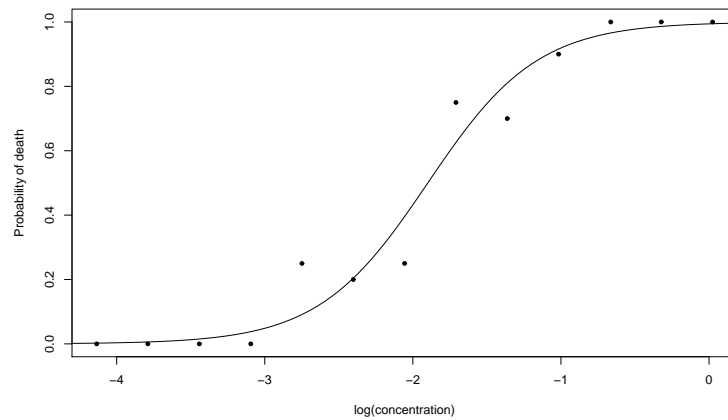
Figure 1.3: The relative frequency of dead flies plotted against the log(concentration) of the insecticide dimethoat together with a curve inferred from the data that provide a relation between the probability of fly death and the concentration

Obviously it would be unethical to carry out a controlled experiment with a lethal outcome on a group of humans. If we are interested in the dose-response relation regarding lethality for humans of a given compound, we can only use *observational data*. That is, we can try to measure the dose that groups of humans are exposed to for other reasons and relate that to their survival. It is important to be aware of the fundamentally different nature of such a data collection strategy compared to the controlled experiment. In particular, for the controlled experiment we design the experiment so that the groups that are exposed to different doses are identical in all other aspects. This allows us to conclude that it is actually the concentration of the compound that is the *cause* of the death. For the observational data there could potentially be a number of other differences between groups that are exposed to different dose levels. Observed differences in response can very well be caused by other – observed or unobserved – alternative differences between the groups. This is a serious weakness of observational data, and substantial amounts of research have been done to facilitate causal conclusions from the analysis of observational data. Though these developments are very useful, the conclusions from a statistical analysis of observational data will never provide strong evidence of a causal explanation unless it is backed up by serious subject matter theory and/or controlled experiments. $\diamond$

Example 1.2.6 (Assay measurements). A very large body of biological or biochemical experiments are known as *assays*. The term covers a wide range of measurement techniques that share the purpose of either measuring quantitatively the amount of a substance in a sample, or detecting qualitatively if the substance is present or not. There is a bewildering terminology associated with assays due to the many different special techniques and areas of application. We cannot cover everything here but only give one concrete example.

ELISA (Enzyme-linked immunosorbent assay) is a so-called immunoassay used to detect or quantify the presence of an antibody or antigen in a sample. A characteristic property of this assay is that the final measurement is a measurement of a light intensity at a given wavelength. Thus it is an indirect measurement of the concentration of the substance of interest given by the resulting light intensity. Changes in experimental conditions result in "noise" in the measurements so that repeated measurements will give different results. Different biological replications will also result in biological variation that will add to the measurement noise. All in all we attempt to capture these variations with a probabilistic model so that we understand the natural variation in our experiment even under identical conditions (up to what we can control). In some situations – with the correct *experimental design* – it may be possible and desirable to decompose the measurement noise from the biological variation. In other situations the two kinds of variation add up to a common variation term where the decomposition is not possible.

From a statistical point of view there is, however, another reason to be interested in assays. Because the measurement is an indirect measurement we will need a method to relate the actual observation to the concentration of the substance. This relation is known as *the standard curve*. In an ideal world there would be a single curve relating the concentration to the measurement plus perhaps some measurement noise. However, in reality such a single, globally valid curve is not likely to be found. Instead it is inferred from measurements involving a *standard solution series* with known concentrations. This leads to a problem known as *regression* where we want to infer the relation between two variables where we know one and measure the other with measurement noise. In the context of e.g. assays it is common to talk about *calibration*, as our measurements are used to calibrate the equipment before doing the experiment of actual interest.   ◇

Example 1.2.7 (Location and scale)**.** A more abstract example of a statistical model is the repeated measurement of a single, quantitative variable. It could be a clinical measure like blood pressure, a biochemical measure like a protein concentration from an assay as described above, but it could also be a computational result like a predicted minimal free energy for a set of (equal length) RNA-molecules. It is customary to summarize such repeated measurements by a typical value, the *location*, together with the spread, or *scale*, of the measurements.

There is an underlying statistical model, the location-scale model, for this kind of data reduction, which is important in itself, but also forms a very important building block for many other situations with multiple different groups of repeated measurements, regression analysis, where the measurements depend on another observed variable, or combinations with a complicated experimental design. Moreover, the use of a normalization of the measurements to so-called z-values or z-scores is perhaps best understood within the framework of the statistical model – especially the assumptions upon which the normalization rests. It is also an integral part of the location-scale model that the normalized z-values have a *residual distribution*, which

is sometimes taken to be the standard normal distribution, but which may in principle be anything. Specification of the residual distribution or clearly announcing the lack of such a specification is a part of the location-scale model. ⋄

Example 1.2.8 (Local alignment)**.** Local alignment of DNA-sequences or proteins is one of the classical tools for bioinformaticians and molecular biologists. Various implementations that solve the algorithmic problem, going back to the seminal paper by Smith and Waterman, of optimally locally aligning two sequences – with a given scoring scheme – have been around for quite some time. We will assume some familiarity with the algorithmic side or at least a knowledge of what a local alignment is. There is also an extensive literature on the probabilistic and statistical aspects of analyzing the outcome of a local alignment. The central problem is to understand the distribution of the optimal local alignment score if we align functionally and/or evolutionary unrelated proteins, say. Otherwise, we have no way of judging whether a given optimal local alignment expresses a true relationship between the proteins or whether it is what we should expect to obtain for unrelated proteins.

If we take a bunch of equal length unrelated proteins, we are essentially in the framework of the location-scale model as discussed. As we will illustrate, there is also a quite good understanding of the residual distribution, which for relevant scoring schemes can be taken to be the *Gumbel distribution*. The location and scale parameters will then reflect such things as residue composition, the specific scoring scheme – do we use BLOSUM60 or PAM250, the gap penalties etc. – and the length of the proteins. We can hardly restrict our attention to aligning proteins of a single length only, so we will be particularly interested in how the parameters depend upon protein length. This turns the statistical model into a regression model where we not only want to understand the location and scale for a single given length but actually how they vary with the lengths of the proteins. More ambitiously we can aim for a model that also captures the dependence upon residue composition and scoring scheme. ⋄

Example 1.2.9 (Gene expression)**.** A gene expression microarray is an experimental technique for measuring the expression level of all genes in a cell (culture) at a given time. It is technically speaking a *multiplex* assay – see the discussion about assays above – as we measure simultaneously the concentration of a large number of substances at the same time.

To make a long story short, one locates on a plate or slide (the microarray) a number of DNA-fragments – either small synthetic oligonucleotide probes or spots of larger DNA sequences obtained by PCR-amplification of cDNA clones. Then one extracts mRNA from cells, which is typically reverse transcribed into cDNA, and the cDNA is poured onto the slide. The cDNA-molecules bind to the DNA-molecules on the slide that are complementary by hydrogen bonds, and if the cDNA has been color labeled in advance, it is possible to read of how much cDNA is attached to a given probe on the plate. This gives an (indirect) measure of the expression level of the corresponding gene in the cell. The initial measurement that comes out of such

an experiment is an image, which shows light intensities measured over the slide. The image is preprocessed to yield a single light intensity measurement for each probe on the slide. It should be emphasized that there are numerous details in the experimental design that can give rise to errors of very different nature. We are not going to focus on these details, but one remark is appropriate. For this technology it is not common to infer standard curves from data for each array. Consequently it is not possible to obtain absolute concentration measurements only relative measurements. In the comparison between measurements on different arrays, elaborate normalization techniques are necessary to replace the lack of knowledge of standard curves. Still we will only be able to discuss relative differences between measurements across arrays and not absolute differences.

The resulting dataset from a microarray experiment is a high-dimensional vector of measurements for each gene represented on the array. A typical experimental design will include several arrays consisting of measurements for several individuals, say, that may be grouped for instance according to different disease conditions. The microarray technology presents us with a lot of statistical challenges. High-dimensional data, few replications per individual (if any), few individuals per group etc., and all this give rise to high uncertainty in the statistical procedures for inference from the data. This is then combined with the desire to ask biologically quite difficult questions. Indeed a challenging task. We will in these notes mostly use microarray data to illustrate various simple models on a gene-by-gene level. That is, where we essentially consider each gene represented on the array independently. This is an important entry to the whole area of microarray data analysis, but these considerations are only the beginning of a more serious analysis.

The use of microarrays for gene expression measurements is a well developed technique. It is, however, challenged by other experimental techniques. At the time of writing the so-called next-generation sequencing techniques promise to replace microarrays in the near future. Though fundamentally different from a technical point of view, the nature of the resulting data offer many of the same challenges as mentioned above – high-dimensional data with few replications per individual, and still the biological questions are difficult. The hope is that the sequencing based techniques have smaller measurement noise and that the need for normalization is less pronounced. In other words, that we get closer to a situation where a single standard curve is globally valid. Whether this is actually the case only time will show.   ⬦

2

Probability Theory

2.1 Introduction

Probability theory provides the foundation for doing statistics. It is the mathematical framework for discussing experiments with an outcome that is uncertain. With probability theory we capture the mathematical essence of the quantification of uncertainty by abstractly specifying what properties such a quantification should have. Subsequently, based on the abstract definition we derive properties about probabilities and give a number of examples. This approach is *axiomatic* and mathematical and the mathematical treatment is self-contained and independent of any interpretation of probabilities we might have. The interpretation is, however, what gives probability theory its special flavor and makes it applicable. We give a mathematical presentation of probability theory to develop a proper language, and to get accustomed to the vocabulary used in probability theory and statistics. However, we cannot and will not try to derive everything we need along the way. Derivations of results are made when they can illustrate how we work with the probabilities and perhaps illuminate the relation between the many concepts we introduce.

2.2 Sample spaces

We will throughout use E to denote a set, called the *sample space*, such that elements $x \in E$ represent the outcome of an experiment we want to conduct. We use small letters like x, y, z to denote elements in E. An *event*, $A \subseteq E$, is a subset of E, and we will use capital letters like A, B, C to denote events. We use the word experiment in a wide sense. We may have a real wet lab experiment in mind or another classical empirical data collection process in mind. But we may also have a database search or some other algorithmic treatment of existing data in mind – or even an experiment

carried out entirely on a computer, that is, a computer simulation.

The notion of a general sample space may seem abstract and therefore difficult. In practice the choice of sample space is often quite natural. If we throw a dice the sample space is the set $\{\boxdot, \boxdot, \boxdot, \boxdot, \boxdot, \boxdot\}$. When modeling random nucleic acids (DNA) the sample space is $\{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$. These are both examples of finite sample spaces. A very common, infinite sample space is the set of real numbers, $\mathbb{R}$, which we use as sample space in most cases where we measure a single, quantitative variable. That could be temperature, pH-value, pressure, concentration of a chemical compound, weight or height of an individual, and many, many other things.

A more complicated sample space is found if we for instance consider mass spectrometry. In mass spectrometry one measures the amount of ions in a sample as a function of the ratio between the mass and the charge (the m/z-ratio) of ions, which produces a *mass spectrum*. The sample space is a set of functions – the set of potential mass spectra. We will typically regard the mass spectrum to be a continuous function of the m/z-ratio, and the sample space is thus the set of continuous functions on an appropriate interval of m/z-values.
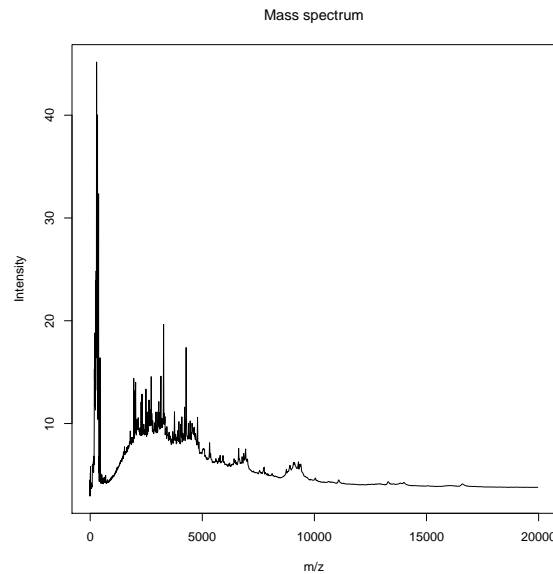
Figure 2.1: A raw mass spectrum. The sample is regarded as a (peaky and oscillating) continuous intensity of molecules as a function of the ratio between the mass and charge of the molecule (the m/z-ratio).

Microarray experiments constitute another class of experiments where the outcome takes its value in a complicated sample space. When the microarray is exposed to a sample of fluorescent RNA-molecules (in fact, cDNA reverse transcribed from the RNA) specific molecules prefer to bind to specific probes, and by subsequently

scanning the array one obtains a light intensity as a function of the position on the slide. The light intensity in the vicinity of a given probe is a measure of the amount of RNA that binds to the probe and thus a measure of the (relative) amount of that particular RNA-molecule in the sample. The sample space is a set of two-dimensional functions – the set of potential light intensities as a function of position on the slide. In most cases the actual outcome of the experiment is not stored, but only some discretized version or representation of the outcome. This is mostly due to technical reasons, but it may also be due to preprocessing of the samples by computer programs. For microarrays, the light intensities are first of all stored as an image of a certain resolution, but this representation is typically reduced even further to a single quantity for each probe on the array.

Though the choice of sample space in many cases is given by the experiment we consider, we may face situations where several choices seem appropriate. The raw data from a microarray experiment is an image, but do we want to model the actual image produced from the scanning? Or are we satisfied with a model of the summarized per probe measurements? In particular when we encounter complicated measurement processes we may in practice need different preprocessing or adaptation steps of the collected data before we actually try to model and further analyze the data. It is always good practice then to clearly specify how these preprocessing steps are carried out and what the resulting final sample space and thus the actual data structure is.

2.3   Probability measures

If we are about to conduct an uncertain experiment with outcome in the sample space E we use probabilities to describe the result of the experiment prior to actually performing the experiment. Since the outcome of the experiment is uncertain we cannot pinpoint any particular element $x \in E$ and say that x will be the outcome. Rather, we assign to any event $A \subseteq E$ a measure of how likely it is that the event will occur, that is, how likely it is that the outcome of the experiment will be an element x belonging to A.

Definition 2.3.1. *A probability measure P assigns to all events $A \subseteq E$ a value $P(A) \in [0, 1]$. The probability measure P must fulfill that:*

(i) $P(E) = 1$,

(ii) *if $A_1, \ldots, A_n$ are disjoint events*

$$P(A_1 \cup \ldots \cup A_n) = P(A_1) + \ldots + P(A_n).$$

Property (ii) above is known as *additivity* of a probability measure. It is crucial that the events are disjoint for additivity to hold.

We will throughout use the name *probability distribution*, or just *distribution*, interchangeably with probability measure. We can immediately derive a number of useful and direct consequences of the definition.

For any event A, we can write $E = A \cup A^c$ with A and A^c disjoint, hence by additivity

$$1 = P(E) = P(A) + P(A^c).$$

From this we obtain

$$P(A^c) = 1 - P(A).$$

In particular, for the empty event $\emptyset$ we have

$$P(\emptyset) = 1 - P(E) = 0.$$

If $A, B \subseteq E$ are two events such that $A \subseteq B$ we have that

$$B = A \cup (B \backslash A)$$

where A and $B \backslash A$ are disjoint. By additivity again we obtain that

$$P(B) = P(A) + P(B \backslash A),$$

hence since $P(B \backslash A) \geq 0$ it follows that if $A \subseteq B$ then

$$P(A) \leq P(B). \tag{2.1}$$

Finally, if $A, B \subseteq E$ are any two events – not necessarily disjoint – then with $C = A \cap B$ we have that $A = (A \backslash C) \cup C$ with $(A \backslash C)$ and C disjoint, thus by additivity

$$P(A) = P(A \backslash C) + P(C).$$

Moreover,

$$A \cup B = (A \backslash C) \cup B$$

with the two sets on the right hand side being disjoint, thus by additivity again

$$\begin{aligned} P(A \cup B) &= P(A \backslash C) + P(B) \\ &= P(A) + P(B) - P(A \cap B). \end{aligned} \tag{2.2}$$

Intuitively speaking, the result states that the probability of the union $A \cup B$ is the sum of the probabilities for A and B, but when the sets are *not* disjoint we have "counted" the probability of the intersection $A \cap B$ twice. Thus we have to subtract it.

We summarize the results derived.

Result 2.3.2. *Properties of probability measures:*

Math Box 2.3.1 (General Probability Measures). In Definition 2.3.1 we only require that the probability measure P should be additive in the sense that for a *finite* sequence $A_1, \ldots, A_n$ of disjoint sets we have that

$$P(A_1 \cup \ldots \cup A_n) = P(A_1) + \ldots + P(A_n).$$

Probability measures are usually required to be σ-*additive*, meaning that for any *infinite* sequence $A_1, A_2, A_3, \ldots$ of disjoint sets it holds that

$$P(A_1 \cup A_2 \cup A_3 \cup \ldots) = P(A_1) + P(A_2) + P(A_3) + \ldots. \tag{2.3}$$

It is a perfectly natural requirement and as it stands, it may seem as a quite innocent extension. If P for instance is a probability measure on a countably infinite set E given by point probabilities $(p(x))_{x \in E}$, it may be observed that (2.3) is fulfilled *automatically* by the definition of P from point probabilities. Requiring that P is σ-additive is, however, a more serious business when dealing with probability measures on non-discrete sample spaces.

It is, moreover, problematic to assign a probability to *every* subset of E. In general one needs to restrict attention to a collection of subsets $\mathcal{E}$, which is required to contain E and to fulfill

- if $A \in \mathcal{E}$ then $A^c \in \mathcal{E}$
- if $A_1, A_2, \ldots \in \mathcal{E}$ then $A_1 \cup A_2 \cup \ldots \in \mathcal{E}$

A collection $\mathcal{E}$ of subsets fulfilling those requirements is called a σ-*algebra* and the sets in $\mathcal{E}$ are called measurable. If one doesn't make this restriction a large number of "natural" probability measures don't exist. The most commonly occurring σ-algebras contain so many sets that it requires sophisticated mathematical arguments to show that there indeed exist sets *not* in the σ-algebra. From a practical, statistical point of view it is unlikely that we ever encounter non-measurable sets, which we really want to assign a probability. However, in some areas of mathematical statistics and probability theory, measurability problems are present all the time – especially problems with verifying that sets of interest are actually measurable.

- *For any event $A \subseteq E$,*
$$P(A^c) = 1 - P(A).$$

- $P(\emptyset) = 0$

- *For two events $A \subseteq B \subseteq E$,*
$$P(B) = P(A) + P(B \backslash A) \quad \text{and} \quad P(A) \leq P(B).$$

- *For any two events $A, B \subseteq E$,*
$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

Note that the abstract definition of a probability measure doesn't say anything about how to *compute* the probability of an event in a concrete case. But we are on the other hand assured that if we manage to come up with a probability measure, it assigns a probability to any event, no matter how weird it may be. Even in cases where there is no chance that we would ever be able to come up with an analytic computation of the resulting probability. Moreover, any general relation or result derived for probability measures, such as those derived above, apply to the concrete situation.

The assignment of a probability $P(A)$ to any event A is a quantification of the uncertainty involved in our experiment. The closer to one $P(A)$ is the more certain it is that the event A occurs and the closer to zero the more uncertain. Some people, especially those involved in gambling, find it useful to express uncertainty in terms of odds. Given a probability measure P we define for any event A the odds that A occurs as

$$\xi(A) = \frac{P(A)}{1 - P(A)} = \frac{P(A)}{P(A^c)}. \tag{2.4}$$

Thus we assign to the event $A \subseteq E$ a value $\xi(A) \in [0, \infty]$, and like the probability measure this provides a quantification of the uncertainty. The larger $\xi(A)$ is the more certain it is that A occurs. A certain event (when $P(A) = 1$) is assigned odds ∞. It follows that we can get the probabilities back from the odds by the formula

$$P(A) = \frac{1}{1 + \xi(A^c)}. \tag{2.5}$$

Odds are used in betting situations because the odds tell how fair[1] bets should be constructed. If two persons, player one and player two, say, make a bet about whether event A or event A^c occurs, how much should the loser pay the winner for this to be a fair bet? If player one believes that A occurs and is willing to bet 1 kroner then for the bet to be fair player two must bet $\xi(A^c)$ kroner on event A^c. For gambling, this is the way British bookmakers report odds – they say that odds for event A are $\xi(A^c)$ to 1 against. With 1 kroner at stake and winning you are paid $\xi(A^c)$ back in addition to the 1 kroner. Continental European bookmakers report the odds as $\xi(A^c) + 1$, which include what you staked.

The frequency interpretation states that the probability of an event A should equal the long run frequency of the occurrence of event A if we repeat the experiment indefinitely. That is, suppose that we perform n independent and identical experiments all governed by the probability measure P and with outcome in the sample space E, and suppose that we observe $x_1, \ldots, x_n$. Then we can compute the relative frequency of occurrences of event A

$$\varepsilon_n(A) = \frac{1}{n} \sum_{i=1}^{n} 1_A(x_i). \tag{2.6}$$

---

[1]Fair means that on average the players should both win (and loose) 0 kroner, cf. the frequency interpretation.

Here 1_A is the indicator function for the event A, so that $1_A(x_i)$ equals 1 if $x_i \in A$ and 0 otherwise. We sometimes also write $1(x_i \in A)$ instead of $1_A(x_i)$. We see that $\varepsilon_n(A)$ is the fraction of experiments in which the event A occurs. As n grows large, the frequency interpretation says that $\varepsilon_n(A)$ must be approximately equal to $P(A)$. Note that this is not a mathematical result! It is the interpretation of what we want the probabilities to mean. To underpin the interpretation we can show that the mathematical theory based on probability measures really is suitable for approximating relative frequencies from experiments, but that is another story. We also need to make a precise definition of what we mean by *independent* in the first place.

The frequency interpretation provides the rationale for using odds in the construction of fair bets. If the two players repeat the same bet n times with n suitably large, and if the bet is fair according to the definition above with player one betting 1 kroner on event A each time, then in the i'th bet, player one wins

$$\xi(A^c)1_A(x_i) - 1_{A^c}(x_i),$$

because this equals $\xi(A^c)$ if event A comes out and -1 if A^c comes out. Considering all n bets, and excluding the case $P(A) = 0$, player one will on average win

$$
\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\left[\xi(A^c)1_A(x_i) - 1_{A^c}(x_i)\right] &= \xi(A^c)\varepsilon_n(A) - \varepsilon_n(A^c) \\
&\simeq \xi(A^c)P(A) - P(A^c) \\
&= \frac{P(A^c)}{P(A)}P(A) - P(A^c) = 0.
\end{aligned}
$$

Likewise, player two will on average win approximately 0 kroner if n is sufficiently large. This is why we say that the bet is fair. Note, however, that the total sum that player one wins/loses is

$$\sum_{i=1}^{n}\left[\xi(A^c)1_A(x_i) - 1_{A^c}(x_i)\right],$$

which may deviate substantively from 0 – and the deviation will become larger as n increases.

The Bayesian interpretation allows *us* to choose the probability measure describing the experiment subjectively. In this interpretation the probability measure is not given objectively by the experiment but reflects our minds and our knowledge of the experiment before conducting it. One theoretical justification of the Bayesian interpretation is that we play a mind game and make everything into a betting situation. Thus we ask ourselves which odds (for all events) *we* believe to be fair in a betting situation before conducting the experiment. Note that this is an entirely subjective process – there is no theory dictating what fairness means – but we are

nevertheless likely to have an opinion about what is fair and what is not. It is possible to show that if we make up our minds about fair odds in a consistent manner[2] we necessarily end up with a probability measure defined by (2.5). This is the so-called *Dutch book argument.* The probabilities don't represent the long term relative frequencies when repeating the experiment. After having conducted the experiment once we have gained new information, which we might want to take into account when deciding what we then believe to be fair odds.

The two interpretations are fundamentally different of nature and this has consequently lead to different opinions in the statistical literature about how to develop a suitable and well-founded statistical theory. The methodologies developed based on either interpretation differ a lot – at least in principle. On the other hand there are many practical similarities and most Bayesian *methods* have a frequency interpretation and vice versa. Discussions about which of the interpretations, if any, is the correct one is of a philosophical and meta mathematical nature – we cannot prove that either interpretation is correct, though pros and cons for the two interpretations are often based on mathematical results. We will not pursue this discussion further. The interested reader can find a balanced an excellent treatment in the book *Comparative Statistical Inference, Wiley,* by Vic Barnett.

Throughout, probabilities are given the frequency interpretation, and methodology is justified from the point of view of the frequency interpretation. This does not by any means rule out the use of Bayesian methodology a priori, but the methodology must then be justified within the framework of the frequency interpretation.

Example 2.3.3 (Coin Flipping)**.** When we flip a coin it lands with either heads or tails upwards. The sample space for describing such an experiment is

$$E = \{\text{heads}, \text{tails}\}$$

with the four events:

$$\emptyset, \ \{\text{heads}\}, \ \{\text{tails}\}, \ E.$$

Any probability measure P on E must fulfill $P(\emptyset) = 0$ and $P(E) = 1$ by definition together with

$$P(\text{tails}) = 1 - P(\text{heads}),$$

hence the probability measure is completely determined by $p = P(\text{heads})$. Note here the convention that when we consider an event consisting of only a single outcome like {heads} we usually drop the curly brackets. This extremely simple experiment is not so interesting in itself, but it serves the purpose of being a fundamental building block for many more interesting sample spaces and probability measures. Maybe it is not coin flipping directly that we use as a building block but some other binary experiment with two possible outcomes. There are several often encountered ways of *coding* or *naming* the outcome of a binary experiment. The most commonly used

---

[2]Not being consistent means that if someone knows our choice of fair odds he can construct a bet in such a way that he will win money with certainty.

sample space for encoding the outcome of a binary experiment is $E = \{0, 1\}$. When using this naming convention we talk about a *Bernoulli experiment*. In the coin flipping context we can let 1 denote heads and 0 tails, then if $x_1, \ldots, x_n$ denote the outcomes of n flips of the coin we see that $x_1 + \ldots + x_n$ is the total number of heads. Moreover, we see that

$$P(x) = p^x(1-p)^{1-x} \tag{2.7}$$

because either $x = 1$ (heads) in which case $p^x = p$ and $(1-p)^{1-x} = 1$, or $x = 0$ (tails) in which case $p^x = 1$ and $(1-p)^{1-x} = 1 - p$. $\diamond$

Exercises

Exercise 2.3.1. If $P(A) = 0.5$, $P(B) = 0.4$ and $P(A \cup B) = 0.6$, compute $P((A \cap B)^c)$.

Exercise 2.3.2. Compute the odds $\xi(A)$ of the event A if $P(A) = 0.5$, if $P(A) = 0.9$ and if $P(A^c) = 0.9$. Compute the probability $P(A)$ of the event A if $\xi(A) = 10$ and if $\xi(A) = 2$.

Exercise 2.3.3. Consider the sample space $E = \{1, \ldots, n\}$ and let $x_i(k) = 1$ if $i = k$ and $= 0$ if $i \neq k$. Show that

$$P(k) = P(1)^{x_1(k)} \cdot P(2)^{x_2(k)} \cdot \ldots \cdot P(n)^{x_n(k)}.$$

2.4 Probability measures on discrete sets

If E is an infinite set we may be able to arrange the elements in E in a sequential way;

$$E = \{x_1, x_2, x_3, \ldots\}.$$

That is, we may be able to give each element in E a positive integer index such that each $i \in \mathbb{N}$ corresponds to exactly one element $x_i \in E$. If we can do so, we say that E is *countably infinite*. One example of a countably infinite set is the positive integers themselves, $\mathbb{N} = \{1, 2, 3, \ldots\}$. Also the integers $\mathbb{Z} = \{\ldots, -2, -1, 0, 1, 2, \ldots\}$ is a countably infinite set, and perhaps more surprising the rational numbers $\mathbb{Q}$ is a countably infinite set. The set of real numbers is, however, not countable. This is proved using the famous diagonal argument by George Cantor.

It is particularly simple to define probability measures on finite and countably infinite sets and this is the subject of the present section. Defining probability measures on the real line is a more subtle problem, which we defer to the following section.

Definition 2.4.1. *We call a set E discrete if it is either finite or countably infinite.*

The DNA, RNA and amino acid alphabets are three examples of finite and hence discrete sets.

Example 2.4.2. Considering any finite set E we can define the set

$$E^* = \{x_1 x_2 \ldots x_n \mid n \in \mathbb{N}, x_i \in E\},$$

which is the set of all sequences of finite length from E. We claim that E^* is discrete. If E is the DNA-alphabet, say, it is clear that there is an infinite number of DNA-sequences of finite length in E^*, but it is no problem to list them as a sequence in the following way:

$$\texttt{A}, \texttt{C}, \texttt{G}, \texttt{T}, \texttt{AA}, \texttt{AC}, \texttt{AG}, \texttt{AT}, \texttt{CA}, \ldots, \texttt{TG}, \texttt{TT}, \texttt{AAA}, \texttt{AAC}, \ldots$$

Hence we first list all sequences of length one, then those of length two, those of length three, four, five and so on and so forth. We can encounter the use of E^* as sample space if we want a probabilistic model of all protein coding DNA-sequences.
$\diamond$

Example 2.4.3. In genetics we find genes occurring in different versions. The versions are referred to as *alleles*. If a gene exists in two versions it is quite common to refer to the two versions as a and A or b and B. Thus if we sample an individual from a population and check whether the person carries allele a or A we have the outcome taking values in the sample space

$$E = \{a, A\}.$$

However, we remember that humans are diploid organisms so we actually carry two copies around. Thus the real outcome is an element in the sample space

$$E = \{aa, Aa, AA\}.$$

The reason that aA is not an outcome also is simply that we cannot distinguish between aA and Aa. The information we can obtain is just that there is an a and an A, but not in any particular order.

If we also consider another gene simultaneously, the sample we get belongs to the sample space

$$E = \{aabb, aaBb, aaBB, Aabb, AaBb, AaBB, AAbb, AABb, AABB\}$$

that contains nine elements.
$\diamond$

To define a probability measure on E in principle requires that we assign a number, $P(A)$, to every event $A \subseteq E$. For a finite set of size n there are 2^n different events so even for sets of quite moderate size, $n = 20$, say, to list $2^{20} = 1048576$ probabilities is not a very practical way of defining a probability measure – not to mention the problems we would get checking that the additivity property is fulfilled. Fortunately there is another and somewhat more tractable way of defining a probability measure on a discrete set.

Definition 2.4.4. *If E is discrete we call $(p(x))_{x \in E}$ a vector of point probabilities indexed by E if $0 \leq p(x) \leq 1$ and*

$$\sum_{x \in E} p(x) = 1. \tag{2.8}$$

We define a probability measure P on E with point probabilities $(p(x))_{x \in E}$ by

$$P(A) = \sum_{x \in A} p(x) \tag{2.9}$$

for all events $A \subseteq E$.

The sum as written above over all $x \in E$ can be understood in the following way: Since E is discrete it is either finite or countably infinite. In the former case it is just a finite sum so we concentrate on the latter case. If E is countable infinite we know that we can order the elements as $x_1, x_2, x_3, \ldots$. Then we simply define

$$\sum_{x \in E} p(x) = \sum_{n=1}^{\infty} p(x_n).$$

It even holds that the infinite sum on the right hand side is a sum of all positive numbers. The careful reader may get worried here, because the arrangement of the elements of E is arbitrary. What if we choose another way of listing the elements will that affect the sum? The answer to this is no, because the terms are all positive, in which case we can live happily with writing $\sum_{x \in E} p(x)$ without being specific on the order of summation. The same does not hold if the terms can be positive as well as negative, but that is not a problem here.

Example 2.4.5 (Random Amino Acids). As mentioned in the introduction a sequence of nucleic acids may be regarded as being random if we don't know any better. Likewise, a sequence of amino acids (a protein) is regarded as being random if we don't know any better. In these notes a protein is simply a finite sequence from the sample space

$$E = \big\{\texttt{A}, \texttt{R}, \texttt{N}, \texttt{D}, \texttt{C}, \texttt{E}, \texttt{Q}, \texttt{G}, \texttt{H}, \texttt{I}, \texttt{L}, \texttt{K}, \texttt{M}, \texttt{F}, \texttt{P}, \texttt{S}, \texttt{T}, \texttt{W}, \texttt{Y}, \texttt{V}\big\}.$$

To specify a probability measure on E we have to specify point probabilities $p(x)$ for $x \in E$ according to 2.4.4. A valid choice is

$$p(x) = 0.05$$

for all $x \in E$. Clearly $p(x) \in [0, 1]$ and since there are 20 amino acids

$$\sum_{x \in E} p(x) = 20 \times 0.05 = 1.$$

R Box 2.4.1 (Vectors)**.** A fundamental data structure in R is a `vector` of e.g. integers or reals. A vector of real valued numbers can be typed in by a command like

```
> x <- c(1.3029,4.2,5.3,3.453,342.34)
```

This results in a vector `x` of length five. Even a single number is always regarded as a vector of length one in R. The `c` above should therefore be understood as a *concatenation* of five vectors of length one. If we define another vector of length two by

```
> y <- c(1.3,2.3)
```

we can concatenate `x` and `y` to give the vector

```
> z <- c(x,y)
```

Then `z` is a vector of length seven containing the numbers $1.3029, 4.2, 5.3, 3.453, 342.34, 1.3, 2.3$. A *probability vector* in R is simply a vector of positive real numbers that sum to one.

Under this probability measure all amino acids are equally likely, and it is known as the uniform distribution on E, cf. the example below. A more reasonable probability measure on E is given by the relative frequencies of the occurrence of the amino acids in real proteins, cf. the frequency interpretation of probabilities. The Robinson-Robinson probabilities come from a survey of a large number of proteins. They read

Amino acid	Probability
A	0.079
R	0.051
N	0.045
D	0.054
C	0.019
E	0.063
Q	0.043
G	0.074
H	0.022
I	0.051
L	0.091
K	0.057
M	0.022
F	0.039
P	0.052
S	0.071
T	0.058
W	0.013
Y	0.032
V	0.064

Using the Robinson-Robinson probabilities, some amino acids are much more probable than others. For instance, Leucine (L) is the most probable with probability 0.091 whereas Tryptophan (W) is the least probable with probability 0.013. ◇

Example 2.4.6 (Uniform distribution)**.** If E is a finite set containing n elements we can define a probability measure on E by the point probabilities

$$p(x) = \frac{1}{n}$$

for all $x \in E$. Clearly $p(x) \in [0, 1]$ and $\sum_{x \in E} p(x) = 1$. This distribution is called the *uniform distribution* on E.

If P is the uniform distribution on E and $A \subseteq E$ is any event it follows by the definition of P that

$$P(A) = \sum_{x \in A} \frac{1}{n} = \frac{|A|}{n}$$

with $|A|$ denoting the number of elements in A. Since the elements in E are all possible but we only regard those in A as favorable, this result gives rise to the formula

$$P(A) = \frac{\text{number of favourable outcomes}}{\text{number of possible outcomes}},$$

which is valid only when P is the uniform distribution. Even though the formula looks innocent, it can be quite involved to apply it in practice. It may be easy to specify the sample space E and the favorable outcomes in the event A, but counting the elements in A can be difficult. Even counting the elements in E can sometimes be difficult too. ◇

R Box 2.4.2 (Biostrings). The package `Biostrings` comes as part of the Bioconductor standard bundle of packages for R. With

```
> library(Biostrings)
```

you get access to some representations of biological sequences in R. For instance

```
> someDNA <- DNAString("ACGTACGTACTT")
```

constructs a DNAString object. You can also make an `RNAString` or a generic biostring using `BString`. There is a `letter` function available to extract letters at specific positions, a `matchPattern` and a `countPattern` function, and there is an `alphabetFrequency` function. There is also a `read-FASTA` function.

The package is developed for handling biological sequences in particular, but may also be suitable in other situations for handling and analyzing large stings in R. The `BString` class is definitely more flexible than the built in R strings, making the manipulation of especially large strings easier and more efficient.
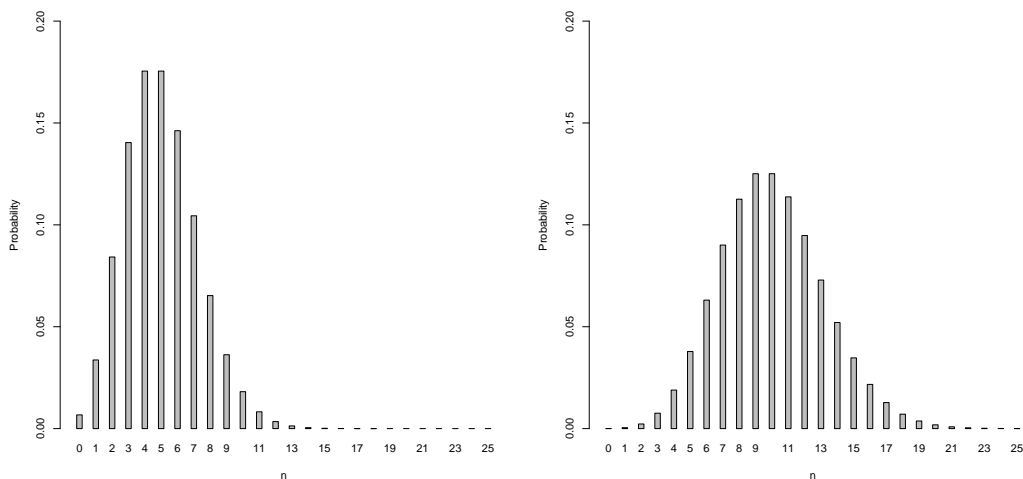

Figure 2.2: The point probabilities for the Poisson distribution with $\lambda = 5$ (left) and $\lambda = 10$ (right).

Example 2.4.7 (The Poisson Distribution). The (infinite) Taylor expansion of the

exponential function is given as

$$\exp(\lambda) = \sum_{n=0}^{\infty} \frac{\lambda^n}{n!}, \quad \lambda \in \mathbb{R}.$$

If $\lambda > 0$ the numbers

$$p(n) = \exp(-\lambda)\frac{\lambda^n}{n!}$$

are positive and

$$\sum_{n=0}^{\infty} \exp(-\lambda)\frac{\lambda^n}{n!} = \exp(-\lambda)\sum_{n=0}^{\infty} \frac{\lambda^n}{n!} = \exp(-\lambda)\exp(\lambda) = 1.$$

Hence $(p(n))_{n \in \mathbb{N}_0}$ can be regarded as point probabilities for a probability measure on the non-negative integers $\mathbb{N}_0$. This distribution is called the *Poisson distribution*.

A typical application of the Poisson distribution is as a model for the number of times a rather unlikely event occurs in a large number of replications of an experiment. One can, for instance, use the Poisson distribution as a model of the number of (non-overlapping) occurrences of a given sequence pattern in a long text string. This has found applications in biological sequence analysis and data mining of text collections.

$\diamond$

2.5 Descriptive methods

If we have observations $x_1, \ldots, x_n$ from a probability measure P on a discrete sample space E, we want to summarize the observations in a form that is somewhat comprehensible. A *tabulation* of the data consists of counting the number of occurrences, n_x, of the outcome $x \in E$ among the n outcomes $x_1, \ldots, x_n$. In other words

$$n_x = \sum_{i=1}^{n} 1(x_i = x)$$

is the frequency of the occurrence of x in the sample. Note the relation between the tabulation of (absolute) frequencies and the computation of the relative frequencies $\varepsilon_n(x)$ is

$$\varepsilon_n(x) = \frac{n_x}{n}.$$

For variables with values in a discrete sample space we have in general only the possibility of displaying the data in a summarized version by tabulations. Depending on the structure of the sample space E the tables can be more or less informative. If E is a product space, that is, if the data are multivariate, we can consider marginal tables for each coordinate as well as multivariate tables where we cross-tabulate
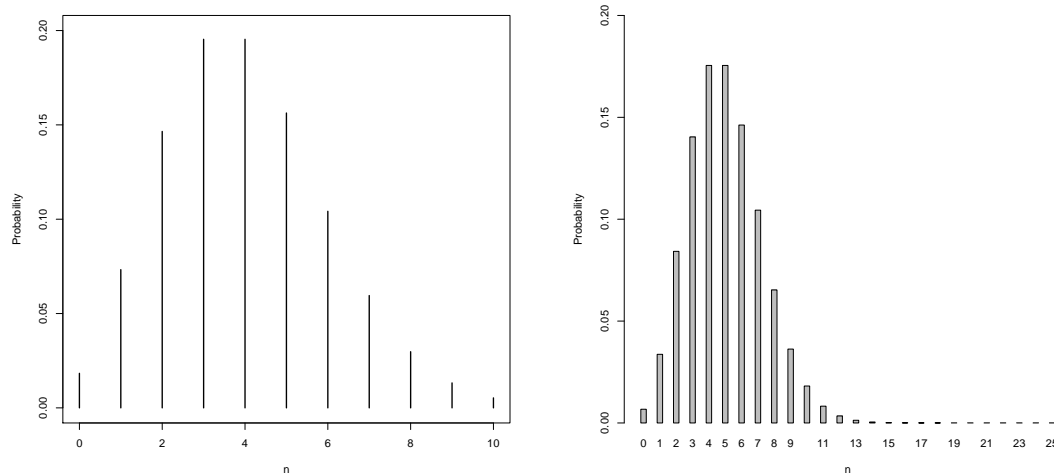
Figure 2.3: Example of a theoretical bar plot. Here we see the Poisson point probabilities ($\lambda = 4$) plotted using either the `plot` function with `type="h"` (left) or the `barplot` function (right).

two or more variables. Tables in dimension three and above are quite difficult to comprehend.

For the special case with variables with values in $\mathbb{Z}$ we can also use bar plots to display the tabulated data. We compare the bar plot with a bar plot of the corresponding point probabilities $p(n)$ as a function of $n \in \mathbb{Z}$ to get a picture of where the measure is concentrated and what the "shape" of the measure is. Usually, the frequencies and point probabilities are plotted as horizontal bars. It is preferable to plot the relative frequencies in the bar plot so that the y-axis becomes comparable to the y-axis for the theoretical bar plot of the point probabilities.

2.5.1 Mean and variance

For probability measures on a discrete sample space $E \subseteq \mathbb{R}$ we can define the mean and variance in terms of the point probabilities.

Definition 2.5.1. *If P is a probability measure on E with point probabilities $(p(x))_{x \in E}$ that fulfill*

$$\sum_{x \in E} |x| p(x) < \infty$$

then we define the mean under P as

$$\mu = \sum_{x \in E} x p(x).$$

R Box 2.5.1 (Bar plots). We can use either the standard `plot` function in R or the `barplot` function to produce the bar plot. The theoretical bar plot of the point probabilities for the Poisson distribution with parameter $\lambda = 4$ is seen in Figure 2.3. The figure is produced by:

```
> plot(0:10, dpois(0:10, 4), type = "h", ylab = "Probability",
+       xlab = "n", ylim = c(0, 0.2), lwd = 2)
> barplot(dpois(0:10, 4), ylab = "Probability", xlab = "n",
+          space = 2, ylim = c(0, 0.2), names.arg = 0:10)
```

We can make R generate the 500 outcomes from the Poisson distribution and tabulate the result using `table`, and then we can easily make a bar plot of the frequencies. Figure 2.4 shows such a bar plot produced by:

```
> tmp <- table(rpois(500, 4))
> plot(tmp, ylab = "Frequency", xlab = "n")
> barplot(tmp, ylab = "Frequency", xlab = "n", space = 2)
```

One should note that the empirical bar plot – besides the y-axis – should look approximately like the theoretical bar plot. One may choose to normalize the empirical bar plot (use relative frequencies) so that it becomes directly comparable with the theoretical bar plot. Normalization can in particular be useful, if one wants to compare two bar plots from two samples of unequal size. Note that if we simulate another 500 times from the Poisson distribution, we get a different empirical bar plot.

If, moreover,

$$\sum_{x \in E} x^2 p(x) < \infty$$

we define the variance under P as

$$\sigma^2 = \sum_{x \in E} (x - \mu)^2 p(x).$$

The variance is on a quadratic scale, and we define the *standard deviation* as

$$\sigma = \sqrt{\sigma^2}.$$

Example 2.5.2. Let P be the uniform distribution on $\{1, \ldots, n\} \subseteq \mathbb{R}$. In this case all the sums that we will need to consider are sums over $\{1, \ldots, n\}$ only, so they are finite and well defined. We find that

$$\mu \;\; = \;\; \sum_{k=1}^{n} \frac{k}{n} = \frac{1}{n} \sum_{k=1}^{n} k = \frac{n+1}{2}$$
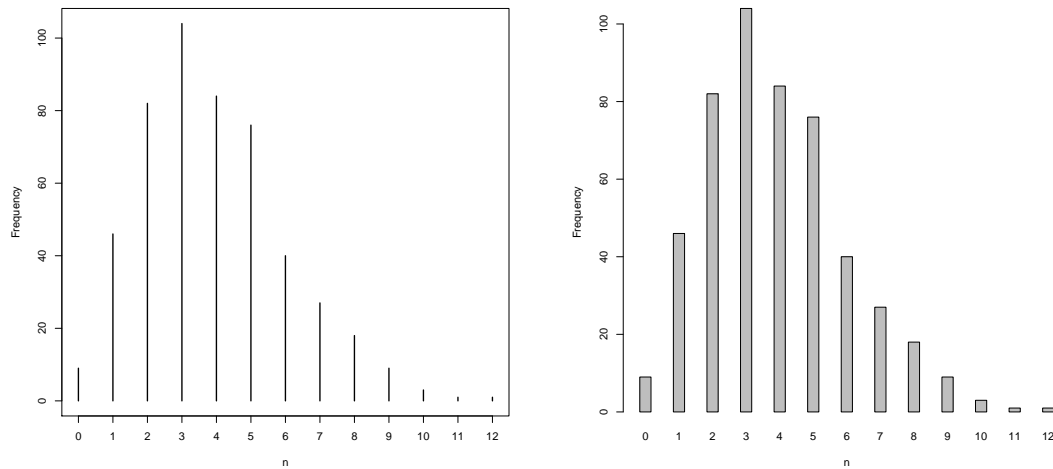
Figure 2.4: Example of an empirical bar plot. This is a plot of the tabulated values of 500 variables from the Poisson distribution ($\lambda = 4$) plotted using either the `plot` function (left) or the `barplot` function (right).

where we use the formula $1 + 2 + 3 + \ldots + (n - 1) + n = (n + 1)n/2$. Note that $\mu = (n+1)/2$ is exactly the midpoint between 1 and n, but note also the μ is not an integer if n is even. Thus the mean can very well be located outside of the sample space.

Regarding the variance we find that

$$
\begin{aligned}
\sigma^2 &= \sum_{k=1}^{n} \frac{(k - \mu)^2}{n} = \frac{1}{n} \sum_{k=1}^{n} \left(k^2 - 2k\mu + \mu^2 \right) \\
&= \frac{1}{n} \left(\sum_{k=1}^{n} k^2 - 2\mu \sum_{k=1}^{n} k + n\mu^2 \right) = \frac{(n + 1)(2n + 1)}{6} - \frac{(n + 1)^2}{4}
\end{aligned}
$$

where we have used the formula $1^2 + 2^2 + 3^2 + \ldots + (n-1)^2 + n^2 = n(n+1)(2n+1)/6$. Rearranging the result into a single fraction yields that the variance for the uniform distribution is

$$
\sigma^2 = \frac{(n + 1)(n - 1)}{12}.
$$

$\diamond$

If have a dataset $x_1, \ldots, x_n$ of observations from E the sample mean or average is

$$
\hat{\mu}_n = \frac{1}{n} \sum_{k=1}^{n} x_k.
$$

If the observations are all realizations from the same probability measure P the sample mean is an estimate of the (unknown) mean under P – provided there is a mean. Likewise, the sample variance

$$\tilde{\sigma}_n^2 = \frac{1}{n} \sum_{k=1}^{n} (x_k - \hat{\mu}_n)^2$$

is an estimate of the variance under P provided there is a variance.

Exercises

Exercise 2.5.1. Figure out what the following R-command does:
```
> tmp <- paste(sample(c("A","C","G","T"),1000,replace=T),
+ sep="",collapse="")
```
Hint: You may want to try it out and to consult the R-help for the functions `sample` and `paste`.

Exercise 2.5.2. An important function of DNA is as a blueprint for proteins. This translation of DNA to a protein works in triples of DNA-letters. A triple of letters from DNA is referred to as a *codon*. DNA-sequences that encode proteins start with the *start codon* `ATG` and stops with one of the three *stop codons* `TAA` , `TAG`, or `TGA`. In between there is a number of complete codons. Figure out how to find the number of codons in `tmp` before the first stop codon.
Hint: You can do *regular pattern matching* in R using `grep` or `regexpr`. With `gregexpr` you obtain a list of positions in the string.

Exercise 2.5.3. Compute for 1000 replications the number of codons in a random DNA-sequence that occur before the first stop codon. Explain why you sometimes get the length -1 (if you do that).
Hint: You may want to consider the `replicate` function.

Exercise 2.5.4. For the 1000 replications, produce a plot showing, as a function of n, the number of times that n codons occur before the first stop codon. That is, produce a bar plot of the relative frequencies of the number of codons before the first stop codon.
Hint: You can use the `table` function to summarize the vector produced in Exercise 2.5.3.

Exercise 2.5.5. How can we simulate under the probability measure on $E = \{A, C, G, T\}$ given by the point probabilities

$$P(A) = 0.3, \ P(C) = 0.1, \ P(G) = 0.4, \ P(T) = 0.2,$$

or

$$P(A) = 0.3, \ P(C) = 0.4, \ P(G) = 0.1, \ P(T) = 0.2.$$

What happens to the number of codons before the occurrence of the first stop codon?

Exercise 2.5.6. An *open reading frame* in a DNA-sequence is a segment of codons between a start and a stop codon. A long open reading frame is an indication that the open reading frame is actually coding for a protein since long open reading frames would be unlikely by chance. Discuss whether you believe that an open reading frame of length more than 33 is likely to be a protein coding gene.

★ **Exercise 2.5.7.** Compute the mean, variance and standard deviation for the uniform distribution on $\{1, \ldots, 977\}$.

★ **Exercise 2.5.8.** Show by using the definition of mean and variance that for the Poisson distribution with parameter $\lambda > 0$

$$\mu = \lambda \quad \text{and} \quad \sigma^2 = \lambda. \tag{2.10}$$

What is the standard deviation?

Exercise 2.5.9. The geometric distribution on $\mathbb{N}_0$ is given by the point probabilities

$$p(k) = p(1-p)^k$$

for $p \in (0,1)$ and $k \in \mathbb{N}_0$. Show that

$$\sum_{k=0}^{\infty} p(k) = 1.$$

Compute the mean under the geometric distribution for $p = 0.1, 0.2, \ldots, 0.9$.
Hint: If you are not able to compute a theoretical formula for the mean try to compute the value of the infinite sum approximately using a finite sum approximation – the computation can be done in R.

Exercise 2.5.10. Plot the point probabilities for the Poisson distribution, `dpois`, with $\lambda = 1, 2, 5, 10, 100$ using the `type="h"`. In all five cases compute the probability of the events $A_1 = \{n \in \mathbb{N}_0 \mid -\sigma \leq n - \mu \leq \sigma\}$, $A_2 = \{n \in \mathbb{N}_0 \mid -2\sigma \leq n - \mu \leq 2\sigma\}$, $A_3 = \{n \in \mathbb{N}_0 \mid -3\sigma \leq n - \mu \leq 3\sigma\}$.

Exercise 2.5.11. Generate a random DNA sequence of length 10000 in R with each letter having probability 1/4. Find out how many times the pattern `ACGTTG` occurs in the sequence.

Exercise 2.5.12. Repeat the experiment above 1000 times. That is, for 1000 sequences of length 10000 find the number of times that the pattern `ACGTTG` occurs in each of the sequences. Compute the average number of patterns occurring per sequence. Make a bar plot of the relative frequencies of the number of occurrences and compare with a theoretical bar plot of a Poisson distribution with λ chosen suitably.

2.6 Probability measures on the real line

Defining a probability measure on the real line $\mathbb{R}$ yields, to an even larger extent than in the previous section, the problem: How are we going to represent the assignment of a probability to all events in a manageable way? One way of doing so is through *distribution functions.*

Definition 2.6.1. *For a probability measure P on $\mathbb{R}$ we define the corresponding distribution function $F : \mathbb{R} \to [0,1]$ by*

$$F(x) = P((-\infty, x]).$$

That is, $F(x)$ is the probability that under P the outcome is less than or equal to x.

We immediately observe that since $(-\infty, y] \cup (y, x] = (-\infty, x]$ for $y < x$ and that the sets $(-\infty, y]$ and $(y, x]$ are disjoint, the additive property implies that

$$F(x) = P((-\infty, x]) = P((-\infty, y]) + P((y, x]) = F(y) + P((y, x]),$$

or in other words

$$P((y, x]) = F(x) - F(y).$$

We can derive more useful properties from the definition. If $x_1 \leq x_2$ then $(-\infty, x_1] \subseteq (-\infty, x_2]$, and therefore from (2.1)

$$F(x_1) = P((-\infty, x_1]) \leq P((-\infty, x_2]) = F(x_2).$$

Two other properties of F are consequences of what is known as *continuity of probability measures*. Intuitively, as x tends to $-\infty$ the set $(-\infty, x]$ shrinks towards the empty set $\emptyset$, which implies that

$$\lim_{x \to -\infty} F(x) = P(\emptyset) = 0.$$

Similarly, when $x \to \infty$ the set $(-\infty, x]$ grows to the whole of $\mathbb{R}$ and

$$\lim_{x \to \infty} F(x) = P(\mathbb{R}) = 1.$$

Likewise, by similar arguments, when $\varepsilon > 0$ tends to 0 the set $(-\infty, x + \varepsilon]$ shrinks towards $(-\infty, x]$ hence

$$\lim_{\varepsilon \to 0, \varepsilon > 0} F(x + \varepsilon) = P((-\infty, x]) = F(x).$$

We collect three of the properties derived for distribution functions.

Result 2.6.2. *A distribution function F has the following three properties:*

(i) *F is increasing: if $x_1 \leq x_2$ then $F(x_1) \leq F(x_2)$.*

(ii) *$\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$.*

(iii) *F is right continuous at any $x \in \mathbb{R}$: $\lim_{\varepsilon \to 0, \varepsilon > 0} F(x + \varepsilon) = F(x)$*

It is of course useful from time to time to know that a distribution function satisfies property (i), (ii), and (iii) in Result 2.6.2, but that these three properties completely characterize the probability measure is more surprising.

Result 2.6.3. *If $F : \mathbb{R} \to [0, 1]$ is a function that has property (i), (ii), and (iii) in Result 2.6.2 there is precisely one probability measure P on $\mathbb{R}$ such that F is the distribution function for P.*
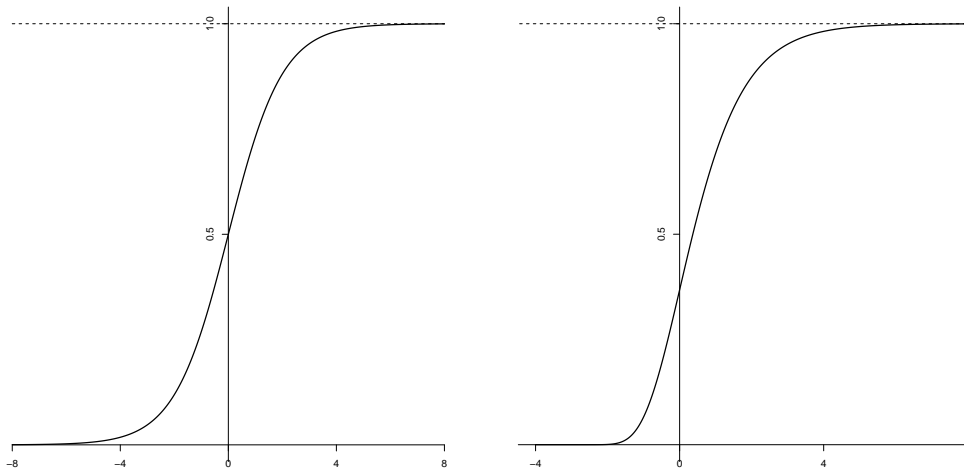
Figure 2.5: The logistic distribution function (left, see Example 2.6.4). The Gumbel distribution function (right, see Example 2.6.5). Note the characteristic S-shape of both distribution functions.

This result not only tells us that the distribution function completely characterizes P but also that we can specify a probability measure just by specifying its distribution function. This is a useful result but also a result of considerable depth, and a formal derivation of the result is beyond the scope of these notes.

Example 2.6.4 (Logistic distribution). The logistic distribution has distribution function

$$F(x) = \frac{1}{1 + \exp(-x)}.$$

The function is continuous, and the reader is encouraged to check that the properties of the exponential function ensure that also (i) and (ii) for a distribution function hold for this function. $\diamond$

Example 2.6.5 (The Gumbel distribution). The distribution function defined by

$$F(x) = \exp(-\exp(-x))$$

defines a probability measure on $\mathbb{R}$, which is known as the *Gumbel* distribution. We leave it for the reader to check that the function indeed fulfills (i), (ii) and (iii). The Gumbel distribution plays a role in the significance evaluation of local alignment scores, see Section 2.12. $\diamond$

If our sample space E is discrete but actually a subset of the real line, $E \subseteq \mathbb{R}$, like $\mathbb{N}$ or $\mathbb{Z}$, we have two different ways of defining and characterizing probability measures on E: through point probabilities or through a distribution function. The connection

is given by

$$F(x) = P((-\infty, x]) = \sum_{y \le x} p(y).$$
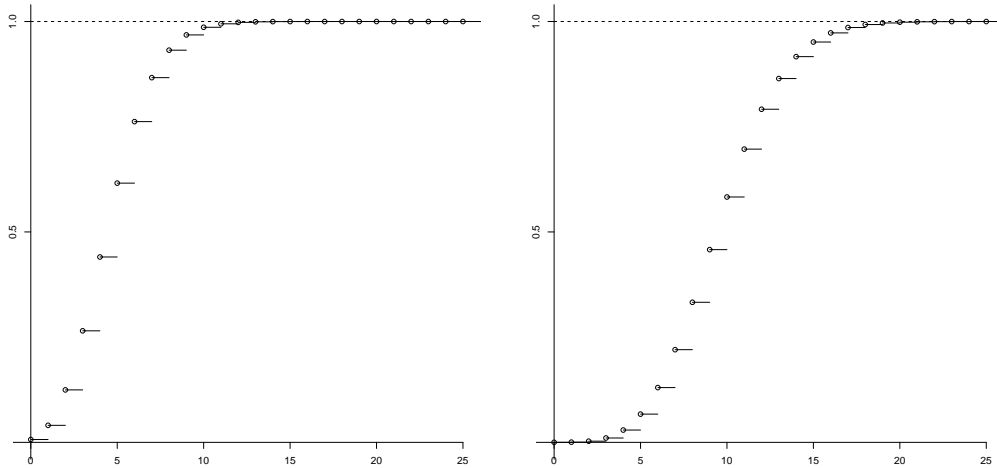
Figure 2.6: The distribution function for the Poisson distribution, with $\lambda = 5$ (left) and $\lambda = 10$ (right).

Example 2.6.6. The distribution function for the Poisson distribution with parameter $\lambda > 0$ is given by

$$F(x) = \sum_{n=0}^{\lfloor x \rfloor} \exp(-\lambda) \frac{\lambda^n}{n!}$$

where $\lfloor x \rfloor$ is the largest integer smaller than x. It is a step function with steps at each of the non-negative integers $n \in \mathbb{N}_0$ and step size at n being $p(n) = \exp(-\lambda) \frac{\lambda^n}{n!}$
$\diamond$

A number of distributions are defined in terms of a *density*. Not all probability measures on $\mathbb{R}$ have densities, e.g. those distributions that are given by point probabilities on $\mathbb{N}$. However, for probability measures that really live on $\mathbb{R}$, densities play to a large extent the same role as point probabilities do for probability measures on a discrete set.

Definition 2.6.7. *A probability measure P is said to have density $f : \mathbb{R} \to [0, \infty)$ if*

$$P(A) = \int_A f(y) \mathrm{d}y$$

for all events $A \subseteq \mathbb{R}$. In particular, for $a < b$,

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

The distribution function for such a probability measure is given by

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

The reader may be unfamiliar with doing integrations over an *arbitrary* event A. If f is a continuous function and $A = [a, b]$ is an interval it should be well known that the integral

$$\int_{a}^{b} f(y)\mathrm{d}y \ \left(= \int_{[a,b]} f(y)\mathrm{d}y \right)$$

is the area under the graph of f from a to b. It is possible for more complicated sets A to assign a kind of generalized area to the set under the graph of f over A. We will not go into any further details. An important observation is that we can specify a *distribution function* F by

$$F(x) = \int_{-\infty}^{x} f(y)\mathrm{d}y \tag{2.11}$$

if $f : \mathbb{R} \to [0, \infty)$ is simply a positive function that fulfills that

$$\int_{-\infty}^{\infty} f(y)\mathrm{d}y = 1. \tag{2.12}$$

Indeed, if the total area from $-\infty$ to ∞ under the graph of f equals 1, the area under f from $-\infty$ to x is smaller (but always positive since f is positive) and therefore

$$F(x) = \int_{-\infty}^{x} f(y)\mathrm{d}y \in [0, 1].$$

When $x \to -\infty$ the area shrinks to 0, hence $\lim_{x \to \infty} F(x) = 0$ and when $x \to \infty$ the area increases to the total area under f, which we assumed to equal 1 by (2.12). Finally, a function given by (2.11) will always be continuous from which the right continuity at any x follows.

That a probability measure P on $\mathbb{R}$ is given by a continuous density f means that the probability of a small interval around x is proportional to the length of the interval with proportionality constant $f(x)$. Thus if $h > 0$ is small, so small that f can be regarded as almost constantly equal to $f(x)$ on the interval $[x - h, x + h]$, then

$$P([x - h, x + h]) = \int_{x-h}^{x+h} f(y)\mathrm{d}y \simeq 2hf(x) \tag{2.13}$$

where $2h$ is the length of the interval $[x - h, x + h]$. Rearranging, we can also write this approximate equality as

$$f(x) \simeq \frac{P([x - h, x + h])}{2h}.$$
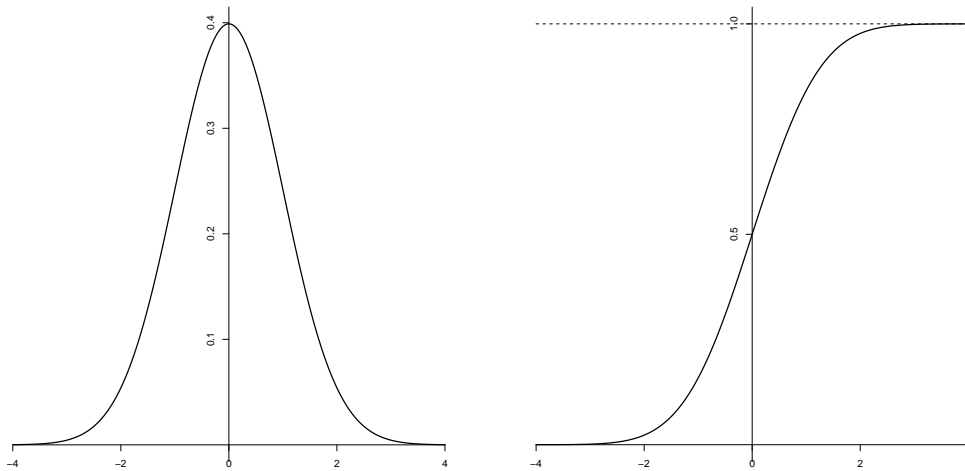
Figure 2.7: The density (left) and the distribution function (right) for the normal distribution.

Example 2.6.8 (The Normal Distribution). The normal or Gaussian distribution on $\mathbb{R}$ is the probability measure with density

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right).$$

It is not entirely trivial to check that $\int_{-\infty}^{\infty} f(x)\mathrm{d}x = 1$, but this is indeed the case. Using that $f(x) = f(-x)$ we can first observe that

$$\int_{-\infty}^{\infty} f(x)\mathrm{d}x = 2\int_{0}^{\infty} f(x)\mathrm{d}x.$$

Using the substitution $y = x^2/2$, and noting that

$$\mathrm{d}x = \frac{1}{x}\mathrm{d}y = \frac{1}{\sqrt{2y}}\mathrm{d}y,$$

we find that

$$
\begin{aligned}
\int_{-\infty}^{\infty} f(x)\mathrm{d}x &= 2\int_{0}^{\infty} f(x)\mathrm{d}x = \frac{2}{\sqrt{2\pi}} \int_{0}^{\infty} \frac{1}{\sqrt{2y}} \exp(-y)\mathrm{d}y \\
&= \frac{1}{\sqrt{\pi}} \int_{0}^{\infty} \frac{1}{\sqrt{y}} \exp(-y)\mathrm{d}y = \frac{\Gamma(1/2)}{\sqrt{\pi}},
\end{aligned}
$$

where $\Gamma(1/2)$ is the Γ-function evaluated in $1/2$. So up to showing that $\Gamma(1/2) = \sqrt{\pi}$, cf. Appendix B, we have showed that f integrates to 1.

The distribution function is by definition

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{y^2}{2}\right) \mathrm{d}y,$$

and it is unfortunately not possible to give a (more) closed form expression for this integral. It is, however, common usage to always denote this particular distribution function with a Φ.

The normal distribution is the single most important distribution in statistics. There are several reasons for this. One reason is that a rich and detailed theory about the normal distribution and a large number of statistical models based on the normal distribution can be developed. Another reason is that the normal distribution actually turns out to be a reasonable approximation of many other distributions of interest – that being a practical observation as well as a theoretical result known as the Central Limit Theorem, see Result 4.7.1. The systematic development of the statistical theory based on the normal distribution is a very well studied subject in the literature. $\diamond$
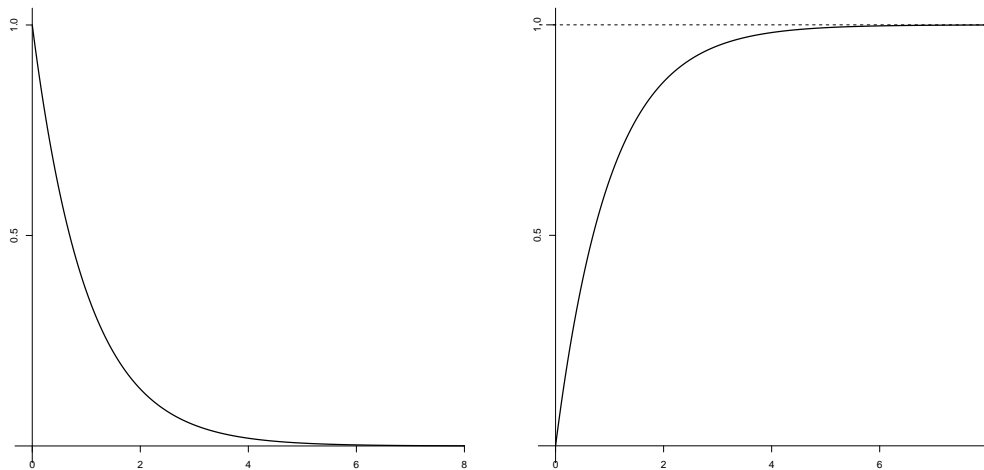
Figure 2.8: The density (left) and the distribution function (right) for the exponential distribution with intensity parameter $\lambda = 1$ (Example 2.6.9).

Example 2.6.9 (The Exponential Distribution)**.** Fix $\lambda > 0$ and define

$$f(x) = \lambda \exp(-\lambda x), \quad x \geq 0.$$

Let $f(x) = 0$ for $x < 0$. Clearly, $f(x)$ is positive, and we find that

$$\begin{aligned} \int_{-\infty}^{\infty} f(x)\mathrm{d}x &= \int_{0}^{\infty} \lambda \exp(-\lambda x)\mathrm{d}x \\ &= \left. -\exp(-\lambda x)\right|_{0}^{\infty} = 1. \end{aligned}$$

For the last equality we use the convention $\exp(-\infty) = 0$ together with the fact that

$\exp(0) = 1$. We also find the distribution function

$$
\begin{aligned}
F(x) &= \int_{-\infty}^{x} f(y)\mathrm{d}y \\
&= \int_{0}^{x} \lambda \exp(-\lambda y)\mathrm{d}y \\
&= -\exp(-\lambda y)\Big|_{0}^{x} = 1 - \exp(-\lambda x)
\end{aligned}
$$

for $x \geq 0$ (and $F(x) = 0$ for $x < 0$). The parameter λ is sometimes called the *intensity parameter*. This is because the exponential distribution is often used to model waiting times between the occurrences of events. The larger λ is, the smaller will the waiting times be, and the higher the *intensity* of the occurrence of the events will be. ◇

It is quite common, as for the exponential distribution above, that we only want to specify a probability measure living on an interval $I \subseteq \mathbb{R}$. By "living on" we mean that $P(I) = 1$. If the interval is of the form $[a, b]$, say, we will usually only specify the density $f(x)$ (or alternatively the distribution function $F(x)$) for $x \in [a, b]$ with the understanding that $f(x) = 0$ for $x \notin [a, b]$ (for the distribution function, $F(x) = 0$ for $x < a$ and $F(x) = 1$ for $x > b$).

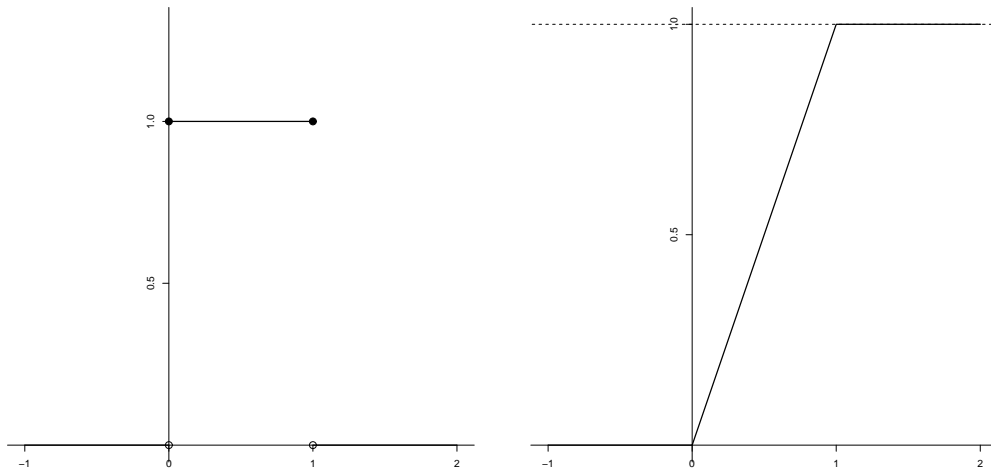Figure 2.9: The density (left) and the distribution function (right) for the uniform distribution on the interval $[0, 1]$ (Example 2.6.10).

Example 2.6.10 (The Uniform Distribution). Let $[a, b] \subseteq \mathbb{R}$ be an interval and define the function $f : \mathbb{R} \to [0, \infty)$ by

$$
f(x) = \frac{1}{b-a}1_{[a,b]}(x).
$$

That is, f is constantly equal to $1/(b-a)$ on $[a, b]$ and 0 outside. Then we find that

$$
\begin{aligned}
\int_{-\infty}^{\infty} f(x)\mathrm{d}x &= \int_{a}^{b} f(x)\mathrm{d}x \\
&= \int_{a}^{b} \frac{1}{b-a}\mathrm{d}x \\
&= \frac{1}{b-a} \times (b-a) = 1.
\end{aligned}
$$

Since f is clearly positive it is a density for a probability measure on $\mathbb{R}$. This probability measure is called the *uniform distribution* on the interval $[a, b]$. The distribution function can be computed (for $a \le x \le b$) as

$$
\begin{aligned}
F(x) &= \int_{-\infty}^{x} f(y)\mathrm{d}y \\
&= \int_{a}^{x} \frac{1}{b-a}\mathrm{d}y \\
&= \frac{x-a}{b-a}.
\end{aligned}
$$

In addition, $F(x) = 0$ for $x < a$ and $F(x) = 1$ for $x > b$. $\diamond$

R Box 2.6.1. Distribution functions and densities for a number of standard probability measures on $\mathbb{R}$ are directly available within R. The convention is that if a distribution is given the R-name `name` then `pname(x)` gives the distribution function evaluated at `x` and `dname(x)` gives the density evaluated at `x`. The normal distribution has the R-name `norm` so `pnorm(x)` and `dnorm(x)` gives the distribution and density function respectively for the normal distribution. Likewise the R-name for the exponential function is `exp` so `pexp(x)` and `dexp(x)` gives the distribution and density function respectively for the exponential distribution. For the exponential distribution `pexp(x,3)` gives the density at `x` with intensity parameter $\lambda = 3$.

Example 2.6.11 (The Γ-distribution)**.** The Γ-distribution with shape parameter $\lambda > 0$ and scale parameter $\beta > 0$ is the probability measure on $[0, \infty)$ with density

$$
f(x) = \frac{1}{\beta^{\lambda}\Gamma(\lambda)}x^{\lambda-1}\exp\left(-\frac{x}{\beta}\right), \quad x > 0
$$

where $\Gamma(\lambda)$ is the Γ-function evaluated in λ, cf. Appendix B. The Γ-distribution with $\lambda = 1$ is the exponential distribution. The Γ-distribution with shape $\lambda = f/2$ for $f \in \mathbb{N}$ and scale $\beta = 2$ is known as the χ^2-distribution with f degrees of freedom. The $\sigma^2\chi^2$-distribution with f degrees of freedom is the χ^2-distribution with f degrees of freedom and scale parameter σ^2, thus it is the Γ-distribution with shape parameter $\lambda = f/2$ and scale parameter $\beta = 2\sigma^2$. $\diamond$
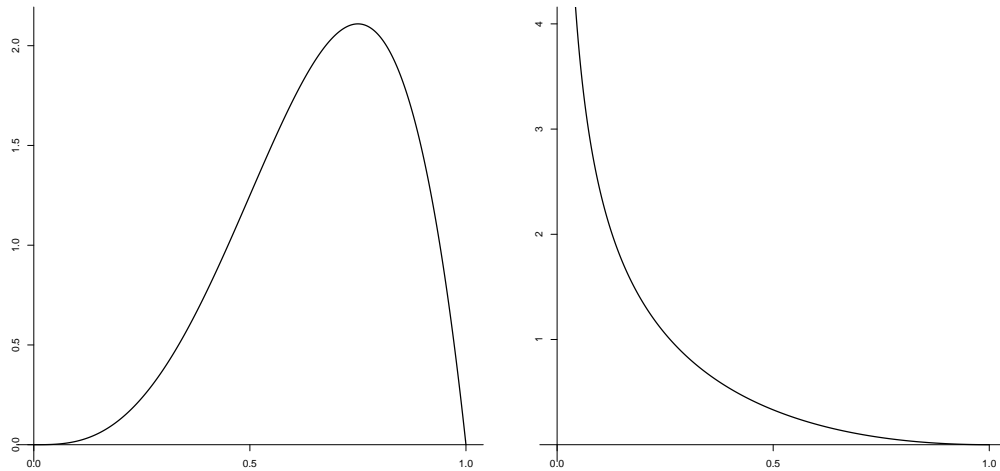
Figure 2.10: The density for the B-distribution (Example 2.6.12) with parameters $(\lambda_1, \lambda_2) = (4, 2)$ (left) and $(\lambda_1, \lambda_2) = (0.5, 3)$ (right)

Example 2.6.12 (The B-distribution). The density for the B-distribution (pronounced β-distribution) with parameters $\lambda_1, \lambda_2 > 0$ is given by

$$f(x) = \frac{1}{B(\lambda_1, \lambda_2)} x^{\lambda_1 - 1} (1 - x)^{\lambda_2 - 1}$$

for $x \in [0, 1]$. Here $B(\lambda_1, \lambda_2)$ is the B-function, cf. Appendix B. This two-parameter class of distributions on the unit interval $[0, 1]$ is quite flexible. For $\lambda_1 = \lambda_2 = 1$ we obtain the uniform distribution on $[0, 1]$, but for other parameters we can get a diverse set of shapes for the density – see Figure 2.10 for two particular examples. Since the B-distribution always lives on the interval $[0, 1]$ it is frequently encountered as a model of a random probability – or rather a random frequency. In population genetics for instance, the B-distribution is found as a model for the frequency of occurrences of one out of two alleles in a population. The shape of the distribution, i.e. the proper values of λ_1 and λ_2, then depends upon issues such as the mutation rate and the migration rates. $\diamond$

From a basic calculus course the intimate relation between integration and differentiation should be well known.

Result 2.6.13. *If F is a differentiable distribution function the derivative*

$$f(x) = F'(x)$$

is a density for the distribution given by F. That is

$$F(x) = \int_{-\infty}^{x} F'(y)\mathrm{d}y.$$
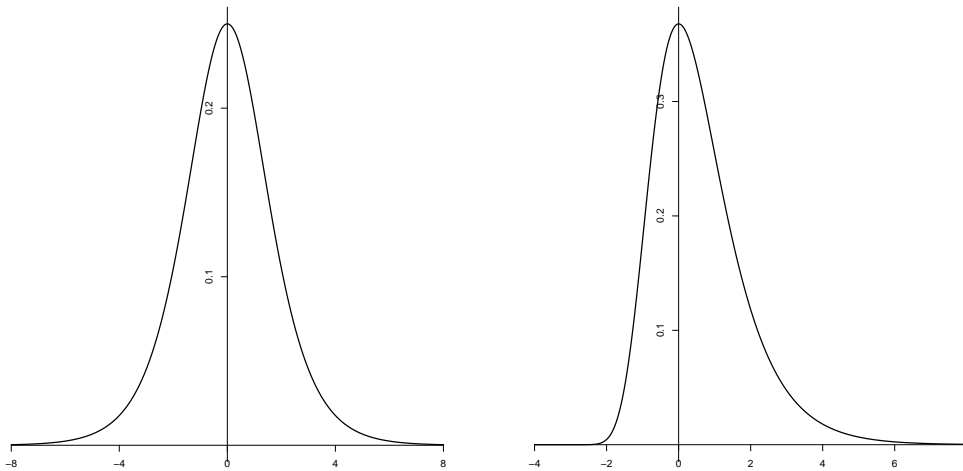
Figure 2.11: The density for the logistic distribution (left, see Example 2.6.14) and the density for the Gumbel distribution (right, see Example 2.6.15). The density for the Gumbel distribution is clearly skewed, whereas the density for the logistic distribution is symmetric and quite similar to the density for the normal distribution.

Example 2.6.14 (Logistic distribution)**.** The density for the logistic distribution is found to be

$$f(x) = F'(x) = \frac{\exp(-x)}{(1 + \exp(-x))^2}.$$

$\diamond$

Example 2.6.15 (Gumbel distribution)**.** The density for the Gumbel distribution is found to be

$$f(x) = F'(x) = \exp(-x)\exp(-\exp(-x)).$$

$\diamond$

Exercises

$\star$

Exercise 2.6.1. Argue that the function

$$F(x) = 1 - \exp(-x^{\beta}), \quad x \geq 0$$

for $\beta > 0$ is a distribution function. It is called the *Weibull distribution* with parameter β. Find the density on the interval $[0, \infty)$ for the Weibull distribution.

★   **Exercise 2.6.2.** Argue that the function

$$F(x) = 1 - x_0^\beta x^{-\beta}, \quad x \geq x_0 > 0$$

for $\beta > 0$ is a distribution function on $[x_0, \infty)$. It is called the *Pareto distribution* on the interval $[x_0, \infty)$. Find the density on the interval $[x_0, \infty)$ for the Pareto distribution.

⌨   **Exercise 2.6.3.** Write two functions in R, `pgumb` and `dgumb`, that computes the distribution function and the density for the Gumbel distribution.

⌨   **Exercise 2.6.4.** Let

$$f_\lambda(x) = \frac{1}{(1 + \frac{x^2}{2\lambda})^{\lambda + \frac{1}{2}}}$$

for $x \in \mathbb{R}$ and $\lambda > 0$. Argue that $f_\lambda(x) > 0$ for all $x \in \mathbb{R}$. Use numerical integration in R, `integrate`, to compute

$$c(\lambda) = \int_{-\infty}^{\infty} f_\lambda(x) \mathrm{d}x$$

for $\lambda = \frac{1}{2}, 1, 2, 10, 100$. Compare the results with π and $\sqrt{2\pi}$. Argue that $c(\lambda)^{-1} f_\lambda(x)$ is a density and compare it, numerically, with the density for the normal distribution.

The probability measure with density $c(\lambda)^{-1} f_\lambda(x)$ is called the *t-distribution* with shape parameter λ, and it is possible to show that

$$c(\lambda) = \sqrt{2\lambda} B(\lambda, \frac{1}{2})$$

where B is the B-function.

2.7   Descriptive methods

In the summary of univariate real observations we are essentially either summarizing the data as a density approximation or via quantiles. The quantiles are often more informative than aiming for the distribution function, say, directly. As encountered in Section 2.3, we introduce

$$\varepsilon_n(A) = \frac{1}{n} \sum_{i=1}^{n} 1_A(x_i)$$

to denote the fraction or *relative frequency* of observations that belong to the event $A \subseteq \mathbb{R}$.

2.7.1   Histograms and kernel density estimation

The underlying assumption in this section is that we have a data set of observations $x_1, \ldots, x_n$ and that these observations are generated by a distribution P that pos-

sesses a density f. A histogram is then an approximation to the density f, thus it is a function

$$\hat{f} : \mathbb{R} \to [0, \infty).$$

When plotting this function the convention is, however, to plot rectangular boxes instead of the step-function that it really is.

Definition 2.7.1. *The histogram with break points $q_1 < q_2 < \ldots < q_k$, chosen so that*

$$q_1 < \min_{i=1,\ldots,n} x_i \le \max_{i=1,\ldots,n} x_i < q_k,$$

is the function $\hat{f}$ given by

$$\hat{f}(x) = \frac{1}{q_{i+1} - q_i} \varepsilon_n \left((q_i, q_{i+1}] \right) \quad \text{for } q_i < x \le q_{i+1}. \tag{2.14}$$

together with $\hat{f}(x) = 0$ for $x \notin (q_1, q_n]$. Usually one plots $\hat{f}$ as a box of height $\hat{f}(q_{i+1})$ located over the interval $(q_i, q_{i+1}]$, and this is what most people associate with a histogram.

The function $\hat{f}$ is constructed so that

$$\begin{aligned}
\int_{-\infty}^{\infty} \hat{f}(x)\mathrm{d}x &= \sum_{i=1}^{k-1} \int_{q_i}^{q_{i+1}} \frac{1}{q_{i+1} - q_i} \varepsilon_n \left((q_i, q_{i+1}] \right) \mathrm{d}x \\
&= \sum_{i=1}^{k-1} \varepsilon_n \left((q_i, q_{i+1}] \right) \\
&= \varepsilon_n((q_1, q_n]) = 1
\end{aligned}$$

where we use the fact that all the data points are contained within the interval $(q_1, q_n]$. Since the function $\hat{f}$ integrates to 1 it is a probability density. The purpose of the histogram is to approximate the density of the true distribution of X – assuming that the distribution has a density.

Sometimes one encounters the *unnormalized* histogram, given by the function

$$\tilde{f}(x) = n\varepsilon_n(q_i, q_{i+1}] = n(q_{i+1} - q_i)\hat{f}(x) \quad \text{for } q_i < x \le q_{i+1}.$$

Here $\tilde{f}(x)$ is constantly equal to the number of observations falling in the interval $(q_i, q_{i+1}]$. Since the function doesn't integrate to 1 it cannot be compared directly with a density, nor is it possible to compare unnormalized histograms directly for two samples of unequal size. Moreover, if the break points are not equidistant, the unnormalized histogram is of little value.

It can be of visual help to add a *rug plot* to the histogram – especially for small or moderate size dataset. A rug plot is a plot of the observations as small "tick marks" along the first coordinate axis.

Example 2.7.2. We consider the histogram of 100 and 1000 random variables whose distribution is $N(0,1)$. The are generated by the computer. We choose the breaks to be equidistant from -4 to 4 with a distance of 0.5, thus the break points are

$$-4 \quad -3.5 \quad -3 \quad -2.5 \quad \dots \quad 2.5 \quad 3 \quad 3.5 \quad 4.$$

We find the histograms in Figure 2.12. Note how the histogram corresponding to the 1000 random variables approximates the density more closely. ◇
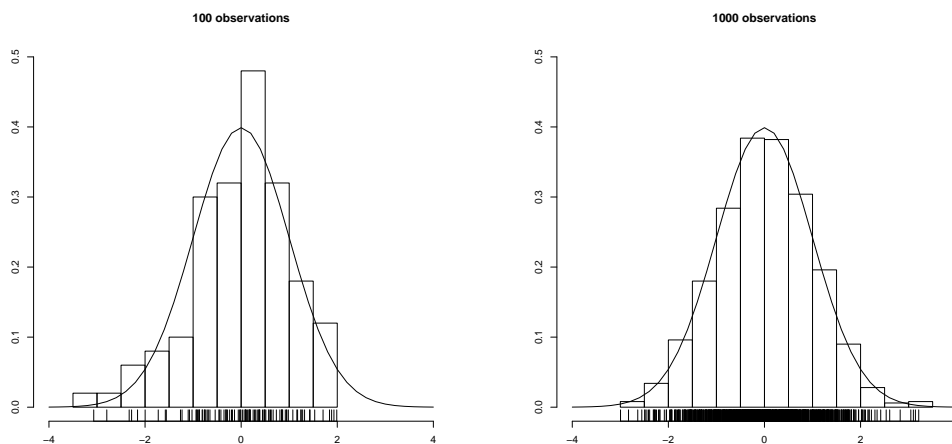
Figure 2.12: The histograms for the realization of 100 (left) and 1000 (right) simulated $N(0,1)$ random variables. A rug plot is added to both histograms, and we compare the histograms with the corresponding density for the normal distribution.

Example 2.7.3. Throughout this section we will consider data from a microarray experiment. It is the so-called ALL dataset (Chiaretti et. al., Blood, vol. 103, No. 7, 2004). It consists of samples from patients suffering from Acute Lymphoblastic Leukemia. We will consider only those patients with B-cell ALL, and we will group the patients according to presence or absence of the BCR/ABL fusion gene.

On Figure 2.13 we see the histogram of the log (base 2) expression levels[3] for six (arbitrary) genes for the group of samples without BCR/ABL.

On Figure 2.14 we have singled out the signal from the gene probe set with the poetic name 1635‗ at, and we see the histograms for the log expression levels for the two groups with or without BCR/ABL. The figure also shows examples of kernel density estimates. ◇

---

[3]Some further normalization has also been done that we will not go into here.

R Box 2.7.1 (Histograms). A histogram of the data in the numeric vector `x` is produced in R by the command

```
> hist(x)
```

This plots a histogram using default settings. The break points are by default chosen by R in a suitable way. It is possible to explicitly set the break points by hand, for instance

```
> hist(x, breaks = c(-3, -2, -1, 0, 1, 2, 3))
```

produces a histogram with break points $-3, -2, -1, 0, 1, 2, 3$. Note that R gives an error if the range of the break points does not contain all the data points in `x`. Note also that the default behavior of `hist` is to plot the *unnormalized* histogram if the break points are equidistant, which they are using the default breakpoints. Specifying non-equidistant breakpoints always gives the normalized histogram. One can make `hist` produce normalized histograms by setting `freq=FALSE` when calling `hist`. Adding a rug plot is done by `rug(x)`.
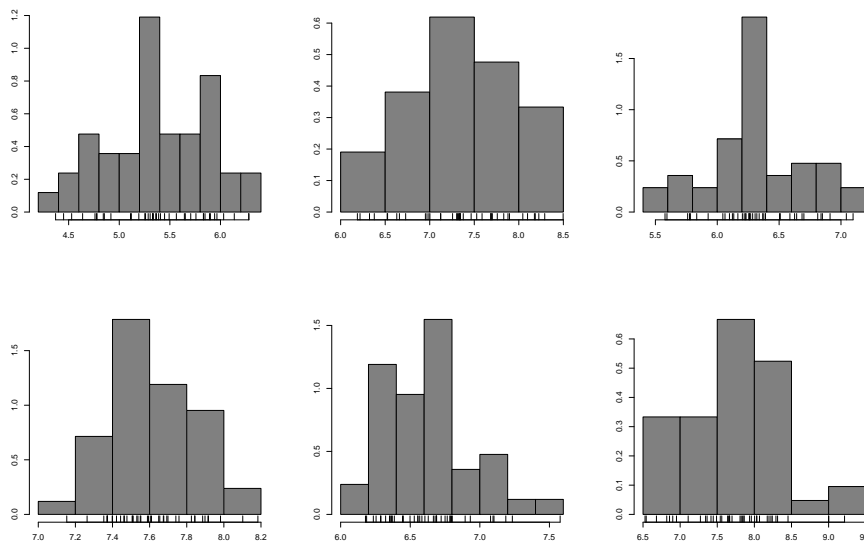
Figure 2.13: Examples of histograms for the log (base 2) expression levels for a random subset of six genes from the ALL dataset (non BCR/ABL fusion gene).
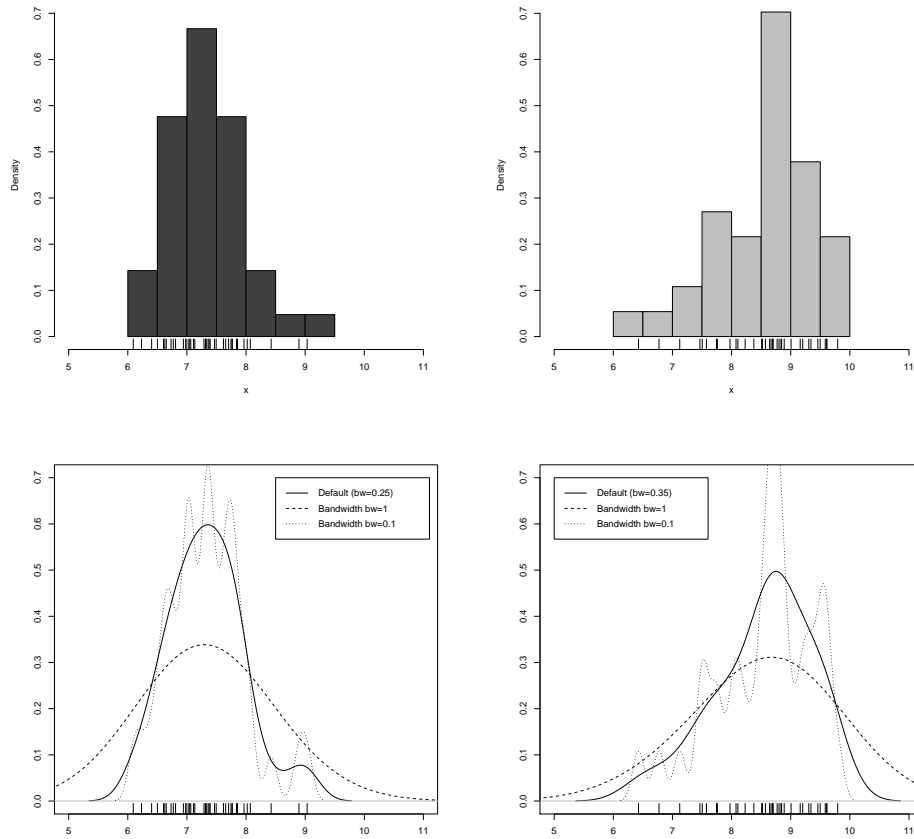
Figure 2.14: Histograms and examples of kernel density estimates (with a Gaussian kernel) of the log (base 2) expression levels for the gene 1635_at from the ALL microarray experiment with (right) or without (left) presence of the BCR/ABL fusion gene.

The histogram is a crude and in some ways unsatisfactory approximation of the density f. The plot we get is sensitive to the choice of break points – the number as well as their location. Moreover, the real density is often thought to be a rather smooth function whereas the histogram by definition is very non-smooth. The use of kernel density estimation remedies some of these problems. Kernel density estimation or smoothing is computationally a little harder than computing the histogram, which for moderate datasets can be done by hand. However, with modern computers, uni-variate kernel density estimation can be done just as rapidly as drawing a histogram – that is, apparently instantaneously.

Kernel density estimation is based on the interpretation of the density as given by

(2.13). Rearranging that equation then says that for small $h > 0$

$$f(x) \simeq \frac{1}{2h} P([x - h, x + h]),$$

and if we then use $P([x - h, x + h]) \simeq \varepsilon_n([x - h, x + h])$ from the frequency interpretation, we get that for small $h > 0$

$$f(x) \simeq \frac{1}{2h} \varepsilon_n([x - h, x + h]).$$

The function

$$\hat{f}(x) = \frac{1}{2h} \varepsilon_n([x - h, x + h])$$

is in fact an example of a kernel density estimator using the *rectangular kernel* . If we define the *kernel* $K_h : \mathbb{R} \times \mathbb{R} \to [0, \infty)$ by

$$K_h(x, y) = \frac{1}{2h} 1_{[x-h,x+h]}(y),$$

we can see by the definition of $\varepsilon_n([x - h, x + h])$ that

$$
\begin{aligned}
\hat{f}(x) &= \frac{1}{n} \sum_{i=1}^{n} \frac{1}{2h} 1_{[x-h,x+h]}(x_i) \\
&= \frac{1}{n} \sum_{i=1}^{n} K_h(x, x_i).
\end{aligned}
$$

One may be unsatisfied with the sharp cut-offs at $\pm h$ – and how to choose h in the first place? We may therefore choose smoother kernel functions but with similar properties as K_h, that is, they are largest for $x = y$, they fall off, but perhaps more smoothly than the rectangular kernel, when x moves away from y, and they integrate to 1 over x for all y.

Definition 2.7.4. *A function*

$$K : \mathbb{R} \times \mathbb{R} \to [0, \infty)$$

is called a kernel if

$$\int_{-\infty}^{\infty} K(x, y) \mathrm{d}x = 1$$

for all y. If K is a kernel we define the kernel density estimate for our dataset $x_1, \ldots, x_n \in \mathbb{R}$ by

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} K(x, x_i).$$

Observe that by the definition of a kernel

$$\int_{-\infty}^{\infty} \hat{f}(x)\mathrm{d}x = \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{\infty} K(x,x_i)\mathrm{d}x = 1.$$

Popular choices of kernels include the Gaussian kernel

$$K_h(x,y) = \frac{1}{h\sqrt{2\pi}}\exp\left(-\frac{(x-y)^2}{2h^2}\right),$$

and the Epanechnikov kernel

$$K_h(x,y) = \frac{3}{h}\left(1 - \frac{(x-y)^2}{h^2}\right)1_{[x-h,y+h]}(y).$$

Both of these kernels as well as the rectangular kernel are examples of choosing a "mother" kernel, $k : \mathbb{R} \to [0,\infty)$, and then taking

$$K_h(x,y) = \frac{1}{h}k\left(\frac{x-y}{h}\right).$$

Note that the mother kernel k needs to be a probability density itself. In this case the optional parameter, $h > 0$, is called the *bandwidth* of the kernel. This parameter determines how quickly the kernel falls off, and plays in general the same role as h does for the rectangular kernel. Indeed, if we take the mother kernel $k(x) = 1_{[-1,1]}(x)$, the kernel K_h defined above is precisely the rectangular kernel. Qualitatively we can say that large values of h gives smooth density estimates and small values gives density estimates that wiggle up and down.

As a curiosity we may note that the definition of the kernel density estimate $\hat{f}$ does not really include the computation of anything! Every computation is carried out at evaluation time. Each time we evaluate $\hat{f}(x)$ we carry out the evaluation of $K(x,x_i)$, $i = 1,\ldots,n$ – we cannot do this before we know which x to evaluate at – and then the summation. So the definition of the kernel density estimate $\hat{f}$ can mostly be viewed as a definition of what we intend to do at evaluation time. However, implementations such as `density` in R do the evaluation of the kernel density estimate once and for all at call time in a number of pre-specified points.

2.7.2   Mean and variance

The mean and variance for probability measures on discrete subsets of $\mathbb{R}$ were defined through the point probabilities, and for probability measures on $\mathbb{R}$ given in terms of a density there is an analogous definition of the mean and variance.

Definition 2.7.5. *If P is a probability measure on $\mathbb{R}$ with density f that fulfills*

$$\int_{-\infty}^{\infty} |x|f(x)\mathrm{d}x < \infty$$

> **R Box 2.7.2** (Kernel density estimation)**.** The `density` function computes the evaluation of the density estimate for a dataset in a finite number of points. Default choice of `kernel` is the Gaussian, and the band width is computed automatically by some rule of thumb. The default number of evaluations is 512.
>
> ```
> > plot(density(rnorm(100), n = 1024, bw = 1))
> ```
>
> produces a plot of the density estimate for the realization of 100 standard normally distributed random variables using bandwidth 1 and 1024 points for evaluation.

then we define the mean under P as

$$\mu = \int_{-\infty}^{\infty} x f(x) \mathrm{d}x.$$

If, moreover,

$$\int_{-\infty}^{\infty} x^2 f(x) \mathrm{d}x < \infty$$

we define the variance under P as

$$\sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) \mathrm{d}x.$$

Example 2.7.6. Consider the exponential distribution with intensity parameter $\lambda > 0$. The density is

$$f(x) = \lambda \exp(-\lambda x)$$

for $x \geq 0$ (and $f(x) = 0$ for $x < 0$). We find by partial integration that

$$
\begin{aligned}
\mu &= \int_0^{\infty} x \lambda \exp(-\lambda x) \mathrm{d}x \\
&= x \exp(-\lambda x) \Big|_0^{\infty} + \int_0^{\infty} \exp(-\lambda x) \mathrm{d}x \\
&= -\frac{1}{\lambda} \exp(-\lambda x) \Big|_0^{\infty} = \frac{1}{\lambda}.
\end{aligned}
$$

$\diamond$

Example 2.7.7. Consider the uniform distribution on $[a, b]$. Then the density is

$$f(x) = \frac{1}{b - a} 1_{[a,b]}(x).$$

We find that the mean is

$$
\begin{aligned}
\mu &= \int_a^b x \frac{1}{b-a} \mathrm{d}x = \frac{1}{2(b-a)} x^2 \Big|_a^b \\
&= \frac{1}{2} \frac{b^2 - a^2}{b-a} = \frac{1}{2} \frac{(b-a)(b+a)}{b-a} = \frac{1}{2}(a+b).
\end{aligned}
$$

We see that μ is the midpoint between a and b. $\diamond$

Example 2.7.8. The Γ-distribution on $[0,\infty)$ with shape parameter $\lambda > 0$ and scale parameter $\beta > 0$ has finite mean

$$
\begin{aligned}
\mu &= \frac{1}{\beta^\lambda \Gamma(\lambda)} \int_0^\infty x x^{\lambda-1} \exp\left(-\frac{x}{\beta}\right) \mathrm{d}x \\
&= \frac{1}{\beta^\lambda \Gamma(\lambda)} \int_0^\infty x^\lambda \exp\left(-\frac{x}{\beta}\right) \mathrm{d}x \\
&= \frac{\beta^{\lambda+1} \Gamma(\lambda+1)}{\beta^\lambda \Gamma(\lambda)} = \beta \frac{\lambda \Gamma(\lambda)}{\Gamma(\lambda)} = \beta\lambda.
\end{aligned}
$$

A similar computation reveals that the variance is $\beta^2 \lambda$. In particular, the mean and variance of the $\sigma^2 \chi^2$-distribution with f degrees of freedom is $\sigma^2 f$ and $2\sigma^4 f$. $\diamond$

Example 2.7.9. A density on $\mathbb{R}$ is called symmetric if $f(x) = f(-x)$. Since $|-x| = |x|$ we have that for a symmetric density

$$
\int_{-\infty}^\infty |x| f(x) \mathrm{d}x = 2 \int_0^\infty x f(x) \mathrm{d}x,
$$

so the mean is defined if and only if

$$
\int_0^\infty x f(x) \mathrm{d}x < \infty.
$$

In that case, using that $(-x)f(-x) = -xf(x)$,

$$
\begin{aligned}
\mu &= \int_{-\infty}^\infty x f(x) \mathrm{d}x = \int_0^\infty x f(x) \mathrm{d}x + \int_{-\infty}^0 x f(x) \mathrm{d}x \\
&= \int_0^\infty x f(x) \mathrm{d}x - \int_0^\infty x f(x) \mathrm{d}x = 0.
\end{aligned}
$$

The mean of a probability measure on $\mathbb{R}$ given in terms of a symmetric density is thus 0 – provided that it is defined. If $\int_0^\infty x f(x) \mathrm{d}x = \infty$ we say that the mean is undefined. $\diamond$

Example 2.7.10. The density for the normal distribution is

$$
f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right),
$$

which is clearly seen to be symmetric, cf. Example 2.7.9. Moreover, with the substitution $y = x^2/2$ we have that $\mathrm{d}y = x\mathrm{d}x$, so

$$
\begin{aligned}
\int_0^\infty xf(x)\mathrm{d}x &= \frac{1}{\sqrt{2\pi}} \int_0^\infty \exp(-y)\mathrm{d}y \\
&= \frac{1}{\sqrt{2\pi}} < \infty.
\end{aligned}
$$

Thus by Example 2.7.9 the normal distribution has mean $\mu = 0$.

Regarding the variance, we use the symmetry argument again and the same substitution to obtain

$$
\begin{aligned}
\sigma^2 &= \int_{-\infty}^\infty x^2 f(x)\mathrm{d}x = 2\int_0^\infty x^2 f(x)\mathrm{d}x \\
&= \frac{2}{\sqrt{2\pi}} \int_0^\infty \sqrt{2y}\exp(-y)\mathrm{d}y = \frac{2\Gamma(3/2)}{\sqrt{\pi}} = 1
\end{aligned}
$$

where we use that that $\Gamma(3/2) = \sqrt{\pi}/2$, cf. Appendix B. $\diamond$

As for discrete probability measures the sample mean and sample variance for a dataset of observations from a distribution P work as estimates of the unknown theoretical mean and variance under P – provided they exist.

2.7.3 Quantiles

While histograms and other density estimates may be suitable for getting an idea about the location, spread and shape of a distribution, other descriptive methods are more suitable for comparisons between two datasets, say, or between a dataset and a theoretical distribution. Moreover, the use of histograms and density estimates builds on the assumption that the distribution is actually given by a density. As an alternative to density estimation we can try to estimate the distribution function directly.

Definition 2.7.11. *The empirical distribution function is the function defined by*

$$
\hat{F}_n(x) = \varepsilon_n((-\infty, x])
$$

for $x \in \mathbb{R}$.

The empirical distribution functions qualifies for the name "distribution function", because it is really a distribution function. It is increasing, with limits 0 and 1 when $x \to -\infty$ and $x \to \infty$ respectively, and it is right continuous. But it is *not* continuous! It has jumps at each x_i for $i = 1, \ldots, n$ and is constant in between. Since distribution functions all have a similar S-shape, empirical distribution functions in themselves are not really ideal for comparisons either. We will instead develop methods based

on the *quantiles*, which are more suitable. We define the quantiles for the dataset first and then subsequently we define quantiles for a theoretical distribution in terms of its distribution function. Finally we show how the theoretical definition applied to the empirical distribution function yields the quantiles for the dataset.

R Box 2.7.3 (Empirical distribution functions)**.** If `x` is a numeric vector in R containing our data we can construct a `ecdf`-object (empirical cumulative distribution function). This requires the `stats` package:

```
> library(stats)
```

Then

```
> edf <- ecdf(x)
```

gives the empirical distribution function for the data in `x`. One can evaluate this function like any other function:

```
> edf(1.95)
```

```
[1] 0.95
```

gives the value of the empirical distribution function evaluated at 1.95. It is also easy to plot the distribution function:

```
> plot(edf)
```

produces a nice plot.

If $x_1, \ldots, x_n \in \mathbb{R}$ are n real observations from an experiment, we can order the observations

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)},$$

thus $x_{(1)}$ denotes the smallest observation, $x_{(n)}$ the largest and $x_{(i)}$ the observation with $i-1$ smaller observations. If $q = i/n$ for $i = 1, \ldots, n-1$, then $x \in \mathbb{R}$ is called a *q-quantile* (for the dataset) if $x_{(i)} \leq x \leq x_{(i+1)}$. In other words, for a given $q = i/n$ there is a whole range of q-quantiles, namely the interval $[x_{(i)}, x_{(i+1)}]$. Informally we can say that x is a q-quantile if the fraction of the observations that are $\leq x$ is q – except that $x_{(i+1)}$ is taken as a q-quantile also.

If $(i-1)/n < q < i/n$ for $i = 1, \ldots, n$ the proper definition of the q-quantile is $x_{(i)}$. This is the only definition that assures *monotonicity* of quantiles in the sense that if x is a q-quantile and y is a p-quantile with $q < p$ then $x < y$.

Some quantiles have special names, e.g. a 0.5-quantile is called a median, and the
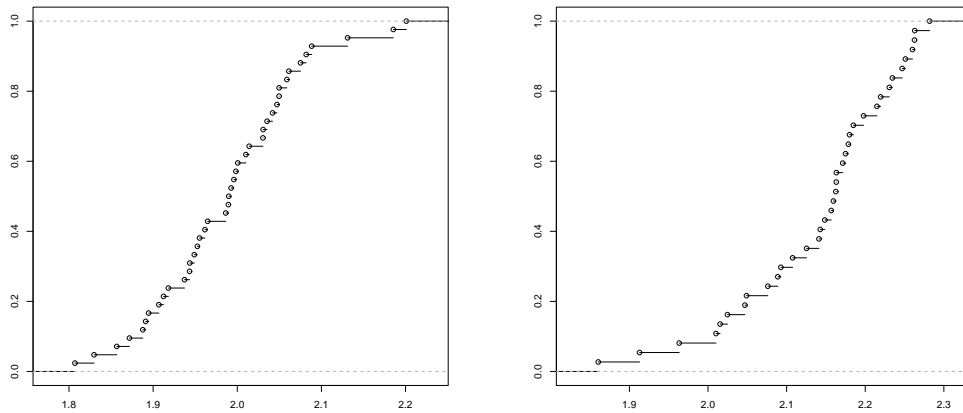
Figure 2.15: The empirical distribution function for the log (base 2) expression levels for the gene 40480_s_at from the ALL microarray experiment with (right) or without (left) presence of the BCR/ABL fusion gene.

upper and lower quartiles are the 0.75- and 0.25-quantiles respectively. Note the ambiguity here. If n is even then *all* the x's in the interval $[x_{(n/2)}, x_{(n/2+1)}]$ are medians, whereas if n is odd the median is uniquely defined as $x_{((n+1)/2)}$. This ambiguity of e.g. the median and other quantiles can be a little annoying in practice, and sometimes one prefers to define a single (empirical) quantile function

$$Q : (0,1) \to \mathbb{R}$$

such that for all $q \in (0,1)$ we have that $Q(q)$ is a q-quantile. Whether one prefers $x_{(n/2)}$, $x_{(n/2+1)}$, or perhaps $(x_{(n/2)} + x_{(n/2+1)})/2$ as the median if n is even is largely a matter of taste.

Quantiles can also be defined for theoretical distributions. We prefer here to consider the definition of a quantile function only.

Definition 2.7.12. *If $F : \mathbb{R} \to [0,1]$ is a distribution function for a probability measure P on $\mathbb{R}$, then $Q : (0,1) \to \mathbb{R}$ is a quantile function for P if*

$$F(Q(y) - \varepsilon) \leq y \leq F(Q(y)) \tag{2.15}$$

for all $y \in (0,1)$ and all $\varepsilon > 0$.

The definition is essentially an attempt to define an inverse of the distribution function. However, the distribution function may not have a real inverse but there will always be a quantile function. This is not at all obvious but we will not pursue the general construction of a quantile function. From a practical point of view it is a

---

R Box 2.7.4 (Quantiles). If `x` is a numeric vector then

```
    > quantile(x)


        0%         25%         50%         75%        100%
-1.2803887 -0.4043956  0.1354512  0.5055087  2.0817652
```

computes the 0%, 25%, 50%, 75%, and 100% quantiles. That is, the minimum, the lower quartile, the median, the upper quartile, and the maximum.

```
    > quantile(x,probs=c(0.1,0.9))


       10%        90%
-0.8547265  0.7652570
```

computes the 0.1 and 0.9 quantile instead, and by setting the `type` parameter to an integer between 1 and 9, one can select how the function handles the non-uniqueness of the quantiles. If `type=1` the quantiles are the generalized inverse of the empirical distribution function, which we will deal with in Section 2.11. Note that with `type` being 4 to 9 the result is not a quantile for the empirical distribution function – though it may still be a reasonable approximation of the theoretical quantile for the unknown distribution function.

---

greater nuisance that the choice of a quantile function is not necessarily unique. As mentioned above the empirical distribution function is an example where the corresponding choice of quantile function is not unique and in R the function `quantile` has in fact 9 different types of quantile computations. Not all of these computations give quantiles for the empirical distribution function, though, but three of them do.

Definition 2.7.13. *If F is a distribution function and Q a quantile function for F the median, or second quartile, of F is defined as*

$$q_2 = \text{median}(F) = Q(0.5).$$

In addition we call $q_1 = Q(0.25)$ and $q_3 = Q(0.75)$ the first end third quartiles of F. The difference

$$\text{IQR} = q_3 - q_1$$

is called the interquartile range.

Note that the definitions of the median and the quartiles depend on the choice of quantile function. If the quantile function is not unique these numbers are not necessarily uniquely defined. The median summarizes in a single number the location

of the probability measure given by F. The interquartile range summarizes how spread out around the median the distribution is.

The definition of a quantile function for any distribution function F and the quantiles defined for a given dataset are closely related. With

$$\hat{F}_n(x) = \varepsilon_n((-\infty, x]) \tag{2.16}$$

the empirical distribution function for the data then any quantile function for the distribution function $\hat{F}_n$ also gives empirical quantiles as defined for the dataset.

We will as mentioned use quantiles to compare two distributions – that being either two empirical distributions or an empirical and a theoretical distribution. In principle, one can also use quantiles to compare two theoretical distributions, but that is not so interesting – after all, we then know whether they are different or not – but the quantiles may tell something about the nature of a difference.

Definition 2.7.14. *If F_1 and F_2 are two distribution functions with Q_1 and Q_2 their corresponding quantile functions a QQ-plot is a plot of Q_2 against Q_1.*

R Box 2.7.5 (QQ-plots). If `x` and `y` are numeric vectors then

> `qqplot(x,y)`

produces a QQ-plot of the empirical quantiles for `y` against those for `x`.

> `qqnorm(x)`

results in a QQ-plot of the empirical quantiles for `x` against the quantiles for the normal distribution.

When making a QQ-plot with one of the distributions, F_2, say, being empirical, it is common to plot

$$\left(Q_1 \left(\frac{2i - 1}{2n} \right), x_{(i)} \right), \quad i = 1, \ldots, n - 1.$$

That is, we compare the smallest observation $x_{(1)}$ with the $1/2n$'th quantile, the second smallest observation $x_{(2)}$ with the $3/2n$'th quantile and so on ending with the largest observation $x_{(n)}$ and the $1 - 1/2n$'th quantile.

If the empirical quantile function Q_1 is created from a dataset with n data points all being realizations from the distribution function F with quantile function Q_1 then the points in the QQ-plot should lie close to a straight line with slope 1 and intercept 0. It can be beneficial to plot a straight line, for instance through suitably chosen quantiles, to be able to visualize any discrepancies from the straight line.
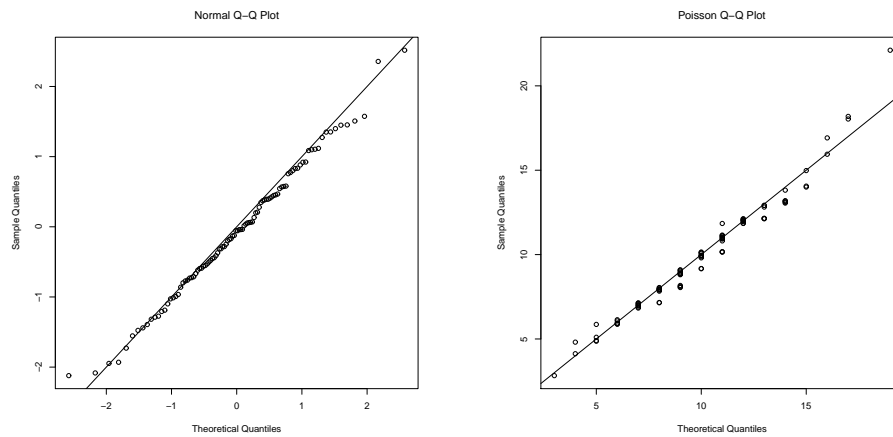
Figure 2.16: A QQ-plot for 100 simulated data points from the standard normal distribution (left) and a QQ-plot for 100 simulated data points from the Poisson distribution with parameter $\lambda = 10$ (right). In the latter case the sample points are for visual reasons "jittered", that is, small random noise is added to visually separate the sample quantiles.
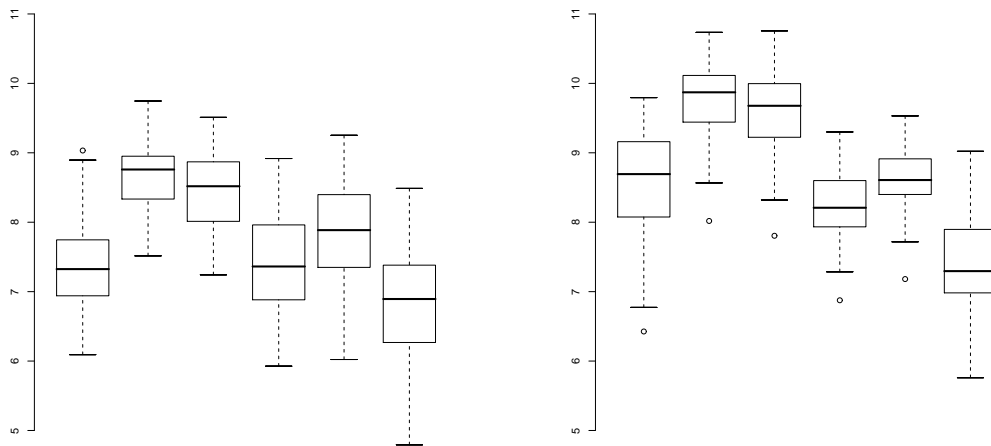
Figure 2.17: Comparing the empirical distributions of the six genes with probe set names 1635_at,1636_g_at,39730_at,40480_s_at, 2039_s_at, 36643_at for those with BCR/ABL (right) and those without (left) using boxplots

Another visualization that is quite useful – specially for comparing three or more empirical distributions – is the *box plot*, which is also based on quantiles. Historically it was also useful for visualizing just a single empirical distribution to get a

rough idea about location and scale, but with current computational power there are more informative plots for single distributions. As a comparative tool for many distributions, box plots are on the other hand quite effective.

R Box 2.7.6 (Box plots). For a numeric vector `x` we get a single box plot by

> `> boxplot(x)`

If `x` is a dataframe the command will instead produce (in one figure) a box plot of each column. By specifying the `range` parameter (= whisker coefficient), which by default equals 1.5, we can change the length of the whiskers.

> `> boxplot(x,range=1)`

produces a box plot with whisker coefficient 1.

A box plot using quantile function Q is given in terms of a five-dimensional vector

$$(w_1, q_1, q_2, q_3, w_2)$$

with $w_1 \leq q_1 \leq q_2 \leq q_3 \leq w_2$. Here

$$q_1 = Q(0.25), \quad q_2 = Q(0.5), \quad q_3 = Q(0.75)$$

are the three quartiles and

$$
\begin{aligned}
w_1 &= \min\{x_i \mid x_i \geq q_1 - c(q_3 - q_1)\} \\
w_2 &= \max\{x_i \mid x_i \leq q_3 + c(q_3 - q_1)\}
\end{aligned}
$$

are called the whiskers. The parameter $c > 0$ is the whisker coefficient. The box plot is drawn as a vertical box from q_1 to q_3 with "whiskers" going out to w_1 and w_2. If data points lie outside the whiskers they are often plotted as points.

Exercises

★   **Exercise 2.7.1.** Let P be the probability measure on $\mathbb{R}$ with density

$$f(x) = cx^2 1_{[0,1]}(x),$$

with c a constant that assures that

$$\int_0^1 f(x)\mathrm{d}x = 1.$$

Argue that f is a density and compute the constant c. Then compute the mean and the variance under P.

⌨   **Exercise 2.7.2.** Compute the mean and variance for the Gumbel distribution.
Hint: You are welcome to try to compute the integrals – it's difficult. Alternatively, you can compute the integrals numerically in R using the `integrate` function.

Exercise 2.7.3. Find the quantile function for the Gumbel distribution.

Exercise 2.7.4. Find the quantile function for the Weibull distribution, cf. Exercise 2.6.1. Make a QQ-plot of the quantiles from the Weibull distribution against the quantiles from the Gumbel distribution.

2.8 Conditional probabilities and independence

If we know that the event A has occurred, but don't have additional information about the outcome of our experiment, we want to assign a *conditional* probability to all other events $B \subseteq E$ – conditioning on the event A. For a given event A we aim at defining a conditional probability measure $P(\ \cdot\ |A)$ such that $P(B|A)$ is the conditional probability of B given A for any event $B \subseteq E$.

Definition 2.8.1. *The conditional probability measure $P(\cdot\ |A)$ for an event $A \subseteq E$ with $P(A) > 0$ is defined by*

$$P(B|A) = \frac{P(B \cap A)}{P(A)} \tag{2.17}$$

for any event $B \subseteq E$.

What we claim here is that $P(\ \cdot\ |A)$ really is a probability measure, but we need to show that. By the definition above we have that

$$P(E|A) = \frac{P(E \cap A)}{P(A)} = \frac{P(A)}{P(A)} = 1,$$

and if $B_1, \ldots, B_n$ are disjoint events

$$
\begin{aligned}
P\big(B_1 \cup \ldots \cup B_n \big| A\big) &= \frac{P\big((B_1 \cup \ldots \cup B_n) \cap A\big)}{P(A)} \\
&= \frac{P\big((B_1 \cap A) \cup \ldots \cup (B_n \cap A)\big)}{P(A)} \\
&= \frac{P(B_1 \cap A)}{P(A)} + \ldots + \frac{P(B_n \cap A)}{P(A)} \\
&= P(B_1|A) + \ldots P(B_n|A),
\end{aligned}
$$

where the third equality follows from the additivity property of P. This shows that $P(\cdot\ |A)$ is a probability measure and we have chosen to call it the conditional probability measure given A. It should be understood that this is a *definition* – though a completely reasonable and obvious one – and not a derivation of what conditional probabilities are. The frequency interpretation is in concordance with this definition: With n repeated experiments, $\varepsilon_n(A)$ is the fraction of outcomes where A occurs and $\varepsilon_n(B \cap A)$ is the fraction of outcomes where $B \cap A$ occurs, hence

$$\frac{\varepsilon_n(B \cap A)}{\varepsilon_n(A)}$$

is the fraction of outcomes where B occurs *among those outcomes where A occurs*. When believing in the frequency interpretation this fraction is approximately equal to $P(B|A)$ for n large.

Associated with conditional probabilities there are two major results known as the *Total Probability Theorem* and *Bayes Theorem*. These results tell us how to compute some probabilities from knowledge of other (conditional) probabilities. They are easy to derive from the definition.

If $A_1, \ldots, A_n$ are *disjoint* events in E and if $B \subseteq E$ is any event then from the definition of conditional probabilities

$$P(B|A_i)P(A_i) = P(B \cap A_i)$$

and since the events $A_1, \ldots, A_n$ are disjoint, so are $B \cap A_1, \ldots, B \cap A_n$, hence by additivity

$$P(B \cap A_1) + \ldots + P(B \cap A_n) = P\left((B \cap A_1) \cup \ldots \cup (B \cap A_n)\right) = P(B \cap A).$$

where $A = A_1 \cup \ldots \cup A_n$. This result tells us how to compute $P(B \cap A)$ from knowledge of the conditional probabilities $P(B|A_i)$ and the probabilities $P(A_i)$ – this is the Total Probability Theorem.

Result 2.8.2 (Total Probability Theorem). *If $A_1, \ldots, A_n$ are disjoint events in E, if $A = A_1 \cup \ldots \cup A_n$, and if $B \subseteq E$ is any event then*

$$P(B \cap A) = P(B|A_1)P(A_1) + \ldots + P(B|A_n)P(A_n). \tag{2.18}$$

If $A = E$ and $P(B) > 0$ we have that

$$P(A_i|B) = \frac{P(A_i \cap B)}{P(B)} = \frac{P(B|A_i)P(A_i)}{P(B)},$$

and if we use the Total Probability Theorem to express $P(B)$ we have derived the Bayes Theorem.

Result 2.8.3 (Bayes Theorem). *If $A_1, \ldots, A_n$ are disjoint events in E with $E = A_1 \cup \ldots \cup A_n$ and if $B \subseteq E$ is any event with $P(B) > 0$ then*

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B|A_1)P(A_1) + \ldots + P(B|A_n)P(A_n)} \tag{2.19}$$

for all $i = 1, \ldots, n$.

The Bayes Theorem is of central importance when calculating with conditional probabilities – whether or not we adopt a frequency or a Bayesian interpretation of probabilities. The example below shows a classical setup where Bayes Theorem is required.

Example 2.8.4. Drug tests are, for instance, used in cycling to test if cyclists take illegal drugs. Suppose that we have a test for the illegal drug EPO, with the property that 99% of the time it reveals (is positive) if a cyclist has taken EPO. This sounds like a good test – or does it? The 0.99 is the *conditional* probability that the test will be positive given that the cyclist uses EPO. To completely understand the merits of the test, we need some additional information about the test and about the percentage of cyclists that use the drug. To formalize, let $E = \{\text{tp}, \text{fp}, \text{tn}, \text{fn}\}$ be the sample space where tp = true positive, fp = false positive, tn = true negative, and fn = false negative. By tp we mean that the cyclist has taken EPO and the test shows that (is positive), by fp that the cyclist hasn't taken EPO but the test shows that anyway (is positive), by fn that the cyclist has taken EPO but the test doesn't show that (is negative), and finally by tn we mean that the cyclist hasn't taken EPO and that the test shows that (is negative). Furthermore, let $A_1 = \{\text{tp}, \text{fn}\}$ (the cyclist uses EPO) and let $A_2 = \{\text{fp}, \text{tn}\}$ (the cyclist does not use EPO). Finally, let $B = \{\text{tp}, \text{fp}\}$ (the test is positive). Assume that the conditional probability that the test is positive given that the cyclist is *not* using EPO is rather low, 0.04 say. Then what we know is:

$$
\begin{aligned}
P(B|A_1) &= 0.99 \\
P(B|A_2) &= 0.04.
\end{aligned}
$$

If we have high thoughts about professional cyclists we might think that only a small fraction, 7%, say, of them use EPO. Choosing cyclists for testing uniformly at random gives that $P(A_1) = 0.07$ and $P(A_2) = 0.93$. From Bayes Theorem we find that

$$
P(A_1|B) = \frac{0.99 \times 0.07}{0.99 \times 0.07 + 0.04 \times 0.93} = 0.65.
$$

Thus, conditionally on the test being positive, there is only 65% chance that he actually did take EPO. If a larger fraction, 30%, say, use EPO, we find instead that

$$
P(A_1|B) = \frac{0.99 \times 0.3}{0.99 \times 0.3 + 0.04 \times 0.7} = 0.91,
$$

which makes us more certain that the positive test actually caught an EPO user.

Besides the fact that "99% probability of revealing an EPO user" is insufficient information for judging whether the test is good, the point is that Bayes Theorem pops up in computations like these. Here we are given some information in terms of conditional probabilities, and we want to know some other conditional probabilities, which can then be computed from Bayes Theorem. $\diamond$

Having defined the conditional probabilities $P(B|A)$ for A and B events in E it is reasonable to say that the event B is *independent* of A if $P(B|A) = P(B)$ – that is, B is independent of A if the probability of B doesn't change even though we know that A has occurred. This implies that

$$
P(B)P(A) = P(B|A)P(A) = P(B \cap A)
$$

from which we see that, if $P(B) > 0$,

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = P(A).$$

Hence if B is independent of A then A is also independent of B. This reasoning leads us to the following definition of independence of two events, which doesn't refer to conditional probabilities explicitly.

Definition 2.8.5. *Two events $A, B \subseteq E$ are said to be independent if*

$$P(A \cap B) = P(A)P(B).$$

2.9   Random variables

It will be a considerable advantage to be able to talk about the *unrealized* outcome of an experiment as a variable, whose value has not yet been disclosed to us. Before the experiment is conducted the outcome is not known and it is described by a probability distribution, but when the experiment is over, a particular value in the sample space will be the result. The notion of a *random variable*, sometimes called a *stochastic variable*, allows us to talk about the outcome, that will take some particular value when the experiment is over, before actually conducting the experiment. We use capital letters like X, Y and Z to denote random variables. In these notes a random variable is *not* a precisely defined mathematical object but rather a useful notational convention. The notion of a random variable is often hard to grasp – how can a variable be random? The formalities can be equally difficult – what is the correct mathematical definition of a random variable? The pragmatic approach take here is that, at the end of the day we should have a useful notation for the variables we intend to measure or observe and the computations we intend to do on these measurements.

What we consider is an experiment with sample space E and governing probability measure P, and we will say that the outcome X is a random variable that takes values in E and has *distribution P*. For an event $A \subseteq E$ we will use the notational convention

$$\mathbb{P}(X \in A) = P(A).$$

It is important to understand that there is always a probability measure associated with a random variable – this measure being the distribution of the random variable. Sometimes we cannot tell or know exactly what the distribution is, but we rather have several *potential* probability measures in mind as candidates for the distribution of the random variable. We have more to say about this in the following chapters on statistics, which deal with figuring out, based on realizations of the experiment, what the distribution of the random variable(s) was.

Note that sometimes (but not in these notes) a random variable is assumed to take values in $\mathbb{R}$. We do not make this restriction, and if X takes values in $\mathbb{R}$ we will usually explicitly write that X is a real valued random variable.

Example 2.9.1. A binary experiment with sample space $E = \{0, 1\}$ is called a *Bernoulli experiment*. A random variable X representing the outcome of such a binary experiment is called a *Bernoulli variable*. The probability

$$p = \mathbb{P}(X = 1)$$

is often called the *success probability*. ◇

2.9.1 Transformations of random variables

A *transformation* is a map from one sample space into another sample space. Random variables can then be transformed using the map, and we are interested in the distribution of resulting random variable. Transformations are the bread-and-butter for doing statistics – it is crucial to understand how random variables and distributions on one sample space give rise to a range of transformed random variables and distributions. Abstractly there is not much to say, but once the reader recognizes how transformations play a key role throughout these notes, the importance of being able to handle transformations correctly should become clear.

If E and E' are two sample spaces, a transformation is a map

$$h : E \to E'$$

that assigns the transformed outcome $h(x)$ in E' to the outcome x in E. If X is a random variable, the h-transformed random variable of X, denoted by $h(X)$, is the random variable whose value is $h(x)$ if $X = x$. We use the notation

$$h^{-1}(A) = \{x \in E \mid h(x) \in A\}$$

for $A \subseteq E'$ to denote the event of outcomes in E for which the transformed outcome ends up in A.

Example 2.9.2. Transformations are done to data all the time. Any computation is essentially a transformation. One of the basic ones is computation of the sample mean. If X_1 and X_2 are two real valued random variables, their sample mean is

$$Y = \frac{1}{2}(X_1 + X_2).$$

The general treatment of transformations to follow will help us understand how we derive the distribution of the transformed random variable Y from the distribution of X_1 and X_2. ◇

Definition 2.9.3. *If P is a probability measure on E, the transformed probability measure, $h(P)$, on E' is given by*

$$h(P)(A) = P(h^{-1}(A)) = P(\{x \in E \mid h(x) \in A\})$$

for any event $A \subseteq E'$. If X is a random variable with distribution P, the distribution of $h(X)$ is $h(P)$.

We observe from the definition that for $A \subseteq E'$

$$\mathbb{P}(h(X) \in A) = h(P)(A) = P(h^{-1}(A)) = \mathbb{P}(X \in h^{-1}(A)).$$

This notation, $\mathbb{P}(h(X) \in A) = \mathbb{P}(X \in h^{-1}(A))$, is quite suggestive – to find the distribution of $h(X)$ we "move" h from the variable to the set by taking the "inverse". Indeed, if h has an inverse, i.e. there is a function $h^{-1} : E' \to E$ such that

$$h(x) = y \iff x = h^{-1}(y)$$

for all $x \in E$ and $y \in E'$, then

$$h(X) = y \iff X = h^{-1}(y).$$

Example 2.9.4 (Indicator random variables)**.** Let X be a random variable taking values in the sample space E and $A \subseteq E$ any event in E. Define the transformation

$$h : E \to \{0,1\}$$

by

$$h(x) = 1_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \in A^c \end{cases}.$$

Thus h is the *indicator function* for the event A. The corresponding transformed random variable

$$Y = h(X) = 1_A(X)$$

is called an indicator random variable. We sometimes also write

$$Y = 1(X \in A)$$

to show that Y indicates whether X takes its value in A or not. Since Y takes values in $\{0, 1\}$ it is a Bernoulli variable with success probability

$$p = \mathbb{P}(Y = 1) = \mathbb{P}(X \in A).$$

$\diamond$

Example 2.9.5. If E is a discrete set and P is a probability measure on E given by the point probabilities $p(x)$ for $x \in E$ then if $h : E \to E'$ the probability measure $h(P)$ has point probabilities

$$q(y) = \sum_{x:h(x)=y} p(x), \quad y \in h(E).$$

Indeed, for $y \in h(E)$ the set $h^{-1}(y)$ is non-empty and contains precisely those points whose image is y. Hence

$$q(y) = h(P)(\{y\}) = P(h^{-1}(y)) = \sum_{x \in h^{-1}(y)} p(x) = \sum_{x:h(x)=y} p(x).$$

As a consequence, if $h : E \to \mathbb{R}$ and if

$$\sum_{x \in E} |h(x)| p(x) < \infty$$

we can compute the mean under $h(P)$ as

$$\mu = \sum_{y \in h(E)} y q(y) = \sum_{y \in h(E)} y \sum_{x:h(x)=y} p(x) = \sum_{y \in h(E)} \sum_{x:h(x)=y} h(x)p(x) = \sum_{x \in E} h(x)p(x),$$

where we have used that the double sum is a sum over every $x \in E$ – just organized so that for each $y \in h(E)$ we first sum all values of $h(x)p(x)$ for those x with $h(x) = y$, and then sum these contributions over $y \in h(E)$.

Likewise, if

$$\sum_{x \in E} h(x)^2 p(x) < \infty$$

we can compute the variance under $h(P)$ as

$$\sigma^2 = \sum_{x \in E} (h(x) - \mu)^2 p(x) = \sum_{x \in E} h(x)^2 p(x) - \mu^2.$$

$\diamond$

Example 2.9.6 (Sign and symmetry). Let X be a real valued random variable whose distribution is given by the distribution function F and consider $h(x) = -x$. Then the distribution function for $Y = h(X) = -X$ is

$$G(x) = \mathbb{P}(Y \le x) = \mathbb{P}(X \ge -x) = 1 - \mathbb{P}(X < -x) = 1 - F(-x) + \mathbb{P}(X = -x).$$

We say that (the distribution of) X is symmetric if $G(x) = F(x)$ for all $x \in \mathbb{R}$. That is, X is symmetric if X and $-X$ have the same distribution. If F has density f we know that $\mathbb{P}(X = -x) = 0$ and it follows that the distribution of $h(X)$ has density $g(x) = f(-x)$ by differentiation of $G(x) = 1 - F(-x)$, and in this case it follows that X is symmetric if $f(x) = f(-x)$ for all $x \in \mathbb{R}$. $\diamond$

Example 2.9.7. Let X be a random variable with values in $\mathbb{R}$ and with distribution given by the distribution function F. Consider the random variable $|X|$ – the absolute value of X. We find that the distribution function for $|X|$ is

$$G(x) = \mathbb{P}(|X| \leq x) = \mathbb{P}(-x \leq X \leq x) = F(x) - F(-x) + \mathbb{P}(X = -x).$$

for $x \geq 0$.

If we instead consider X^2, we find the distribution function to be

$$G(x) = \mathbb{P}(X^2 \leq x) = \mathbb{P}(-\sqrt{x} \leq X \leq \sqrt{x}) = F(\sqrt{x}) - F(-\sqrt{x}) + \mathbb{P}(X = -\sqrt{x})$$

for $x \geq 0$. $\diamond$

Example 2.9.8 (Median absolute deviation)**.** We have previously defined the interquartile range as a measure of the spread of a distribution given in terms of quantiles. We introduce here an alternative measure. If X is a real valued random variable whose distribution has distribution function F and median q_2 we can consider the transformed random variable

$$Y = |X - q_2|,$$

which is the absolute deviation from the median. The distribution function for Y is

$$
\begin{aligned}
F_{\text{absdev}}(x) &= \mathbb{P}(Y \leq x) = \mathbb{P}(X \leq q_2 + x) - \mathbb{P}(X < q_2 - x) \\
&= F(q_2 + x) - F(q_2 - x) + \mathbb{P}(X = q_2 - x).
\end{aligned}
$$

The *median absolute deviation* is defined as

$$\text{MAD} = \text{median}(F_{\text{absdev}}).$$

The median absolute deviation is like the interquartile range a number that represents how spread out around the median the distribution is.

For a symmetric distribution we always have the median $q_2 = 0$ and therefore we have $F_{\text{absdev}} = 2F - 1$. Using the definition of the median we get that if x is MAD then

$$2F(x - \varepsilon) - 1 \leq \frac{1}{2} \leq 2F(x) - 1$$

from which is follows that $F(x - \varepsilon) \leq 3/4 \leq F(x)$. In this case it follows that MAD is in fact equal to the upper quartile q_3. Using symmetry again one can observe that the lower quartile $q_1 = -q_3$ and hence for a symmetric distribution MAD equals half the interquartile range. $\diamond$

Example 2.9.9 (Location and Scale)**.** Let X denote a real valued random variable with distribution given by the distribution function $F : \mathbb{R} \rightarrow [0, 1]$. Consider the transformation $h : \mathbb{R} \rightarrow \mathbb{R}$ given by

$$h(x) = \sigma x + \mu$$

for some constants $\mu \in \mathbb{R}$ and $\sigma > 0$. Then the distribution of

$$Y = h(X) = \sigma X + \mu$$

has distribution function

$$G(x) = \mathbb{P}(h(X) \le x) = \mathbb{P}\left(X \le \frac{x - \mu}{\sigma}\right) = F\left(\frac{x - \mu}{\sigma}\right).$$

If F is differentiable we know that the distribution of X has density

$$f(x) = F'(x).$$

We observe that G is also differentiable and applying the chain rule for differentiation we find that the distribution of Y has density

$$g(x) = G'(x) = \frac{1}{\sigma} f\left(\frac{x - \mu}{\sigma}\right).$$

The random variable Y is a translated and scaled version of X. The parameter μ is known as the location parameter and σ as the scale parameter. We observe that from a single distribution given by the distribution function F we obtain a two-parameter family of distributions from F by translation and scaling.

If the distribution of X has density f, mean 0 and variance 1 we can compute the mean and variance of the distribution of Y. From Definition 2.7.5 it follows, using the substitution $y = \frac{x - \mu}{\sigma}$, that the mean is

$$
\begin{aligned}
\int_{-\infty}^{\infty} x \frac{1}{\sigma} f\left(\frac{x - \mu}{\sigma}\right) \mathrm{d}x &= \int_{-\infty}^{\infty} (\sigma y + \mu) f(y) \mathrm{d}y \\
&= \sigma \int_{-\infty}^{\infty} y f(y) \mathrm{d}y + \mu \int_{-\infty}^{\infty} f(y) \mathrm{d}y = \mu.
\end{aligned}
$$

Likewise, using the same substitution, the variance is

$$
\begin{aligned}
\int_{-\infty}^{\infty} (x - \mu)^2 \frac{1}{\sigma} f\left(\frac{x - \mu}{\sigma}\right) \mathrm{d}x &= \sigma^2 \int_{-\infty}^{\infty} \left(\frac{x - \mu}{\sigma}\right)^2 f\left(\frac{x - \mu}{\sigma}\right) \frac{1}{\sigma} \mathrm{d}x \\
&= \sigma^2 \int_{-\infty}^{\infty} y^2 f(y) \mathrm{d}y = \sigma^2.
\end{aligned}
$$

It follows that the normal distribution with location parameter μ and scale parameter σ has density

$$\frac{1}{\sigma \sqrt{2\pi}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right). \tag{2.20}$$

The abbreviation

$$X \sim N(\mu, \sigma^2).$$

is often used to denote a random variable X, which is normally distributed with location parameter μ and scale parameter σ. In the light of Example 2.7.10 the normal distribution $N(0,1)$ has mean 0 and variance 1, thus the $N(\mu, \sigma^2)$ normal distribution with location parameter μ and scale parameter σ has mean μ and variance σ^2.

$\diamond$

R Box 2.9.1. For some of the standard distributions on the real line $\mathbb{R}$, one can easily supply additional parameters specifying the location and scale in R. For the normal distribution `pnorm(x,1,2)` equals the density at x with location parameter $\mu = 1$ and scale parameter $\sigma = 2$. Similarly, `plogis(x,1,2)` gives the density for the logistic distribution at `x` with location parameter $\mu = 1$ and scale parameter $\sigma = 2$.

Example 2.9.10 (QQ-plots). When we want to compare an empirical distribution to a theoretical distribution we are most often interested in a comparison where we just know the *shape* of the distribution but not the location and scale. If X has distribution with distribution function F, quantile function Q and our dataset is $x_1, \ldots, x_n$ each being realizations of a random variable with the same distribution as

$$\mu + \sigma X$$

for some unknown scale $\sigma > 0$ and location $\mu \in \mathbb{R}$. If we make a QQ-plot of the empirical quantile function against Q it will still result in points that are close to a straight line, but with slope σ and intercept μ. This is because $Q_{\mu,\sigma}(y)$ defined by

$$Q_{\mu,\sigma}(y) = \mu + \sigma Q(y)$$

is actually a quantile function for the distribution of $\mu + \sigma X$. To see this, recall that the distribution function for $\mu + \sigma X$ is

$$F_{\mu,\sigma}(x) = F\left(\frac{x - \mu}{\sigma}\right),$$

hence for any $y \in (0,1)$ and $\varepsilon > 0$

$$
\begin{aligned}
F_{\mu,\sigma}(Q_{\mu,\sigma}(y) - \varepsilon) &= F\left(\frac{\sigma Q(y) + \mu - \mu - \varepsilon}{\sigma}\right) = F\left(Q(y) - \varepsilon/\sigma\right) \\
&\leq y \\
&\leq F(Q(y)) = F_{\mu,\sigma}(Q_{\mu,\sigma}(y)).
\end{aligned}
$$

This adds considerable value to the QQ-plot as it can now be used to justify the choice of the shape of the distribution without having to discuss how the unknown scale and location parameters are chosen or perhaps estimated from data. $\diamond$
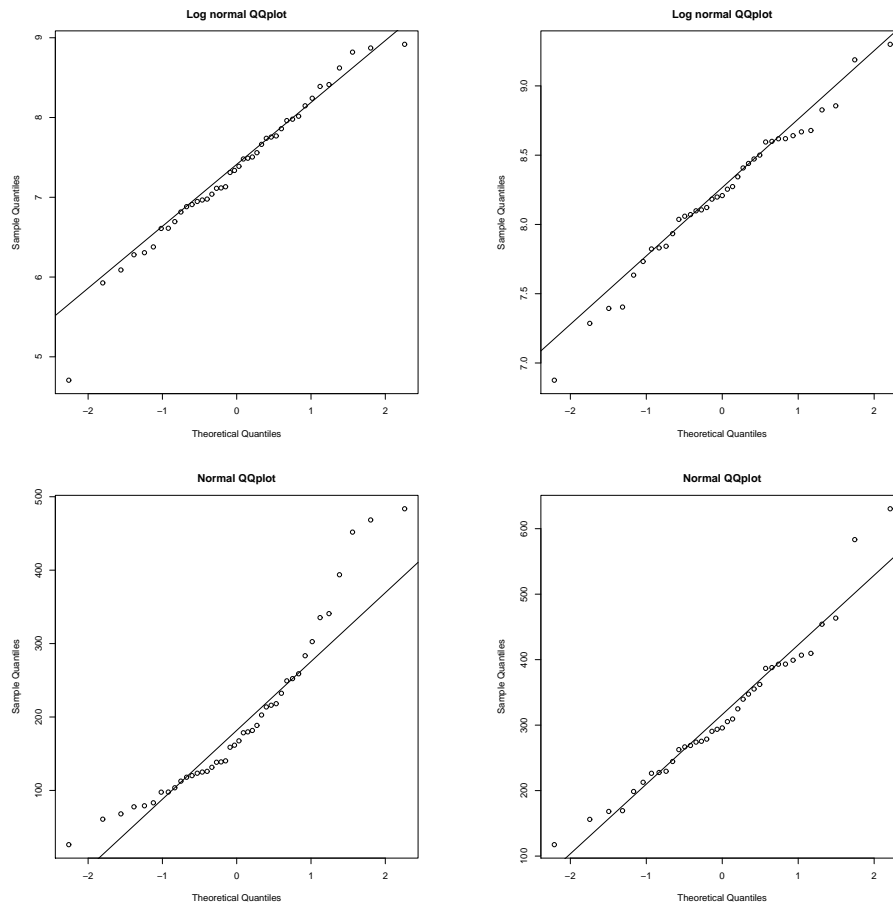
Figure 2.18: QQplots for gene 1635_at from the ALL dataset. Here we see the expression levels and log (base 2) expression levels against the normal distribution with (right) or without (left) presence of the BCR/ABL fusion gene.

Exercises

Exercise 2.9.1. Compute the interquartile range and the median absolute deviation for the $N(0, 1)$ standard normal distribution.

Exercise 2.9.2. Find the density for the distribution of $\exp(X)$ if X is normally distributed. This is the *log-normal distribution*. Make a QQ-plot of the normal distribution against the log-normal distribution.

★   **Exercise 2.9.3.** Assume that X is a real valued random variable, whose distribution has density f and distribution function F. Let $h : \mathbb{R} \to [0, \infty)$ be the transformation $h(x) = x^2$. Show that

$$\mathbb{P}(h(X) \leq y) = 1 - F(-\sqrt{y}) + F(\sqrt{y}),$$

and then show that the distribution of $h(X)$ has density

$$g(y) = \frac{f(-\sqrt{y}) + f(\sqrt{y})}{2\sqrt{y}}$$

for $y > 0$. Argue that if the distribution of X is symmetric then this expression reduces to

$$g(x) = \frac{f(\sqrt{y})}{\sqrt{y}}.$$

★   **Exercise 2.9.4.** Assume that the distribution of X is the normal distribution. Show that the distribution of X^2 has density

$$g(y) = \frac{1}{\sqrt{2\pi y}} \exp\left(-\frac{y}{2}\right)$$

for $y > 0$.

2.10   Joint distributions, conditional distributions and independence

For all practical purposes we have more than one random variable in the game when we do statistics. A single random variable carries a probability distribution on a sample space, but if we hope to infer unknown aspects of this distribution, it is rarely sufficient to have just one observation. Indeed, in practice we have a data set consisting of several observations as in all the examples previously considered. The question, we need to ask, is which assumptions we want to make on the interrelation between the random variables representing the entire data set?

If we have n random variables, $X_1, X_2, \ldots, X_n$, they each have a distribution, which we refer to as their *marginal distributions*. Are the marginal distributions enough? No, not at all! Lets consider the following example.

Example 2.10.1. Let X and Y are two random variables representing a particular DNA letter in the genome in two different but related organisms. Lets assume for simplicity that each letter has the uniform distribution as the marginal distribution. What is the probability of observing $(X = \texttt{A}, Y = \texttt{A})$? If the events $(X = \texttt{A})$ and $(Y = \texttt{A})$ were *independent*, we know that the correct answer is $0.25 \times 0.25 = 0.125$. However, we claimed that the organims are related, and the interpretation is that the occurrence of a given letter in one of the organims, at this particular position, gives us information about the letter in the other organism. What the correct probability is depends on many details, e.g. the evolutionary distance between the organims. In

Subsection 2.10.2 we consider two examples of such models, the Jukes-Cantor model and the Kimura model.   ◇

The probability measure the governs the combined behaviour of the pair (X, Y) above is often referred to as the *joint distribution* of the variables. We need to know the joint distribution to compute probabilities involving more than one the random variable. The marginal probability distributions, given in general as

$$P_i(A) = \mathbb{P}(X_i \in A),$$

are sufficient if we only want to compute probabilities of events involving single random variables only.

Joint distributions can be arbitrarily complicated, but we defer the general treatment to Section 2.13 and focus here on the concept of independence.

2.10.1   Random variables and independence

Whenever we have observations of variables we believe are *not* interrelated, we will use the *independence assumption*, which is formally a specification of the joint distribution of the variables from the knowledge of their marginal distributions only.

Definition 2.10.2 (Independence). *Let $X_1, X_2, \ldots, X_n$ be n random variables such that X_i takes values in a sample space E_i for $i = 1, \ldots, n$. We say that $X_1, \ldots, X_n$ are independent if*

$$\mathbb{P}(X_1 \in A_1, \ldots, X_n \in A_n) = \mathbb{P}(X_1 \in A_1) \cdot \ldots \cdot \mathbb{P}(X_n \in A_n)$$

for all events $A_1 \subseteq E_1, \ldots, A_n \subseteq E_n$. Moreover, for given marginal distributions of the X_i's, the right hand side above specifies a unique probability distribution that defines a joint distribution of the X_i's.

With this definition we can specify the distribution of $X_1, \ldots, X_n$ by specifying only the marginal distributions and say the magic words; the X_i's are *independent*. If the variables all represent the same experiment, hence they all take values in the same sample space E, we can in addition assume that the marginal distributions are identical. In that case we say that the random variables $X_1, \ldots, X_n$ are *independent* and *identically distributed*, which is typically abbreviated iid.

If we consider n real valued random variables $X_1, X_2, \ldots, X_n$ such that the marginal density for the distribution of X_i is f_i, then if the joint density, f, factorizes as

$$f(x_1, \ldots, x_n) = f_1(x_1) \cdot \ldots \cdot f_n(x_n)$$

we get that

$$
\begin{aligned}
\mathbb{P}(X_1 \in A_1, \ldots, X_n \in A_n) &= \int_{A_1} \cdots \int_{A_n} f_1(x_1) \cdot \ldots \cdot f_n(x_n) \mathrm{d}x_n \ldots \mathrm{d}x_1 \\
&= \int_{A_1} f_1(x_1) \mathrm{d}x_1 \cdot \ldots \cdot \int_{A_n} f_n(x_n) \mathrm{d}x_n \\
&= \mathbb{P}(X_1 \in A_1) \cdot \ldots \cdot \mathbb{P}(X_n \in A_n).
\end{aligned}
$$

A similar computation holds with point probabilities, which replaces the densities on a discrete sample space, and where the integrals are replaced with sums. We summarize this as follows.

Result 2.10.3. *Let $X_1, X_2, \ldots, X_n$ be n random variables. If the sample spaces E_i are discrete, if the marginal distributions have point probabilities $p_i(x)$, $x \in E_i$ and $i = 1, \ldots, n$, and the point probabilities for the distribution of $(X_1, \ldots, X_n)$ factorize as*

$$
P(X_1 = x_1, \ldots, X_n = x_n) = p(x_1, \ldots, x_n) = p_1(x_1) \cdot \ldots \cdot p_n(x_n)
$$

then the X_i's are independent.

If $E_i = \mathbb{R}$, if the marginal distributions have densities $f_i : \mathbb{R} \to [0, \infty)$, $i = 1, \ldots, n$, then if the (n-dimensional) density for the distribution of $(X_1, \ldots, X_n)$ factorizes as

$$
f(x_1, \ldots, x_n) = f_1(x_1) \cdot \ldots \cdot f_n(x_n)
$$

the X_i's are independent.

The result above is mostly used as follows. If we want to construct a joint distribution of $X_1, \ldots, X_n$ and we want the marginal distribution of X_i to be $N(\mu, \sigma^2)$, say, for $i = 1, \ldots, n$ and we want $X_1, \ldots, X_n$ to be independent, the theorem above says that this is actually what we obtain by taking the joint distribution to have the density that is the product of the marginal densities. Using the properties of the exponential function we see that the density for the joint distribution of n iid $N(\mu, \sigma^2)$ distributed random variables is then

$$
f(x_1, \ldots, x_n) = \frac{1}{\sigma^n (\sqrt{2\pi})^n} \exp\left(-\sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2} \right).
$$

Example 2.10.4. One example of a computation where we need the joint distribution is the following. If X_1 and X_2 are two real valued random variables, we are interested in computing the distribution function for their sum, $Y = X_1 + X_2$. If the joint density is f this computation is

$$
F(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(X_1 + X_2 \leq y) = \int \int_{\{(x_1, x_2) \mid x_1 + x_2 \leq y\}} f(x_1, x_2) \mathrm{d}x_1 \mathrm{d}x_2.
$$

If X_1 and X_2 are *independent* we have that $f(x_1, x_2) = f_1(x_1)f_2(x_2)$ and the double integral can we written as

$$F(y) = \int_{-\infty}^{\infty} \int_{-\infty}^{y-x_2} f_1(x_1)\mathrm{d}x_1 f_2(x_2)\mathrm{d}x_2.$$

Whether we can compute this double integral in practice is another story, which depends on the densities f_1 and f_2.   ◇

Example 2.10.5. We consider the random variables X and Y that take values in $E_0 \times E_0$ with $E_0 = \{\texttt{A}, \texttt{C}, \texttt{G}, \texttt{T}\}$ and with point probabilities given by the matrix

			Y		
		A	C	G	T	
	A	0.0401	0.0537	0.0512	0.0400	0.1850
	C	0.0654	0.0874	0.0833	0.0651	0.3012
X	G	0.0634	0.0848	0.0809	0.0632	0.2923
	T	0.0481	0.0643	0.0613	0.0479	0.2215
		0.217	0.2901	0.2766	0.2163	

The marginal distribution of X and Y respectively are also given as the row sums (right column) and the column sums (bottom row). It is in fact this matrix representation of discrete, multivariate distributions with "the marginal distributions in the margin" that caused the name "marginal distributions". One verifies easily – or by using R – that all entries in this matrix are products of their marginals. For instance,

$$P(X = \texttt{G}, Y = \texttt{C}) = 0.0848 = 0.2923 \times 0.2901.$$

◇

Example 2.10.6 (Hardy-Weinberg equilibrium)**.** When diploid organisms like humans reproduce, the offspring receives one copy of each chromosome from each of the parents. Thus if the father carries the *aa* allele combination of a particular gene and the mother carries the *Aa* allele combination, then the offspring can get both of the allele combinations *aa* and *Aa*. The *Hardy-Weinberg equilibrium* is an assumption about *independence* between the alleles that are passed on from the mother and from the father.

In a large population we sample uniformly a random individual and let X_m and X_f denote the random variables representing the allele that was passed on to the individual from the mother and the father, respectively. We cannot observe X_m and X_f, as we cannot from the individual determine which of the observed alleles is from the mother, and which is from the father.

If we assume that the male and female subpopulations have the same proportion $p \in [0, 1]$ of allele a, the marginal distribution of X_m and X_f is given by

$$\mathbb{P}(X_m = a) = \mathbb{P}(X_f = a) = p$$

R Box 2.10.1 (Outer products). To compute all 16 products of pairs of entries in the vectors `rmp <- c(0.185, 0.3012, 0.2923, 0.2215)` and `cmp <- c(0.217,0.2901,0.2766,0.2163)` is perhaps a little tedious. A convenient operation in R is the *outer product*, which does exactly this:

```
> rmp %o% cmp
        [,1]    [,2]    [,3]    [,4]
[1,]   0.0401  0.0537  0.0512  0.0400
[2,]   0.0654  0.0874  0.0833  0.0651
[3,]   0.0634  0.0848  0.0809  0.0632
[4,]   0.0481  0.0643  0.0613  0.0479
```

Alternatively, `rmp %*% t(cmp)` or `outer(rmp,cmp,"*")` does the same job in this case. The former works correctly for vectors only. The binary operator `%o%` is a wrapper for the latter, which works for any function `f` of two variables as well. If we want to compute `f(rmp[i],cmp[j])`, say, for all combinations of `i` and `j`, this is done by `outer(rmp,cmp,f)`. Note that when we use `outer` with the arithmetic operators like multiplication, `*`, we need the quotation marks.

---

and

$$\mathbb{P}(X_m = A) = \mathbb{P}(X_f = A) = 1 - p.$$

If X_f and X_m are *independent* the offspring allele combination, Y, which we can observe, has the distribution given by

$$
\begin{aligned}
\mathbb{P}(Y = aa) &= \mathbb{P}(X_f = a, X_m = a) = \mathbb{P}(X_f = a)\mathbb{P}(X_m = a) = p^2 \\
\mathbb{P}(Y = Aa) &= \mathbb{P}(X_f = A, X_m = a) + \mathbb{P}(X_f = a, X_m = A) = 2p(1-p) \\
\mathbb{P}(Y = AA) &= \mathbb{P}(X_f = A, X_m = A) = (1-p)^2.
\end{aligned}
$$

This distribution of the allele combination given in terms of the proportions of the two alleles in the population is the Hardy-Weinberg equilibrium. In reality this is not really a question of whether the population is in any sort of equilibrium. It is a distributional consequence of an independence assumption.   ◇

Example 2.10.7 (Linkage). The alleles of two genes, or markers for that matter, that are located on the same chromosome may occur in an associated way. For example, consider a gene that occurs as allele a or A and another gene that occurs as allele b and B, and we let X_1 and X_2 be random variables that represent the allele we find on (one of) the chromosomes in a random individual. The marginal distribution of X_1 and X_2 are given by

$$\mathbb{P}(X_1 = a) = p \quad \text{and} \quad \mathbb{P}(X_2 = b) = q$$

where $p, q \in [0, 1]$ are the proportions of alleles a and b, respectively, in the population. If X_1 and X_2 are independent, we have that the distribution of (X_1, X_2) is

given by

$$
\begin{aligned}
\mathbb{P}(X_1 = a, X_2 = b) &= pq \\
\mathbb{P}(X_1 = a, X_2 = B) &= p(1-q) \\
\mathbb{P}(X_1 = A, X_2 = b) &= (1-p)q \\
\mathbb{P}(X_1 = A, X_2 = B) &= (1-p)(1-q).
\end{aligned}
$$

If the distribution of (X_1, X_2) deviates from this we have *linkage*. This is another way of saying that we have dependence. $\diamond$

When can we assume independence? This is a very important question, and it is not easy to answer. In some cases there are good arguments that can justify the independence assumption. In the example above on the Hardy-Weinberg equilibrium the independence assumption is often refered to as *random mating*. This may or may not be questionable, and we have to think about what possible alternatives there are. A model of non-random mating will require some more thoughts, and how will it affect the allele frequencies? Instead of arguing for or against, we could also accept the random mating assumption, and simply try to check if the resulting allele model fits the data. After all, the derived allele frequencies form the important part of the model. Checking the model in this case amounts to fitting the more general model where the three allele combinations have an arbitrary distribution, and then we compare the more general model with the model obtained under the Hardy-Weinberg equilibrium assumption.

In general terms, whenever a model assumption is made it should be justifiable or at least not contrary to actual belief. This holds for the independence assumption as well. When this is said there are many situations where the data at hand are analyzed under an independence assumption that is not valid. The reasons can be many – from ignorance or lack of skills over lack of better alternatives to deliberate choices for efficiency reasons, say. In general, to use a model that does not fit the data actually analyzed can potentially lead to biased or downright wrong results. If one is in doubt about such things, the best advice is to seek assistance from a more experienced person. There are no magic words or spells that can make problems with lack of model fit go away.

2.10.2 Random variables and conditional distributions

If X is a random variable with distribution P, taking values in the sample space E, and if $A \subseteq E$, then the *conditional distribution* of X given that $X \in A$ is the probability measure $P(\,\cdot\mid A)$. We write

$$
\mathbb{P}(X \in B \mid X \in A) = P(B \mid A)
$$

for $B \subseteq E$.

For two random variables we make the following definition.

Definition 2.10.8. *If X and Y are two random variables taking values in the sample spaces E_1 and E_2 respectively and if $A \subseteq E_1$, then the conditional distribution of Y given that $X \in A$ is defined as*

$$\mathbb{P}(Y \in B | X \in A) = \frac{\mathbb{P}(Y \in B, X \in A)}{\mathbb{P}(X \in A)}$$

for $B \subseteq E_2$ provided that $\mathbb{P}(X \in A) > 0$.

Note that if $\mathbb{P}(X \in A) = 0$, the conditional distribution of Y given that $X \in A$ is not defined. Note also that

$$\mathbb{P}(Y \in B, X \in A) = \mathbb{P}(Y \in B | X \in A)\mathbb{P}(X \in A), \qquad (2.21)$$

and that this equality holds no matter how we define $\mathbb{P}(Y \in B | X \in A)$ in cases where $\mathbb{P}(X \in A) = 0$ (if $\mathbb{P}(X \in A) = 0$ the equality reads that $0 = 0$ irrespectively of the definition of the conditional probability).

Considering discrete sample spaces E_1 and E_2 the conditional distribution of Y given $X = x$ can be given in terms of point probabilities $p(y|x)$, $y \in E_2$. If the joint distribution of X and Y has point probabilities $p(x, y)$ for $x \in E_1$ and $y \in E_2$, then by definition

$$p(y|x) = \mathbb{P}(Y = y | X = x) = \frac{\mathbb{P}(Y = y, X = x)}{\mathbb{P}(X = x)} = \frac{p(x, y)}{\sum_{y \in E_2} p(x, y)} \qquad (2.22)$$

where the third equality follows from the fact that the marginal distribution of X has point probabilities

$$p_1(x) = \sum_{y \in E_2} p(x, y).$$

With $A = \{x\}$ and $B = \{y\}$ the formula given by (2.21) reads

$$p(x, y) = p(y|x)p_1(x). \qquad (2.23)$$

Example 2.10.9. We consider the random variables X and Y as in Example 2.13.3, thus the point probabilities are given by the matrix

		A	C	G	T	
		A	C	G	T	
	A	0.1272	0.0063	0.0464	0.0051	0.1850
X	C	0.0196	0.2008	0.0082	0.0726	0.3012
	G	0.0556	0.0145	0.2151	0.0071	0.2923
	T	0.0146	0.0685	0.0069	0.1315	0.2215
		0.2170	0.2901	0.2766	0.2163	

> **Math Box 2.10.1** (More about conditional distributions). If P denotes the joint distribution of (X, Y) on $E_1 \times E_2$, then the conditional probability measure given the event $A \times E_2$ for $A \subseteq E_1$ takes the form
>
> $$P(B_1 \times B_2 | A \times E_2) = \frac{P(B_1 \times B_2 \cap A \times E_2)}{P(A \times E_2)} = \frac{P((B_1 \cap A) \times B_2))}{P_1(A)}$$
>
> for $B_1 \subseteq E_1$ and $B_2 \subseteq E_2$. Here we use $P_1(A)$ to denote the marginal distribution of X. This conditional probability measure is the conditional distribution of the bundled variable (X, Y) given that $X \in A$. The conditional distribution of Y given that $X \in A$, as defined above, is recognized as the second marginal of the conditional distribution of (X, Y) given that $X \in A$. By this we mean that if $P(\cdot | E_1 \times A)$ denotes the distribution of (X, Y) conditionally on $X \in A$, which is a probability measure on $E_1 \times E_2$, then the second marginal of this measure, which is a probability measure on E_2, coincides with the conditional distribution of Y given $X \in A$.

Here the additional right column and the bottom row shows the marginal distributions of X and Y respectively. Note that these marginals are the same as considered in Example 2.10.5, but that the point probabilities for the joint distribution of X and Y certainly differ from the product of the marginals. Thus X and Y are not independent. We can compute the conditional distribution of Y given X as having point probabilities

$$\mathbb{P}(Y = y | X = x) = \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(X = x)} = \frac{p(x, y)}{\sum_{y \in E} p(x, y)}.$$

Note that we have to divide by precisely the row sums above. The resulting matrix of conditional distributions is

		A	C	G	T
		Y
X	A	0.6874	0.0343	0.2507	0.0276
	C	0.0649	0.6667	0.0273	0.2411
	G	0.1904	0.0495	0.7359	0.0242
	T	0.0658	0.3093	0.0311	0.5938

The rows in this matrix are conditional point probabilities for the distribution of Y conditionally on X being equal to the letter on the left hand side of the row. Each row sums by definition to 1. Such a matrix of conditional probabilities is called a *transition probability matrix*. In terms of mutations (substitutions) in a DNA-sequence the interpretation is that fixing one nucleic acid (the X) in the first sequence we can read of from this matrix the probability of finding any of the four nucleic acids at the corresponding position in a second evolutionarily related sequence. Note that the nomenclature in molecular evolution traditionally is that a *transition* is a

change from a pyrimidine to a pyrimidine or a purine to a purine, whereas a change from a pyrimidine to a purine or vice versa is called a *transversion*. ◇

In probability theory the conditional probabilities are computed from the definition and formulas above. In statistics, however, conditional probabilities are often used as the fundamental building blocks when specifying a probability model. We consider two concrete examples from molecular evolution and then the general concept of structural equations.

Example 2.10.10 (The Jukes-Cantor model)**.** It is possible to systematically specify models of molecular evolution where there is a time parameter $t \geq 0$ that dictates how dependent two nucleotides are. The simplest such model is the *Jukes-Cantor* model. There is one parameter $\alpha > 0$ in addition to the time parameter that tells how many mutations occur per time unit. The conditional probabilities on the DNA alphabet are

$$
\begin{array}{rcl}
P^t(x,x) &=& 0.25 + 0.75 \times \exp(-4\alpha t) \\
P^t(x,y) &=& 0.25 - 0.25 \times \exp(-4\alpha t), \quad \text{if } x \neq y,
\end{array}
$$

where $P^t(x,y)$ is the conditional probability $\mathbb{P}(Y = y | X = x)$ when X and Y are separated by the evolutionary time t. ◇

Example 2.10.11 (The Kimura model)**.** Another slightly more complicated model of molecular evolution is the *Kimura model*. The model captures the observed fact that a substitution of a purine with a purine or a pyrimidine with a pyrimidine (a transition) is happening in the course of evolution with a different rate than a substitution of a purine with pyrimidine or pyrimidine with purine (a transversion). If we let $\alpha > 0$ be a parameter determining the rate of transitions and $\beta > 0$ a parameter determining the rate of a transversion the Kimura model is given by

$$
\begin{array}{rcl}
P^t(x,x) &=& 0.25 + 0.25 \exp(-4\beta t) + 0.5 \exp(-2(\alpha+\beta)t) \\
P^t(x,y) &=& 0.25 + 0.25 \exp(-4\beta t) - 0.5 \exp(-2(\alpha+\beta)t), \quad \text{for a transversion} \\
P^t(x,y) &=& 0.25 - 0.25 \exp(-4\beta t), \quad \text{for a transition}
\end{array}
$$

with $P^t(x,y)$ as above is the conditional probability $\mathbb{P}(Y = y | X = x)$ when X and Y are separated by the evolutionary time t. ◇

The two models, the Jukes-Cantor model and the Kimura model, may seem a little arbitrary when encountered as given above. They are not, but they fit into a more general and systematic model construction that is deferred to Section 2.15. There it is shown that the models arise as solutions to a system of differential equations with specific interpretations.
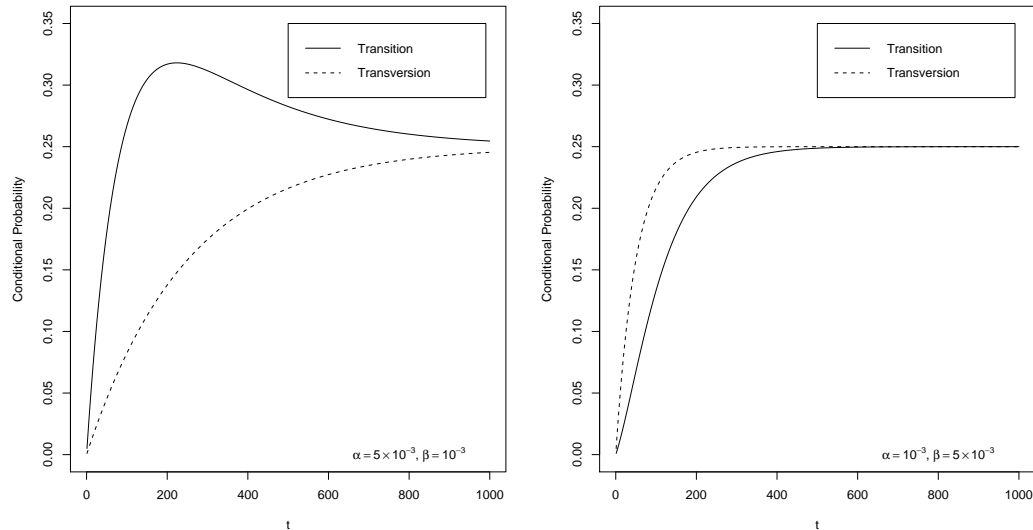
Figure 2.19: Two examples of transition and transversion probabilities for the Kimura model as a function of evolutionary distance in units of time.

Both of the previous examples gave the conditional distributions directly in terms of a formula for the conditional point probabilities. The following examples use a different strategy. Here the conditional probabilities are given by a transformation.

Example 2.10.12. If X and ε are real valued random variables, a *structural equation* defining Y is an equation of the form

$$Y = h(X, \varepsilon)$$

where h is any function taking values in a sample space E. The interpretation of the structural equation is as a conditional probability distribution of Y given X. That is, the conditional probability distribution of Y given $X = x$ is $h(x, \varepsilon)$ – a transformation of the distribution of ε by $h(x, \cdot)$ where x is fixed.

Structural equations are very useful for model specifications and are used quite extensively but often more or less implicitly in the literature. One thing to keep in mind is that the combination of the choice of h and the distribution of ε is *not* uniquely deducible from the conditional distributions they specify. Hence the structural equation itself is deducible from data alone.   ⋄

From a statistical point of view there are at least two practical interpretations of the structural equation model. One is that the values of X are fixed by us – we design the experiment and choose the values. We could, for instance, administer a

(toxic) compound at different doses to a bunch of flies and observe what happens, cf. Example 1.2.5. The variable ε captures the uncertainty in the experiment – at a given dose only a fraction of the flies will die. Another possibility is to observe (X, Y) jointly where, again, Y may be the death or survival of an insect but X is the concentration of a given compound as measured from the dirt sample where the insect is collected. Thus we do not decide the concentration levels. In the latter case the structural equation model may still provide a useful model of the observed, conditional distribution. However, we should be careful with the interpretation. If we intervene and force X to take a particular value, it may not have the expected effect, as predicted by the structural equation model, on the outcome Y. A explanation, why this is so, is that X and ε may be *dependent*. Thus if we want to interpret the structural equation correctly when we make interventions, the X and ε variables should be independent. One consequence of this is, that when we design an experiment it is essential to break any relation between the dose level X and the noise variable ε. The usual way to achieve this is to *randomize*, that is, to assign the dose levels randomly to the flies.

The discussion about interpretations of structural equations is only a scratch in the surface on the whole discussion of causal inference and causal conclusions from statistical data analysis. We also touched upon this issue in Example 1.2.5. Briefly, probability distributions are descriptive tools that are able to capture association, but by themselves they provide no explanations about the causal nature of things. Additional assumptions or controlled experimental designs are necessary to facilitate causal conclusions.

We treat two concrete examples of structural equation models, the *probit regression model* and the *linear regression model*.

Example 2.10.13. If Φ denotes the distribution function for the normal distribution and ε is uniformly distributed on $[0, 1]$ we define

$$Y = 1(\varepsilon \leq \Phi(\alpha + \beta X)).$$

Thus the sample space for Y is $\{0, 1\}$ and $h(x, \varepsilon) = 1(\varepsilon \leq \Phi(\alpha + \beta x))$ is the indicator function that ε is smaller than or equal $\Phi(\alpha + \beta x)$. We find that

$$\mathbb{P}(Y = 1) = \mathbb{P}(\varepsilon \leq \Phi(\alpha + \beta x)) = \Phi(\alpha + \beta x).$$

The resulting model of the conditional distribution of the Bernoulli variable Y given $X = x$ is known as the *probit regression model*. $\diamond$

Example 2.10.14. If $\varepsilon \sim N(0, \sigma^2)$ and

$$Y = \alpha + \beta X + \varepsilon$$

we have a structural equation model with

$$h(x, \varepsilon) = \alpha + \beta x + \varepsilon.$$

This model is the *linear regression model*, and is perhaps the single most important statistical model. For fixed x this is a location transformation of the normal distribution, thus the model specifies that

$$Y \mid X = x \sim N(\alpha + \beta x, \sigma^2).$$

In words, the structural equation says that there is a linear relationship between the Y variable and the X variable plus some additional "noise" as given by ε. One important observation is that there is an embedded asymmetry in the model. Even though we could "solve" the equation above in terms of X and write

$$X = -\frac{\alpha}{\beta} + \frac{1}{\beta}Y - \frac{1}{\beta}\varepsilon$$

this formula does not qualify for being a structural equation too. The explanation is that if X and ε are independent then Y and ε are not!   ◇

To illustrate the point with causality and the asymmetry of structural equations lets elaborate a little on the saying that "mud does not cause rain". Assume that the measurable level of mud is Y and that the measurable amount of rain is X and that the model

$$Y = \alpha + \beta X + \varepsilon$$

is a good model of how the mud level increases ($\beta > 0$) when it rains. Thus if we perform the rain dance and it starts raining, then the mud level increases accordingly, and by making it rain we have caused more mud. However, we cannot turn this around. It won't start raining just because we add more mud. The point is that the correct interpretation of a structural equation is tied closely together with the quantities that we model, and the equality sign in a structural equation should be interpreted as an assignment from right to left. Arguably, the more suggestive notation

$$Y \leftarrow \alpha + \beta X + \varepsilon$$

could be used – similarly to the assignment operator used in R.

2.10.3   Transformations of independent variables

The first thing to observe is a somewhat "obvious" fact. If $h_1 : E_1 \rightarrow E_1'$ and $h_2 : E_2 \rightarrow E_2'$ are two transformations, if X_1 and X_2 are two *independent* random variables taking values in E_1 and E_2, respectively, and if $A_1 \subseteq E_1'$ and $A_2 \subseteq E_2'$ are two events

$$
\begin{aligned}
\mathbb{P}(h_1(X_1) \in A_1, h_2(X_2) \in A_2) &= \mathbb{P}(X_1 \in h^{-1}(A_1), X_2 \in h^{-1}(A_2)) \\
&= \mathbb{P}(X_1 \in h^{-1}(A_1))\mathbb{P}(X_2 \in h^{-1}(A_2)) \\
&= \mathbb{P}(h_1(X_1) \in A_1)P(h_2(X_2) \in A_2).
\end{aligned}
$$

Thus we have the following result.

Result 2.10.15. *If X_1 and X_2 are independent random variables taking values in E_1 and E_2 respectively, and if $h_1 : E_1 \to E_1'$ and $h_2 : E_2 \to E_2'$ are two transformations then the random variables $h_1(X_1)$ and $h_2(X_2)$ are independent.*

The result should be intuitive and good to have in mind. In words it says that marginal transformations of *independent* random variables result in independent random variables. Intuitive as this may sound it was worth a derivation, since the definition of independence is a purely mathematical one. The result is therefore just as much a reassurance that our concept of independence is not counter intuitive in the sense that we cannot introduce dependence by marginally transforming independent variables.

Other typical transformations that we will deal with are summation and taking the maximum or taking the minimum of *independent* random variables.

We start with the summation and consider the transformation $h : \mathbb{Z} \times \mathbb{Z} \to \mathbb{Z}$ given by

$$h(x_1, x_2) = x_1 + x_2.$$

We can use Result 2.8.2 to obtain that

$$p(y) = \mathbb{P}(Y = y) = \mathbb{P}(X_1 + X_2 = y) = \sum_{x \in \mathbb{Z}} \mathbb{P}(X_1 + X_2 = y, X_2 = x)$$

Due to independence of X_1 and X_2 we find that

$$\mathbb{P}(X_1 + X_2 = y, X_2 = x) = \mathbb{P}(X_1 = y - x, X_2 = x) = p_1(y - x)p_2(x),$$

hence

$$p(y) = \sum_{x \in \mathbb{Z}} p_1(y - x)p_2(x). \tag{2.24}$$

Result 2.10.16. *Let X_1 and X_2 be two independent random variables each taking values in $\mathbb{Z}$ and with $(p_1(x))_{x \in \mathbb{Z}}$ and $(p_2(x))_{x \in \mathbb{Z}}$ denoting the point probabilities for the distribution of X_1 and X_2 respectively. Then the distribution of the random variable $Y = X_1 + X_2$, which also takes values in $\mathbb{Z}$, has point probabilities*

$$p(y) = \sum_{x \in \mathbb{Z}} p_1(y - x)p_2(x). \tag{2.25}$$

Remark 2.10.17. Finding the distribution of the sum of three or more random variables can then be done iteratively. That is, if we want to find the distribution of $X_1 + X_2 + X_3$ where X_1, X_2 and X_3 are independent then we rewrite

$$X_1 + X_2 + X_3 = (X_1 + X_2) + X_3$$

and first find the distribution (i.e. the point probabilities) of $Y = X_1 + X_2$ using Result 2.10.16. Then the distribution of $Y + X_3$ is found again using Result 2.10.16 (and independence of Y and X_3, cf. Result 2.10.15). Note that it doesn't matter how we place the parentheses.

Example 2.10.18 (Sums of Poisson Variables). Let X_1 and X_2 be independent Poisson distributed with parameter λ_1 and λ_2 respectively. The point probabilities for the distribution of $X_1 + X_2$ are then given by

$$
\begin{aligned}
p(y) &= \sum_{x=0}^{y} \exp(-\lambda_1) \frac{\lambda_1^{y-x}}{(y-x)!} \exp(-\lambda_2) \frac{\lambda_2^x}{x!} \\
&= \exp(-(\lambda_1 + \lambda_2)) \sum_{x=0}^{y} \frac{\lambda_1^{y-x} \lambda_2^x}{(y-x)! x!} \\
&= \exp(-(\lambda_1 + \lambda_2)) \frac{1}{y!} \sum_{x=0}^{y} \frac{y!}{x!(y-x)!} \lambda_2^x \lambda_1^{y-x}
\end{aligned}
$$

Using the *binomial formula* for the last sum we obtain that

$$
p(y) = \exp(-(\lambda_1 + \lambda_2)) \frac{(\lambda_1 + \lambda_2)^y}{y!}.
$$

This shows that the distribution of $X_1 + X_2$ is a Poisson distribution with parameter $\lambda_1 + \lambda_2$. $\diamond$

A derivation similar to the derivation for variables taking integer values is possible for real valued random variables whose distributions are given by densities f_1 and f_2, respectively. The differences being that the sum is replaced by an integral and the point probabilities by densities.

Result 2.10.19. *If X_1 and X_2 are independent real valued random variables with distributions having density f_1 and f_2, respectively, the density, g, for the distribution of $Y = X_1 + X_2$ is given by*

$$
g(y) = \int_{-\infty}^{\infty} f_1(y-x) f_2(x) \mathrm{d}x.
$$

Example 2.10.20. If $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$ are two *independent* normally distributed random variables then

$$
Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2), \tag{2.26}
$$

thus the sum is also normally distributed. Indeed, to simplify computations assume that $\mu_1 = \mu_2 = 0$ and $\sigma_1 = \sigma_2 = 1$, in which case the claim is that $Y \sim N(0, 2)$. The trick is to make the following observation

$$
(y-x)^2 + x^2 = y^2 - 2xy + 2x^2 = \frac{y^2}{2} + 2\left(x - \frac{y}{2}\right)^2,
$$

> **Math Box 2.10.2** (Sums of continuous variables)**.** There is, in fact, a technical problem just copying the derivation for integer valued variables to real valued variables when we want to compute the density for the sum of two variables. The problem is that all outcomes x have probablity zero of occuring. An alternative derivation is based on Example 2.10.4. A substitution of x_1 with $x_1 - x_2$ and an interchange of the integration order yields the formula
>
> $$F(y) = \int_{-\infty}^{y} \int_{-\infty}^{\infty} f_1(x_1 - x_2) f_2(x_2) \mathrm{d}x_2 \mathrm{d}x_1$$
>
> for the distribution function of the sum. From this we read of directly that the density is the inner integral
>
> $$\int_{-\infty}^{\infty} f_1(y - x_2) f_2(x_2) \mathrm{d}x_2$$
>
> as a function of y.

which leads to

$$
\begin{aligned}
g(y) &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{(y-x)^2}{2}} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \mathrm{d}x \\
&= \frac{1}{\sqrt{4\pi}} e^{-\frac{y^2}{4}} \underbrace{\int_{-\infty}^{\infty} \frac{1}{\sqrt{\pi}} e^{-\left(x - \frac{y}{2}\right)^2} \mathrm{d}x}_{=1} \\
&= \frac{1}{\sqrt{4\pi}} e^{-\frac{y^2}{4}}.
\end{aligned}
$$

We used that the integral above for fixed y is the integral of the density for the $N\left(\frac{y}{2}, \frac{1}{2}\right)$ distribution, which is thus 1. Then observe that g is the density for the $N(0, 2)$ distribution as claimed. For general μ_1, μ_2, σ_1 and σ_2 we can make the same computation with a lot of bookkeeping, but this is not really the interesting message here. The interesting message is that the sum of two normals is normal. ◇

Continuing adding normally distributed random variables we just end up with normally distributed random variables with their means and variances added up accordingly. If $X_1, \ldots, X_n$ are iid $N(\mu, \sigma)$ then we find that

$$S = X_1 + \ldots + X_n \sim N(n\mu, n\sigma^2).$$

Thus from Example 2.9.9 the distribution of the sample mean

$$Y = \frac{1}{n} S = \frac{1}{n}(X_1 + \ldots + X_n)$$

is $N(\mu, \frac{\sigma^2}{n})$. This property of the normal distribution is quite exceptional. It is so exceptional that the sample mean of virtually any collection of n iid variables strives

towards being normally distributed. The result explaining this fact is one of the germs, if not the biggest then at least among the really great ones, in probability theory. This is the *Central Limit Theorem* or CLT for short.

Result 2.10.21. *If $X_1, \ldots, X_n$ are iid with mean μ and variance σ^2 then*

$$\frac{1}{n}(X_1 + \ldots + X_n) \overset{approx}{\sim} N\left(\mu, \frac{\sigma^2}{n}\right). \tag{2.27}$$

The precise meaning of the formulation is that

$$\mathbb{P}\left(\frac{1}{\sigma\sqrt{n}}(X_1 + \ldots + X_n - n\mu) \leq x\right) \to \Phi(x)$$

for $n \to \infty$. The content is that the distribution of the sample mean is well approximated by the normal distribution if n is sufficiently large. How large n needs to be depends on the actual distribution of the X_i's, but the approximation is often surprisingly good in practice even for n as small as 20. A formal derivation of the central limit theorem is beyond the scope of these notes, and it does, moreover, not shed much light on its usefulness.

Then we turn to taking the maximum or taking the minimum of real valued random variables. If $X_1, \ldots, X_n$ are n real value random variables and

$$Y = \max(X_1, X_2, \ldots, X_n)$$

then $Y \leq x$ if and only if all the X's are $\leq x$. Hence

$$\mathbb{P}(Y \leq x) = \mathbb{P}(X_1 \leq x, X_2 \leq x, \ldots, X_n \leq x).$$

If $X_1, \ldots, X_n$ are iid with common distribution having distribution function F the right hand side factorizes into

$$\mathbb{P}(X_1 \leq x)\mathbb{P}(X_2 \leq x) \cdot \ldots \cdot \mathbb{P}(X_n \leq x)$$

and therefore, using that the common distribution function for the X's is F, we get

$$\mathbb{P}(Y \leq x) = \mathbb{P}(X_1 \leq x)\mathbb{P}(X_2 \leq x) \cdot \ldots \cdot \mathbb{P}(X_n \leq x) = F(x)^n.$$

A similar derivation for the minimum is possible, which gives us the following result.

Result 2.10.22. *If $X_1, \ldots, X_n$ are n iid real value random variables with distribution function F then the distribution function, G, for*

$$Y_1 = \max(X_1, X_2, \ldots, X_n)$$

is given as

$$G_1(x) = F(x)^n$$

and the distribution function for

$$Y_2 = \min(X_1, X_2, \ldots, X_n)$$

is given as

$$G_2(x) = 1 - (1 - F(x))^n$$

Example 2.10.23. Let $X_1, \ldots, X_n$ be n iid Gumbel distributed random variables. Their common distribution function is

$$F(x) = \exp(-\exp(-x)).$$

From Result 2.10.22 we find that the distribution function for $Y = \max\{X_1, \ldots, X_n\}$ is

$$
\begin{aligned}
G(x) & = F(x)^n = \exp(-\exp(-x))^n \\
& = \exp(-n\exp(-x)) \\
& = \exp(-\exp(-(x - \log n))).
\end{aligned}
$$

From this we see that the distribution of Y is also a Gumbel distribution with location parameter $\log n$. $\diamond$

Example 2.10.24. Let $X_1, \ldots, X_n$ be n iid exponentially distributed random variables with intensity parameter λ. Their common distribution function is

$$F(x) = 1 - \exp(-\lambda x).$$

From Result 2.10.22 we find that the distribution function for $Y = \min\{X_1, \ldots, X_n\}$ is

$$
\begin{aligned}
G(x) & = 1 - (1 - F(x))^n = 1 - \exp(-\lambda x)^n \\
& = 1 - \exp(-n\lambda x).
\end{aligned}
$$

Thus the distribution of the minimum Y is also an exponential distribution with intensity parameter $n\lambda$. $\diamond$

2.11 Simulations

The use of computer simulations is an omnipresent tool in current scientific research ranging from issues such as forecasting of the weather or the economic developments to understanding the mechanisms working in biology, chemistry and physics. A (computer) simulation consists of an implementation of our favorite experiment as a mathematical model on a computer and then "we run the experiment on the computer". If the model consists of a set of ordinary differential equations, say, this amounts to solving the equations numerically using the computer. This provides a very useful *supplement* to the mathematical analysis of the model, which is typically so complex that we have no hope of writing down an easily interpretable solution on paper. We are not interested in solving differential equations but instead in simulating the outcome of an experiment with a random component. Redoing a simulation several times should lead to different results reflecting the probability measure that models the random phenomenon. This is in principle somewhat of a

problem. Whereas the computer is entirely deterministic – logical and deterministic behavior is what makes the computer useful in the first place – we ask for outcomes of running identical simulations that should differ. Solving deterministic differential equations is directly suitable for a computer implementation, but simulating random variables is not.

The problem is essentially always solved by the following two step procedure.

- The computer *emulates* the generation of independent, identically distributed random variables with the uniform distribution on the unit interval $[0, 1]$.

- The emulated uniformly distributed random variables are by *transformation* turned into variables with the desired distribution.

If we had access to a stream of truly independent, identically distributed random variables with the uniform distribution on $[0, 1]$, perhaps arriving to the computer from an external truly random source, and if we implement the second step above correctly we would indeed obtain simulations of random variables with the desired distribution. In practice we don't use externally generated random numbers but rely on a program generating *pseudo random numbers*. It is even questionable if there exists such a thing as an external, truly random source of random numbers. The pseudo random number generator emulates the generation of truly random numbers, and the philosophy is that if we can't statistically *detect* the difference, there is, from a practical point of view, no difference! But the pseudo random number generator is certainly, like all other computer programs, a deterministic program.

In these notes we will not discuss in any further details how to generate pseudo random numbers. This will be regarded as a black box tool that we leave for others to think about. Two things are, however, important to keep in mind. First of all the pseudo random number generator may not provide sufficiently good pseudo random numbers, which can lead to wrong results. If you use the standard generator provided by R you are not likely to run into problems, but it is not guaranteed that all programming languages provide a useful standard pseudo random number generator. If in doubt you should seek qualified assistance. Secondly, the pseudo random number generator is always initialized with a *seed* – a number that tells the generator how to start. Providing the same seed will lead to the same sequence of numbers. If we need to run independent simulations we should be cautious not to restart the pseudo random number generator with the same seed. It is, however, an advantage when debugging programs that you can always provide the same sequence of random numbers. Most pseudo random number generators have some default way of setting the seed if no seed is provided by the user.

For the rest of this section we will assume that we have access to a sequence $U_1, U_2, \ldots, U_n$ of iid random variables uniformly distributed on $[0, 1]$, which in practice are generated using a pseudo random number generator. This means that

$$\mathbb{P}(U_1 \in A_1, U_2 \in A_2, \ldots, U_n \in A_n) = \mathbb{P}(U_1 \in A_1)\mathbb{P}(U_2 \in A_2) \cdot \ldots \cdot \mathbb{P}(U_n \in A_n)$$

R Box 2.11.1 (Pseudo random numbers). In R we can generate pseudo random numbers using the command `runif`. For instance

> u <- runif(100)

produces a vector `u` containing 100 pseudo random numbers. In general `runif(n)` produces n pseudo random numbers. Use `set.seed(m)` to set the seed for the generator to be m. Every time R is opened a new seed is set based on the current time, so `set.seed` is only used if one needs to reproduce the same results several times.

---

for all events $A_1, \ldots, A_n \subseteq [0, 1]$, and that

$$\mathbb{P}(U_i \in [a, b]) = b - a$$

for $0 \le a \le b \le 1$. It then follows by Result 2.10.15 that if $X_i = h(U_i)$ for a given transformation h then the X_i's are independent. Thus we have the following result.

Result 2.11.1. *Let P_0 denote the uniform distribution on $[0, 1]$ and $h : [0, 1] \to E$ a map with the transformed probability measure on E being $P = h(P_0)$. Then $X_1, X_2, \ldots, X_n$ defined by*

$$X_i = h(U_i)$$

are n iid random variables each with distribution P.

Algorithm 2.11.1 (Simulating discrete random variables). If we want to simulate random variables taking values in a *discrete* space E we can proceed as follows: Assume that the distribution that we want to simulate from has point probabilities $p(x)$, $x \in E$, and choose for each $x \in E$ an interval

$$I(x) = (a(x), b(x)] \subseteq [0, 1]$$

such that

- the length, $b(x) - a(x)$, of $I(x)$ equals $p(x)$,

- and the intervals $I(x)$ are mutually disjoint:

$$I(x) \cap I(y) = \emptyset \text{ for } x \neq y.$$

Letting $u_1, \ldots, u_n$ be generated by a pseudo random number generator we define

$$x_i = x \quad \text{if } u_i \in I(x)$$

for $i = 1, \ldots, n$. Then $x_1, \ldots, x_n$ is a realization of n iid random variables with distribution having point probabilities $p(x)$, $x \in E$.

Simulating discrete random variables basically amounts to Algorithm 2.11.1. There the implicit transformation used is $h : [0,1] \to E$ given by

$$h(u) = x \quad \text{if } u \in I(x) = (a(x), b(x)].$$

If U is uniformly distributed on $[0,1]$

$$\mathbb{P}(h(U) = x) = \mathbb{P}(U \in I(x)) = b(x) - a(x) = p(x).$$

This shows that the algorithm indeed simulates random variables with the desired distribution.

In practice we need to choose the intervals suitably. If the random variable, X, that we want to simulate takes values in $\mathbb{N}$ a possible choice of $I(x)$ is

$$I(x) = (\mathbb{P}(X \leq x - 1), \mathbb{P}(X \leq x)].$$

We see that the length of $I(x)$ equals

$$\mathbb{P}(X \leq x) - \mathbb{P}(X \leq x - 1) = \mathbb{P}(X = x) = p(x)$$

as required, and the intervals are clearly disjoint. To easily compute these intervals we need easy access to the distribution function

$$F(x) = \mathbb{P}(X \leq x).$$

For the simulation of real valued random variables there is also a generic solution that relies on the knowledge of the distribution function.

Definition 2.11.2. *Let* $F : \mathbb{R} \to [0,1]$ *be a distribution function. A function*

$$F^{\leftarrow} : (0,1) \to \mathbb{R}$$

that satisfies

$$F(x) \geq y \;\Leftrightarrow\; x \geq F^{\leftarrow}(y) \tag{2.28}$$

for all $x \in \mathbb{R}$ *and* $y \in (0,1)$ *is called a generalized inverse of* F.

There are a few important comments related to the definition above. First of all suppose that we can solve the equation

$$F(x) = y$$

for all $x \in \mathbb{R}$ and $y \in (0,1)$, yielding an inverse function $F^{-1} : (0,1) \to \mathbb{R}$ of F that satisfies

$$F(x) = y \;\Leftrightarrow\; x = F^{-1}(y). \tag{2.29}$$

Then the inverse function is also a generalized inverse function of F. However, not all distribution functions have an inverse, but all distribution functions have a *generalized* inverse and this generalized inverse is in fact unique. We will not show this

although it is not particularly difficult. What matters in practice is whether we can find the (generalized) inverse of the distribution function. Note also, that we do not really want to define the value of $F^{\leftarrow}$ in 0 or 1. Often the only possible definition is $F^{\leftarrow}(0) = -\infty$ and $F^{\leftarrow}(1) = +\infty$.

At this point the generalized inverse is useful because it is, in fact, a quantile function. To see this, first observe that with $x = F^{\leftarrow}(y)$ then

$$F^{\leftarrow}(y) \leq x \Leftrightarrow y \leq F(x) = F(F^{\leftarrow}(y))$$

by the definition of $F^{\leftarrow}$. On the other hand, suppose that there exists a $y \in (0, 1)$ and an $\varepsilon > 0$ such that $F(F^{\leftarrow}(y) - \varepsilon) \geq y$ then again by the definition of $F^{\leftarrow}$ it follows that

$$F^{\leftarrow}(y) - \varepsilon \geq F^{\leftarrow}(y),$$

which cannot be the case. Hence there is no such $y \in (0, 1)$ and $\varepsilon > 0$ and

$$F(F^{\leftarrow}(y) - \varepsilon) < y$$

for all $y \in (0, 1)$ and $\varepsilon > 0$. This shows that $F^{\leftarrow}$ is a quantile function.

Result 2.11.3. *The generalized inverse distribution function $F^{\leftarrow}$ is a quantile function.*

There may exist other quantile functions besides the generalized inverse of the distribution function, which are preferred from time to time. If F has an inverse function then the inverse is the only quantile function, and it is equal to the generalized inverse.

We will need the generalized inverse to transform the uniform distribution into any distribution we would like. The uniform distribution on $[0, 1]$ has distribution function

$$G(x) = x$$

for $x \in [0, 1]$. By the definition of $F^{\leftarrow}$ we have that

$$F^{\leftarrow}(U) \leq x \iff U \leq F(x)$$

for all $x \in \mathbb{R}$ and hence

$$\mathbb{P}(F^{\leftarrow}(U) \leq x) = \mathbb{P}(U \leq F(x)) = G(F(x)) = F(x).$$

Note that it doesn't matter if we have defined $F^{\leftarrow}$ on $(0, 1)$ only as the uniform random variable U with probability 1 takes values in $(0, 1)$. We have derived the following result.

Result 2.11.4. *If $F^{\leftarrow} : (0, 1) \to \mathbb{R}$ is the generalized inverse of the distribution function $F : \mathbb{R} \to [0, 1]$ and if U is uniformly distributed on $[0, 1]$ then the distribution of*

$$X = F^{\leftarrow}(U)$$

has distribution function F.

The result above holds for all quantile functions, but it is easier and more explicit just to work with the generalized inverse, as we have done.

Algorithm 2.11.2 (Simulating real valued random variables)**.** If we want to simulate random variables taking values in $\mathbb{R}$ with distribution function F we can proceed as follows: First we find the generalized inverse, $F^{\leftarrow} : (0, 1) \to \mathbb{R}$, of F.

Then we let $u_1, \ldots, u_n$ be generated by a pseudo random number generator and we define
$$x_i = F^{\leftarrow}(u_i)$$
for $i = 1, \ldots, n$. Then $x_1, \ldots, x_n$ is a realization of n iid random variables with distribution having distribution function F.

Example 2.11.5. The exponential distribution with parameter $\lambda > 0$ has distribution function
$$F(x) = 1 - \exp(-\lambda x), \quad x \geq 0.$$

The equation
$$F(x) = 1 - \exp(-\lambda x) = y$$

is solved for $y \in (0, 1)$ by

$$F^{-1}(y) = -\frac{1}{\lambda} \log(1 - y).$$

Thus the simulation of exponentially distributed random variables can be based on the transformation
$$h(y) = -\frac{1}{\lambda} \log(1 - y).$$

$\diamond$

R Box 2.11.2 (Inverse distribution functions)**.** Like distribution functions and densities the (generalized) inverse distribution functions are directly available in R for a number of standard distributions. The general convention is that `qname` is the inverse distribution function for a distribution called `name`. The inverse distribution function for the normal distribution evaluated at x is therefore given by `qnorm(x)`.

Exercises

Exercise 2.11.1. Write three R functions: `pgumbel`, `dgumbel`, and `qgumbel` taking three arguments `x,location,scale` and returning the value (in `x`) of the distribution function, the density, and the inverse distribution function (the quantile function) respectively for the Gumbel distribution with location and scale parameters given by `location` and `scale`.

⌨   **Exercise 2.11.2.** Write a fourth function, `rgumbel`, that takes one integer argument, n, and the `location,scale` arguments, and returns a vector of length n, which is a simulated realization of **n** independent and identically Gumbel distributed random variables.

★   **Exercise 2.11.3.** Let X have the geometric distribution with success probability p. Show that the distribution function for X is

$$F(x) = 1 - (1-p)^{\lfloor x \rfloor + 1},$$

where $\lfloor x \rfloor \in \mathbb{Z}$ is the integer fulfilling that $x - 1 < \lfloor x \rfloor \leq x$ for $x \in \mathbb{R}$. Define likewise $\lceil x \rceil \in \mathbb{Z}$ as the integer fulfilling that $x \leq \lceil x \rceil < x + 1$ for $x \in \mathbb{R}$.
Argue that

$$\lfloor z \rfloor \geq x \iff z \geq \lceil x \rceil$$

for all $x, z \in \mathbb{R}$ use this to show that

$$F^{\leftarrow}(y) = \left\lceil \frac{\log(1-y)}{\log(1-p)} - 1 \right\rceil$$

is the generalized inverse for the distribution function for the geometric distribution.

⌨   **Exercise 2.11.4.** Write an R-function, `my.rgeom`, that takes two arguments, `(n,p)`, and returns a simulation of **n** iid geometrically distributed variables with success parameter p. Note that the operations $\lfloor \cdot \rfloor$ and $\lceil \cdot \rceil$ are known as *floor* and *ceiling*.

2.12   Local alignment - a case study

As a case study of some of the concepts that have been developed up to this point in the notes, we present in this section some results about the distribution of the score of optimally locally aligned random amino acid sequences.

This is a classical problem and a core problem in biological sequence analysis whose solution has to be found in the realms of probability theory and statistics. Moreover, even though the problem may seem quite specific to biological sequence analysis it actually holds many of the general issues related to the extraction of data from large databases and how the extraction procedure may need a careful analysis for a correct interpretation of the results.

The local alignment problem is essentially not a biological problem but a computer science problem of finding substrings of two strings that match well. Given two strings of letters can we compute two substrings of letters – one from each – that maximize the length of matching letters? In the words `ABBA` and `BACHELOR` we can find the substring `BA` in both but no substrings of length 3. The stringent requirement of exact matching is often to hard. If we try to match `BACHELOR` and `COUNCILLOR` we could get an exact match of length 3 of `LOR` but allowing for two mismatches we can match `CHELOR` and `CILLOR` of length 6. It may also be an idea to allow for gaps in the matching. If we try to match `COUNCILLOR` with `COUNSELOR` we can match `COUN--CILLOR` with `COUNSE---LOR` where we have introduced two gaps in the former

string and three in the latter.

When we introduce mismatches and gaps we will penalize their occurrences and formally we end up with a combinatorial optimization problem where we optimize a score over all selections of substrings. The score is a sum of (positive) match contributions and (negative) mismatch and gap constributions. The computational solution known as the Smith-Waterman algorithm is a so-called dynamic programming algorithm, which can solve the optimization problem quite efficiently.

In the context of biological sequences, e.g. proteins regarded as a string of letters from the amino acid alphabet, there are many implementations of the algorithm and it is in fact not awfully complicated to write a simple implementation yourself. The actual algorithm is, however, outside of the scope of these notes. In R there is an implementation of several alignment algorithms, including the Smith-Waterman algorithm for local alignment, in the `pairwiseAlignment` function from the package `Biostrings`. The widely used *BLAST* program and friends offer a heuristic that improves the search time tremendously and makes it feasible to locally align a protein to the entire database of known protein sequences.

The purpose of the computation of the local alignment is to find parts of the protein that share evolutionary or functional relations to other proteins in the database. The algorithm will always produce a best optimal local alignment disregarding whether the best alignment has any meaning or is just a coincidence due to the fact that a database contains a large number of different sequences. Thus we need to be able to tell if a local alignment with a given score is something we would expect to see by chance or not.

Assume that $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ are in total $n + m$ random variables with values in the 20 letter amino acid alphabet

$$E = \big\{ \mathtt{A, R, N, D, C, E, Q, G, H, I, L, K, M, F, P, S, T, W, Y, V} \big\}.$$

We regard the first n letters $X_1, \ldots, X_n$ as one sequence (one protein or a protein sequence database) and $Y_1, \ldots, Y_m$ as another sequence (another protein or protein sequence database). The joint distribution of the $n+m$ random variables is specified by assuming that they are all independent and identically distributed. The distribution of each single letter in the sequences may be given by the Robinson-Robinson frequencies from Example 2.4.5 – or any other distribution on the 20 letter amino acid alphabet. The joint distribution is a probability measure on the product space E^{n+m}.

The optimal local alignment score can be regarded as a transformation

$$h : E^{n+m} \to \mathbb{R}$$

where $h(x_1, \ldots, x_n, y_1, \ldots, y_m)$ denotes the score obtained by optimally aligning the first n letters $x_1, \ldots, x_n$ with the last m letters $y_1, \ldots, y_m$. Let

$$S_{n,m} = h(X_1, \ldots, X_n, Y_1, \ldots, Y_m)$$

denote the random variable we obtain by transforming the $n + m$ random variables by the function h. This gives a distribution on $\mathbb{R}$, but it is really only a distribution living on a discrete – even finite – subset of $\mathbb{R}$. Indeed, the function h can at most take 20^{n+m} different values as this is the size of the sample space E^{n+m}. Although one in principle can compute the distribution of $S_{n,m}$ by going through each of the different 20^{n+m} possible values of h and compute the corresponding probability, this is a futile task for realistic values of n and m.

There are (at least) two alternative approaches. One is to compute the distribution of $S_{n,m}$ by simulations as discussed in Section 2.11. Simulations are a universally applicable tool for studying the distribution of transformations. The downside is first of all that it can be time-consuming in particular in the setup of local alignment, and second that you get no or little theoretical insight about the distribution of $S_{n,m}$. Another approach is to find an approximating distribution with a few parameters, whose values depend on n, m, the distribution of the letters, and the choice of scoring mechanism for the local alignment.

For optimal local alignment scores a frequently used approximation is a location scale transformation of a Gumbel distribution. Under certain conditions on the scoring mechanism and the letter distribution, and for n and m sufficiently large,

$$\mathbb{P}(S_{n,m} \leq x) \simeq \exp(-Knm \exp(-\lambda x)) \tag{2.30}$$

with $\lambda, K > 0$ two parameters not depending upon n and m. We use the notation $\simeq$ to denote approximately equal without specifying what we mean by that in any detail. In practice, we work and argue as if $\simeq$ means equality, but we keep in mind that it is only an approximation. Defining the variable

$$S'_{n,m} = \lambda S_{n,m} - \log(Knm)$$

we see that $S'_{n,m}$ is a location-scale transformation of $S_{n,m}$, hence

$$
\begin{aligned}
\mathbb{P}(S'_{n,m} \leq x) &\simeq \exp\left(-Knm \exp\left(-\frac{\lambda(x + \log(Knm))}{\lambda}\right)\right) \\
&= \exp(-\exp(-x)).
\end{aligned}
$$

In other words, the distribution of $S_{n,m}$ is approximately a Gumbel distribution with location parameter $\log(Knm)/\lambda$ and scale parameter $1/\lambda$.

On Figure 2.20 we see the density for the location-scale transformed Gumbel distributions, taking $n = 1,000$ and $m = 100,000$, that approximate the distribution of the maximal local alignment score using either BLOSUM62 or BLOSUM50 together with the affine gap penalty function. We observe that for the BLOSUM50 matrix the distribution is substantially more spread out.

gap open	gap ext.	BLOSUM62		BLOSUM50		PAM250	
		λ	K	λ	K	λ	K
12	3	0.300	0.10	0.178	0.028	0.170	0.022
12	2	0.300	0.09	0.158	0.019	0.145	0.012
12	1	0.275	0.05	−	−	−	−
11	3	0.301	0.09	0.167	0.028	0.153	0.017
11	2	0.286	0.07	0.130	0.009	0.122	0.009
11	1	0.255	0.035	−	−	−	−
10	3	0.281	0.06	0.139	0.013	0.129	0.012
10	2	0.266	0.04	0.099	0.007	−	−
10	1	0.216	0.014	−	−	−	−
9	3	0.273	0.06	0.107	0.008	0.102	0.010
9	2	0.244	0.030	−	−	−	−

Table 2.1: A selection of values for the λ and K parameters for different choices of scoring schemes in the local alignment algorithm. The entries marked by − are scoring schemes where the approximation (2.30) breaks down. The values presented here are from *Altschul, S.F. and Gish, W. Local Alignment Statistics. Methods in Enzymology, vol 266, 460-480.* They are computed (estimated) on the basis of alignments of simulated random amino acid sequences using the Robinson-Robinson frequencies.
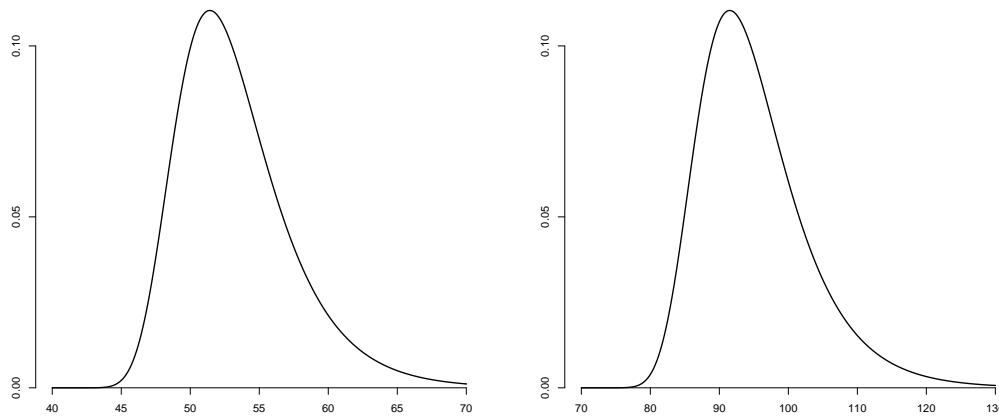
Figure 2.20: The density for the Gumbel distribution that approximates the maximal local alignment score with $n = 1,000$ and $m = 100,000$ using an affine gap penalty function with gap open penalty 12 and gap extension penalty 2 together with a BLOSUM62 (left) or BLOSUM50 (right) scoring matrix.

If $n = n_1 + n_2$, the approximation given by (2.30) implies that

$$
\begin{aligned}
\mathbb{P}(S_{n,m} \leq x) &\simeq \exp(-Knm\exp(-\lambda x)) \\
&= \exp(-K(n_1 + n_2)m\exp(-\lambda x)) \\
&= \exp(-Kn_1 m\exp(-\lambda x))\exp(-Kn_2 m\exp(-\lambda x)).
\end{aligned}
$$

If (2.30) holds for $S_{n_1,m}$ and $S_{n_2,m}$ too, then

$$\mathbb{P}(S_{n_1,m} \leq x) \simeq \exp(-Kn_1 m \exp(-\lambda x)) \quad \text{and} \quad \mathbb{P}(S_{n_2,m} \leq x) \simeq \exp(-Kn_2 m \exp(-\lambda x)),$$

which implies that

$$\mathbb{P}(S_{n,m} \leq x) \simeq \mathbb{P}(S_{n_1,m} \leq x)\mathbb{P}(S_{n_2,m} \leq x).$$

This (approximate) equality says that the distribution of $S_{n,m}$ behaves (approximately) as if $S_{n,m} = \max\{S_{n_1,m}, S_{n_2,m}\}$ and that $S_{n_1,m}$ and $S_{n_2,m}$ are independent.

The justification for the approximation (2.30) is quite elaborate. There are theoretical (mathematical) results justifying the approximation for *ungapped* local alignment. When it comes to gapped alignment, which is of greater practical interest, there are some theoretical results suggesting that (2.30) is not entirely wrong, but there is no really satisfactory theoretical underpinning. On the other hand, a number of detailed simulation studies confirm that the approximation works well – also for gapped local alignment. For gapped, as well as ungapped, local alignment there exist additional corrections to (2.30) known as *finite size corrections*, which reflect the fact that (2.30) works best if *n and m* are large – and to some extent of the same order. Effectively, the corrections considered replace the product nm by a smaller number $n'm'$ where $n' < n$ and $m' < m$. We will not give details here. Another issue is how the parameters λ and K depend upon the scoring mechanism and the distribution of the letters. This is not straight forward – even in the case of ungapped local alignment, where we have analytic formulas for computing λ and K. In the gapped case, the typical approach is to estimate values of λ and K from simulations.

Exercises

Alice and Bob work in Professor A.K.'s laboratory on a putative serine protease:

```
>gi|21220235|ref|NP_626014.1| putative secreted serine protease [Streptomyces coelicolor A3(2)]
MKHRRIPRRRVAVVGAGITALVAAGVTFQTANASEAPKTAAPETLSVSAAGELASTLLGDLGADAAGTYY
DAQAKSLVVNVLDQSAAQTVEEAGAKARVVENSLADLKSARTALTKDATIPGTSWATDPTTNKVVVTADR
TVSEAELAKLTKVVDGLGAKAELKRTKGEYKPFVAGGDAITGGGRCSLGFNVTKGGEPYFITAGHCTES
ISTWSDSSGNVIGENAASSFPDNDYGLVKYTADVDHPSEVNLYNGSSQAISGAAEATVGMQVTRSGSTTQ
VHDGTVTGLDATVNYGNGDIVNGLIQTDVCAEPGDSGGSLFSGDQAIGLTSGGSGDCTSGGETFFQPVTE
ALSATGTQIG
```

Bob is convinced that the protein is related to the Human protein

```
>gi|4506141|ref|NP_002766.1| HtrA serine peptidase 1 [Homo sapiens]
MQIPRAALLPLLLLLLAAPASAQLSRAGRSAPLAAGCPDRCEPARCPPQPEHCEGGRARDACGCCEVCGA
PEGAACGLQEGPCGEGLQCVVPFGVPASATVRRRAQAGLCVCASSEPVCGSDANTYANLCQLRAASRRSE
RLHRPPVIVLQRGACGQGQEDPNSLRHKYNFIADVVEKIAPAVVHIELFRKLPFSKREVPVASGSGFIVS
EDGLIVTNAHVVTNKHRVKVELKNGATYEAKIKDVDEKADIALIKIDHQGKLPVLLLGRSSELRPGEFVV
AIGSPFSLQNTVTTGIVSTTQRGGKELGLRNSDMDYIQTDAIINYGNSGGPLVNLDGEVIGINTLKVTAG
ISFAIPSDKIKKFLTESHDRQAKGKAITKKKYIGIRMMSLTSSKAKELKDRHRDFPDVISGAYIIEVIPD
TPAEAGGLKENDVIISINGQSVVSANDVSDVIKRESTLNMVVRRGNEDIMITVIPEEIDP
```

Bob does a local alignment of the two proteins using the Smith-Waterman algorithm with the BLOSUM50 scoring matrix and affine gap penalty function with gap open penalty 10 and gap extension penalty 3. He uses the an implementation called SSearch and finds the (edited) alignment:

```
 >>QUERY sequence                                                      (480 aa)
Smith-Waterman score: 68;  29.333% identity (64.000% similar) in 75 aa overlap (263-331:270-344)
Entrez lookup  Re-search database  General re-search
>QUERY   263- 331:                                            ------------     :

          230       240       250       260       270       280       290       300
QUERY  GENAASSFPDNDYGLVKYTADVDHPSEVNLYNGSSQAISGAAEATVGMQVTRSGSTTQVHDGTVTGLDATVNYGNGDI--
                                          ..:    :  :.  ::  ....  ..::. .:.. :. ..
QUERY  VELKNGATYEAKIKDVDEKADIALIKIDHQGKLPVLLLGRSSELRPGEFVVAIGSPFSLQNTVTTGIVSTTQRGGKELGL
    230       240       250       260       270       280       290       300

            310       320       330       340       350       360
QUERY  VNG---LIQTDVCAEPGDSGGSLFSGD-QAIGLTSGGSGDCTSGGETFFQPVTEALSATGTQIG
        :.     ::::.  . :.::: : . : ..::...
QUERY  RNSDMDYIQTDAIINYGNSGGPLVNLDGEVIGINTLKVTAGISFAIPSDKIKKFLTESHDRQAKGKAITKKKYIGIRMMS
    310       320       330       340       350       360       370       380
```

Note that the output shows in the upper right corner that the length of one of the sequences in 480. The other one has length 360.

Alice is not so keen on Bobs theory and she does a database search against the set of Human proteins using BLAST with default parameter settings (Blosum62, gap open 11, gap extension 1, Human protein database contains at the time aprox. 13.000.000 amino acids). She finds the best alignment of the serine protease with the Human serine peptidase far down the list

```
 Score = 61
 Identities = 26/95 (27%), Positives = 40/95 (42%), Gaps = 15/95 (15%)

Query  243  DVDHPSEVNLYNGSSQA------ISGAAEATVGMQVTRSGSTTQVHDGTVTGLDATVNYG  296
            DVD +++ L    Q       + ++E  G  V  GS  + +   TG+ +T   G
Sbjct  244  DVDEKADIALIKIDHQGKLPVLLLGRSSELRPGEFVVAIGSPFSLQNTVTTGIVSTTQRG  303

Query  297  NGDIVNGL-------IQTDVCAEPGDSGGSLFSGD  324
               ++  GL      IQTD    G+SGG L + D
Sbjct  304  GKEL--GLRNSDMDYIQTDAIINYGNSGGPLVNLD  336
```

Alice believes that this provides her with evidence that Bobs theory is wrong. Who is A. K. going to believe?

Exercise 2.12.1. Compute the probability – using a suitable model – for each of the local alignments above of getting a local alignment with a score as large or larger than the given scores. Pinpoint the information you need to make the computations. Discuss the assumptions upon which your computations rest. What is the conclusion?

2.13   Multivariate distributions

If $X_1, \ldots, X_n$ are n random variables taking values in the sample spaces $E_1, \ldots, E_n$, we can *bundle* the variables into a single random variable. The bundled variable $X = (X_1, \ldots, X_n)$ takes values in the product space

$$E_1 \times \ldots \times E_n = \{(x_1, \ldots, x_n) | x_1 \in E_1, \ldots, x_n \in E_n\}.$$

The product space is the set of n-tuples with the i'th coordinate belonging to E_i for $i = 1, \ldots, n$. Each of the variables can represent the outcome of an experiment, and we want to consider all the experiments simultaneously. To do so, we need to define the distribution of the bundled variable X that takes values in the product space. Thus we need to define a probability measure on the product space. In this case we talk about the *joint distribution* of the random variables $X_1, \ldots, X_n$, and we often call the distribution on the product space a *multivariate distribution* or a *multivariate probability measure*. We do not get the joint distribution automatically from the distribution of each of the random variables – something more is needed. We need to capture how the variables interact, and for this we need the joint distribution. If the joint distribution is P and A is an event having the product form

$$A = A_1 \times \ldots \times A_n = \{(x_1, \ldots, x_n) \mid x_1 \in A_1, \ldots, x_n \in A_n\}$$

we use the notation

$$P(A) = \mathbb{P}(X \in A) = \mathbb{P}((X_1, \ldots, X_n) \in A_1 \times \ldots \times A_n) = \mathbb{P}(X_1 \in A_1, \ldots, X_n \in A_n).$$

The right hand side is particularly convenient, for if some set A_i equals the entire sample space E_i, it is simply left out. Another convenient, though slightly technical point, is that knowledge of the probability measure on product sets specifies the measure completely. This is conceptually similar to the fact that the distribution function uniquely specifies a probability measure on $\mathbb{R}$.

Example 2.13.1. Let X_1, X_2, and X_3 be three real valued random variables, that is, $n = 3$ and $E_1 = E_2 = E_3 = \mathbb{R}$. Then the bundled variable $X = (X_1, X_2, X_3)$ is a three dimensional vector taking values in $\mathbb{R}^3$. If $A_1 = A_2 = [0, \infty)$ and $A_3 = \mathbb{R}$ then

$$\mathbb{P}((X_1, X_2, X_3) \in [0, \infty) \times [0, \infty) \times \mathbb{R}) = \mathbb{P}(X_1 \in [0, \infty), X_2 \in [0, \infty)).$$

Moreover, in this case we rewrite the last expression as

$$\mathbb{P}(X_1 \in [0, \infty), X_2 \in [0, \infty)) = \mathbb{P}(X_1 \geq 0, X_2 \geq 0).$$

If $A_1 = [a, b]$, $A_2 = \mathbb{R}$ and $A_3 = [c, d]$ for $a, b, c, d \in \mathbb{R}$ then

$$\begin{aligned}
\mathbb{P}((X_1, X_2, X_3) \in [a, b] \times \mathbb{R} \times [c, d]) &= \mathbb{P}(X_1 \in [a, b], X_3 \in [c, d]) \\
&= \mathbb{P}(a \leq X_1 \leq b, c \leq X_3 \leq d).
\end{aligned}$$

$\diamond$

Definition 2.13.2 (Marginal distribution)**.** *If the bundled variable $X = (X_1, \ldots, X_n)$ has distribution P, the marginal distribution, P_i, of X_i is given by*

$$P_i(A) = \mathbb{P}(X_i \in A) = P(E_1 \times \ldots \times E_{i-1} \times A \times E_{i+1} \times \ldots \times E_n)$$

for $A \subseteq E_i$.

If the sample spaces that enter into a bundling are discrete, so is the product space – the sample space of the bundled variable. The distribution can therefore in principle be defined by point probabilities.

Example 2.13.3. If two DNA-sequences that encode a protein (two genes) are evolutionary related, then typically there is a pairing of each nucleotide from one sequence with an identical nucleotide from the other with a few exceptions due to mutational events (an alignment). We imagine in this example that the only mutational event occurring is *substitution* of nucleic acids. That is, one nucleic acid at the given position can mutate into another nucleic acid. The two sequences can therefore be aligned in a letter by letter fashion without gaps, and we are going to consider just a single aligned position in the two DNA-sequences. We want a probabilistic model of the pair of letters occurring at that particular position. The sample space is going to be the product space

$$E = E_0 \times E_0 = \{\texttt{A}, \texttt{C}, \texttt{G}, \texttt{T}\} \times \{\texttt{A}, \texttt{C}, \texttt{G}, \texttt{T}\},$$

and we let X and Y denote the random variables representing the two aligned nucleic acids. To define the joint distribution of X and Y, we have to define point probabilities $p(x, y)$ for $(x, y) \in E$. It is convenient to organize the point probabilities in a matrix (or array) instead of as a vector. Consider for instance the following matrix

	A	C	G	T
A	0.1272	0.0063	0.0464	0.0051
C	0.0196	0.2008	0.0082	0.0726
G	0.0556	0.0145	0.2151	0.0071
T	0.0146	0.0685	0.0069	0.1315

As we can see the probabilities occurring in the diagonal are (relatively) large and those outside the diagonal are small. If we let $A = \{(x, y) \in E \mid x = y\}$ denote the event that the two nucleic acids are identical, then

$$\mathbb{P}(X = Y) = P(A) = \sum_{(x,y) \in A} p(x, y) = 0.1272 + 0.2008 + 0.2151 + 0.1315 = 0.6746.$$

This means that the probability of obtaining a pair of nucleic acids with a mutation is $\mathbb{P}(X \neq Y) = P(A^c) = 1 - P(A) = 0.3254.$ ◇

NA

Compared to discrete sample spaces, the situation seems more complicated if we bundle real valued variables. If we bundle n real valued random variables $X_1, \ldots, X_n$, the bundled variable takes values in $\mathbb{R}^n$ – the set of n-dimensional real vectors. To define and handle distributions on $\mathbb{R}^n$ easily becomes quite technical. There exists a multivariate analog of the distribution function, but it is a clumsy object to work with. The best situation arises when the joint distribution is given by a density.

Definition 2.13.4. *If $X_1, \ldots, X_n$ are n real valued random variables we say that the distribution of $X = (X_1, \ldots, X_n)$ has density*

$$f : \mathbb{R}^n \to [0, \infty)$$

if

$$\mathbb{P}(X_1 \in A_1, \ldots, X_n \in A_n) = \int_{A_1} \cdots \int_{A_n} f(x_1, \ldots, x_n) \mathrm{d}x_n \ldots \mathrm{d}x_1.$$

Note that from the fact that $\mathbb{P}(X \in \mathbb{R}^n) = 1$, a density f must fulfill that

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, \ldots, x_n) \mathrm{d}x_n \ldots \mathrm{d}x_1 = 1.$$

Remember also the convention about our formulations: Every time we talk about random variables, we are in fact only interested in their distribution, which is a probability measure on the sample space. The content of the definition above is therefore that the distribution – the probability measure – is given simply by specifying a density. One of the deeper results in probability theory states that if a function f, defined on $\mathbb{R}^n$ and with values in $[0, \infty)$, integrates to 1 as above, then it defines a probability measure on $\mathbb{R}^n$.

If you feel uncomfortable with the integration, remember that if $A_i = [a_i, b_i]$ with $a_i, b_i \in \mathbb{R}$ for $i = 1, \ldots, n$ then

$$\mathbb{P}(a_1 \le X_1 \le b_1, \ldots, a_n \le X_n \le b_n) = \int_{a_1}^{b_1} \cdots \int_{a_n}^{b_n} f(x_1, \ldots, x_n) \mathrm{d}x_n \ldots \mathrm{d}x_1$$

can be computed as n successive ordinary integrals.

The integrations can be carried out in any order, so carrying out the integration over the i'th variable as the last ourter integration we can compute the distribution function for the distribution of X_i as follows:

$$F_i(x) = \mathbb{P}(X_i \le x) = \int_{-\infty}^{x} \underbrace{\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty}}_{n-1} f(x_1, \ldots, x_n) \mathrm{d}x_n \ldots \mathrm{d}x_{i+1} \mathrm{d}x_{i-1} \ldots \mathrm{d}x_1 \mathrm{d}x_i.$$

Consequently we can see that the marginal distribution of X_i also has a density that is given by the function

$$x_i \mapsto \underbrace{\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty}}_{n-1} f(x_1, \ldots, x_n) \mathrm{d}x_n \ldots \mathrm{d}x_{i+1} \mathrm{d}x_{i-1} \ldots \mathrm{d}x_1.$$

We state this as a result.

Result 2.13.5. *If* $f : \mathbb{R}^n \to [0, \infty)$ *is the density for the joint distribution of* $X_1, \ldots, X_n$ *then the marginal distribution of* X_i *has density*

$$f_i(x_i) = \underbrace{\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty}}_{n-1} f(x_1, \ldots, x_n) \mathrm{d}x_n \ldots \mathrm{d}x_{i+1} \mathrm{d}x_{i-1} \ldots \mathrm{d}x_1. \tag{2.31}$$

In a similar manner, one can compute the density for the distribution of any subset of coordinates of an n-dimensional random variable, whose distribution has a density, by integrating out over the other coordinates.
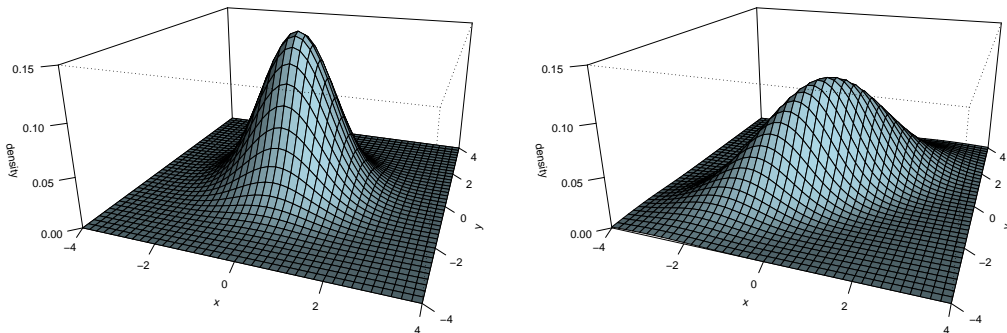
Figure 2.21: Two examples of the density for the bivariate normal distribution as considered in Example 2.13.6 with $\rho = 0$ (left) and $\rho = 0.75$ (right).

Example 2.13.6 (Bivariate normal distribution)**.** Consider the function

$$f(x, y) = \frac{\sqrt{1 - \rho^2}}{2\pi} \exp\left(-\frac{x^2 - 2\rho xy + y^2}{2}\right)$$

for $\rho \in (-1, 1)$. We will first show that this is a density for a probability measure on $\mathbb{R}^2$, thus that it integrates to 1, and then find the marginal distributions. The numerator in the exponent in the exponential function can be rewritten as

$$x^2 - 2\rho xy + y^2 = (x - \rho y)^2 + (1 - \rho^2)y^2,$$

hence

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp\left(-\frac{x^2 - 2\rho xy + y^2}{2}\right) \mathrm{d}y\mathrm{d}x$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp\left(-\frac{(x-\rho y)^2}{2} - \frac{(1-\rho^2)y^2}{2}\right) \mathrm{d}y\mathrm{d}x$$

$$= \int_{-\infty}^{\infty}\left\{\int_{-\infty}^{\infty} \exp\left(-\frac{(x-\rho y)^2}{2}\right) \mathrm{d}x\right\} \exp\left(-\frac{(1-\rho^2)y^2}{2}\right) \mathrm{d}y.$$

The inner integral can be computed for fixed y using substitution and knowledge about the one-dimensional normal distribution. With $z = x - \rho y$ we have $\mathrm{d}z = \mathrm{d}x$, hence

$$\int_{-\infty}^{\infty} \exp\left(-\frac{(x-\rho y)^2}{2}\right) \mathrm{d}x = \int_{-\infty}^{\infty} \exp\left(-\frac{z^2}{2}\right) \mathrm{d}z = \sqrt{2\pi},$$

where the last equality follows from the fact that the density for the normal distribution on $\mathbb{R}$ is

$$\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right),$$

which integrates to 1. We see that the inner integral does not depend upon y and is constantly equal to $\sqrt{2\pi}$, thus another substitution with $z = y\sqrt{1-\rho^2}$ (using that $\rho \in (-1, 1)$) such that $\mathrm{d}z = \sqrt{1-\rho^2}\mathrm{d}y$ gives

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp\left(-\frac{x^2 - 2\rho xy + y^2}{2}\right) \mathrm{d}y\mathrm{d}x = \sqrt{2\pi} \int_{-\infty}^{\infty} \exp\left(-\frac{(1-\rho^2)y^2}{2}\right) \mathrm{d}y$$

$$= \frac{\sqrt{2\pi}}{\sqrt{1-\rho^2}} \int_{-\infty}^{\infty} \exp\left(-\frac{z^2}{2}\right) \mathrm{d}z$$

$$= \frac{2\pi}{\sqrt{1-\rho^2}}$$

where we once more use that the last integral equals $\sqrt{2\pi}$. This shows that the (positive) function f integrates to 1, and it is therefore a density for a probability measure on $\mathbb{R}^2$.

Suppose that the distribution of (X, Y) has density f on $\mathbb{R}^2$, then we have almost computed the marginal distributions of X and Y by the integrations above. The marginal distribution of Y has by Result 2.13.5 density

$$f_2(y) = \frac{\sqrt{1-\rho^2}}{2\pi} \int_{-\infty}^{\infty} \exp\left(-\frac{x^2 - 2\rho xy + y^2}{2}\right) \mathrm{d}x$$

$$= \frac{1}{\sqrt{2\pi(1-\rho^2)^{-1}}} \exp\left(-\frac{y^2}{2(1-\rho^2)^{-1}}\right).$$

Here we used, as argued above, that the integration over x gives $\sqrt{2\pi}$, and then we have rearranged the expression a little. From (2.20) in Example 2.9.9 we recognize

Math Box 2.13.1 (Multivariate normal distributions)**.** We can define the family of n-dimensional regular normal or Gaussian distributions via their densities on $\mathbb{R}^n$ for $n \geq 1$. The measures are characterized in terms of a vector $\xi \in \mathbb{R}^n$ and a *positive definite* symmetric $n \times n$ matrix Λ. Positive definite means that for any vector $x \in \mathbb{R}^n$ we have $x^t \Lambda x > 0$. Consequently one can show that the positive function $x \mapsto \exp(-(\frac{1}{2}x - \xi)^t \Lambda(x - \xi))$ have a finite integral over $\mathbb{R}^n$ and that

$$\int_{\mathbb{R}^n} \exp(-\frac{1}{2}(x - \xi)^t \Lambda(x - \xi)) \mathrm{d}x = \sqrt{\frac{(2\pi)^n}{\det\Lambda}}.$$

Here $\det\Lambda$ is the determinant of the matrix Λ. The density for the n-dimensional regular normal distribution with parameters ξ and Λ is then

$$f(x) = \sqrt{\frac{\det\Lambda}{(2\pi)^n}} \exp\left(-\frac{1}{2}(x - \xi)^t \Lambda(x - \xi)\right).$$

The full bivariate normal distribution considered in Example 2.13.6 as given by (2.32) corresponds to

$$\Lambda = \begin{pmatrix} \lambda_1 & -\rho \\ -\rho & \lambda_2 \end{pmatrix},$$

for which we can verify that $\det\Lambda = \lambda_1\lambda_2 - \rho^2$.

That Λ is positive definite implies that $\det\Lambda > 0$ and thus that Λ is invertible. We denote the inverse by $\Sigma = \Lambda^{-1}$, and using that $\det\Sigma = (\det\Lambda)^{-1}$ we can also write the density as

$$f(x) = \frac{1}{\sqrt{(2\pi)^n \det\Sigma}} \exp\left(-\frac{1}{2}(x - \xi)^t \Sigma^{-1}(x - \xi)\right).$$

One can then show that the i'th marginal distribution is the normal distributions with location parameter ξ_i and scale parameter Σ_{ii}.

It is possible to define a multivariate normal distribution for *positive semi-definite* matrices Λ that only fulfill that $x^t \Lambda x \geq 0$, but it requires a little work. If we don't have strict inequality, there is no density, and the multivariate normal distribution is called singular.

this density as the density for the normal distribution with scale parameter $\sigma^2 = (1 - \rho^2)^{-1}$ and location parameter $\mu = 0$. Note that for $\rho = 0$ we have $\sigma^2 = 1$ and for $\rho \to \pm 1$ we have $\sigma^2 \to \infty$. The density is entirely symmetric in the x- and y-variables, so the marginal distribution of X is also $N(0, (1 - \rho^2)^{-1})$.

This probability measure is called a bivariate normal or Gaussian distribution. The example given above contains only a single parameter, ρ, whereas the general bivariate normal distribution is given in terms of five parameters. The general bivariate

density is given as

$$f(x,y) = \frac{\sqrt{\lambda_1\lambda_2 - \rho^2}}{2\pi} \exp\left(-\frac{1}{2}\left(\lambda_1(x - \xi_1)^2 - 2\rho(x - \xi_1)(y - \xi_2) + \lambda_2(y - \xi_2)^2\right)\right)$$

(2.32)

where $\xi_1, \xi_2 \in \mathbb{R}$, $\lambda_1, \lambda_2 \in (0, \infty)$ and $\rho \in (-\lambda_1\lambda_2, \lambda_1\lambda_2)$. For this density it is possible to go through similar computations as above, showing that it integrates to 1. If the distribution of (X, Y) is given by f it is likewise possible to compute the marginal distributions, where $X \sim N\left(\xi_1, \frac{\lambda_2}{\lambda_1\lambda_2 - \rho^2}\right)$ and $Y \sim N\left(\xi_2, \frac{\lambda_1}{\lambda_1\lambda_2 - \rho^2}\right)$. $\diamond$

It is possible to define n-dimensional normal distributions in a similar way to the bivariate normal distribution considered in Example 2.13.6; see Math Box 2.13.1. The theory for the multivariate normal distribution is very rich and well developed, and it has played and still plays an important role in theoretical as well as applied statistics. The widespread use of analysis of variance (ANOVA) and classical linear regression is founded on results about the multivariate normal distribution, and these classical models are also important tools in bioinformatics, for instance in the modeling and analysis of gene expression data. The theory is, however, covered so well in the literature, that we will pursue any systematic development in these notes.

Example 2.13.7. The marginal distributions of X and Y computed in Example 2.13.6 were both found to be $N(0, (1 - \rho^2)^{-1})$. The product of the corresponding marginal densities is

$$f_1(x)f_2(y) = \frac{1 - \rho^2}{2\pi} \exp\left(-\frac{x^2 + y^2}{2(1 - \rho^2)^{-1}}\right),$$

which is seen to be equal to $f(x, y)$ from Example 2.13.6 if and only if $\rho = 0$. Thus if the distribution of (X, Y) is given by the density f in Example 2.13.6, then X and Y are independent if and only if the parameter ρ equals 0. $\diamond$

Example 2.13.8. (Bivariate Dirichlet distribution) Let $\lambda_1, \lambda_2, \lambda > 0$ be given. Define for $(x, y) \in \mathbb{R}^2$ with $x, y > 0$ and $x + y < 1$ the function

$$f(x, y) = \frac{\Gamma(\lambda_1 + \lambda_2 + \lambda)}{\Gamma(\lambda_1)\Gamma(\lambda_2)\Gamma(\lambda)} x^{\lambda_1 - 1} y^{\lambda_2 - 1} (1 - x - y)^{\lambda - 1}.$$

The function f is a priori defined on the open 2-dimensional *unit simplex*

$$U_2 = \{(x, y) \in \mathbb{R}^2 \mid x, y > 0, x + y < 1\}.$$

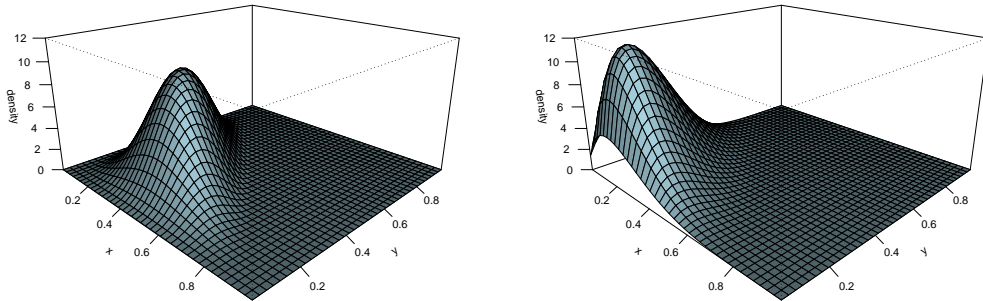By defining $f(x, y) = 0$ outside U_2, we can show that f integrates to 1. We only

Figure 2.22: Two examples of the density for the bivariate Dirichlet distribution as considered in Example 2.13.8 with $\lambda_1 = \lambda_2 = \lambda = 4$ (left) and $\lambda_1 = \lambda_2 = 2$ together with $\lambda = 6$ (right).

need to integrate over the set U_2 as f is defined to be 0 outside, and we find that

$$\int_0^1 \int_0^{1-x} x^{\lambda_1-1} y^{\lambda_2-1} (1-x-y)^{\lambda-1} \mathrm{d}y \mathrm{d}x$$

$$= \int_0^1 \left\{ \int_0^{1-x} y^{\lambda_2-1} (1-x-y)^{\lambda-1} \mathrm{d}y \right\} x^{\lambda_1-1} \mathrm{d}x$$

$$= \int_0^1 \left\{ \int_0^1 (1-x)^{\lambda_2-1} z^{\lambda_2-1} (1-x)^{\lambda-1} (1-z)^{\lambda-1} (1-x) \mathrm{d}z \right\} x^{\lambda_1-1} \mathrm{d}x$$

$$= \int_0^1 \left\{ \int_0^1 z^{\lambda_2-1} (1-z)^{\lambda-1} \mathrm{d}z \right\} (1-x)^{\lambda_2+\lambda-1} x^{\lambda_1-1} \mathrm{d}x$$

$$= \int_0^1 z^{\lambda_2-1} (1-z)^{\lambda-1} \mathrm{d}z \int_0^1 x^{\lambda_1-1} (1-x)^{\lambda_2+\lambda-1} \mathrm{d}x$$

where we have used the substitution $(1-x)z = y$ in the first inner integral, for which $(1-x)\mathrm{d}z = \mathrm{d}y$. Both of the last two integrals can be recognized as B-integrals, cf. (B.7), and according to (B.6) we find that

$$\int_0^1 z^{\lambda_2-1} (1-z)^{\lambda-1} \mathrm{d}z \int_0^1 x^{\lambda_1-1} (1-x)^{\lambda_2+\lambda-1} \mathrm{d}x = \frac{\Gamma(\lambda_2)\Gamma(\lambda)}{\Gamma(\lambda_2+\lambda)} \frac{\Gamma(\lambda_1)\Gamma(\lambda_2+\lambda)}{\Gamma(\lambda_1+\lambda_2+\lambda)}$$

$$= \frac{\Gamma(\lambda_1)\Gamma(\lambda_2)\Gamma(\lambda)}{\Gamma(\lambda_1+\lambda_2+\lambda)}.$$

We see that the integral of the positive function f is 1, and f is the density for the *bivariate Dirichlet distribution* on $\mathbb{R}^2$. Since f is 0 outside the 2-dimensional unit simplex U_2, the distribution is really a distribution living on the unit simplex. $\diamond$

> **Math Box 2.13.2** (Multivariate Dirichlet distribution)**.** There is an n-dimensional Dirichlet distribution living on the n-dimensional unit simplex
>
> $$U_n = \{(x_1, \ldots, x_n) \mid x_1, \ldots, x_n > 0, x_1 + \ldots + x_n < 1\},$$
>
> with density
>
> $$f(x_1, \ldots, x_n) = \frac{\Gamma(\lambda_1 + \ldots + \lambda_n + \lambda)}{\Gamma(\lambda_1) \cdots \Gamma(\lambda_n)\Gamma(\lambda)} x_1^{\lambda_1 - 1} x_2^{\lambda_2 - 1} \cdots x_n^{\lambda_n - 1} (1 - x_1 - \ldots - x_n)^{\lambda - 1}$$
>
> for $(x_1, \ldots, x_n) \in U_n$ and with the parameters $\lambda_1, \ldots, \lambda_n, \lambda > 0$. We refrain from showing that the density f really integrates to 1 over the unit simplex U_n.
>
> We note that if the joint distribution of $X_1, \ldots, X_n$ is an n-dimensional Dirichlet distribution, then
> $$(X_1, \ldots, X_n, 1 - X_1 - \ldots - X_n)$$
> is an $n + 1$-dimensional probability vector. Therefore the Dirichlet distribution is often encountered as a model of random probability vectors (e.g. frequencies).

Exercises

Exercise 2.13.1. Consider the pair of random variables (X, Y) as in Example 2.13.3 in the notes. Thus (X, Y) denotes the pair of aligned nucleotides in an alignment of two DNA-sequences and the joint distribution of this pair is given by the point probabilities

	A	C	G	T
A	0.12	0.03	0.04	0.01
C	0.02	0.27	0.02	0.06
G	0.02	0.01	0.17	0.02
T	0.05	0.03	0.01	0.12

- Compute the point probabilties for the marginal distributions, P_1 and P_2, of the variables X and Y respectively.

- Compute the point probabilities for the distribution, P, that make X and Y independent with marginal distribution P_1 and P_2 respectively.

- Compute the probability under P as given above for the event $X = Y$. Compare with the probability for this event as computed in Example 2.13.3.

2.13.1 Conditional distributions and conditional densities

If X and Y are real valued random variables with joint distribution P on $\mathbb{R}^2$ having density f, then the probability of X being equal to x is 0 for all $x \in \mathbb{R}$. Thus we cannot use Definition 2.10.8 to find the conditional distribution of Y given $X = x$. It is, however, possible to define a conditional distribution in a sensible way as long as $f_1(x) > 0$ where f_1 denotes the density for the *marginal* distribution of X. This

is even possible if we consider not just two real valued random variables, but in fact if we consider two random variables X and Y with values in $\mathbb{R}^n$ and $\mathbb{R}^m$ such that the joint distribution P on $\mathbb{R}^{n+m}$ has density f.

Definition 2.13.9 (Conditional densities). *If f is the density for the joint distribution of two random variables X and Y taking values in $\mathbb{R}^n$ and $\mathbb{R}^m$, respectively, then with*

$$f_1(x) = \int_{\mathbb{R}^m} f(x,y)\mathrm{d}y$$

we define the conditional distribution of Y given $X = x$ to be the distribution with density

$$f(y|x) = \frac{f(x,y)}{f_1(x)}.$$

for all $y \in \mathbb{R}^m$ and $x \in \mathbb{R}^n$ with $f_1(x) > 0$.

Note that from this definition we find the formula

$$f(x,y) = f(y|x)f_1(x), \tag{2.33}$$

which reads that the density for the joint distribution is the product of the densities for the marginal distribution of X and the conditional distribution of Y given X. Note the analogy to (2.21) and especially (2.23) for point probabilities. It is necessary to check that the definition really makes sense, that is, that $f(y|x)$ actually defines a density for all $x \in \mathbb{R}$ with $f_1(x) > 0$. It is positive by definition, and we also see that

$$\int_{\mathbb{R}^m} f(y|x)\mathrm{d}y = \frac{\int_{\mathbb{R}^m} f(x,y)\mathrm{d}y}{f_1(x)} = 1$$

by (2.31).

If we let P_x denote the probability measure with density $y \mapsto f(y|x)$ for given $x \in \mathbb{R}^n$ with $f_1(x) > 0$, i.e. the conditional distribution of Y given $X = x$, then for $B \subseteq \mathbb{R}^m$,

$$\mathbb{P}(Y \in B | X = x) = P_x(B) = \int_B f(y|x)\mathrm{d}y.$$

If moreover $A \subseteq \mathbb{R}^n$

$$
\begin{aligned}
\mathbb{P}(X \in A, Y \in B) = P(A \times B) \ &= \ \int_A \int_B f(x,y)\mathrm{d}y\mathrm{d}x \\
&= \ \int_A \int_B f(y|x)f_1(x)\mathrm{d}y\mathrm{d}x \\
&= \ \int_A \left\{ \int_B f(y|x)\mathrm{d}y \right\} f_1(x)\mathrm{d}x \\
&= \ \int_A P_x(B)f_1(x)\mathrm{d}x.
\end{aligned}
$$

The interpretation is that the realization of X and Y can always be thought of as a two step procedure. First X is realized according to the probability measure with density f_1, and we determine whether X belongs to the event A or not (the outer integration over A). Then conditionally on $X = x$ the realization of Y is given according to the conditional probability measure P_x.

A common usage of conditional densities is to *define* multivariate probability measures via (2.33). Thus if we are given a marginal density and a conditional density, then we can define f by (2.33). By successive integration, first over y then over x, we see that this defines a density on $\mathbb{R}^{n+m}$.

Example 2.13.10 (Linear regression)**.** Let X be a real valued random variable with distribution $N(0, \sigma_1^2)$. Define

$$Y = \alpha + \beta X + \varepsilon \tag{2.34}$$

where $\varepsilon \sim N(0, \sigma_2^2)$ is independent of X. We read this equation as a structural equation, see Example 2.10.12 that defines Y as follows. Given $X = x$ the conditional distribution of Y equals the distribution of $\alpha + \beta x + \varepsilon$. This is a location transformation of the normal distribution, thus the conditional distribution of Y given $X = x$ is $N(\alpha + \beta x, \sigma_2^2)$. We can then use (2.33) to compute the joint distribution of (X, Y) as

$$
\begin{aligned}
f(x, y) &= \frac{1}{2\pi\sigma_1\sigma_2} \exp\left(-\frac{(y - \alpha - \beta x)^2}{2\sigma_2^2} - \frac{x^2}{2\sigma_1^2} \right) \\
&= \frac{1}{2\pi\sigma_1\sigma_2} \exp\left(-\frac{1}{2} \left(\frac{\sigma_1^2\beta^2 + \sigma_2^2}{\sigma_1^2\sigma_2^2} x^2 - 2\frac{\beta}{\sigma_2^2} x(y - \alpha) + \frac{1}{\sigma_2^2}(y - \alpha)^2 \right) \right) \tag{2.35}
\end{aligned}
$$

One alternative interpretation of the defining equation (2.34) of Y is that Y is a transformation of X and ε. In Exercise 2.13.5 it is shown that this interpretation leads to the same result. That is, the same joint distribution of X and Y. This is fortunate as it shows that the distributional interpretation of equation (2.34) is not ambiguous. $\diamond$

Exercises

Exercise 2.13.2. In this exercise we consider a dataset of amino acid pairs (x, y). We think of the data as representing the outcome of a random variable (X, Y). Here X and Y represent an amino acid at a given position in two evolutionary related proteins (same protein, but from two different species, say). The dataset may be obtained from a (multiple) alignment of (fractions) of proteins. The only mutational event is substitution of one amino acid for another. In this exercise you are allowed to think of the different positions as independent, but X and Y are dependent.

- Load the `aa` dataset into R with `data(aa)` and use `table` to cross-tabulate the data according to the values of x and y. Compute the matrix of relative frequencies for the occurrence of all pairs (x, y) for $(x, y) \in E_0 \times E_0$ where E_0 denotes the amino acid alphabet.

Exercise 2.13.3. Continuing with the setup from the previous exercise we denote by P the probability measure with point probabilities, $p(x, y)$, being the relative frequencies computed above. It is a probability measure on $E_0 \times E_0$

- Compute the point probabilities, $p_1(x)$ and $p_2(y)$, for the marginal distributions of P and show that X and Y are not independent.

- Compute the *score* matrix defined as

$$S_{x,y} = \log \frac{p(x, y)}{p_1(x) p_2(y)}.$$

  Discuss the interpretation of the values.

- If (X, Y) have distribution P, $S_{X,Y}$ can be regarded as a random variable with values in a finite sample space (why?). Compute the mean of $S_{X,Y}$.

- Assume instead that X and Y are in fact independent with distributions given by the point probabilities $p_1(x)$ and $p_2(y)$ respectively and compute then the mean of $S_{X,Y}$.

$\star$

Exercise 2.13.4. Show that if the distribution of (X, Y) is given by the density f in (2.35) then the marginal distribution of Y is $N(\alpha, \beta^2 \sigma_1^2 + \sigma_2^2)$. Then show that X and Y are independent if and only if $\beta = 0$.
Hint: See Example 2.13.6, formula (2.32).

$\star\,\star$

Exercise 2.13.5. Let X and ε be independent real valued random variables with distribution $N(0, \sigma_1^2)$ and $N(0, \sigma_2^2)$ respectively. Show that the density for their joint distribution is
$$f(x, \varepsilon) = \frac{1}{2\pi \sigma_1 \sigma_2} \exp\left(-\frac{x^2}{2\sigma_1^2} - \frac{\varepsilon^2}{2\sigma_2^2}\right).$$
Define the transformation
$$h : \mathbb{R}^2 \to \mathbb{R}$$
by $h(x, \varepsilon) = (x, \alpha + \beta y + \varepsilon)$ and find the density for the distribution of $h(X, \varepsilon)$. Compare with Example 2.13.10.
Hint: First find $h^{-1}(I \times J)$ for $I, J \subseteq \mathbb{R}$. You may think of these sets as intervals. Then use the definitions of transformed distributions and densities to compute

$$\mathbb{P}(X \in I, \alpha + \beta X + \varepsilon \in J) = \mathbb{P}((X, \varepsilon) \in h^{-1}(I \times J)).$$

2.14 Descriptive methods

Multivariate datasets from the sample space $\mathbb{R}^d$ with $d \geq 2$ are more difficult to visualize than a one-dimensional dataset. It is the same problem as we have with tabulations. Two-dimensional tables are basically the limit for our comprehension. Bivariate datasets ($d = 2$) can also be visualized for instance via scatter plots and bivariate kernel density estimation, but for multivariate datasets we usually have to rely on visualizing univariate or bivariate transformations.

If $x_i = (x_{i1}, x_{i2}) \in \mathbb{R}^2$ is bivariate a simple plot of x_{i2} against x_{i1} for $i = 1, \ldots, n$
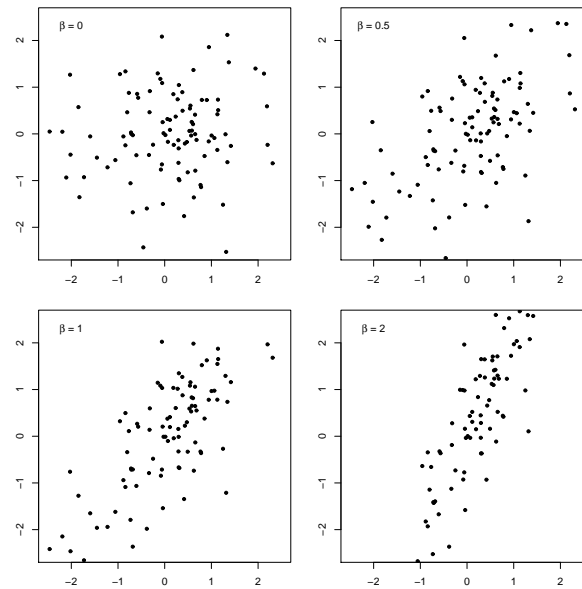
Figure 2.23: Scatter plot of 100 simulated variables as described in Example 2.14.1. The variables have a linear dependence structure determined by the parameter β that ranges from 0 to 2 in these plots.

is called a *scatter plot*. It corresponds to a one-dimensional rug plot, but the two dimensions actually gives us a better visualization of the data in this case. The scatter plot may highlight some simple dependence structures in the data. For instance, if the conditional distribution of X_{i2} given $X_{i1} = x_{i1}$ is $N(\alpha + \beta x_{i1}, \sigma_2^2)$ as in Example 2.13.10 with $\alpha, \beta \in \mathbb{R}$, then this will show up on the scatter plot (if $\beta \neq 0$) as a tendency for the points to lie around a line with slope β. How obvious this is depends on how large σ_2^2 is compared to β and how spread out the distribution of X_{i1} is.

Example 2.14.1. We simulate $n = 100$ bivariate iid random variables (X_{i1}, X_{i2}) for $i = 1, \ldots, 100$ where the distribution of $X_{i1} \sim N(0, 1)$ and the *conditional* distribution of X_{i2} given $X_{i1} = x_{i1}$ is $N(\beta x_{i1}, 1)$ with $\beta = 0, 0.5, 1, 2$. The resulting scatter plots are shown in Figure 2.23. The dependence between X_{i1} and X_{i2} is determined by β, and as shown in Exercise 2.13.4, the variables are independent if and only if $\beta = 0$. From the figure we observe how the dependence through β influences the scatter plot. As β becomes larger, the points clump around a line with slope β. For small β, in this case $\beta = 0.5$ is quite small, the scatter plot is not so easy to distinguish from the scatter plot with $\beta = 0$ as compared to the larger values of β. Thus weak dependence is not so easy to spot by eye. $\diamond$

Example 2.14.2 (Iris data)**.** A classical dataset that is found in numerous textbooks contains characteristic measurements of the Iris flower for three different species, *Iris*
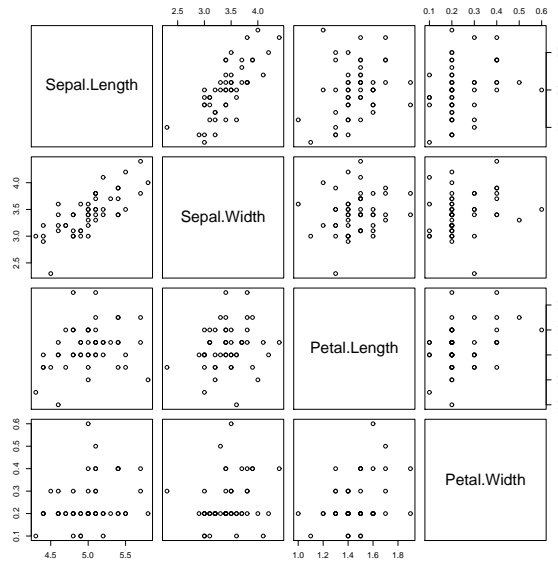
Figure 2.24: Considering the species Setosa in the Iris dataset we see a bivariate scatter plot of all the pair-wise combinations of the four variables in the dataset. This is obtained in R simply by calling `plot` for a dataframe containing the relevant variables.

Setosa, *Iris Virginica* and *Iris Versicolor*. The data were collected in 1935 by Edgar Anderson, and is today available in R via `data(iris)`. The date are organized in a dataframe as follows

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5.0	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa
4.6	3.4	1.4	0.3	setosa
5.0	3.4	1.5	0.2	setosa
4.4	2.9	1.4	0.2	setosa
4.9	3.1	1.5	0.1	setosa
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

There are 50 measurements for each species, thus a total of 150 rows in the data frame. In Figure 2.24 we see two-dimensional scatter plots of the different measurements against each other for the species *Setosa*. Restricting our attention to the

two variables *Petal length* and *Petal width*, we see in Figure 2.25 scatter plots, two-dimensional kernel density estimates and corresponding contour curves. The kernel density estimates are produced by `kde2d` from the `MASS` package. ◇

> **R Box 2.14.1** (Bivariate kernel density estimation)**.** The package `MASS` contains a function, `kde2d`, for bivariate kernel density estimation. It uses a bivariate Gaussian kernel, where you can change the bandwidth along each of the coordinate axes separately. It returns the evaluation of the kernel density estimate in a quadratic grid. You can use the plotting functions `persp`, `contour` and `filled.contour` to visualize the estimated density.

2.15 Transition probabilities

Recall that if we have two random variables X and Y taking values in the discrete sample space E we have the relation

$$\mathbb{P}(X = x, Y = y) = \mathbb{P}(Y = y | X = x)\mathbb{P}(X = x).$$

If we think about the two step interpretation of the joint distribution the relation above says that the joint distribution is decomposed into a first step where X is sampled according to its marginal distribution and then Y is sampled conditionally on X. With such a dynamic interpretation of the sampling process it is sensible to call $\mathbb{P}(Y = y | X = x)$ the *transition probabilities*. Note the the entire dependence structure between X and Y is captured by the conditional distribution, and X and Y are independent if $\mathbb{P}(Y = y | X = x) = \mathbb{P}(Y = y)$ for all $y \in E$.

It is possible to specify whole families of probability measures on $E \times E$ where E is a finite sample space, e.g. $E = \{\texttt{A}, \texttt{C}, \texttt{G}, \texttt{T}\}$, which are indexed by a time parameter $t \geq 0$ and perhaps some additional parameters that capture the dependence structure. This is done by specifying the transition probabilities via a system of differential equations. We introduce P_t as the distribution of the pair of random variables (X, Y) on $E \times E$ indexed by the time parameter $t \geq 0$. We are going to assume that the *marginal distribution* of X, which is given by the point probabilities

$$\pi(x) = \sum_{y \in E} P_t(x, y),$$

does not depend on t, and we are going to define the matrix P^t for each $t \geq 0$ by
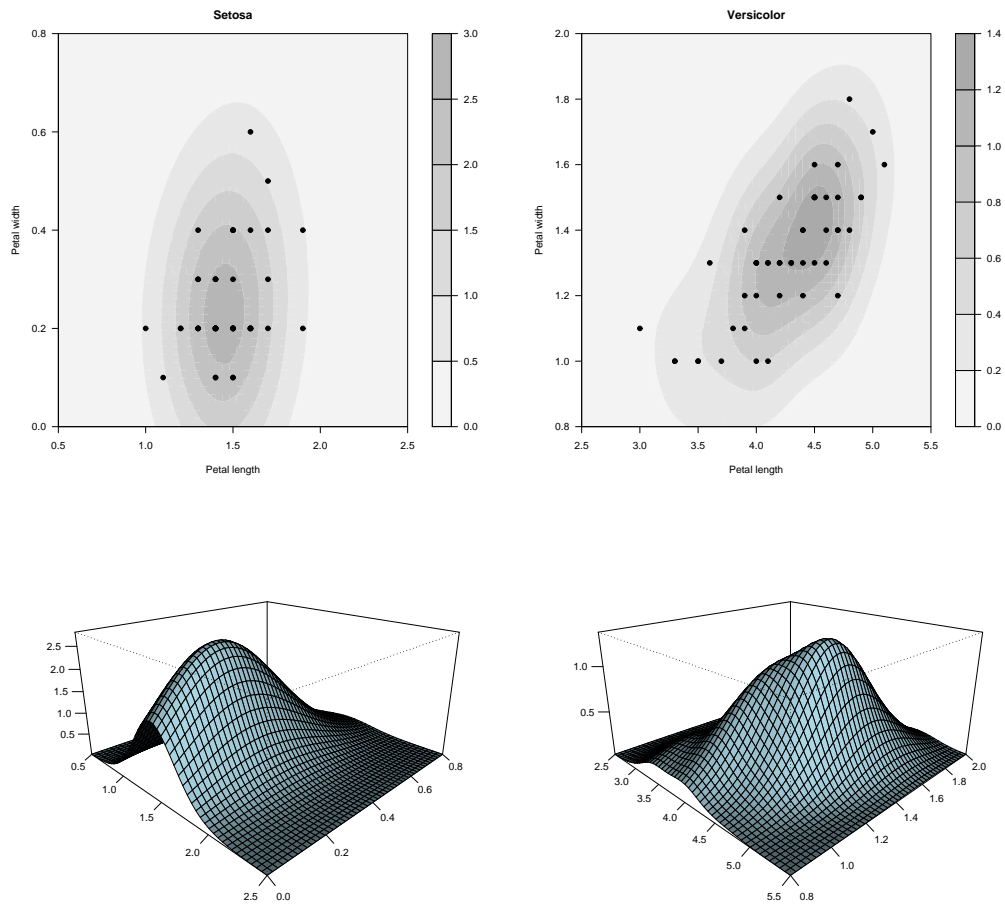
$$P^t(x, y) = \frac{P_t(x, y)}{\pi(x)}.$$

Figure 2.25: For the Iris Setosa and Iris Versicolor we have estimated a two-dimensional density for the simultaneous distribution of petal length and width. In both cases we used a bivariate Gaussian kernel with bandwidth 0.8 in both directions.

That is, $P^t(x, y)$ is the conditional probability that x changes into y over the time interval t (note that we may have $x = y$). There is a natural *initial condition* on P^t for $t = 0$, since for $t = 0$ we will assume that $X = Y$, thus

$$P^0(x, y) = \begin{cases} 1 & \text{if } x = y \\ 0 & \text{otherwise} \end{cases}$$

This is the only reasonable assumption if t is to be interpreted as a time parameter. With this initial condition we specify the matrix of conditional probabilities by the

differential equation

$$\frac{\mathrm{d}P^t(x,y)}{\mathrm{d}t} = \sum_{z \in E} P^t(x,z)\lambda(z,y) \qquad (2.36)$$

for $x,y \in E$ and $(\lambda(z,y))_{z,y \in E}$ a matrix of *mutation intensities* satisfying $\lambda(z,y) \geq 0$ for $z \neq y$ and

$$\lambda(z,z) = -\sum_{y \neq z} \lambda(z,y). \qquad (2.37)$$

Thus to specify P^t for all $t \geq 0$ by (2.36) we need only to specify a single matrix of intensities. In a very short time interval, Δ, the interpretation of the differential equation is that

$$\begin{aligned} \frac{P^{t+\Delta}(x,y) - P^t(x,y)}{\Delta} &\simeq \sum_{z \in E} P^t(x,z)\lambda(z,y) \\ &= P^t(x,y)\lambda(y,y) + \sum_{z \neq y} P^t(x,z)\lambda(z,y). \end{aligned}$$

Rearranging yields

$$P^{t+\Delta}(x,y) \simeq P^t(x,y)(1 + \Delta\lambda(y,y)) + \sum_{z \neq y} P^t(x,z)\Delta\lambda(z,y).$$

This equation reads that the conditional probability that x changes into y in the time interval $t + \Delta$ is given as a sum of two terms. The first term is the probability $P^t(x,y)$ that x changes into y in the time interval t times the factor $1 + \Delta\lambda(y,y)$. This factor has the interpretation as the probability that y doesn't mutate in the short time interval Δ. The second term is a sum of terms where x changes into some $z \neq y$ in the time interval t and then in the short time interval Δ it changes from z to y with probability $\Delta\lambda(z,y)$. In other words, for $z \neq y$ the entries in the intensity matrix have the interpretation that $\Delta\lambda(z,y)$ is approximately the probability that z changes into y in the short time interval Δ, and $1 + \Delta\lambda(y,y)$ is approximately the probability that y doesn't change in the short time interval Δ.

In Exercises 2.15.1 and 2.15.2 it is verified, as a consequence of $\lambda(x,y) \geq 0$ for $x \neq y$ and (2.37), that the solution to the system of differential equations (2.36) is indeed a matrix of conditional probabilities, that is, all entries of P^t are ≥ 0 and the row sums are always equal to 1.

It is a consequence of the general theory for systems of linear differential equations that there exists a unique solution to (2.36) with the given initial condition. In general the solution to (2.36) can, however, not easily be given a simple analytic representation, unless one accepts the exponential of a matrix as a simple analytic expression; see Math Box 2.15.1. There we also show that the solution satisfies the *Chapman-Kolmogorov* equations; for $s,t \geq 0$ and $x,y \in E$

$$P^{t+s}(x,y) = \sum_{z \in E} P^t(x,z)P^s(z,y). \qquad (2.38)$$

Math Box 2.15.1 (Transition probability matrices). With $E = \{x_1, \ldots, x_n\}$ define the matrix

$$
\Lambda = \begin{pmatrix}
\lambda(x_1, x_1) & \lambda(x_1, x_2) & \ldots & \lambda(x_1, x_n) \\
\lambda(x_2, x_1) & \lambda(x_2, x_2) & \ldots & \lambda(x_2, x_n) \\
\vdots & \vdots & & \vdots \\
\lambda(x_n, x_1) & \lambda(x_n, x_2) & \ldots & \lambda(x_n, x_n)
\end{pmatrix},
$$

then the set of differential equations defined by (2.36) can be expressed in matrix notation as

$$
\frac{\mathrm{d}P^t}{\mathrm{d}t} = P^t \Lambda.
$$

If Λ is a real number, this differential equation is solved by $P^t = \exp(t\Lambda)$, and this solution is also the correct solution for matrix Λ, although it requires a little more work to understand how the exponential function works on matrices. One can take as a definition of the exponential function the usual Taylor expansion

$$
\exp(t\Lambda) = \sum_{n=0}^{\infty} \frac{t^n \Lambda^n}{n!}.
$$

It is possible to verify that this infinite sum makes sense. Moreover, one can simply differentiate w.r.t. the time parameter t term by term to obtain

$$
\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t} \exp(t\Lambda) &= \sum_{n=1}^{\infty} n \frac{t^{n-1} \Lambda^n}{n!} \\
&= \left(\sum_{n=1}^{\infty} \frac{t^{n-1} \Lambda^{n-1}}{(n-1)!} \right) \Lambda \\
&= \left(\sum_{n=0}^{\infty} \frac{t^n \Lambda^n}{n!} \right) = \exp(t\Lambda)\Lambda.
\end{aligned}
$$

Many of the usual properties of the exponential function carry over to the exponential of matrices. For instance, $\exp(t\Lambda + s\Lambda) = \exp(t\Lambda) \exp(s\Lambda)$, which shows that

$$
P^{t+s} = P^t P^s.
$$

This is the *Chapman-Kolmogorov* equations, (2.38), in their matrix version.

For some special models we are capable of obtaining closed form expressions for the solution. This is for instance the case for the classical examples from molecular evolution with $E = \{\texttt{A}, \texttt{C}, \texttt{G}, \texttt{T}\}$ that lead to the Jukes-Cantor model and the Kimura model introduced earlier via their the conditional probabilities.

Example 2.15.1 (The Jukes-Cantor model). The *Jukes-Cantor* model is given by

assuming that

$$\lambda(x, y) = \begin{cases} -3\alpha & \text{if } x = y \\ \alpha & \text{if } x \neq y \end{cases}$$

That is, the matrix of intensities is

	A	C	G	T
A	-3α	α	α	α
C	α	-3α	α	α
G	α	α	-3α	α
T	α	α	α	-3α

The parameter $\alpha > 0$ tells how many mutations that occur per time unit. The solution to (2.36) is

$$\begin{aligned} P^t(x, x) &= 0.25 + 0.75 \times \exp(-4\alpha t) \\ P^t(x, y) &= 0.25 - 0.25 \times \exp(-4\alpha t), \quad \text{if } x \neq y, \end{aligned}$$

which is verified by checking that P^t fulfills the differential equation. $\diamond$

Example 2.15.2 (The Kimura model). Another slightly more complicated model defined in terms of the differential equations (2.36), that admits a relatively nice closed form solution, is the *Kimura model*. Here the intensity matrix is assumed to be

	A	C	G	T
A	$-\alpha - 2\beta$	β	α	β
C	β	$-\alpha - 2\beta$	β	α
G	α	β	$-\alpha - 2\beta$	β
T	β	α	β	$-\alpha - 2\beta$

with $\alpha, \beta > 0$. The interpretation is that a substitution of a purine with a purine or a pyrimidine with a pyrimidine (a transition) is happening with another intensity than a substitution of a purine with pyrimidine or pyrimidine with purine (a transversion). We see that the intensity for transversions is $\lambda(x, y) = \beta$ and the intensity for transitions is $\lambda(x, y) = \alpha$. The solution is

$$\begin{aligned} P^t(x, x) &= 0.25 + 0.25 \exp(-4\beta t) + 0.5 \exp(-2(\alpha + \beta)t) \\ P^t(x, y) &= 0.25 + 0.25 \exp(-4\beta t) - 0.5 \exp(-2(\alpha + \beta)t), \quad \text{if } \lambda(x, y) = \alpha \\ P^t(x, y) &= 0.25 - 0.25 \exp(-4\beta t), \quad \text{if } \lambda(x, y) = \beta, \end{aligned}$$

which is verified by checking that the system of differential equations is fulfilled. It is tedious but in principle straight forward. $\diamond$

Exercises

★

Exercise 2.15.1. Let $P^t(x, y)$ for $x, y \in E$ satisfy the system of differential equations

$$\frac{\mathrm{d}P^t(x, y)}{\mathrm{d}t} = \sum_{z \in E} P^t(x, z)\lambda(z, y)$$

with $\lambda(x, y) \geq 0$ for $x \neq y$ and $\lambda(x, x) = -\sum_{y \neq y} \lambda(x, y)$. Define

$$s_x(t) = \sum_y P^t(x, y)$$

as the "row sums" of P^t for each $x \in E$. Show that $\mathrm{d}s_x(t)/\mathrm{d}t = 0$ for $t \geq 0$. Argue that with the initial condition, $P^0(x, y) = 0$, $x \neq y$ and $P^0(x, x) = 1$, then $s_x(0) = 1$ and conclude that $s_x(t) = 1$ for all $t \geq 0$.

★ ★

Exercise 2.15.2. We consider the same system of differential equations as above with the same initial condition. Assume that $\lambda(x, y) > 0$ for all $x \neq y$. Use the sign of the resulting derivative

$$\frac{\mathrm{d}P^0(x, y)}{\mathrm{d}t} = \lambda(x, y)$$

at 0 to argue that for a sufficiently small $\varepsilon > 0$, $P^t(x, y) \geq 0$ for $t \in [0, \varepsilon]$. Use this fact together with the Chapman-Kolmogorov equations, (2.38), to show that $P^t(x, y) \geq 0$ for all $t \geq 0$.

3

Statistical models and inference

Given one or more realizations of an experiment with sample space E, which probability measure P on E models – or describes – the experiment adequately? This is the crux of statistics – the *inference* of a suitable probability measure or aspects of a probability measure from empirical data. What we consider in the present chapter is the problem of *estimation* of one or more parameters that characterize the probability measure with the main focus on using the maximum likelihood estimator.

3.1   Statistical Modeling

A statistical model is a collection of probability measures on the sample space E. Before we can estimate the probability measure – the unknown parameters – we need to decide upon which class of probability measures we believe are reasonable models of the data that we want to model. In other words, we need to set up the framework that we will be working within. In different contexts there are different ways to specify the statistical model, but the goal is to find a suitable class of probability measures. We give four examples of different scientific problems and corresponding datasets, which give rise to four different ways of specifying a statistical model.

Example 3.1.1 (Neurons)**.** Neuron cells generate, as explained in Example 1.2.1, electrical signals, and an observable quantity is the time between such signals – the so-called interspike times. The activity of the neuron – that is, the rate by which the electrical signals are fired – depends on a range of different things including the input signals from other neuron cells, and the whole system is influenced by external stimuli. The activity of auditory neuron cells is, for instance, influenced by sound.

In this example we consider a dataset with 312 measurements of the interspike times in seconds for an auditory neuron cell. The measurements are taken in a steady state situation where there is no external stimuli to influence the activity of the cell.
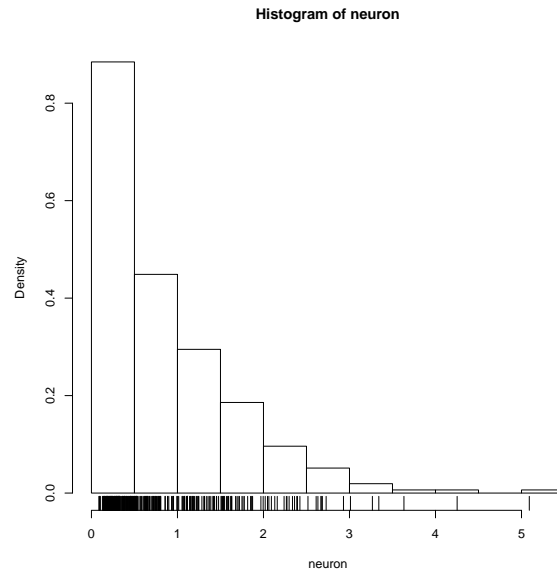
Figure 3.1: The histogram of the neuron interspike times.

Our model assumptions will be that the $n = 312$ measurements, $x_1, \ldots, x_{312}$, of interspike times are the realization of 312 iid positive, real valued random variables $X_1, \ldots, X_{312}$ and the objective is to estimate the common distribution of the interspike times.

Figure 3.1 shows a histogram of the data. The histogram resembles the density for the exponential distribution, so we will restrict our attention further to the model where the common distribution is an exponential distribution with density $\lambda \exp(-\lambda x)$ for $x \geq 0$ and an unknown parameter $\lambda > 0$.

We know that the theoretical mean of the exponential distribution with parameter $\lambda > 0$ is $1/\lambda$, and we can obtain an ad hoc estimate of λ by equating the theoretical mean equal to the empirical mean and solve for λ. That is, we get the estimate

$$\hat{\lambda} = \frac{312}{\sum_{i=1}^{312} x_i} = 1.147.$$

Figure 3.2 shows a QQ-plot of the observations against the quantiles for the exponential distribution, which seems to confirm that the exponential distribution is a suitable choice.

A more careful analysis will, however, reveal that there is a small problem. The problem can be observed by scrutinizing the rug plot on the histogram, which shows that there is a clear gap from 0 to the smallest observations. For the exponential distribution one will find a gap of roughly the same order of magnitude as the
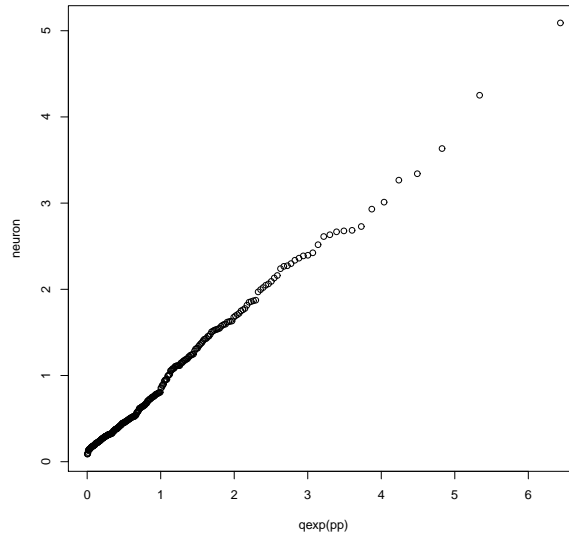
Figure 3.2: The QQ-plot of the neuron interspike times against the quantiles for the exponential distribution.

distances between the smallest observations – for the dataset the gap is considerably larger.

It is possible to take this into account in the model by including an extra, unknown parameter $\mu \geq 0$ such that $X_i - \mu$ is exponentially distributed with parameter $\lambda > 0$. This means that X_i has a distribution with density

$$f_{\mu,\lambda}(x) = \lambda \exp(-\lambda(x - \mu)), \quad \text{for } x \geq \mu.$$

The exponential distribution with parameter λ is a scale transformation with scale parameter $\sigma = 1/\lambda$ of the exponential distribution with $\lambda = 1$. The extended model above is seen to be a scale-location transformation with scale parameter $\sigma = 1/\lambda$ and location parameter μ. At this point we will not pursue any ad hoc methods for parameter estimation in the extended two-parameter model, but we will make two remarks. First, the points in the QQ-plot seem to fall on a straight line, which confirms the choice of the exponential distribution – but only up to a scale-location transformation. Second, a location transformation of the exponential distribution, or any other distribution living on the positive half-line for that matter, does not give a particularly pleasant model.

Let us elaborate. If $\mu > 0$ the model dictates that interspike times are all greater than μ. It is hardly conceivable that a single dataset can provide evidence for a fixed, absolute lower bound > 0 on the interspike times let alone provide a sufficiently good

estimate should it be the case that such a bound exists. It is well documented in the literature that after a spike there is a *refractory period* where the cell cannot fire, which can explain why we see the gap between 0 and the smallest observations. However, if we compute an estimate $\hat{\mu} > 0$ based on the dataset, we obtain a resulting model where a future observation of an interspike time smaller than $\hat{\mu}$ is in direct conflict with the model. Unless we have physiological substantive knowledge that supports an absolute lower bound we must be skeptical about an estimate. There is in fact evidence for a lower bound, whose value is of the order of one millisecond – too small to explain the gap seen in the data. In conclusion, it is desirable to construct a refined model, which can explain the gap and the exponential-like behavior without introducing an absolute lower bound.   ◇

Example 3.1.2 (Evolutionary Distance)**.** In this example we consider a dataset

$$(x_1, y_1), \ldots, (x_n, y_n)$$

consisting of n pairs of aligned nucleotides from two homologous DNA-sequences. The sample space we consider is therefore $E = \{A, C, G, T\} \times \{A, C, G, T\}$. We will assume that the only evolutionary event that occurs is substitution and that the pairs of aligned variables are realizations of n iid random variables $(X_1, Y_1), \ldots, (X_n, Y_n)$. In this case we need only to specify a family of probability measures on E.

We have in Section 2.10 introduced two examples of probability measures for such dependent pairs of nucleotides indexed by a time parameter; the Jukes-Cantor model and the Kimura model. One of the parameters that enter in all these models is the time parameter, and we will need to distinguish between two different situations. We may actually know the evolutionary distance in calendar time between the two homologous sequences, like for the Hepatitis C dataset considered in Example 1.2.4. This situation is encountered when we observe an evolving organism, like a virus, over time. In many applications, for instance in phylogenetics, where we want to trace back the evolution at the molecular level from present sequences, we do not know the evolutionary distance in calendar time. Indeed, one of the main objectives is in this case to *estimate* the time to a common forefather – at least relatively to other sequences. In such a case the time parameter enters on an equal footing with other unknown parameters.

If we consider the Jukes-Cantor model from Example 2.10.10, the conditional probabilities for the second (y) nucleotide given the first (x) are for $t, \alpha > 0$ given by the formulas

$$\begin{aligned} P^t(x, x) &= 0.25 + 0.75 \times \exp(-4\alpha t) \\ P^t(x, y) &= 0.25 - 0.25 \times \exp(-4\alpha t), \quad \text{if } x \neq y. \end{aligned}$$

We also need to specify the marginal distribution of the first x-nucleotide. We may here simply take an arbitrary probability measure on $\{A, C, G, T\}$ given by point probabilities $(p(A), p(C), p(G), p(T))$. This specifies the full distribution of (X_1, Y_1) by

the point probabilities

$$\mathbb{P}(X_1 = x, Y_1 = y) = p(x)P^t(x, y) = \begin{cases} p(x)(0.25 + 0.75 \times \exp(-4\alpha t)) & \text{if } x = y \\ p(x)(0.25 - 0.25 \times \exp(-4\alpha t)) & \text{if } x \neq y \end{cases}.$$

The point probabilities using the Kimura model from Example 3.3.18 are given the same way just by replacing the conditional probabilities by those from Example 3.3.18, which also introduces the extra parameter $\beta > 0$.

If we dwell for a moment on the interpretation of the Jukes-Cantor model, the conditional probability $P^t(x, x)$ is the probability that the a nucleotide does not change over the time period considered. For the hepatitis C virus dataset considered in Example 1.2.4, Table 1.2 shows for instance that out of the 2610 nucleotides in segment A there are 78 that have mutated over the period of 13 years leaving 2532 unchanged. With reference to the frequency interpretation we can try to estimate α in the Jukes-Cantor model by equation the formula for $P^t(x, x)$ equal to 2532/2610. This gives

$$\hat{\alpha} = -\frac{\log((\frac{2532}{2610} - \frac{1}{4})\frac{4}{3})}{4 \times 13} = 7.8 \times 10^{-4}.$$

$\diamond$

Example 3.1.3 (Forensic Statistics)**.** We want to construct a model of the occurrence of repeat counts for STRs, as used in forensics, in a given population; see Example 1.2.3. For simplicity, we assume first that we only consider a single short tandem repeat, which we take to be TH01. The fundamental repeat pattern for TH01 is `AATG`, and our experiment consists of obtaining a sample from the population and then counting the number of times this pattern is repeated at a particular position in the genome for the individuals in the sample. We have to remember that since humans are diploid organisms, we get a pair of counts for each tandem repeat. Since we cannot distinguish between which of the chromosomes the repeats are located on, the observations are naturally given as *unordered* pairs of counts. Observing the counts 7 and 8, say, from an individual is the same as observing the counts 8 and 7. We make the convention that we will report the counts with the smallest first. Our sample space is

$$E = \{(x, y) \mid x \leq y, \ x, y \in A\}$$

where A denotes m possible repeat counts observable for TH01. The situation is the same as with genes where we may have two alleles and we can observe only three allele combinations. Here the number of different alleles is m and we can observe $m(m + 1)/2$ different allele combinations. For TH01 we can take $A = \{5, 6, 7, 8, 9, 9.3, 10, 11\}$ for the purpose of this example. We note that 9.3 is also a possibility, which corresponds to a special variant of the repeat pattern with 9 full occurrences of the fundamental repeat pattern `AATG` but with the partial pattern `ATG` between the 6th and 7th occurrence.

		y							
		5	6	7	8	9	9.3	10	11
	5	0	0	1	0	0	0	0	0
	6		19	31	7	16	48	0	0
	7			10	9	11	43	0	0
x	8				3	9	17	3	0
	9					6	19	1	1
	9.3						47	1	0
	10							0	0
	11								0

Table 3.1: Tabulation of the repeat counts for the short tandem repeat TH01 for the Caucasian population included in the NIST STR dataset.

For the NIST dataset we can tabulate the occurrences of the different repeat counts for TH01. Table 3.1 shows the tabulation for the Caucasian population in the dataset. The dataset also includes data for Hispanics and African-Americans.

To build a model we regard the observations above as being realizations of iid random variables all with the same distribution as a pair (X, Y) that takes values in E. The full model – sometimes called the saturated model – is given as the family of all probability measures on E. The set of point probabilities

$$\Theta = \{(p(x,y))_{(x,y)\in E} \mid \sum_{x,y} p(x,y) = 1\}$$

on E provides a *parametrization* of all the probability measures. It is always possible in principle to work with the full model on a finite sample space – but the number of parameters easily grows out of control – there is $m(m+1)/2$ parameters in general and 36 for the TH01 case above[1]. One possible, simplifying assumption is that the distribution is given by the Hardy-Weinberg equilibrium. This assumptions enforces that there is a probability measure P_0 on A with point probabilities p_0 such that the point probabilities for the distribution of (X, Y) is

$$p(x,y) = \begin{cases} 2p_0(x)p_0(y) & \text{if } x \neq y \\ p_0(x)^2 & \text{if } x = y \end{cases}$$

This will bring us down to a model with only 8 parameters. It is an important question if such a simplifying assumption on the model is appropriate – or if the data at hand are actually in conflict with the model.

Another important question is what to do when we consider several short tandem repeats. If we in addition to TH01 consider TPOX, say, we can have another model for the marginal distribution of the repeat counts (Z, W). A table corresponding

---

[1]Since the point probabilities sum to 1 there are only m(m+1)/2-1 free parameters

		w							
		5	6	8	9	10	11	12	13
	5	0	0	1	0	0	0	0	0
	6		0	0	1	0	0	0	0
	8			82	45	20	78	15	0
z	9				3	2	16	2	0
	10					0	10	2	0
	11						19	5	0
	12							0	1
	13								0

Table 3.2: Tabulation of the repeat counts for the short tandem repeat TPOX for the Caucasian population included in the NIST STR dataset.

to the TPOX counts in the NIST dataset for the Caucasian population is found in table 3.2. The question is whether we can assume that (X, Y) and (Z, W) are independent? ◇

Example 3.1.4 (Gene expression)**.** We have in Example **??** considered data from the ALL gene expression dataset. There are thousands of genes whose expression levels are measured for different individuals giving rise to a huge dataset. Additional information about the individuals, who all suffered from leukemia, is also collected and one scientific objective is to identify genes for different groups of individuals that are systematically differently expressed between the two groups. One especially interesting grouping is to divide the individuals into two groups according to the BCR/ABL fusion gene. Some of the individuals have this fusion gene, which is a join of two genes – the BCR and ABL genes – which is a result of a large scale mutation. The gene is well known to be an oncogene, that is, a gene associated with cancer, and in this particular case with leukemia.

The objective is indeed a challenge. The technology behind gene expression data is described in more detail in Example 1.2.9 and there are several sources of variation that does not make it straight forward to detect such genes in the two groups by eye. In this example we will focus on a less ambitious goal, which is just to establish a suitable model of the gene expression measurement for a single gene. Our favorite gene is the one with the probe set label 1635_at. There is in total 79 individuals in the dataset, which we divide into two groups, which we call group 1 (BCR/ABL present) and group 2 (BCR/ABL not present) with 37 and 42 individuals in the two groups, respectively.

The basic model assumptions will be that the our observations,

$$x_{1,1}, \ldots, x_{1,37}$$
$$x_{2,1}, \ldots, x_{2,42}$$

are the realization of 79 independent random variables

$$X_{1,1}, \ldots, X_{1,37}$$
$$X_{2,1}, \ldots, X_{2,42}.$$

However, we will not insist that all the variables have the same distribution. On the contrary, the purpose is to figure out if there is a difference between the two groups. We will, however, assume that within either of the two groups the variables do have the same distribution.

The measurement of the gene expression level can be viewed as containing two components. One component is *the signal*, which is the deterministic component of the measurement, and the other component is *the noise*, which is a random component. In the microarray setup we measure a light intensity as a distorted representation of the true concentration of a given RNA-molecule in the cell. If we could get rid of all experimental uncertainty and, on top of that, the biological variation what would remain is the raw signal – an undistorted measurement of the expression level of a particular RNA-molecule under the given experimental circumstances. Even if the technical conditions were perfect giving rise to no noise at all, the biological variation would still remain. Indeed, this variation can also be regarded as part of the insight into and understanding of the function of the biological cell. The bottom line is that we have to deal with the signal as well as the noise. We are going to consider two modeling paradigms for this signal and noise point of view; the additive and the multiplicative noise model.

Let X denote a random variable representing the light intensity measurement of our favorite gene. The additive noise model says that

$$X = \mu + \sigma\varepsilon \tag{3.1}$$

where $\mu \in \mathbb{R}$, $\sigma > 0$ and ε is a real valued random variable whose distribution has mean value 0 and variance 1. This is simply a scale-location model and it follows from Example 2.9.9 that the mean value of the distribution of X is μ and the variance is σ^2. Thus μ is the mean expression of the gene – the signal – and $\sigma\varepsilon_i$ captures the noise that makes X fluctuate around μ. It is called the additive noise model for the simple fact that the noise is added to the expected value μ of X. It is quite common to assume in the additive noise model that the distribution of ε is $N(0,1)$. The model of X is thus $N(\mu, \sigma^2)$, where μ and σ are the *parameters*. The parameters (μ, σ) can take any value in the parameter space $\mathbb{R} \times (0, \infty)$.

For the multiplicative noise model we assume instead that

$$X = \mu\varepsilon \tag{3.2}$$

with ε a *positive* random variable. We will in this case insist upon $\mu > 0$ also. If the distribution of ε has mean value 1 and variance σ^2 it follows from Example 2.9.9 that the distribution of X has mean value μ and variance $\mu^2\sigma^2$. If we specify

the distribution of ε this is again a two-parameter model just as the additive noise model. The most important difference from the additive noise is that the standard deviation of X is $\mu\sigma$, which scales with the mean value μ. In words, the noise gets larger when the signal gets larger.

The formulation of the multiplicative noise model in terms of mean and variance is not so convenient. It is, however, easy to transform from the multiplicative noise model to the additive noise model by taking logarithms. Since $X = \mu\varepsilon$ then

$$\log X = \log \mu + \log \varepsilon,$$

which is an additive noise model for $\log X$. If one finds a multiplicative noise model most suitable it is quite common to transform the problem into an additive noise model by taking logarithms. Formally, there is a minor coherence problem because if ε has mean 1 then $\log \varepsilon$ does not have mean value 0 – it can be shown that it will always have mean value smaller than 0. If we on the other hand assume that $\log \varepsilon$ has mean value 0 it holds that ε will have mean value greater than 1. However, this does not interfere with the fact that if the multiplicative noise model is a suitable model for X then the additive noise model is suitable for $\log X$ and vice versa – one should just remember that if the mean of X is μ then the mean of $\log X$ is not equal to $\log \mu$ but it is in fact smaller than $\log \mu$.

Returning to our concrete dataset we choose to take logarithms (base 2), we assume an additive noise model for the logarithms with error distribution having mean 0 and variance 1, and we compute the ad hoc estimates of μ (the mean of $\log X$) and σ^2 (the variance of $\log X$) within the groups as the empirical means and the empirical variances.

group 1	group 2
$\hat{\mu}_1 = \frac{1}{37}\sum_{i=1}^{37}\log x_{1,i} = 8.54$	$\hat{\mu}_2 = \frac{1}{42}\sum_{i=1}^{42}\log x_{2,i} = 7.33$
$\tilde{\sigma}_1^2 = \frac{1}{37}\sum_{i=1}^{37}(\log x_{1,i} - \hat{\mu}_1)^2 = 0.659$	$\tilde{\sigma}_2^2 = \frac{1}{42}\sum_{i=1}^{42}(\log x_{2,i} - \hat{\mu}_2)^2 = 0.404$

To deal with our objective – is there a difference between the groups – it is of relevance to compute the difference of the estimated means

$$\hat{\mu}_1 - \hat{\mu}_2 = 1.20.$$

However, the size of this number does in reality not tell us anything. We have to interpret the number relatively to the size of the noise as estimated by $\tilde{\sigma}_1^2$ and $\tilde{\sigma}_2^2$. The formalities are pursued later.

One remark is appropriate. If we don't take logarithms the empirical means

$$\tilde{\mu}_1 = \frac{1}{37}\sum_{i=1}^{37} x_{1,i} = 426.6, \quad \tilde{\mu}_2 = \frac{1}{42}\sum_{i=1}^{42} x_{2,i} = 179.1$$

are reasonable estimates of the means (of X) within the two groups. For the multiplicative mode it is best to compute the ratio

$$\frac{\tilde{\mu}_1}{\tilde{\mu}_2} = 2.38$$

for comparing the mean values instead of the difference. Again any interpretation of this number needs to be relative to the size of the noise. The downside of using ad hoc estimates and ad hoc approaches shows it face here, because $\hat{\mu}_1 \neq \log \tilde{\mu}_1$ we have two procedures. Either we take logarithms, compute the means and compute their difference or we compute the means, take logarithms and compute their difference and we end up with two different results. Which is the most appropriate? The model based, likelihood methods introduced later will resolve the problem. $\diamond$

All four examples above specify a family of probability distributions on the relevant sample space by a set of *parameters*. For the exponential distribution used as a model of the interspike times the parameter is the intensity parameter $\lambda > 0$. For the additive model (or multiplicative in the log-transformed disguise) the parameters for the gene expression level are $\mu \in \mathbb{R}$ and $\sigma > 0$. For the model in the forensic repeat counting example the parameters are the point probabilities on the finite sample space E, or if we assume Hardy-Weinberg equilibrium the point probabilities on the set A. For the Jukes-Cantor model the parameters are the marginal point probabilities for the first nucleotide and then $\alpha > 0$ and time $t > 0$ that determines the (mutation) dependencies between the two nucleotides.

Three of the examples discussed above are all examples of *phenomenological* models. By this we mean models that attempt to *describe* the observed phenomena (the empirical data) and perhaps *relate* observables. For the neuronal interspike times we describe the distribution of interspike times. For the gene expression data we relate the expression level to presence or absence of the BCR/ABL fusion gene, and for the evolution of molecular sequences we relate the mutations to the time between observations. Good phenomenological modeling involves a interplay between the modeling step, the data analysis step and the subject matter field. We do not, however, attempt to explain or derive the models completely from fundamental theory. It seems that once we step away from the most fundamental models of physics there is either no derivable, complete theory relating observables of interest or it is mathematically impossible to derive the exact model. The boundaries between theory, approximations and phenomenology are, however, blurred.

Most statistical models are phenomenological of nature – after all, the whole purpose of using a probabilistic model is to capture randomness, or uncontrollable variation, which by nature is difficult to derive from theory. Often the need for a probabilistic model is due to the fact that our knowledge about the quantities we model is limited. A few classical probabilistic models are, however, derived from fundamental *sampling principles*. We derive some of these models in the next section.

3.2 Classical sampling distributions

The classical probability models, the binomial, multinomial, geometric and hyper-geometric distributions, that we consider in this section are all distributions on the positive integers that arise by *sampling and counting*, and which can be understood naturally in terms of transformations of an underlying probability measure.

Example 3.2.1 (Binomial distribution)**.** Let $X_1, \ldots, X_n$ denote n iid Bernoulli variables with success parameter $p \in [0, 1]$. The fundamental sample space is $E = \{0, 1\}$, the bundled variable $X = (X_1, \ldots, X_n)$ takes values in $\{0, 1\}^n$, and the distribution of X is

$$P(X = x) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\sum_{i=1}^{n} x_i}(1-p)^{n - \sum_{i=1}^{n} x_i},$$

where $x = (x_1, \ldots, x_n)$. Let

$$h : E \to E' = \{0, 1, \ldots, n\}$$

be given as

$$h(x_1, \ldots, x_n) = \sum_{i=1}^{n} x_i.$$

The distribution of

$$Y = h(X) = \sum_{i=1}^{n} X_i$$

is called the *binomial distribution* with probability parameter p and size parameter n. We use the notation $Y \sim \text{Bin}(n, p)$ to denote that Y is binomially distributed with size parameter n and probability parameter p. We find the point probabilities of the distribution of $h(X)$ as follows: For any vector $x = (x_1, \ldots, x_n)$ with $\sum_{i=1}^{n} x_i = k$ it follows that

$$\mathbb{P}(X = x) = p^k(1-p)^{n-k}.$$

Thus all outcomes that result in the same value of $h(X)$ are equally probable. So

$$\mathbb{P}(h(X) = k) = \sum_{x:h(x)=k} P(X = x) = \binom{n}{k} p^k(1-p)^{n-k}$$

where $\binom{n}{k}$ denotes the number of elements $x \in E$ with $h(x) = \sum_{i=1}^{n} x_i = k$. From Section B.2 we get that

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}.$$

$\diamond$

Example 3.2.2 (Multinomial distribution)**.** If our fundamental sample space contains m elements instead of just 2 we can derive an m-dimensional version of the binomial distribution. We will assume that the fundamental sample space is $\{1, 2, \ldots, m\}$

but the m elements could be labeled any way we like. Let $X_1, \ldots, X_n$ be n iid random variables, let $X = (X_1, \ldots, X_n)$ denote the bundled variable with outcome in $E = \{1, 2, \ldots, m\}^n$, and consider the transformation

$$h : E \to E' = \{0, 1, 2, \ldots, n\}^m$$

defined by

$$h(x) = (h_1(x), \ldots, h_m(x)), \quad h_j(x) = \sum_{i=1}^{n} 1(x_i = j).$$

That is, $h_j(x)$ counts how many times the outcome j occurs. The distribution of

$$h(X) = (h_1(X), \ldots, h_m(X))$$

is called the *multinomial distribution* with probability parameter $p = (p_1, \ldots, p_m)$. We observe that for $x \in E^n$ with $h(x) = (k_1, \ldots, k_m)$ then

$$\mathbb{P}(X = x) = p_1^{k_1} p_2^{k_2} \ldots p_m^{k_m},$$

hence

$$\mathbb{P}(h(X) = (k_1, \ldots, k_m)) = \sum_{x : h(x) = (k_1, \ldots, k_m)} P(X = x) = \binom{n}{k_1 \ldots k_m} p_1^{k_1} p_2^{k_2} \ldots p_m^{k_m},$$

where $\binom{n}{k_1 \ldots k_m}$ denotes the number of ways to label n elements such that k_1 are labeled 1, k_2 are labeled 2, etc. From Section B.2 we get that

$$\binom{n}{k} = \frac{n!}{k_1! k_2! \cdots k_m!}.$$

$\diamond$

Example 3.2.3 (Geometric distribution)**.** Let X denote a Bernoulli variable taking values in $E = \{0, 1\}$ such that $P(X = 1) = p$. We can think of X representing the outcome of flipping a (skewed) coin with 1 representing heads (success). We then let $X_1, X_2, X_3, \ldots$ be independent random variables each with the same distribution as X corresponding to repeating coin flips independently and indefinitely. Let T denote the first throw where we get heads, that is $X_1 = X_2 = \ldots = X_{T-1} = 0$ and $X_T = 1$. Due to independence the probability of $T = t$ is $p(1-p)^{t-1}$. If we introduce the random variable

$$Y = T - 1,$$

which is the number of tails (failures) we get before we get the first head (success) we see that

$$P(Y = k) = p(1-p)^k.$$

This distribution of Y with point probabilities $p(k) = p(1-p)^k$, $k \in \mathbb{N}_0$, is called the *geometric distribution* with *success probability* $p \in [0, 1]$. It is a probability distribution on the non-negative integers $\mathbb{N}_0$. $\diamond$

The distributions derived above can be interpreted as sampling distributions in the following sense. If we have a population of individuals that have one of two properties (ill or healthy, male or female, liberal or conservative) and we sample n individuals *completely at random* meaning that we sample from the uniform distribution on the population n times *with replacement*, the number of individuals among the n sampled individuals with property one (being ill, say) follows a binomial distribution $B(n, p)$ with p being the fraction of individuals with property one. And the number of times we have to sample before we get the first individual with property one is a geometric distribution. If there are more than two properties we get the multinomial distribution instead of the binomial distribution. We might for instance divide people according to a more refined view on their political observance – as socialists, liberals and conservatives, say – giving rise to a multinomial distribution with three property labels.

In some cases with finite population sampling the samples are taken *without replacement*, which result in a different distribution.

Example 3.2.4 (Hypergeometric distribution)**.** We consider here the situation where we sample individuals or objects from a finite population, and where these individuals can have a finite number of different *properties*, which we for convenience label $\{1, \ldots, m\}$. The actual sample space E consists of the entire population and each individual has precisely one of the properties. Let N_j denote the total number of individuals with property j that are in E – with $j \in \{1, \ldots, m\}$ – and

$$N = \sum_{j=1}^{m} N_j$$

is then the total number of individuals in the population. If we independently sample $X_1, \ldots, X_n$ from E completely at random *without replacement* the probability of getting $X_1 = x_1$ the first time is $1/N$ and the probability for getting $X_2 = x_2 \neq x_1$ the second time is $1/(N-1)$ etc. Since we assume that we make independent samples the probability of getting $x_1, \ldots, x_n \in E$ in n samples is

$$p(x_1, \ldots, x_n) \;=\; \frac{1}{N(N-1)(N-2) \ldots (N-n)} = \frac{1}{N^{(n)}}.$$

We define the transformation

$$h : E^n \to \{0, 1, \ldots, n\}^m$$

given by

$$h(x) = (h_1(x), \ldots, h_m(x)), \quad h_j(x) = \sum_{i=1}^{n} 1(\text{property}(x_i) = j)$$

for $x = (x_1, \ldots, x_n) \in E^n$. Thus $h_j(x)$ is the number of elements in $x = (x_1, \ldots, x_n)$ that have property j. The distribution of $h(X_1, \ldots, X_n)$ is the *hypergeometric distribution*. Since the point probabilities for all $(x_1, \ldots, x_n)$ are $1/N^{(n)}$ the transformation

problem is, like for the binomial and multinomial distribution, a combinatorial problem. How many outcomes $(x_1, \ldots, x_n)$ in E^n result in $h(x_1, \ldots, x_n) = (k_1, \ldots, k_m)$ where $k_1 \leq N_1, \ldots, k_m \leq N_m$ and $k_1 + \ldots k_m = n$? As derived in Section B.2 there are

$$\binom{N_j}{k_j}$$

different ways to pick out k_j different elements in E with property j. Thus there are

$$\binom{N_1}{k_1}\binom{N_2}{k_2}\cdots\binom{N_m}{k_m}$$

ways to pick out k_1 elements with property 1, k_2 with property 2 etc. There are, moreover, $n!$ different ways to order these variables, and we find that the point probabilities for the hypergeometric distribution are given by

$$
\begin{aligned}
\mathbb{P}(h(X) = (k_1, \ldots, k_m)) &= \frac{\text{number of } (x_1, \ldots, x_n) \text{ with } h(x_1, \ldots, x_n) = (k_1, \ldots, k_m)}{N^{(n)}} \\
&= \frac{n!\binom{N_1}{k_1}\binom{N_2}{k_2}\cdots\binom{N_m}{k_m}}{N^{(n)}} \\
&= \frac{\binom{N_1}{k_1}\binom{N_2}{k_2}\cdots\binom{N_m}{k_m}}{\binom{N}{n}} \\
&= \binom{n}{k_1 \ldots k_m}\frac{N_1^{(k_1)}N_2^{(k_2)}\cdots N_m^{(k_m)}}{N^{(n)}}.
\end{aligned}
$$

Here the second equality follows by the definition of the binomial coefficients and the third by the definition of the multinomial coefficients.

The last formula above for the point probabilities looks similar to the formula for the multinomial distribution. Indeed, for $N_1, \ldots, N_m$ sufficiently large compared to n it holds that

$$\frac{N_1^{(k_1)}N_2^{(k_2)}\cdots N_m^{(k_m)}}{N^{(n)}} \simeq \left(\frac{N_1}{N}\right)^{k_1}\left(\frac{N_2}{N}\right)^{k_2}\cdots\left(\frac{N_m}{N}\right)^{k_m}$$

and the hypergeometric distribution is well approximated by the multinomial distribution with $p_j = N_j/N$. This makes a lot of sense intuitively. If we sample *with replacement* the variables $X_1, \ldots, X_n$ are iid with

$$\mathbb{P}(\text{property}(X_i) = j) = N_j/N.$$

If $N_1, \ldots, N_m$ are large compared to n it hardly makes a difference whether we sample with or without replacement, and in such cases we will usually prefer to work with the multinomial distribution. $\diamond$

Exercises

Exercise 3.2.1. Consider as in Example 3.2.3 the iid Bernoulli variables $X_1, X_2, X_3, \ldots$ with success probability p. Denote by T the number of variables before the n'th 1 occurs. That is, $X_T = 1$ and the number of 1's among the first variables $X_1, \ldots, X_{T-1}$ is precisely $n - 1$. Show that the number of 0's, $Y = T - n$, before the n'th 1 has point probabilities

$$\mathbb{P}(Y = k) = \binom{k + n - 1}{k} p^n (1 - p)^k$$

for $k \in \mathbb{N}_0$. The distribution is known as the *negative binomial distribution.*

3.3 Statistical Inference

3.3.1 Parametric Statistical Models

In the previous sections we have introduced various examples of *parametrized families* $(P_\theta)_{\theta \in \Theta}$ of probability measures on a sample space E. The set Θ is called the parameter space and θ the parameter. These are the abstract notations that we will use. The interpretation is that we consider an experiment taking values in E and the measures P_θ for $\theta \in \Theta$ as candidate models of the outcome of the experiment. We just don't know the right one. Based on a sample $x \in E$ we are trying to figure out which of the measures are good candidates for having produced x. In many cases the sample x is actually a realization $x_1, \ldots, x_n$ of n independent (and perhaps identically distributed) random variables $X_1, \ldots, X_n$ taking values in a sample space E_0. In this case $E = E_0^n$ and P_θ is given in terms of the marginal distributions of the X_i's.

Example 3.3.1 (Multinomial model)**.** Let $E_0 = \{1, 2, \ldots, m\}$, let $\theta = (p(1), \ldots, p(m))$ denote a vector of point probabilities on E_0 and let

$$\Theta = \left\{ (p_1, \ldots, p_m) \mid p(i) \geq 0, \sum_{i=1}^m p_i = 1 \right\}.$$

If P_θ denotes the distribution of n iid random variables $X_1, \ldots, X_n$, with each of them taking values in E_0 and having distribution with point probabilities θ, then $(P_\theta)_{\theta \in \Theta}$ is a statistical model on $E = E_0^n$. Note that this statistical model is parametrized by *all* probability measures on E_0 in terms of the point probabilities θ. From Example 3.2.2 we observe that by defining

$$N_j = \sum_{i=1}^n 1(X_i = j)$$

the distribution of the vector

$$(N_1, \ldots, N_m)$$

is the multinomial distribution with probability parameters $\theta = (p_1, \ldots, p_m)$ when the distribution of $X_1, \ldots, X_n$ is P_θ. The model is known as the multinomial model.

<div align="right">◇</div>

Example 3.3.2. Let $E_0 = [0, \infty)$, let $\lambda \in (0, \infty)$, and let P_λ be the distribution of n iid exponentially distributed random variables $X_1, \ldots, X_n$ with intensity parameter λ, that is, the distribution of X_i has density

$$f_\lambda(x) = \lambda \exp(-\lambda x)$$

for $x \geq 0$. With $\Theta = (0, \infty)$ the family $(P_\lambda)_{\lambda \in \Theta}$ of probability measures is a statistical model on $E = E_0^n = (0, \infty)^n$.

Note that we most often use other names than θ for the concrete parameters in concrete models. For the exponential distribution we usually call the parameter λ.

<div align="right">◇</div>

As the two previous examples illustrate the parameter space can take quite different shapes depending on the distributions that we want to model. It is, however, commonly the case that $\Theta \subseteq \mathbb{R}^d$ for some d. In the examples $d = m$ and $d = 1$ respectively.

As mentioned, we search for probability measures in the statistical model that fit a dataset well. The purpose of estimation is to find a *single* $\hat{\vartheta} \in \Theta$, an estimate of θ, such that $P_{\hat{\vartheta}}$ is "the best" candidate for having produced the observed dataset. We have up to this point encountered several ad hoc procedures for computing estimates for the unknown parameter. We will in the following formalize the concept of an estimator as the procedure for computing estimates and introduce a systematic methodology – the *maximum likelihood method*. The maximum likelihood method will in many cases provide quite reasonable estimates and it is straight forward to implement the method, since the likelihood function that we will maximize is defined directly in terms of the statistical model. Moreover, the maximum likelihood method is the de facto standard method for producing estimates in a wide range of models.

3.3.2 Estimators and Estimates

As indicated by the title of this section there is a distinction between an estimator and an estimate.

Definition 3.3.3. *An estimator is a map*

$$\hat{\theta} : E \to \Theta.$$

For a given observation $x \in E$ the value of $\hat{\theta}$ at x,

$$\hat{\vartheta} = \hat{\theta}(x),$$

is called the estimate of θ.

The estimator is a map from E into Θ. It is the procedure by which we compute the estimate once we have data. This implies that if X is a random variable taking values in E and having distribution P_θ, then we can consider the transformed random variable $\hat{\theta}(X)$. This random variable is usually also called $\hat{\theta}$, thus $\hat{\theta}$ means a map *as well as* a random variable and in both cases we refer to $\hat{\theta}$ as an estimator. When we regard $\hat{\theta}$ as a random variable the estimate $\hat{\vartheta} = \hat{\theta}(x)$ becomes the realization of $\hat{\theta}$ when x is the realization of X. When $E = E_0^n$ the observation $x \in E$ is a vector of realizations $x_1, \ldots, x_n$ of the random variables $X_1, \ldots, X_n$. The estimator is then the transformation $\hat{\theta}(X_1, \ldots, X_n)$ of the n random variables and the estimate is the realization $\hat{\vartheta} = \hat{\theta}(x_1, \ldots, x_n)$.

When we view $\hat{\theta}$ as a random variable it has a distribution. Finding and understanding this distribution is an important problem, since it is the distribution of the estimator that determines the properties of the estimator. In fact, a property of an estimator is always going to mean a property about the distribution of that estimator. Note that the distribution of $\hat{\theta}$ depends upon which measure P_θ we consider. A good estimator is loosely speaking an estimator whose distribution under P_θ is concentrated around θ for all $\theta \in \Theta$.

Before we can try to identify a single parameter in Θ it is necessary to deal with the following identifiability problem.

Definition 3.3.4. *The parameter θ is said to be identifiable if the map $\theta \mapsto P_\theta$ is one-to-one. That is, for two different parameters θ_1 and θ_2 the corresponding measures P_{θ_1} and P_{θ_2} differ.*

If we don't have identifiability we have no chance of distinguishing between parameters for which the corresponding probability measures are identical. This can be a serious problem, which should be carefully investigated for every model under consideration.

Example 3.3.5. If P_λ denotes the exponential distribution with $\lambda \in (0, \infty)$ the intensity parameter, then λ is identifiable since for $\lambda_1 \neq \lambda_2$ we have $\exp(-\lambda_1 x) \neq \exp(-\lambda_2 x)$ for all $x > 0$, thus the two corresponding distribution functions differ.

We can choose another way to parametrize the exponential distribution, where we let $\Theta = \mathbb{R}\backslash\{0\}$ and P_λ be the exponential distribution with intensity parameter λ^2. That is, the density for P_λ is

$$f_\lambda(x) = \lambda^2 \exp(-\lambda^2 x)$$

for $\lambda \in \mathbb{R}$. We observe that λ and $-\lambda$ give rise to the same probability measure namely the exponential distribution with intensity parameter λ^2. $\diamond$

The example above illustrates the problem with non-identifiability of the parameter. We will never be able to say whether λ or $-\lambda$ is the "true" parameter since they both give rise to the same probability measure. The example is on the other hand

a little stupid because we would probably not choose such a silly parametrization. But in slightly more complicated examples it is not always so easy to tell whether the parametrization we choose makes the parameter identifiable. In some cases the natural way to parametrize the statistical model leads to non-identifiability.

Example 3.3.6 (Jukes-Cantor). If we consider the Jukes-Cantor model from Example 3.1.2 and take both $\alpha > 0$ and $t > 0$ as unknown parameters, then we have non-identifiability. This is easy to see, as the two parameters always enter the conditional probabilities via the product αt. So $\kappa \alpha$ and t/κ gives the same conditional probabilities for all $\kappa > 0$.

This does make good sense intuitively since t is the time parameter (calendar time) and α is the mutation intensity (the rate by which mutations occur). A model with large t and small α is of course equivalent to a model where we scale the time down and the intensity up.

If we fix either of the parameters we get identifiability back, so there is a hope that we can estimate α from a present, observable process where we know t, and then with fixed α we may turn everything upside down and try to estimate t. This argument is based on the "molecular clock", that is, that the mutation rate is constant over time. $\diamond$

Example 3.3.7 (ANOVA). Consider the additive noise model

$$X_i = \mu_i + \sigma_i \varepsilon_i$$

for the vector $X = (X_1, \ldots, X_n)$ of random variables where $\varepsilon_i \sim N(0,1)$ are iid. The sample space is in this case $E = \mathbb{R}^n$. We regard the scale parameters $\sigma_i > 0$ as known and fixed. The parameter space for this model is then $\Theta = \mathbb{R}^n$ and

$$\theta = (\mu_1, \ldots, \mu_n).$$

The parameter is identifiable.

In Example 3.1.4 we considered a situation where X_i was the log-expression level for a gene, and where we had two groups. This grouping can be coded by a *factor*, which is a map from the index set

$$g : \{1, \ldots, n\} \to \{1, 2\}$$

such that $g(i) = 1$ denotes that the i'th observation belongs to group 1 and $g(i) = 2$ otherwise. The model we considered in Example 3.1.4 correspond to saying that

$$\mu_i = \alpha_{g(i)}$$

where $(\alpha_1, \alpha_2) \in \mathbb{R}^2$. We have reduced the parameter space to a 2-dimensional parameter space.

The grouping indicates whether a particular fusion gene was present or absent for the individual. There are, however, other potentially important groupings in the

dataset. It is for instance also known whether the individual is a male or a female. If we define another factor

$$b : \{1, \ldots, n\} \rightarrow \{1, 2\}$$

such that $b(i) = 1$ if the i'th individual is a female and $b(i) = 2$ otherwise we can consider the model

$$\mu_i = \alpha_{g(i)} + \beta_{b(i)}$$

where μ_i is broken down into the addition of an $\alpha_{g(i)}$-component and a $\beta_{b(i)}$-component where $(\alpha_1, \alpha_2) \in \mathbb{R}^2$ as above and similarly $(\beta_1, \beta_2) \in \mathbb{R}^2$. The parameter space is

$$\Theta = \mathbb{R}^4$$

and the parameter is $\theta = (\alpha_1, \alpha_2, \beta_1, \beta_2)$. This is a convenient and intuitive way to parametrize the model, but the parameter for this model is *not* identifiable as given here. This is because for all $\kappa \in \mathbb{R}$

$$\begin{aligned} \mu_i &= \alpha_{g(i)} + \beta_{b(i)} \\ &= (\alpha_{g(i)} + \kappa) + (\beta_{b(i)} - \kappa). \end{aligned}$$

We can make the additional restriction that

$$\beta_1 + \beta_2 = 0$$

and then the parameter becomes identifiable. This restriction effectively reduces the number of free parameters by 1 and gives a model with only 3 free parameters. An alternative, identifiable parametrization is given by $(\alpha, \beta, \gamma) \in \mathbb{R}^3$ and

$$\mu_i = \gamma + \alpha 1(g(i) = 1) + \beta 1(b(i) = 1).$$

The interpretation of this model is simple. There is a common mean value γ and then if the fusion gene is present the level changes with α and if the individual is female the level changes with β.

The additive model above should be compared with a (fully identifiable) model with four free parameters:

$$\mu_i = \nu_{g(i),b(i)}$$

where the parameter is $(\nu_{1,1}, \nu_{1,2}, \nu_{2,1}, \nu_{2,2}) \in \mathbb{R}^4$. $\diamond$

The estimation of the unknown parameter for a given statistical model only makes sense if the parameter is identifiable. If we try to estimate the parameter anyway and come up with an estimate $\hat{\vartheta}$, then the estimate itself does not have any interpretation. Only the corresponding probability measure $P_{\hat{\vartheta}}$ can be interpreted as an approximation of the true probability measure. Without identifiability, we cannot really define or discuss properties of the estimators directly in terms of the parameter. One will need to either reparameterize the model or discuss only aspects/transformations of

the parameter that are identifiable. Identifiability is on the other hand often a messy mathematical problem. In subsequent examples we will only discuss identifiability issues if there are problems with the parameter being identifiable.

Though identifiability is important for interpretations, probability models with non-identifiable parametrization and corresponding estimators are routinely studied and applied to certain tasks. One example is the neural networks or more generally many so-called machine learning models and techniques. The primary purpose of such models is to work as approximation models primarily for prediction purposes – often referred to as black box prediction. The estimated parameters that enter in the models are not in themselves interpretable.

3.3.3   Maximum Likelihood Estimation

We remind the reader about two ways of representing a probability measure on E. If E is discrete we can give a probability measure in terms of point probabilities, thus if P_θ for $\theta \in \Theta$ is a parametrized family of probability measures on a discrete sample space E, P_θ is defined by the point probabilities $p_\theta(x)$ for $x \in E$. Likewise, if $E \subseteq \mathbb{R}^n$ for some n we can give a parametrized family P_θ, $\theta \in \Theta$, of probability measures on E by densities $f_\theta(x)$ for $x \in E$.

Definition 3.3.8. *Assume that $x \in E$ is a given observation. The likelihood function is the function*

$$\mathscr{L}_x : \Theta \to \mathbb{R}$$

defined as follows: If E is discrete and P_θ has point probabilities $(p_\theta(x))_{x \in E}$ we define

$$\mathscr{L}_x(\theta) = p_\theta(x). \tag{3.3}$$

If $E \subseteq \mathbb{R}^n$ and P_θ has density $f_\theta : E \to [0, \infty)$ we define

$$\mathscr{L}_x(\theta) = f_\theta(x). \tag{3.4}$$

We have two different but similar definitions of the likelihood function depending on whether E is a discrete or a continuous sample space. In fact, there is a more abstract framework, *measure and integration theory*, where these two definitions are special cases of a single unifying definition. It is on the other hand clear that despite the two different definitions, $\mathscr{L}_x(\theta)$ is for both definitions a quantification of how likely the observed value of x is under the probability measure P_θ.

We will often work with the minus-log-likelihood function

$$l_x(\theta) = -\log \mathscr{L}_x(\theta) \tag{3.5}$$

instead of the likelihood function itself. There are several reasons for this. For practical applications the most notable reason is that the likelihood function often turns

out to be a product of a large number of probabilities (or densities). Since probabilities are less than 1 such a product becomes very small and the risk of running into *underflow* problems on computers becomes substantial. Taking the logarithm turns products into sums and the problem essentially disappears. From a theoretical point of view the minus-log-likelihood function is simply nicer to work with for a number of standard models, and it also plays an important role in the further analysis of the maximum likelihood estimator to be defined below.

Definition 3.3.9 (Tentative)**.** *The maximum likelihood estimator, abbreviated MLE, is the function*

$$\hat{\theta} : E \to \Theta$$

such that $\hat{\theta}(x)$ is the value of θ at which the likelihood function attains its global maximum (the minus-log-likelihood function attains its global minimum). We write

$$\hat{\theta}(x) = \operatorname*{argmax}_{\theta \in \Theta} \mathscr{L}_x(\theta) = \operatorname*{argmin}_{\theta \in \Theta} l_x(\theta). \tag{3.6}$$

The definition is tentative because there are a number of problems that we need to take into account. In a perfect world Definition 3.3.9 would work, but in reality the likelihood function $\mathscr{L}_x$ may not attain a global maximum (in which case $\hat{\theta}(x)$ is not defined) and there may be several θ's at which $\mathscr{L}_x$ attains its global maximum (in which case the choice of $\hat{\theta}(x)$ is ambiguous). The problem with non-uniqueness of the global maximum is not a real problem. The real problem is that there may be several *local* maxima and when searching for the global maximum we are often only able to justify that we have found a local maximum. The problem with non-existence of the a global maximum is also a real problem. In some situations there exists a unique $\hat{\theta}(x)$ for all $x \in E$ such that

$$\mathscr{L}_x(\theta) < \mathscr{L}_x(\hat{\theta}(x)) \quad \text{for all } \theta \neq \hat{\theta}(x),$$

but quite frequently there is only a subset $A \subseteq E$ such that there exists a unique $\hat{\theta}(x)$ for all $x \in A$ with

$$\mathscr{L}_x(\theta) < \mathscr{L}_x(\hat{\theta}(x)) \quad \text{for all } \theta \neq \hat{\theta}(x).$$

If $\mathbb{P}(X \in A^c)$ is small we can with high probability maximize the likelihood function to obtain an estimate. We just have to remember that under some unusual circumstances we cannot. When studying the properties of the resulting estimator we therefore need to consider two things: (i) how $\hat{\theta}(x)$ behaves on the set A where it is defined, and (ii) how probable it is that $\hat{\theta}(x)$ is defined.

If the parameter space is continuous we can use calculus to find the MLE analytically. We differentiate the (minus-log-) likelihood function w.r.t. the parameter and try to find stationary points, that is, θ's where the derivative is zero. If $\Theta = \mathbb{R}$ or an open interval and l_x is twice differentiable it holds that $\tilde{\theta} \in \Theta$ is a *local* minimizer for

$l_x(\theta)$ if (and only if)

$$\frac{\mathrm{d}l_x}{\mathrm{d}\theta}(\tilde{\theta}) \;=\; 0 \tag{3.7}$$

$$\frac{\mathrm{d}^2 l_x}{\mathrm{d}\theta^2}(\tilde{\theta}) \;>\; 0. \tag{3.8}$$

From this we can conclude that if there is a *unique* solution $\tilde{\theta} \in \Theta$ to (3.7) that also fulfills (3.8) then $\hat{\theta}(x) = \tilde{\theta}$. To see way, assume that there is a $\theta_0 \in \Theta$ such that $l_x(\theta_0) < l_x(\tilde{\theta})$. The second condition, (3.8), assures that the derivative is > 0 for $\theta \in (\tilde{\theta}, \theta_1)$, say, and if $\theta_0 > \tilde{\theta}$ there is a $\theta_2 \in (\theta_1, \theta_0)$ where the derivative is $<$ 0. Somewhere in (θ_1, θ_2) the derivative must take the value 0 – a local maximum – which contradicts that $\tilde{\theta}$ is the unique solution to (3.7).

The equation

$$\frac{\mathrm{d}l_x}{\mathrm{d}\theta}(\theta) = 0$$

is known as the *likelihood equation*, and if the MLE, $\hat{\theta}(x)$, exists it must satisfy this equation[2]. So as a starting point one can always try to solve this equation to find a candidate for the MLE. If $\Theta \subseteq \mathbb{R}^d$ is a multidimensional parameter space there exists a similar approach, see Math Box 3.3.1, in which case the likelihood equation becomes a set of equations. Having found a candidate for $\hat{\theta}(x)$ that solves the likelihood equation(s), there can, however, be substantial problems when $d \geq 2$ with ensuring that the solution is a global maximum. There are some special classes of models where it is possible to show that when there is a solution to the likelihood equation then it is the global minimizer for l_x and hence the MLE. Typically such arguments rely on *convexity* properties of l_x. For many other models there are no such result and it is hardly ever possible to solve the likelihood equations analytically. One therefore needs to rely on numerical methods. To this end there are several ways to proceed. The *Newton-Raphson* algorithm solves the likelihood equation numerically whereas the *gradient descent* algorithm and its variants are direct numerical minimization algorithms of l_x. The Newton-Raphson algorithm requires that we can compute the second derivative of l_x whereas gradient descent requires only the first derivative. See *Numerical Optimization* by Jorge Nocedal and Stephen J. Wright for a authoritative treatment of optimization methods.

These algorithms are general purpose optimization algorithms that do not take into account that we try to maximize a likelihood function. In special cases some specific algorithms exists for maximizing the likelihood, most notably there are a number of models where one will encounter the so-called EM-algorithm. It is possible that the reader of these notes will never have to actually implement a numerical optimization algorithm, and that is not the subject of these notes anyway. A fairly large and ever growing number of models are available either in R or via alternative programs. Multivariate numerical optimization is a specialist's job! It is good to know, though,

---

[2] except in cases where a global maximum is attained at the boundary of the parameter set Θ

Algorithm 3.3.1 (Newton-Raphson)**.** We consider the case $\Theta \subseteq \mathbb{R}$ and want to solve the likelihood equation

$$\frac{\mathrm{d}l_x}{\mathrm{d}\theta}(\theta) = 0.$$

A first order Taylor expansion of the derivative of l_x from θ_0 amounts to

$$\frac{\mathrm{d}l_x}{\mathrm{d}\theta}(\theta) \simeq \frac{\mathrm{d}l_x}{\mathrm{d}\theta}(\theta_0) + \frac{\mathrm{d}^2 l_x}{\mathrm{d}^2\theta}(\theta_0)(\theta - \theta_0).$$

For a given $\theta_0 \in \Theta$ we solve the *linear* equation

$$\frac{\mathrm{d}l_x}{\mathrm{d}\theta}(\theta_0) + \frac{\mathrm{d}^2 l_x}{\mathrm{d}^2\theta}(\theta_0)(\theta - \theta_0) = 0$$

instead of the likelihood equation. The solution is

$$\theta_1 = \theta_0 - \left(\frac{\mathrm{d}^2 l_x}{\mathrm{d}^2\theta}(\theta_0) \right)^{-1} \frac{\mathrm{d}l_x}{\mathrm{d}\theta}(\theta_0).$$

Taylor expanding from θ_1 instead and solving the corresponding linear equation leads to a θ_2 and we can iteratively continue this procedure, which result in a sequence $(\theta_n)_{n \geq 0}$ defined by

$$\theta_n = \theta_{n-1} - \left(\frac{\mathrm{d}^2 l_x}{\mathrm{d}^2\theta}(\theta_{n-1}) \right)^{-1} \frac{\mathrm{d}l_x}{\mathrm{d}\theta}(\theta_{n-1}).$$

If the *initial guess* θ_0 is sufficiently close to $\hat{\vartheta}$ (the MLE) then θ_n converges rapidly towards $\hat{\vartheta}$. Note that a *fixed point* of the algorithm, that is, a θ such that if we put in θ in the formula above we get back θ, is a solution to the likelihood equation.

what kind of optimization is going on when one computes the MLE in practice, and it is also good to know that all numerical algorithms, whether they are general purpose or problem specific, typically rely on an initial guess θ_0 of the parameter. The algorithm will then iteratively "improve" upon the guess. It is good practice to choose a number of different initial guesses or come up with qualified initial guesses to prevent that the algorithm either diverges or converges towards a wrong *local* minimum close to a "bad" initial guess. Note that there is no way to assure in general that an algorithm has found a global minimizer.

Example 3.3.10 (Exponential distribution)**.** We consider the model with $E = [0, \infty)^n$, $\Theta = (0, \infty)$, and P_λ the probability measure under which $X_1, \ldots, X_n$ are iid exponentially distributed with intensity parameter λ. The density for the exponential distribution with intensity parameter λ is

$$f_\lambda(x) = \lambda \exp(-\lambda x).$$

Math Box 3.3.1 (Multivariate optimization). If $\Theta \subseteq \mathbb{R}^d$ we have $\theta = (\theta_1, \dots, \theta_d)$ and a multivariate analog of the likelihood equation. If $l_x(\theta)$ is differentiable the derivative is

$$D l_x(\theta) = \left(\frac{\partial l_x}{\partial \theta_1}(\theta), \dots, \frac{\partial l_x}{\partial \theta_d}(\theta) \right)$$

and the likelihood equation reads

$$D l_x(\theta) = 0.$$

In reality this is a system of d equations, and a solution is a stationary point. A global minimizer in the interior of Θ of the minus-log-likelihood function is necessarily a stationary point.

The matrix of second derivatives is

$$D^2 l_x(\theta) = \left\{ \begin{array}{ccc} \frac{\partial^2 l_x}{\partial \theta_1^2}(\theta) & \cdots & \frac{\partial^2 l_x}{\partial \theta_d \partial \theta_1}(\theta) \\ \vdots & & \vdots \\ \frac{\partial^2 l_x}{\partial \theta_1 \partial \theta_d}(\theta) & \cdots & \frac{\partial^2 l_x}{\partial \theta_d^2}(\theta) \end{array} \right\},$$

and if $\tilde{\theta}$ is a solution of the likelihood equation, $D l_x(\tilde{\theta}) = 0$, then if $D^2 l_x(\tilde{\theta})$ is *positive definite* the point $\tilde{\theta}$ is a *local* minimizer of the minus-log-likelihood function. In general it is very difficult to say anything about global properties like existence and uniqueness of the MLE unless $D^2 l_x(\theta)$ is positive definite for *all* $\theta \in \Theta$. In this case $l_x(\theta)$ is a *convex* function and a solution to the likelihood equation is a global minimizer.

The gradient is $\nabla l_x(\theta) = D l_x(\theta)^T$, and the multivariate analog of the Newton-Raphson algorithm for numerically solving the likelihood equation reads

$$\theta_n = \theta_{n-1} - D^2 l_x(\theta_{n-1})^{-1} \nabla l_x(\theta_{n-1}).$$

For an initial guess θ_0 it yields a sequence $(\theta_n)_{n \geq 0}$ that – hopefully – converges to the MLE $\hat{\vartheta} = \hat{\theta}(x)$.

Having observed $x_1, \dots, x_n$ the likelihood function is

$$\mathscr{L}_x(\lambda) = \prod_{i=1}^{n} \lambda \exp(-\lambda x_i) = \lambda^n \exp\left(-\lambda \sum_{i=1}^{n} x_i \right)$$

and the minus-log-likelihood function is

$$l_x(\lambda) = \lambda \sum_{i=1}^{n} x_i - n \log \lambda.$$

If we differentiate we obtain the likelihood equation

$$\frac{d l_x}{d \lambda}(\lambda) = \sum_{i=1}^{n} x_i - \frac{n}{\lambda} = 0.$$
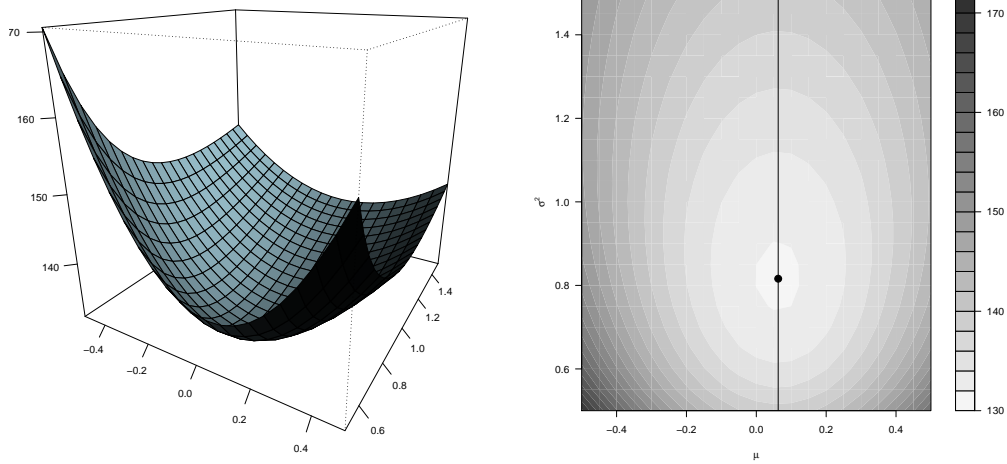
Figure 3.3: The minus-log-likelihood function (left) and a contour plot (right) for the scale-location parameters (μ, σ^2) based on $n = 100$ simulations of iid $N(0,1)$-distributed variables. The MLE is $\hat{\mu} = 0.063$ and $\tilde{\sigma}^2 = 0.816$. The profile minus-log-likelihood function of σ^2 is given by evaluating the minus-log-likelihood function along the straight line, as shown on the contour plot, given by $\mu = \hat{\mu}$.

There is a unique solution to the likelihood equation,

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} x_i},$$

and since $\frac{\mathrm{d}^2 l_x}{\mathrm{d}^2 \lambda}(\lambda) = \frac{n}{\lambda^2} > 0$ together with uniqueness of the solution $\hat{\lambda}$ to the likelihood equation, $\hat{\lambda}$ is a global minimizer of the minus-log-likelihood function and hence the MLE. $\diamond$

The example above shows that the ad hoc estimate of λ considered in Example 3.1.1 for neuronal interspike data is in fact also the maximum likelihood estimate.

Example 3.3.11 (Normal distribution). Let $X_1, \ldots, X_n$ be n iid random variables with the $N(\mu, \sigma^2)$ distribution. The statistical model is given by $E = \mathbb{R}^n$, $\Theta = \mathbb{R} \times (0, \infty)$, $\theta = (\mu, \sigma^2)$ and P_θ the probability measure that has density

$$
\begin{aligned}
f_\theta(x_1, \ldots, x_n) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right) \\
&= \frac{1}{(\sqrt{2\pi\sigma^2})^n} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2\right).
\end{aligned}
$$

The minus-log-likelihood function for observing $x = (x_1, \ldots, x_n) \in E$ is therefore

$$l_x(\mu, \sigma^2) = \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 + \frac{n}{2} \log \sigma^2 + n \log \sqrt{2\pi}.$$

We want to minimize this function of the two-dimensional parameter (μ, σ^2). We first fix σ^2 and regard $l_x(\mu, \sigma^2)$ as a function of μ only. Then we have a one dimensional parameter and we find the likelihood equation to be

$$\frac{\mathrm{d}l_x}{\mathrm{d}\mu}(\mu, \sigma^2) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu) = 0$$

which is a simple, linear equation in μ with the solution

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n}x_i.$$

Differentiating once more gives that

$$\frac{\mathrm{d}^2l_x}{\mathrm{d}^2\mu}(\mu, \sigma^2) = \frac{n}{2\sigma^2} > 0$$

so $\hat{\mu}$ is a global minimizer of $l_x(\mu, \sigma^2)$ for fixed σ^2. We conclude that no matter what the value of σ^2 is the likelihood is maximized as a function of μ at $\hat{\mu}$. If we therefore fix $\mu = \hat{\mu}$ and consider the minus-log-likelihood function

$$l_x(\hat{\mu}, \sigma^2) = \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \hat{\mu})^2 + \frac{n}{2}\log\sigma^2 + n\log\sqrt{2\pi}$$

as a function of σ^2 only, the corresponding likelihood equation becomes (note that we differentiate w.r.t. the parameter σ^2)

$$\frac{\mathrm{d}l_x}{\mathrm{d}\sigma^2}(\hat{\mu}, \sigma^2) = -\frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i - \hat{\mu})^2 + \frac{n}{2\sigma^2} = 0.$$

This is again a quite simple equation, and by multiplying with $\sigma^4 > 0$ we can rearrange the equation into

$$\sigma^2 n = \sum_{i=1}^{n}(x_i - \hat{\mu})^2.$$

The solution is

$$\tilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2.$$

Similarly we may show that the derivative $\frac{\mathrm{d}l_x}{\mathrm{d}\sigma^2}(\hat{\mu}, \sigma^2)$ is < 0 for $\sigma^2 < \tilde{\sigma}^2$ and > 0 for $\sigma^2 > \tilde{\sigma}^2$. Thus $l_x(\hat{\mu}, \sigma^2)$ is monotonely decreasing up to $\tilde{\sigma}^2$ and then monotonely increasing thereafter. Thus the solution $\tilde{\sigma}^2$ is the unique global minimizer for $l_x(\hat{\mu}, \sigma^2)$ and in conclusion $(\hat{\mu}, \tilde{\sigma}^2)$ is the unique global maximizer for the likelihood function.

$\diamond$
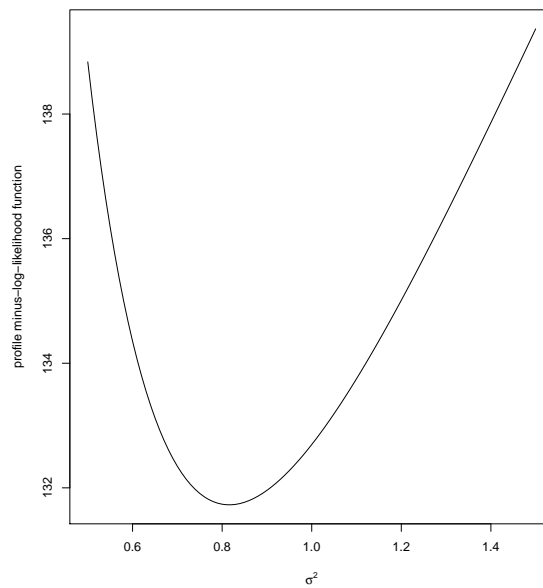
Figure 3.4: Example of the profile minus-log-likelihood function for σ^2 with 100 simulated $N(0,1)$ variables.

For the statistical model above – the normal distribution with unknown mean and variance – where we have two unknown parameters, we were able to derive the MLE by reducing the two-dimensional optimization problem to successive one-dimensional optimization problems. The technique is useful – analytically as well as numerically. Instead of throwing ourselves into a difficult optimization problem of several variables, we may try to solve the problem one variable at a time. The likelihood as a function of one parameter optimized over all other parameters is known as the *profile likelihood*. For instance,

$$l_x(\hat{\mu}, \sigma^2) = \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \hat{\mu})^2 + \frac{n}{2} \log \sigma^2 + n \log \sqrt{2\pi}$$

considered above is the profile minus-log-likelihood of σ^2.

Example 3.3.12 (Neuron interspike time models revisited)**.** As remarked after Example 3.3.10 we actually used the maximum likelihood estimator for estimation of λ in the simple model based on the exponential distribution of the neuronal interspike times considered in Example 3.1.1. We consider here the extended location-scale model also suggested in Example 3.1.1.

Having observed $x_1, \ldots, x_n$ the likelihood function is

$$\mathscr{L}_x(\mu, \lambda) = \prod_{i=1}^{n} \lambda \exp(-\lambda(x_i - \mu)) = \lambda^n \exp\left(-\lambda \sum_{i=1}^{n} (x_i - \mu)\right)$$

for $\mu < \min\{x_1, \ldots, x_n\}$. For $\mu \geq \min\{x_1, \ldots, x_n\}$ the likelihood function is 0 because at least one of the factors is 0. This effectively restricts the full parameter space $[0, \infty) \times (0, \infty)$ to $[0, \min\{x_1, \ldots, x_n\}) \times (0, \infty)$.

Fixing $\mu \in [0, \min\{x_1, \ldots, x_n\})$ computations identical to those in Example 3.3.10 show that

$$\hat{\lambda}(\mu) = \frac{n}{\sum_{i=1}^{n}(x_i - \mu)}$$

is the unique maximizer of the likelihood function. The profile likelihood function then becomes

$$\mathscr{L}_x(\mu, \hat{\lambda}(\mu)) = \hat{\lambda}(\mu)^n \exp(-n) = \left(\frac{n}{\sum_{i=1}^{n}(x_i - \mu)}\right)^n \exp(-n)$$

and we see that it is *monotonely increasing in μ*.

The conclusion is that this is a nasty model. The idea behind the likelihood method is to find parameters such that the probability measure fits the data in the best possible way. Here it turns out that the closer μ gets to the upper bound $\min\{x_1, \ldots, x_n\}$, the better the model fits the data. We cannot, however, take μ equal to $\min\{x_1, \ldots, x_n\}$ as this would make the likelihood drop to the all time minimum of 0! There is no maximum likelihood estimator for the model and we take this as a bad sign. Just as there are conceptual problems with $\mu > 0$, as discussed in Example 3.1.1, there are also inferential problems, and we cannot produce a sensible estimator of μ. $\diamond$

In the next example we use another trick – a *reparameterization* – which makes it possible to find the minimum of the minus-log-likelihood function quite easily.

Example 3.3.13 (Multinomial model). Consider the multinomial model from Example 3.3.1. The likelihood function for observing $x = (x_1, \ldots, x_n) \in \{1, \ldots, m\}^n$ is

$$\mathscr{L}_x(\theta) = \prod_{i=1}^{n} p(x_i) = \prod_{j=1}^{m} p_j^{n_j}$$

with

$$n_j = \sum_{i=1}^{n} 1(n_i = j)$$

since the product $\prod_{i=1}^{n} p(x_i)$ contains the p_j-factor precisely n_j times.

We will here consider the parameter space Θ consisting of $\theta = (p_1, \ldots, p_m)$ where $p_j > 0$ and $p_1 + \ldots + p_m = 1$ so that none of the point probabilities can be equal to 0. We make a reparameterization in terms of the parameter $\beta = (\beta_1, \ldots, \beta_m) \in (0, \infty)^m$ via

$$p_j(\beta) = \frac{e^{\beta_j}}{\sum_{i=1}^{m} e^{\beta_i}}$$

for $j = 1, \ldots, m$. We may note that the β parameter is in fact *not* identifiable though the θ parameter is. It just turns out that the optimization in the β-parametrization is simpler.

The minus-log-likelihood function in terms of the new parametrization reads

$$l_x(\beta) = n \log \left(\sum_{j=1}^{m} e^{\beta_j} \right) - \log \prod_{j=1}^{m-1} e^{\beta_j n_j} = n \log \left(\sum_{j=1}^{m} e^{\beta_j} \right) - \sum_{j=1}^{m} \beta_j n_j.$$

If we fix all parameters but β_j we find by differentiation that β_j must fulfill that

$$\frac{n e^{\beta_j}}{\sum_{j=1}^{m} e^{\beta_j}} - n_j = 0$$

or that

$$p_j(\beta) = \frac{e^{\beta_j}}{\sum_{j=1}^{m} e^{\beta_j}} = \frac{n_j}{n},$$

which has a solution in β_j if and only if $n_j > 0$. This shows that in the θ-parametrization there is a unique minimizer of the minus-log-likelihood equation given by

$$\hat{p}_j = \frac{n_j}{n}$$

if $n_1, \ldots, n_m > 0$. Strictly speaking we have showed that this solution is the only possible minimizer – we haven't showed that it actually is a minimizer[3]

Using the reparameterization we have found that if $n_1, \ldots, n_m > 0$ the likelihood function $\mathscr{L}_x(\theta)$ attains a unique maximum over the set Θ in

$$\left(\frac{n_1}{n}, \ldots, \frac{n_m}{n} \right),$$

which is therefore the maximum likelihood estimate. The vector $(n_1, \ldots, n_m)$ is the realization of the random variable $(N_1, \ldots, N_m)$ and the maximum likelihood estimator is

$$\hat{\theta} = (\hat{p}_1, \ldots, \hat{p}_m) = \left(\frac{N_1}{n}, \ldots, \frac{N_m}{n} \right)$$

This maximum likelihood estimator is very reasonable, since the estimator for p_j,

$$\hat{p}_j = \frac{N_j}{n},$$

is the relative frequency of observations being equal to j. ⋄

The previous examples gave explicit expressions for the maximum likelihood estimator. It is not always the case that we can come up with a nice analytic solution to the optimization problem – not even by clever reparameterizations or using the profile method. In fact, it is rather the exception than the rule that the maximum

---

[3]If one is familiar with convexity concepts it is possible to show that l_x is convex in the β-parametrization, and convexity implies that a solution to the likelihood equation is a global minimizer.
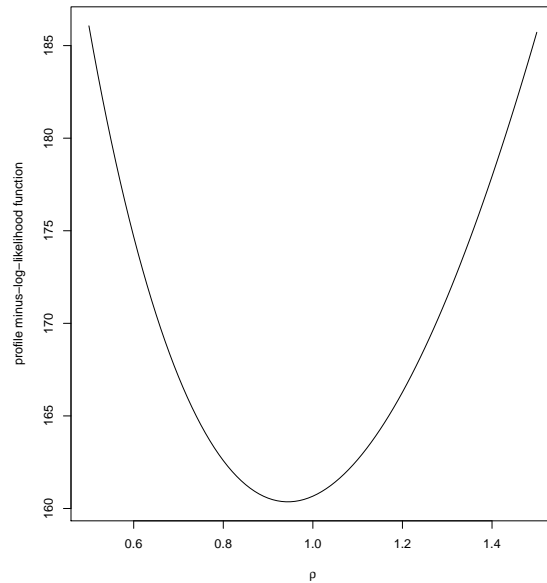
Figure 3.5: The profile minus-log-likelihood function for ρ with 100 simulated Gumbel variables.

likelihood estimator is given by a closed form analytic expression. The next example will show an analysis where we use almost all of the techniques considered previously and though we cannot find a complete analytic solution we are able to derive precise conditions for the existence of a unique maximum to the likelihood function and we can derive a single equation in a one-dimensional parameter that we need to solve numerically.

Example 3.3.14 (Gumbel distribution)**.** If $X_1, \ldots, X_n$ are iid random variables with distribution being a location-scale transformation of the Gumbel distribution, then the statistical model is given by the sample space $E = \mathbb{R}^n$, parameter space $\Theta = \mathbb{R} \times (0, \infty)$, and parameters $\theta = (\mu, \sigma)$. The probability measure P_θ has density

$$f_{\mu, \sigma}(x_1, \ldots, x_n) = \prod_{i=1}^{n} \frac{1}{\sigma} \exp\left(-\frac{x_i - \mu}{\sigma} - \exp\left(-\frac{x_i - \mu}{\sigma} \right) \right).$$

The minus-log-likelihood function for observing $x = (x_1, \ldots, x_n)$ is therefore

$$l_x(\mu, \sigma) = n \log \sigma + \sum_{i=1}^{n} \frac{x_i - \mu}{\sigma} + \sum_{i=1}^{n} \exp\left(-\frac{x_i - \mu}{\sigma} \right). \tag{3.9}$$

It is not so easy to find the minimum of this function analytically let alone to show that it has a minimum. But there is a way around dealing with $l_x(\mu, \sigma)$ directly via

a *reparameterization*. Introduce the parameters $(\eta, \rho) \in \mathbb{R} \times (0, \infty)$ given by

$$(\eta, \rho) = \left(\frac{\mu}{\sigma}, \frac{1}{\sigma} \right).$$

Clearly then $(\mu, \sigma) = \left(\frac{\eta}{\rho}, \frac{1}{\rho} \right)$, so there is a one-to-one correspondence between the parameters (μ, σ) and the parameters (η, ρ). One of the beautiful and useful properties of the maximum likelihood estimation principle is, that it is invariant under reparameterizations. This means that we can just as well minimize l_x in this new parametrization, which turns out to be more convenient.

We find that in the (η, ρ)-parametrization

$$
\begin{aligned}
l_x(\eta, \rho) &= \sum_{i=1}^{n} (\rho x_i - \eta) + \sum_{i=1}^{n} \exp(-\rho x_i + \eta) - n \log \rho \\
&= \rho \sum_{i=1}^{n} x_i + \exp(\eta) \sum_{i=1}^{n} \exp(-\rho x_i) - n\eta - n \log \rho \\
&= \rho n \overline{x} + \exp(\eta) \sum_{i=1}^{n} \exp(-\rho x_i) - n\eta - n \log \rho
\end{aligned}
$$

where we have introduced the notation $\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$, and one can already see the simplification of the expression for the minus-log-likelihood function. Moreover, if we fix the ρ parameter and differentiate w.r.t. η, we obtain

$$\frac{\mathrm{d} l_x}{\mathrm{d}\eta}(\eta, \rho) = \exp(\eta) \sum_{i=1}^{n} \exp(-\rho x_i) - n.$$

To find the minimum of the minus-log-likelihood function for any fixed ρ we equate the derivative to 0. This gives the equation

$$\exp(\eta) \sum_{i=1}^{n} \exp(-\rho x_i) = n,$$

whose solution is

$$\hat{\eta}(\rho) = -\log \left(\frac{1}{n} \sum_{i=1}^{n} \exp(-\rho x_i) \right).$$

Differentiation once more w.r.t. η yields that

$$\frac{\mathrm{d}^2 l_x}{\mathrm{d}\eta^2}(\eta, \rho) = \exp(\eta) \sum_{i=1}^{n} \exp(-\rho x_i) > 0,$$

which shows that not only does the minus-log-likelihood function attain a local minimum at $\hat{\eta}(\rho)$ as a function of η for given ρ but actually a global minimum.
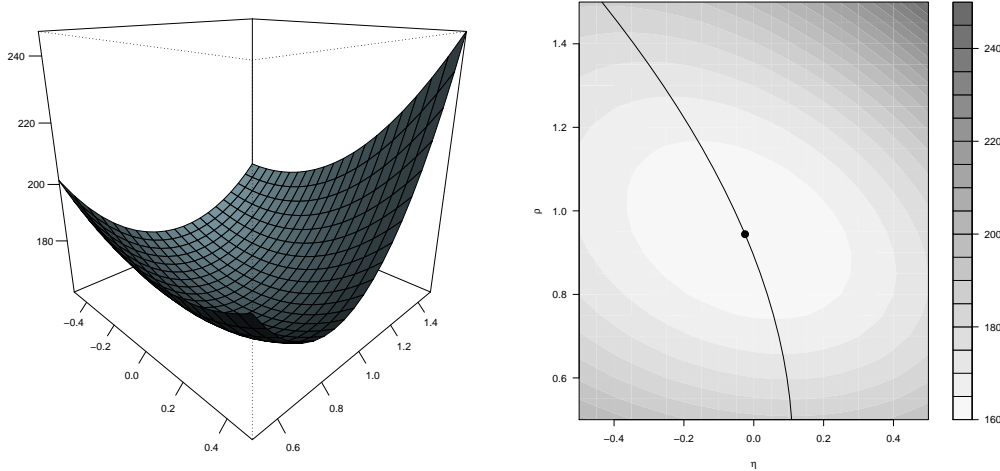
Figure 3.6: The minus-log-likelihood function (left) and a contour plot (right) for the (η, ρ) reparameterization of the scale-location model based on $n = 100$ simulations of iid Gumbel distributed variables. The MLE is $\hat{\eta} = 0.025$ and $\hat{\rho} = 0.944$. The profile minus-log-likelihood function of ρ is given by evaluating the minus-log-likelihood function along the curved line, as shown on the contour plot, given by the equation $\hat{\eta}(\rho) = \eta$.

The minimizer $\hat{\eta}(\rho)$ depends upon ρ, which makes things more complicated as compared to the normal distribution where the minimizer of μ for fixed σ^2 does not depend upon σ^2. To get any further we plug the expression of $\hat{\eta}(\rho)$ back into the minus-log-likelihood function giving the profile minus-log-likelihood function

$$l_x(\hat{\eta}(\rho), \rho) = \rho n \overline{x} + n + n \log \left(\frac{1}{n} \sum_{i=1}^n \exp(-\rho x_i) \right) - n \log \rho,$$

which is a function of ρ alone. Differentiation gives

$$\frac{\mathrm{d}l_x}{\mathrm{d}\rho}(\hat{\eta}(\rho), \rho) = n \overline{x} - n \frac{\sum_{i=1}^n x_i \exp(-\rho x_i)}{\sum_{i=1}^n \exp(-\rho x_i)} - \frac{n}{\rho}.$$

Equating this derivative to 0 yields the equation

$$\frac{\sum_{i=1}^n x_i \exp(-\rho x_i)}{\sum_{i=1}^n \exp(-\rho x_i)} + \frac{1}{\rho} = \overline{x}, \tag{3.10}$$

whose solution does not seem to have a nice analytic expression. We can note that for ρ approaching 0, the left hand side behaves as $\overline{x} + 1/\rho > \overline{x}$ and for ρ approaching ∞ the left hand side behaves as $\min\{x_1, \ldots, x_n\}$. If there are at least two different observations the latter is strictly smaller than $\overline{x}$ and since the left hand side is continuous in ρ there is in this case always a solution to the equation.

A second differentiation of the profile minus-log-likelihood function – and some algebraic manipulations – give

$$\frac{\mathrm{d}^2 l_x}{\mathrm{d}\rho^2}(\hat{\eta}(\rho), \rho) = n \sum_{i=1}^{n} \left(x_i - \frac{\sum_{i=1}^{n} x_i \exp(-\rho x_i)}{\sum_{i=1}^{n} \exp(-\rho x_i)} \right)^2 \frac{\exp(-\rho x_i)}{\sum_{i=1}^{n} \exp(-\rho x_i)} + \frac{n}{\rho^2} > 0.$$

Since this shows that the derivative is strictly increasing there can be only one solution to the equation above.

The conclusion of our analysis is, that if $n \geq 2$ and at least two of the observations are different there is precisely one solution to the equation above, hence there is a unique global minimum for the profile minus-log-likelihood function, and consequently there is a unique global minimizer for the full minus-log-likelihood function.

From a practical point of view we need to solve (numerically) the equation (3.10) and then plug this solution into $\hat{\eta}(\rho)$. This gives the maximum likelihood estimate of (η, ρ). The Newton-Raphson algorithm for solving the equation reads

$$\rho_n = \rho_{n-1} - \frac{\mathrm{d}^2 l_x}{\mathrm{d}\rho^2}(\hat{\eta}(\rho_{n-1}), \rho_{n-1})^{-1} \frac{\mathrm{d} l_x}{\mathrm{d}\rho}(\hat{\eta}(\rho_{n-1}), \rho_{n-1})$$

for $n \geq 1$ and an initial guess ρ_0. $\diamond$

The previous example elaborated on the idea of using the profiled minus-log-likelihood function. Combined with a reparameterization the profile method reduced the bivariate optimization to the solution of a univariate equation. In the next example we show how a reparameterization can help simplify the computations considerably though they can be carried out in the original parametrization if desired.

Example 3.3.15 (Evolutionary models)**.** We consider the model from Example 3.1.2 with $(X_1, Y_1), \ldots, (X_n, Y_n)$ being n iid random variables each taking values in $E_0 \times E_0$ with $E_0 = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$. We take the distribution of (X_i, Y_i), when the evolutionary distance in time is t, to be given by the point probabilities

$$P_{t,p,\theta}(x, y) = p(x) P_\theta^t(x, y).$$

Here p is a vector of point probabilities on E_0, so that $p(x)$ is the probability of x, and $P_\theta^t(x, y)$ is the conditional probability that x mutates into y in time t. These conditional probabilities depend on the additional parameter $\theta \in \Theta$. The main focus is on the estimation of θ. Having observed $z = ((x_1, y_1), \ldots, (x_n, y_n)) \in E = (E_0 \times E_0)^n$ the full likelihood function becomes

$$\mathscr{L}_z(t, p, \theta) = \prod_{i=1}^{n} P_{t,p,\theta}(x_i, y_i) = \prod_{i=1}^{n} p(x_i) P_\theta^t(x_i, y_i)$$

and the minus-log-likelihood function is

$$l_z(t, p, \theta) = - \sum_{i=1}^{n} \log p(x_i) - \sum_{i=1}^{n} \log P_\theta^t(x_i, y_i).$$

We observe that the first term depends upon p only and the second term on (t, θ). We can therefore optimize each term separately to find the maximum likelihood estimator. In addition we see that the first term is simply the minus-log-likelihood function for the multinomial distribution, and naturally the MLE of the marginal point probability $p(x)$ is the relative frequency of x among the observations $x_1, \ldots, x_n$.

Turning to the second term we know from Example 3.3.6 that there may be a problem with identifiability of t and the additional parameters, and we consider here the situation where we know (or fix) t. Then the second term,

$$\tilde{l}_z(\theta) = -\sum_{i=1}^{n} \log P_\theta^t(x_i, y_i),$$

is regarded as a function of θ only. We also introduce the variables

$$n_z(x, y) = \sum_{i=1}^{n} 1(x_i = x, y_i = y),$$

thus $n_z(x, y)$ is the number of observed mutations of x with y. Then we can rewrite

$$\tilde{l}_z(\theta) = -\sum_{x,y} n_z(x, y) \log P_\theta^t(x, y)$$

since there are exactly $n_z(x, y)$ terms in the sum $\sum_{i=1}^{n} \log P_\theta^t(x_i, y_i)$ that equals $\log P_\theta^t(x, y)$.

For the rest of this example we will consider the special model, the Jukes-Cantor model, where

$$\begin{aligned} P_\alpha^t(x, x) &= 0.25 + 0.75 \times \exp(-4\alpha t) \\ P_\alpha^t(x, y) &= 0.25 - 0.25 \times \exp(-4\alpha t), \quad \text{if } x \neq y, \end{aligned}$$

for some (known) $t > 0$ and $\alpha > 0$ the unknown additional parameter. Introducing

$$n_1 = \sum_x n_z(x, x) \quad \text{and} \quad n_2 = \sum_{x \neq y} n_z(x, y)$$

we find that

$$\tilde{l}_z(\alpha) = -n_1 \log(0.25 + 0.75 \times \exp(-4\alpha t)) - n_2 \log(0.25 - 0.25 \times \exp(-4\alpha t)).$$

If we differentiate we obtain

$$\begin{aligned} \frac{\mathrm{d}\tilde{l}_z}{\mathrm{d}\alpha}(\alpha) &= \frac{3n_1 t \exp(-4\alpha t)}{0.25 + 0.75 \times \exp(-4\alpha t)} - \frac{n_2 t \exp(-4\alpha t)}{0.25 - 0.25 \times \exp(-4\alpha t)} \\ &= 4t \exp(-4\alpha t) \left(\frac{3n_1}{1 + 3\exp(-4\alpha t)} - \frac{n_2}{1 - \exp(-4\alpha t)} \right) \end{aligned}$$

and the likelihood equation $\frac{d\tilde{l}_z}{d\alpha}(\alpha) = 0$ is equivalent to the equation

$$3n_1(1 - \exp(-4\alpha t)) = n_2(1 + 3\exp(-4\alpha t)).$$

This equation has a (unique) solution if and only if $3n_1 > n_2$ in which case

$$\hat{\alpha} = \frac{1}{4t}\log\frac{3(n_1 + n_2)}{3n_1 - n_2} = \frac{1}{4t}\log\frac{3n}{3n_1 - n_2}$$

is the maximum likelihood estimator. Moreover, we see from the expression of the derivative of $\tilde{l}_z(\alpha)$ that (given $3n_1 > n_2$) then $\frac{d\tilde{l}_z}{d\alpha}(\alpha) < 0$ if $\alpha < \hat{\alpha}$ and $\frac{d\tilde{l}_z}{d\alpha}(\alpha) > 0$ if $\alpha > \hat{\alpha}$. This shows that $\tilde{l}_z(\alpha)$ is monotonely decreasing up to $\hat{\alpha}$ and monotonely increasing thereafter. Hence $\hat{\alpha}$ is the global minimum of the minus-log-likelihood function.

Working with the minus-log-likelihood and in particular differentiation in the α-parametrization is hideous. It is much easier to make a reparameterization by

$$\gamma = \gamma(\alpha) = 0.25 - 0.25\exp(-4\alpha t)$$

such that

$$\alpha = \alpha(\gamma) = \frac{1}{4t}\log(1 - 4\gamma)$$

for $\gamma \in (0, 0.25)$. In the γ-parameter the minus-log-likelihood becomes

$$l_z(\gamma) = -n_1\log(1 - 3\gamma) - n_2\log\gamma$$

for $\gamma \in (0, 0.25)$. The differentiation is easier yielding the derivative

$$l'_z(\gamma) = \frac{3n_1}{1 - 3\gamma} - \frac{n_2}{\gamma},$$

and solving the likelihood equation is always possible for $\gamma \in (0, 1/3)$ and gives

$$\hat{\gamma} = \frac{n_2}{3n}.$$

The solution is in $(0, 0.25)$ if and only if $3n_1 > n_2$ in which case we get

$$\hat{\alpha} = \alpha(\hat{\gamma}) = \frac{1}{4t}\log\left(1 - 4\frac{n_2}{3n}\right) = \frac{1}{4t}\log\frac{3n}{3n_1 - n_2}.$$

$\diamond$

Example 3.3.16. We turn to the data from the Hepatitis C virus evolution, as considered in Example 1.2.4, and we want to estimate the α parameter in the Jukes-Cantor model. The two quantities that enter in the estimator are n_2, the total number of mutations, and n_1, the remaining number of non-mutated nucleotide pairs. Observe that $n_1 = n - n_2$. We will consider two situations. Either we pool all of the three segments of the virus and make one estimate, or we estimate α separately for the segments A, B, and C. The following table shows the result.
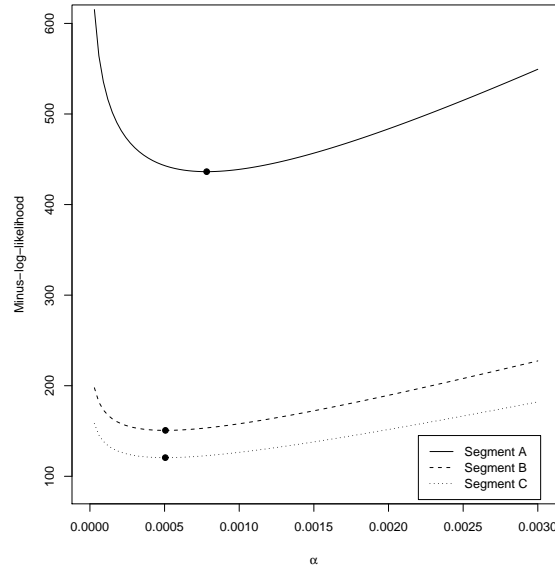
Figure 3.7: The (partial) minus-log-likelihood functions $\tilde{l}_z(\alpha)$ for the hepatitis C virus data, as considered in Example 3.3.16, using the Jukes-Cantor model. We estimate α separately for the three segments A, B, and C of the genome. The corresponding maximum likelihood estimates are marked on the plots.

	Segment A	B	C	A+B+C
n_1	2532	1259	1009	4800
n_2	78	25	20	123
$\hat{\alpha}$	7.8×10^{-4}	5.0×10^{-4}	5.0×10^{-4}	6.5×10^{-4}

The time is measured in years so that $t = 13$ and the estimated mutation rates are thus per year mutation rates. Seemingly, segment A shows a different mutation rate than segments B and C do – but conclusions like this have to be based on knowledge about the uncertainty of the estimates. We deal with this in Chapter **??**. Plugging the estimated mutation rate $\hat{\alpha}_A = 7.8 \times 10^{-4}$ for segment A into the expression for $P_\alpha^t(x, y)$, we find for instance that for $x \neq y$

$$P_{\hat{\alpha}}^1(x, y) = 0.25 - 0.25 \times \exp(-4 \times 7.8 \times 10^{-4}) = 0.000779.$$

This is an estimate of the chance that any specific single nucleotide in segment A will mutate within a year.

Still dividing the data according to the three segments, the frequency vectors $n_A, n_B, n_C \in \mathbb{N}_0^4$ and relative frequency vectors $p_A, p_B, p_C \in \mathbb{R}^4$ (MLEs of p) are

	A	C	G	T	Total
n_A	483	786	763	578	2610
p_A	0.185	0.301	0.292	0.221	1
n_B	257	398	350	279	1284
p_B	0.200	0.310	0.273	0.217	1
n_C	233	307	286	203	1029
p_C	0.226	0.298	0.278	0.197	1

To investigate whether the model is actually a good model for the data, we reconsider the table of observed mutations for segment A together with the expected number of mutations. Since all mutations are equally probable for the Jukes-Cantor model – the probability being $0.25 - 0.25 \times \exp(-4 \times 13 \times 7.8 \times 10^{-4}) = 0.0099$ for segment A – the expected number of mutations from nucleotide x is $n_A(x) \times 0.0099$.

		H90		
	A	C	G	T
A		1	11	1
C	4		1	20
G	13	3		1
T	3	19	1	

H77 (left table)

		H90		
	A	C	G	T
A		4.8	4.8	4.8
C	7.8		7.8	7.8
G	7.6	7.6		7.6
T	5.7	5.7	5.7	

H77 (right table)

Observed mutations - segment A   **Expected mutations - segment A**

Looking at these two tables we are suspicious about whether the Jukes-Cantor model actually is adequate. There are many more *transitions* than *transversions*, where the Jukes-Cantor model predicts the same number.   ◇

Example 3.3.17 (Logistic regression)**.** This example introduces a class of models for analyzing the effect of a continuous factor/covariate on the distribution of a single Bernoulli variable. This is a quite classical situation with a *dichotomous dose-response experiment*. The response is a 0-1 variable, like dead/alive or red/green, but the probability of the variable depends upon a continuous dose, for instance the concentration of a toxic compound, pesticide or insecticide.

In an experiment, carried out by Jørgen Jespersen at Statens Skadedyrslaboratorium, 260 flies (*Musca domestica*) were exposed to the insecticide dimethoat in varying concentrations. The experiment was organized with 13 groups of 20 flies, and flies in the same group were given the same concentration. The results can be summarized in the following table:

Concentration	log(concentration)	Deaths	Survivors
0.016	−4.135	0	20
0.0226	−3.790	0	20
0.032	−3.442	0	20
0.0453	−3.094	0	20
0.064	−2.749	5	15
0.0905	−2.402	4	16
0.128	−2.056	5	15
0.181	−1.709	15	5
0.256	−1.363	14	6
0.362	−1.016	18	2
0.515	−0.664	20	0
0.724	−0.323	20	0
1.024	0.024	20	0

Thus we have observed 260 Bernoulli variables $X_1, \ldots, X_{260}$ (death = 1 and survival = 0), which we can safely assume independent. But it would obviously be wrong to assume that they have the same distribution. On the contrary, we are interested in figuring out the effect of the concentration of dimethoat on the death rate of the flies, that is, on the distribution of the Bernoulli variables.

We will parametrize the probability of death as a function of dose (concentration), and we introduce

$$p(y) = \frac{\exp(\alpha + \beta y)}{1 + \exp(\alpha + \beta y)},$$

where $p(y) \in (0, 1)$ and $\alpha, \beta > 0$. The *logistic regression model* is then defined by letting the probability of $X_i = 1$ given the dose y_i be $p(y_i)$. The function $y \mapsto p(y)$ is known as the logistic function, which is why this model of the probability as a function of dose level is known as logistic regression. It is common not to use the concentration directly as the dose level but instead use the log(concentration). Thus $y = $ log(concentration). The observation consists in general of a vector $x = (x_1, \ldots, x_n)$ of 0-1 variables, with $n = 260$ in the fly death example, and the statistical model of logistic regression has sample space $E = \{0, 1\}^n$, parameter $\theta = (\alpha, \beta)$ and parameter space $\Theta = \mathbb{R}^2$.

Observing that

$$1 - p(y) = \frac{1}{1 + \exp(\alpha + \beta y)}$$

we can rewrite to find

$$\log \frac{p(y)}{1 - p(y)} = \alpha + \beta y.$$

The left hand side is the logarithm of the odds that the fly die, so the model says that the log odds for dying depends linearly upon the dose level (log(concentration) for the flies).

The likelihood function for observing $x = (x_1, \ldots, x_n) \in E$ is

$$\mathscr{L}_x(\alpha, \beta) = \prod_{i=1}^{n} p(y_i)^{x_i} (1 - p(y_i))^{1-x_i} = \prod_{i=1}^{n} \frac{\exp(\alpha x_i + \beta y_i x_i)}{1 + \exp(\alpha + \beta y_i)}.$$

The minus-log-likelihood function becomes

$$
\begin{aligned}
l_x(\alpha, \beta) &= \sum_{i=1}^{n} \left\{ \log(1 + \exp(\alpha + \beta y_i)) - \alpha x_i - \beta y_i x_i \right\} \\
&= \sum_{i=1}^{n} \log(1 + \exp(\alpha + \beta y_i)) - \alpha S - \beta SS,
\end{aligned}
$$

where

$$S = \sum_{i=1}^{n} x_i \quad \text{and} \quad SS = \sum_{i=1}^{n} y_i x_i.$$

Note that you can compute the likelihood function from the table above. You do not need the observations, only the summary given in the table. It is actually sufficient to know just $S = 121$, $SS = -151.9$ and the log(concentrations).

We fix α and differentiate w.r.t. β and find

$$
\begin{aligned}
\frac{dl_x}{d\beta}(\alpha, \beta) &= \sum_{i=1}^{n} \frac{y_i \exp(\alpha + \beta y_i)}{1 + \exp(\alpha + \beta y_i)} - SS = \sum_{i=1}^{n} y_i p(y_i) - SS \\
\frac{d^2 l_x}{d\beta^2}(\alpha, \beta) &= \sum_{i=1}^{n} \frac{y_i^2 \exp(\alpha + \beta y_i)}{(1 + \exp(\alpha + \beta y_i))^2} = \sum_{i=1}^{n} y_i^2 p(y_i)(1 - p(y_i)).
\end{aligned}
$$

Likewise, we fix β and differentiate w.r.t. α

$$
\begin{aligned}
\frac{dl_x}{d\alpha}(\alpha, \beta) &= \sum_{i=1}^{n} \frac{\exp(\alpha + \beta y_i)}{1 + \exp(\alpha + \beta y_i)} - S = \sum_{i=1}^{n} p(y_i) - S \\
\frac{d^2 l_x}{d\alpha^2}(\alpha, \beta) &= \sum_{i=1}^{n} \frac{\exp(\alpha + \beta y_i)}{(1 + \exp(\alpha + \beta y_i))^2} = \sum_{i=1}^{n} p(y_i)(1 - p(y_i)).
\end{aligned}
$$

Since both of the second derivatives are > 0, we conclude that $l_x(\alpha, \beta)$ as a function of one of the parameters (and the other fixed) can have at most one local minimum, which is then a global minimum. This *does not* prove the uniqueness of minima of l_x regarded as a function of two variables.

We can approach the bivariate optimization by general purpose optimization algorithms, and use for instance the Newton-Raphson algorithm as discussed in Math Box 3.3.1. As an alternative, which illustrates the use of the one-dimensional Newton-Raphson algorithm along the coordinate axis, consider the following *alternating*
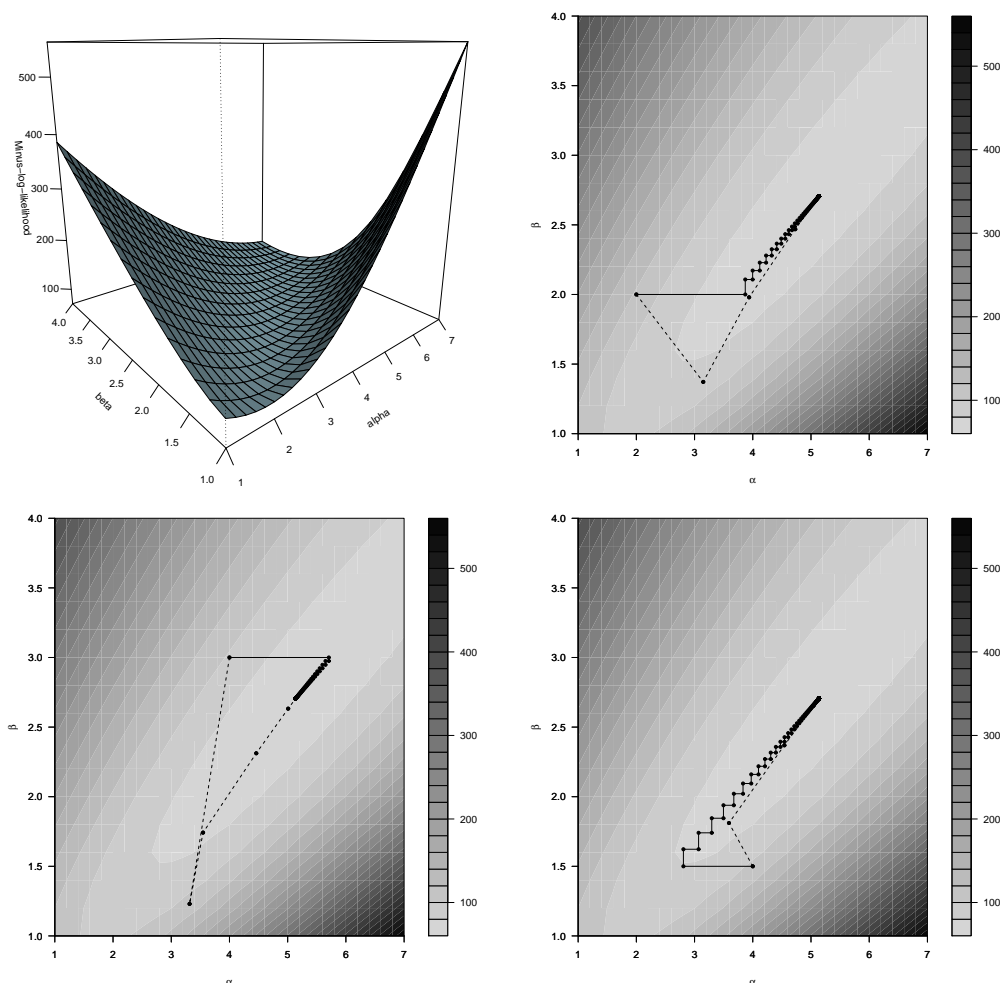
Figure 3.8: The minus-log-likelihood function (upper left) using the logistic regression model for the fly data. It does not have a pronounced minimum, but rather a long valley running diagonally in the α-β-parametrization. The contour plots (upper right and bottom) show how two different algorithms converge towards the minimizer (the MLE) $(\hat{\alpha}, \hat{\beta}) = (5.137, 2.706)$ for different starting points $(\alpha_0, \beta_0) = (2, 2), (4, 3)$ and $(4, 1.5)$. The broken line is the bivariate Newton-Raphson algorithm and the black line is the alternating Newton-Raphson algorithm, as discussed in Example 3.3.17, which is slower an moves in zig-zag along the α and β axes.

Newton-Raphson algorithm. With initial guess (α_0, β_0) we define the sequence $(\alpha_n, \beta_n)_{n \geq 0}$ by

$$\alpha_n = \alpha_{n-1} - \left(\sum_{i=1}^{n} p_{n-1}(y_i)(1 - p_{n-1}(y_i)) \right)^{-1} \left(\sum_{i=1}^{n} p_{n-1}(y_i) - S \right)$$

and then

$$\beta_n = \beta_{n-1} - \left(\sum_{i=1}^{n} y_i^2 p'_{n-1}(y_i)(1 - p'_{n-1}(y_i))\right)^{-1} \left(\sum_{i=1}^{n} y_i p'_{n-1}(y_i) - SS\right)$$

where

$$p_{n-1}(y) = \frac{\exp(\alpha_{n-1} + \beta_{n-1}y)}{1 + \exp(\alpha_{n-1} + \beta_{n-1}y)} \quad p'_{n-1}(y) = \frac{\exp(\alpha_n + \beta_{n-1}y)}{1 + \exp(\alpha_n + \beta_{n-1}y)}.$$

The algorithm amounts to making a one-dimensional Newton-Raphson step first along the α axis, then along the β axis and so on and so forth. It is not a particular fast algorithm – if the sequence of parameters converge they do it slowly – but curiously for this example the alternating Newton-Raphson algorithm is more stable than the raw bivariate Newton-Raphson. This means that it will actually converge for a large set of initial guesses. Convergence of the raw bivariate Newton-Raphson is quite sensitive to making a good initial guess. There are ways to deal with such problems – via *moderations* of the steps in the Newton-Raphson algorithm. But that is beyond the scope of these notes. If any of the algorithms converge then the resulting point is, for the logistic regression model, always a global minimizer. For the flies the MLE is

$$(\hat{\alpha}, \hat{\beta}) = (5.137, 2.706).$$
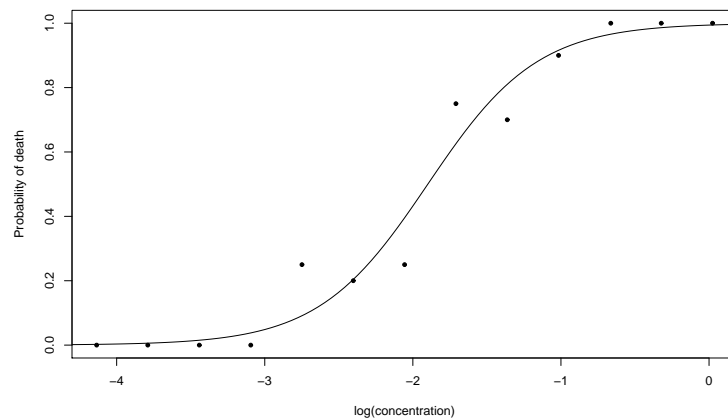
Figure 3.9: The logistic curve for the probability of fly death as a function of log(concentration) of dimethoat using the MLE parameters $(\hat{\alpha}, \hat{\beta}) = (5.137, 2.706)$

Sometimes the likelihood function does not attain a maximum, and neither algorithm will converge. This happens if there are two concentrations, $c_0 < c_1$, such that all flies that received a dose below c_0 survived and all flies that received a dose above c_1 died, and we have no observations with dose levels in between. This is not really a problem with the model, but rather a result of a bad experimental design. $\diamond$

Example 3.3.18 (Kimura model)**.** Instead of considering the Jukes-Cantor model in Example 3.3.15 we can consider the Kimura model given by the conditional probabilities

$$
\begin{aligned}
P^t(x,x) &= 0.25 + 0.25\exp(-4\beta t) + 0.5\exp(-2(\alpha+\beta)t) \\
P^t(x,y) &= 0.25 + 0.25\exp(-4\beta t) - 0.5\exp(-2(\alpha+\beta)t), \quad \text{if } \lambda(x,y) = \alpha \\
P^t(x,y) &= 0.25 - 0.25\exp(-4\beta t), \quad \text{if } \lambda(x,y) = \beta.
\end{aligned}
$$

We regard t to be fixed and $(\alpha,\beta) \in (0,\infty) \times (0,\infty)$ as the unknown parameters. If we introduce the three numbers

$$
\begin{aligned}
n_1 &= \sum_x n_z(x,x) = n_z(\mathtt{A},\mathtt{A}) + n_z(\mathtt{C},\mathtt{C}) + n_z(\mathtt{G},\mathtt{G}) + n_z(\mathtt{T},\mathtt{T}) \\
n_2 &= \sum_{x,y:\lambda(x,y)=\alpha} n_x(x,y) = n_z(\mathtt{A},\mathtt{G}) + n_z(\mathtt{C},\mathtt{T}) + n_z(\mathtt{G},\mathtt{A}) + n_z(\mathtt{T},\mathtt{C}) \\
n_3 &= \sum_{x,y:\lambda(x,y)=\beta} n_z(x,y) = n - n_1 - n_2
\end{aligned}
$$

being the number of nucleotide pairs with no mutations, the number of transitions and the number of transversions respectively, then the (partial) minus-log-likelihood function becomes

$$
\begin{aligned}
\tilde{l}_z(\alpha,\beta) &= -n_1 \log(0.25 + 0.25\exp(-4\beta t) + 0.5\exp(-2(\alpha+\beta)t)) \\
&\quad -n_2 \log(0.25 + 0.25\exp(-4\beta t) - 0.5\exp(-2(\alpha+\beta)t)) \\
&\quad -n_3 \log(0.25 - 0.25\exp(-4\beta t)).
\end{aligned}
$$

Direct computations with the minus-log-likelihood function are even more hideous in the (α,β)-parametrization for the Kimura model than for the Jukes-Cantor model. A reparameterization is possible but we refrain from the further theoretical analysis of the model. Instead we turn to the Hepatitis C virus data, which we analyzed in Example 3.3.16 using the Jukes-Cantor model, we apply a standard multivariate numerical optimization algorithm, for instance `optim` in R, for computing the maximum likelihood estimates of α and β. The results for this example are summarized in the following table:

	Segment		
	A	B	C
n_1	2532	1259	1009
n_2	63	20	14
n_3	15	5	6
$\hat{\alpha}$	1.9×10^{-3}	1.2×10^{-3}	1.2×10^{-3}
$\hat{\beta}$	2.2×10^{-4}	1.5×10^{-4}	2.3×10^{-4}
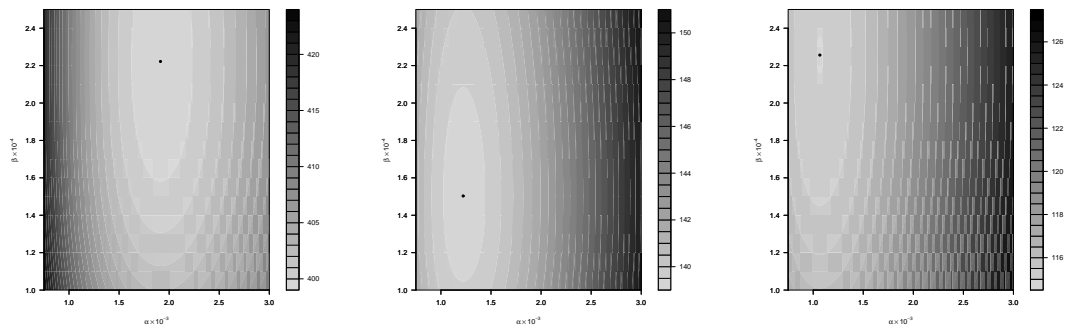
Figure 3.10: The contour plots of the minus-log-likelihood function for the hepatitis C data using the Kimura model. The three plots represent the three different segments A, B, and C (from left to right). The corresponding MLEs are found by numerical optimization using `optim` in R.

The MLEs of the p_A, p_B and p_C vectors do not change, and we can compute the table of expected mutations

		H90
		A	C	G	T
	A		1	11	1
H77	C	4		1	20
	G	13	3		1
	T	3	19	1	

Observed mutations - segment A

		H90
		A	C	G	T
	A		1.4	11.7	1.4
H77	C	2.3		2.3	19.0
	G	18.4	2.2		2.2
	T	1.7	14.0	1.7	

Expected mutations - segment A

Compared to the Jukes-Cantor model, the expectations based on the estimated Kimura model seem to be in greater concordance with the observed expectations.

$\diamond$

Exercises

Exercise 3.3.1. Consider the Jukes-Cantor model in Example 3.3.15. Define the variable $Z_i = 1(X_i \neq Y_i)$. Argue that $N_1 = \sum_{i=1}^{n} Z_i$ follows a binomial distribution with success parameter

$$p(\alpha) = 0.25 - 0.25 \exp(-4\alpha t).$$

For the binomial distribution we know that the MLE is

$$\hat{p} = \frac{1}{n} N_1.$$

Show that the MLE of α in Example 3.3.15 exists if and only if there is a solution to the equation $p(\alpha) = \hat{p}$ and show that the resulting solution is then precisely $\hat{\alpha}$.

Color blindness is a so-called X-linked recessive trait. This means that the gene responsible for color blindness is located on the X chromosome and that color blindness is recessive. Females who have two X chromosomes are thus color blind if the allele resulting in color blindness (the CB-allele in the following) is present on both X chromosomes whereas males, who have only a single X chromosome, are color blind whenever the CB-allele is present on their X chromosome.

We denote the proportion of males in a population that have the CB-allele by $p \in (0, 1)$. This is the parameter we want to estimate.

We assume in the following that we have observations from m randomly selected males, $x_1, \ldots, x_m \in \{0, 1\}$, such that $x_i = 1$ if male number i is color blind (has the CB-allele). We assume that the observations are independent.

Exercise 3.3.2. Argue that the probability of observing $x = (x_1, \ldots, x_m)$ equals

$$p^{m_b}(1-p)^{m_B}$$

where $m_b = \sum_{i=1}^{m} x_i$ is the number of males with the CB-allele and $m_B = m - m_b$ is the number of males without the CB-allele, find the likelihood function and the minus-log-likelihood function for p and show that the MLE equals

$$\hat{p} = \frac{m_b}{m}.$$

Assume in addition that we have observations from f randomly selected females, $y_1, \ldots, y_f \in \{0, 1\}$, where $y_i = 1$ if female number i is color blind, that is, if she has two CB-alleles. We will assume that the allele distribution in the total population satisfies the *Hardy-Weinberg equilibrium*, which means that the proportion of females with 2 CB-alleles is p^2, the proportion with 1 is $2p(1-p)$ and the proportion with 0 is $(1-p)^2$. We assume that the observations are independent and also independent of the male observations above.

Exercise 3.3.3. Argue that the probability that $y_i = 1$ equals p^2 and the probability that $y_i = 0$ equals $2p(1-p) + (1-p)^2 = (1-p)(1+p)$. Letting $y = (y_1, \ldots, y_f)$ argue that the probability of observing (x, y) equals

$$p^{m_b + 2f_b}(1+p)^{f_B}(1-p)^{m_B + f_B}$$

where $f_b = \sum_{i=1}^{f} y_i$ is the number of females with two CB-alleles (that are color blind) and $f_B = f - f_b$ is the number of females with at most one CB-allele. Having observed (x, y) find the likelihood function and the minus-log-likelihood function for p and show that the MLE equals

$$\hat{p} = \frac{-m_B + \sqrt{m_B^2 + 4n(m_b + 2f_b)}}{2n}$$

where $n = 2(f_b + f_B) + m_b + m_B = 2f + m$ is the total number of X chromosomes in the sample of males and females. In a study we find $f_b = 40$, $f_B = 9032$, $m_b = 725$ and $m_B = 8324$. Compute the MLE of p.

For the following exercises we consider n independent random variables $X_1, \ldots, X_n$ and we assume that the distribution of X_i is the Poisson distribution with mean value parameter

$$\lambda_i = e^{\beta y_i + \alpha}$$

where $y_1, \ldots, y_n \in \mathbb{R}$ are known but $\alpha, \beta \in \mathbb{R}$ are unknown. We have the observation $x = (x_1, \ldots, x_n) \in \mathbb{N}_0^n$ and the objective is to estimate α, β.

Exercise 3.3.4. Show that the minus-log-likelihood function is

$$l_x(\alpha, \beta) = \sum_{i=1}^{n} e^{\beta y_i + \alpha} - \beta x_i y_i - \alpha x_i + \log x_i!$$

Exercise 3.3.5. Fix β and show that for fixed β the minimum of the minus-log-likelihood function in α is

$$\hat{\alpha}(\beta) = \log \frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} e^{\beta y_i}}.$$

Exercise 3.3.6. Show that the profile minus-log-likelihood function in β is

$$l_x(\hat{\alpha}(\beta), \beta) = \sum_{i=1}^{n} x_i - \beta x_i y_i - \log \left(\frac{\sum_{j=1}^{n} x_i}{\sum_{j=1}^{n} e^{\beta y_j}} \right) x_i + \log x_i!$$

and that the minimizer of the profile minus-log-likelihood solves the equation

$$\frac{\sum_{i=1}^{n} y_i e^{\beta y_i}}{\sum_{i=1}^{n} e^{\beta y_i}} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i}.$$

Exercise 3.3.7. Implement the Newton-Raphson algorithm for solving the equation in β above and implement then a function for estimation of (α, β) for a given dataset $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$.

3.4   Hypothesis testing

If we have a statistical model on a sample space E given by a parametrized family of probability measures $(P_\theta)_{\theta \in \Theta}$ and if $\Theta_0 \subseteq \Theta$ is a subset of the parameter space we might naturally ask if the smaller model $(P_\theta)_{\theta \in \Theta_0}$ is in fact an adequate model?

A *statistical test* is a procedure so that we, for a given observation $x \in E$, can decide whether it is conceivable that the unknown parameter, θ, is to be found in Θ_0 or whether we have to use the entire parameter space Θ.

This is formalized as follows. We call the *hypothesis* that $\theta \in \Theta_0$ the *null-hypothesis* and write

$$H_0 : \theta \in \Theta_0.$$

We call $\theta \in \Theta \backslash \Theta_0$ the *alternative.*

A procedure for deciding whether the null-hypothesis is true can be stated as a division of the sample space E into two disjoint events A and R such that $E = A \cup R$. We call A the *acceptance region* and R the *rejection region*. If we observe $x \in R$ then we say that we *reject* the null-hypothesis and if $x \in A$ we say that we *accept* the null-hypothesis. The *power* for $\theta \in \Theta$ of the test is defined as

$$\beta(\theta) = P_\theta(R).$$

The *level* of the test is defined as

$$\alpha = \max_{\theta \in \Theta_0} \beta(\theta) = \max_{\theta \in \Theta_0} P_\theta(R)$$

as the maximal probability for rejecting the null-hypothesis over all possible choices of θ from the null-hypothesis. That is, α gives the largest probability of rejecting the null-hypothesis by mistake. For $\theta \in \Theta \backslash \Theta_0$ the power, $\beta(\theta)$, is the probability of correctly rejecting the null-hypothesis under the specific alternative θ.

A good test has small level α and large power $\beta(\theta)$ for all $\theta \in \Theta \backslash \Theta_0$. However, these two requirements are at odds with one another. If we enlarge the acceptance set, say, the level as well as the power goes down and if we enlarge the rejection set, the level as well as the power goes up.

In practice we always specify the acceptance and rejection regions via a *test statistic*. A test statistic is a function $h : E \to \mathbb{R}$, such that with a given choice of threshold $c \in \mathbb{R}$ the acceptance region is defined as

$$A = \{x \in E \mid h(x) \leq c\}.$$

This is called a one-sided test. Sometimes we encounter two-sided tests defined as

$$A = \{x \in E \mid c_1 \leq h(x) \leq c_2\}$$

for $c_1, c_2 \in \mathbb{R}$. Often the two-sided tests are symmetric with $c_1 = -c_2$ in which case we can just as well rephrase the test as a one-sided test with the test statistic $|h|$. Note that for the one-sided test we have $A = h^{-1}((-\infty, c])$, hence

$$P_\theta(A) = P_\theta(h^{-1}((-\infty, c])) = h(P_\theta)((-\infty, c])$$

by the definition of what the transformed probability measure $h(P_\theta)$ is. The point is that to compute the power and level of a test based on a test statistic h we need to know the transformed probability measure $h(P_\theta)$.

If we have a one-sided test statistic h we reject if $h(x)$ is too large. If we have settled for a level α-test we use the distribution of the test statistic to compute the corresponding threshold, which is the $(1-\alpha)$-quantile for the distribution, and which we often refer to as the level α *critical value* or just the critical value for short. If we reject a hypothesis we often say that the conclusion is *statistically significant* and the critical value is sometimes called the *significance level*.

Example 3.4.1. In Section 2.12 we considered local alignment of protein sequences. The problem can be formalized as a problem of testing a hypothesis. The setup is that the two sequences of letters $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ are sampled from some (joint) distribution. Our null hypothesis is that the two sequences are *independent*. The optimal local alignment score is our test statistic and large values are critical, thus it is a one-sided test. The problem is to find the distribution of the test statistic so that we can compute the critical value. To get a level α test we need to compute the $1 - \alpha$ quantile for the distribution of the test statistic.

In Section 2.12 we discussed the use of the Gumbel distribution as an approximation to the distribution of the optimal local alignment score. In this case the critical value becomes

$$c(\alpha) = -\frac{1}{\lambda} \log \left(-\frac{\log(1 - \alpha)}{Knm} \right).$$

$\diamond$

3.4.1 Two sample t-test

We develop in this section a classical test statistic, the two sample t-test, where we use Example 3.1.4 as inspiration. In that example we consider gene expression of a particular gene for two different groups of individuals. The sample space is $\mathbb{R}^{79}$ and the observations are the log-expression measurements.

The full model consists of specifying that the measurements are all assumed independent and that the distribution of the log-expressions are $N(\mu_1, \sigma_1^2)$ in group 1 and $N(\mu_2, \sigma_2^2)$ in group 2. Equivalently we can specify the model by saying that we observe the realization of $X_{i,j}$ for $i = 1$ and $j = 1, \ldots, 37$ or $i = 2$ and $j = 1, \ldots, 42$, and that

$$X_{i,j} = \mu_i + \sigma_i \varepsilon_{i,j}$$

where $\varepsilon_{i,j}$ are 79 iid random variables with the $N(0,1)$-distribution.

In a slightly more general framework there are n observations in the first group and m in the second and thus $n + m$ observations in total. The full model has a 4-dimensional parameter vector

$$\theta = (\mu_1, \sigma_1^2, \mu_2, \sigma_2^2) \in \Theta = \mathbb{R} \times (0, \infty) \times \mathbb{R} \times (0, \infty).$$

The interesting null-hypothesis is

$$H_0 : \mu_1 = \mu_2,$$

that is,

$$\Theta_0 = \{(\mu_1, \sigma_1^2, \mu_2, \sigma_2^2) \in \Theta \mid \mu_1 = \mu_2\}.$$

As in Example 3.1.4 we consider the estimators

$$\hat{\mu}_1 = \frac{1}{n} \sum_{j=1}^{n} X_{1,j} \quad \hat{\mu}_2 = \frac{1}{m} \sum_{j=1}^{m} X_{2,j},$$

which are also the maximum-likelihood estimators in this model according to Example 3.3.11.

If we consider the difference of the estimators we get that

$$\begin{aligned}
\hat{\mu}_1 - \hat{\mu}_2 &= \frac{1}{n} \sum_{j=1}^{n} X_{1,j} - \frac{1}{m} \sum_{j=1}^{m} X_{2,j} \\
&= \frac{1}{n} \sum_{j=1}^{n} (\mu_1 + \sigma_1 \varepsilon_{1,j}) - \frac{1}{m} \sum_{j=1}^{m} (\mu_2 + \sigma_2 \varepsilon_{2,j}) \\
&= \mu_1 - \mu_2 + \frac{\sigma_1}{n} \sum_{j=1}^{n} \varepsilon_{1,j} + \frac{\sigma_2}{m} \sum_{j=1}^{m} -\varepsilon_{2,j}.
\end{aligned}$$

Since all the $\varepsilon_{i,j}$'s are assumed independent and $N(0,1)$-distribution we can, using the result in Math Box 2.10.20, find that

$$\hat{\mu}_1 - \hat{\mu}_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}\right).$$

If we choose $\hat{\mu}_1 - \hat{\mu}_2$ as the test statistic – using a symmetric, two-sided test so that we reject if $|\hat{\mu}_1 - \hat{\mu}_2| > c$ – the normal distribution above tells us how to compute the power of the test and in particular the level. Under the null-hypothesis it holds that $\mu_1 - \mu_2 = 0$, but the distribution of the test statistic still depends upon the unknown parameters σ_1^2 and σ_2^2. If we thus want to choose the critical value c such that the probability of wrongly rejecting the null-hypothesis is $\leq \alpha$ we choose the

$1 - \alpha/2$-quantile for the normal distribution with mean 0 and variance $\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}$. Since this is a scale transformation of the $N(0, 1)$-normal distribution we can compute the quantile by Example 2.9.10 as

$$c(\alpha) = \sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}} z_{1-\alpha/2}$$

where $z_{1-\alpha/2}$ is the $1 - \alpha/2$-quantile for the $N(0, 1)$-distribution. For $\alpha = 0.05$ the quantile is 1.96 and for $\alpha = 0.01$ the quantile is 2.58.

We encounter a practical problem. Even if we have decided that a level 5% test is the relevant test to use we cannot compute the corresponding threshold for the given test statistic because it depends upon the unknown parameters σ_1^2 and σ_2^2. A widespread solution is to use the *plug-in principle* and simply plug in the estimators of the unknown parameters in the formula for the threshold. The resulting threshold becomes

$$\tilde{c}_\alpha = \sqrt{\frac{\tilde{\sigma}_1^2}{n} + \frac{\tilde{\sigma}_2^2}{m}} z_{1-\alpha/2},$$

and we reject the null-hypothesis if $|\hat{\mu}_1 - \hat{\mu}_2| > \tilde{c}_\alpha$.

Returning to the computations in Example 3.1.4 and taking $\alpha = 0.05$ the threshold becomes

$$\tilde{c}_\alpha = \sqrt{\frac{0.659}{37} + \frac{0.404}{42}} 1.96 = 0.32$$

and since the test statistic in this case equals 1.20 we reject the hypothesis that the group means are equal. Had we taken $\alpha = 0.01$ the threshold would be 0.46 and we would still reject. Another way of phrasing this is that the difference in the estimated means is large compared to the variance in the data – so large that it is very unlikely that it will be this large by chance if the mean value parameters are equal.

Because we use estimated values for the unknown variance parameters the computations above are not exact. Though we aim for a level of the test being α it is not necessarily precisely α. The problem is most pronounced when n and m are small, 2-5, say, in which case one can make a gross mistake.

We will show later that the empirical variance,

$$\tilde{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \hat{\mu})^2,$$

which is also the maximum-likelihood estimator in the normal model, systematically underestimates the variance parameter by a factor $(n - 1)/n$. This is something that will be made more precise later on. For the present discussion it leads to the alternative variance estimator

$$\hat{\sigma}^2 = \frac{1}{n - 1} \sum_{i=1}^{n} (X_i - \hat{\mu})^2,$$

which is by far the most used estimator of the variance in practice. This is for instance the estimator computed by `var` in R. When $n = 5$ the new estimator is a factor $5/4 = 1.25$ larger – a considerable amount. For large n the difference between the two estimators becomes negligible. Effectively, using the larger variance estimator increases the threshold and consequently our conclusions will be more conservative – we are less likely to reject the hypothesis that the mean value parameters are equal.

R Box 3.4.1 (T-test)**.** The two sample t-test can be carried out using the `t.test` function in R. There are two essentially different ways for using the function. If the data for the two groups are stored in the two vectors `x` and `y` you compute the t-test by

```
> t.test(x,y)
```

If the two vectors contain the gene expression measurements from the two groups considered in Example 3.1.4 the output is

```
        Welch Two Sample t-test

data:  x and y
t = 7.1679, df = 67.921, p-value = 7.103e-10
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.8678568 1.5374938
sample estimates:
mean of x mean of y
 8.536492  7.333816
```

Alternatively, the data for the two groups may be stored in a data frame `myData` with one column labeled `value`, say, containing the measurements/observations and another column labeled `groups`, say, which is a factor with two levels. Then we can compute the t-test by

```
> t.test(value~group,data=myData)
```

The default setting for `t.test` is to compute the t-test without assuming equal variances. We can specify equal variances by setting `var.equal=TRUE`.

There might, however, still be problems with using the approximation. In the attempt to remedy the problems we can choose to consider the t-test statistic

$$T = \frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{\frac{\hat{\sigma}_1^2}{n} + \frac{\hat{\sigma}_2^2}{m}}}.$$

Unfortunately we cannot obtain a useful, exact representation of the distribution of

this test statistic. A common approximation of the distribution of the t-test statistic *under the null-hypothesis* that $\mu_1 = \mu_2$ is as a t-distribution with *degrees of freedom*

$$\text{df} = \frac{(\hat{\sigma}_1^2/n + \hat{\sigma}_2^2/m)^2}{(\hat{\sigma}_1^2/n)^2/(n-1) + (\hat{\sigma}_2^2/m)^2/(m-1)}. \tag{3.11}$$

The t-distribution was considered in Exercise 2.6.4. The degrees of freedom equals twice the shape parameter of the t-distribution, that is $\text{df} = 2\lambda$ in the notation of Exercise 2.6.4. The t-distribution does not have other parameters and once we fix the degrees of freedom we can compute the $(1 - \alpha/2)$-quantile $w_{1-\alpha/2}$ and reject the hypothesis if

$$|T| > w_{1-\alpha/2},$$

which is equivalent to rejecting if

$$|\hat{\mu}_1 - \hat{\mu}_2| > \sqrt{\frac{\hat{\sigma}_1^2}{n} + \frac{\hat{\sigma}_2^2}{m}} \, w_{1-\alpha/2}.$$

For $\alpha = 0.05$ the $1 - \alpha/2 = 0.975$-quantile for the t-distribution with 2 degrees of freedom is 4.3, with 5 degrees of freedom it is 2.57 and with 20 degrees of freedom it is 2.09. It always holds that $w_{1-\alpha/2} \geq z_{1-\alpha/2}$ and the effective change of using the approximation with the t-distribution is once again that our conclusions get more conservative – we are less likely to reject the hypothesis that the mean value parameters are equal.

If we are willing to make one extra assumption, it is possible to derive exact results. If $\sigma_1^2 = \sigma_2^2$ and if we denote this common parameter by σ^2 we have reduced the model to a model with three parameters $(\mu_1, \mu_2, \sigma^2) \in \mathbb{R} \times \mathbb{R} \times (0, \infty)$. Under this assumption

$$\hat{\mu}_1 - \hat{\mu}_2 \sim N\left(\mu_1 - \mu_2, \sigma^2 \left(\frac{1}{n} + \frac{1}{m}\right)\right).$$

We choose to estimate σ^2 using the so-called *pooled variance estimator* defined as

$$\hat{\sigma}^2 = \frac{1}{n+m-2} \left(\sum_{j=1}^{n} (X_{1,j} - \hat{\mu}_1)^2 + \sum_{j=1}^{m} (X_{2,j} - \hat{\mu}_2)^2\right). \tag{3.12}$$

In this setup we introduce the t-test statistic

$$T = \frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{\hat{\sigma}^2 \left(\frac{1}{n} + \frac{1}{m}\right)}} = \sqrt{\frac{n+m}{nm}} \left(\frac{\hat{\mu}_1 - \hat{\mu}_2}{\hat{\sigma}}\right),$$

It can be shown that under the null-hypothesis that $\mu_1 = \mu_2$ the distribution of T is exactly a t-distribution with degrees of freedom $n + m - 2$.

If the degrees of freedom for the tests – with or without the assumption of equal variances – is large, the quantiles for the t-distribution are close to the corresponding

quantiles for the normal distribution. For doing a single test the practical difference by using the normal distribution is minor for $n \geq 20$.

To summarize the conclusions from the derivations above, we consider the null-hypothesis

$$H_0 : \mu_1 = \mu_2.$$

If we assume *equal* variances we compute the two-sample t-test

$$T = \sqrt{\frac{nm}{n+m}} \left(\frac{\hat{\mu}_1 - \hat{\mu}_2}{\hat{\sigma}} \right), \tag{3.13}$$

where $\hat{\sigma}^2$ is the pooled variance estimate given by (3.12) and we reject the hypothesis if $|T| > w_{1-\alpha/2}$ where $w_{1-\alpha/2}$ is the $1 - \alpha/2$ quantile for the t-distribution with $n + m - 2$ degrees of freedom.

If we do not assume equal variances we compute the two-sample t-test

$$T = \frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{\frac{\hat{\sigma}_1^2}{n} + \frac{\hat{\sigma}_2^2}{m}}} \tag{3.14}$$

and we reject the hypothesis if $|T| > w_{1-\alpha/2}$ where $w_{1-\alpha/2}$ is the $1 - \alpha/2$ quantile for the t-distribution with degrees of freedom given by (3.11). In this case where we do not assume equal variances the test is often referred to as the Welch two-sample t-test.

The relevant quantiles can be computed in R using the quantile function `qt` for the t-distribution. However, the t-test can also be computed using the `t.test` function. This function reports the conclusion of the test in terms of a *p-value*. If the computed t-test statistic is equal to t for the concrete dataset then the p-value is

$$p = \mathbb{P}(|T| > |t|)$$

where T has the t-distribution with df degrees of freedom. Alternatively, because the t-distribution is symmetric

$$p = 2(1 - F_{\mathrm{df}}(|t|))$$

where F_{df} is the distribution function for the t-distribution with df degrees of freedom. The null-hypothesis is rejected at level α if and only if the p-value is $\leq \alpha$.

There is one natural question left. Should we choose the Welch t-test or should we assume equal variances? Technically there is no problem in using the Welch t-test though the t-distribution used is not exact. If the estimated variances are close to each other there will only be minor differences between the Welch t-test and the equal variance t-test, and if they are not, the equal variance t-test is not appropriate. Should we actually make a formal statistical test of the hypothesis that the variance parameters σ_1^2 and σ_2^2 are equal? If we reject, then we use the Welch t-test, and otherwise we use the equal variance t-test. Such a procedure has been criticized in
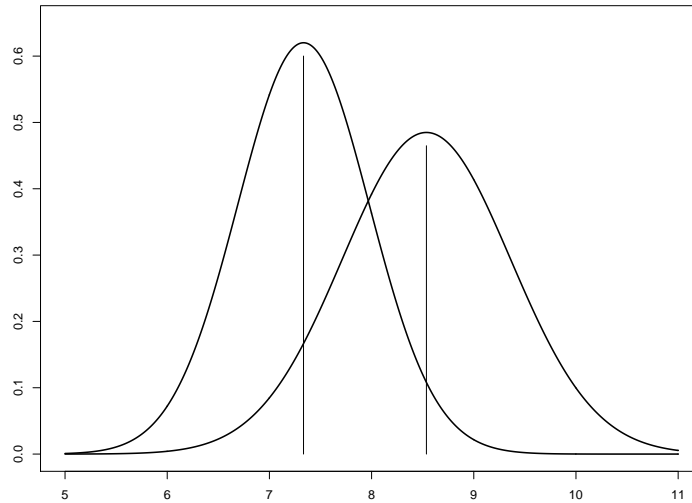
Figure 3.11: The densities for the normal distributions with estimated parameters for the gene expression data for gene 1635_at. We reject the hypothesis that the mean value parameters are equal.

the literature. An introduction of a test of equal variances, which dictates which t-test to use destroys properties of the subsequent t-tests and it turns out to be a better idea simply to use the Welch t-test. We should, however, be careful with the conclusions if the variances really are different. If we reject the hypothesis of equal means it is not so clear how we interpret that one of the means is larger than the other if also the variances differ. For the gene expression example the estimated means for group 1 is larger than for group 2 and we have rejected the hypothesis that they are equal. However, the estimated variance for group 1 is also larger than for group 2, which implies that the larger mean alone does not imply that the normal distribution we have fitted for group 1 is unequivocally "to the right" of the normal distribution fitted to group 2.

3.4.2 Likelihood ratio tests

In the following definition we suppose that X is a random variable with distribution P_θ for some $\theta \in \Theta_0$ and that we have a likelihood function $\mathcal{L}_x(\theta)$ for $\theta \in \Theta$ when observing $X = x$.

Definition 3.4.2. *Observing $X = x$*

$$Q(x) = \frac{\max_{\theta \in \Theta_0} \mathcal{L}_x(\theta)}{\max_{\theta \in \Theta} \mathcal{L}_x(\theta)}$$

is called the likelihood ratio test statistic. Since $\Theta_0 \subseteq \Theta$, we have $Q(x) \in (0, 1]$. Small values of $Q(x)$ are critical.

To use the test statistic above we need to know its distribution under P_θ for $\theta \in \Theta_0$. It is in general impossible to find this distribution, but in many situations of practical importance we can find a useful approximation. We state this as a theorem, though it is not precise in terms of the prerequisites required for the approximation to be valid.

Result 3.4.3. *If Θ is a d-dimensional parameter space and Θ_0 d_0-dimensional, the distribution of*

$$-2 \log Q(X)$$

can be approximated by a χ^2-distribution with $d - d_0$ degrees of freedom. Large values of $-2 \log Q(x)$ are critical.

Remark 3.4.4. The "dimension" of Θ and Θ_0 is a little too involved to define in a precise mathematical sense. It essentially covers the more intuitive idea of "the number of free parameters". In practice, it is often easy to compute the dimension drop $d - d_0$ as this is simply the number of (different) restrictions that we put on the parameters in Θ to get Θ_0. For instance, in the example above with the 2-sample t-test the dimension of Θ is 3 (or 4), the dimension of Θ_0 is 2 (or 3) and the dimension drop is 1.

If $\hat{\theta}(x)$ denotes the maximum-likelihood estimate and $\hat{\theta}_0(x)$ the maximum-likelihood estimate under the null hypothesis it follows that

$$-2 \log Q(x) = 2(-\log \mathcal{L}_x(\hat{\theta}_0(x)) + \log \mathcal{L}_x(\hat{\theta}(x))) = 2(l_x(\hat{\theta}_0(x)) - l_x(\hat{\theta}(x)))$$

Having computed $-2 \log Q(x)$ we often report the test by computing a p-value. If F_{df} denotes the distribution function for the χ^2-distribution with df degrees of freedom the p-value is

$$p = 1 - F_{df}(-2 \log Q(x)).$$

This is the probability for observing a value of $-2 \log Q(X)$ under the null-hypothesis that is as large or larger than the observed value $-2 \log Q(x)$.

Example 3.4.5. For the Kimura model in Example 3.3.18 we observe that the null-hypothesis

$$H_0 : \alpha = \beta$$

is equivalent to the Jukes-Cantor model. If we consider Segment A of the virus genome we can compute $\tilde{l}_x(\hat{\alpha}, \hat{\beta}) = 399.2$ and under the null-hypothesis $\tilde{l}_x(\hat{\alpha}_0, \hat{\alpha}_0) = 436.3$. We find that

$$-2 \log Q(x) = 74.2.$$

Under the null-hypothesis we make a single restriction on the parameters, and the p-value using a χ^2-distribution with 1 degree of freedom is 7.0×10^{-18}, which is

effectively 0. This means that we will by all standards reject the null-hypothesis over the alternative. That is to say, we reject that the Jukes-Cantor model is adequate for modeling the molecular evolution of Segment A. ◇

One word of warning. We computed $\tilde{l}_x$ above instead of the full minus-log-likelihood. We did so because the remaining part of the full minus-log-likelihood does not involve the parameters α and β and is unaffected by the hypothesis. All terms that remain constant under the full model and the hypothesis can always be disregarded as the difference in the computation of $-2\log Q$ is unaffected by these terms. However, be careful always to disregard the same terms when computing $l_x(\hat{\theta}_0(x))$ as when computing $l_x(\hat{\theta}(x))$.

Example 3.4.6. Continuing Example 3.3.18 we will investigate if the three segments can be assumed to have the same parameters. Thus for the full model we have six parameters $\alpha_A, \beta_A, \alpha_B, \beta_B, \alpha_C, \beta_C$, two for each segment. We set up the null-hypothesis

$$H_0 : \alpha_A = \alpha_B = \alpha_C, \quad \beta_A = \beta_B = \beta_C.$$

We find that $-2\log Q(x) = 6.619$, and since the full model has 6 free parameters and the model under the null-hypothesis has 2, we compute the p-value using the χ^2-distribution with 4 degrees of freedom. The p-value is 0.157 and we do not reject the hypothesis that all three segments have the same parameters. ◇

The methodology of statistical testing is very rigid. Formally we have to set up the hypothesis prior to considering the data, since testing a hypothesis that is formulated based on the data almost automatically demolishes the assumptions that are used to derive the distribution of the test statistics. This makes statistical testing most appropriate for confirmatory analyses where we know what to expect prior to the actual data analysis and want to confirm and document that our beliefs are correct. On the other hand, exploratory data analysis where we don't know in advance what to expect is an important part of applied statistics. Statistical testing is used anyway as an exploratory tool to investigate a range of different hypotheses, and there are numerous algorithms and ad hoc procedures for this purpose. The merits of such procedures are extremely difficult to completely understand. It is, however, always important to understand how to correctly interpret a hypothesis test and the conclusions we can draw. Disregarding whether we use test statistics for a less formal, exploratory analysis or a formal confirmatory analysis we have to remember that if we accept a null-hypotheses we do in fact have little evidence that the hypothesis is true. What we can conclude is that there is no evidence in the data for concluding that the hypothesis is false, which is a considerably vaguer conclusion! A hypothesis may be screamingly wrong even though we are unable to document it. If the test we use has little power against a particular alternative, it will be very difficult to detect such a deviation from the null-hypothesis. On the other hand, if we reject a null-hypothesis we may be rather certain that the null-hypothesis is false, but "how false" is it? If we have a large dataset we may be able to statistically

detect small differences that are of little practical relevance. Statistical significance provides documentation for rejecting a hypothesis, e.g. that there is a difference of the mean values for two groups, but does not in itself document that the conclusion is important or significant in the usual sense of the word, e.g. that the difference of the mean values is of any importance.

3.4.3 Multiple testing

One of the problems with formal statistical testing is that the more tests we do the less reliable are our conclusions. If we make a statistical test at a 5%-level there is 5% chance the we by mistake reject the hypothesis even though it is true. This is not a negligible probability but 5% has caught on in the literature as a suitable rule-of-thumb level. The problem is that if we carry out 100 tests at a 5%-level then we expect that 1 out of 20 tests, that is, 5 in total, reject the null-hypothesis even if it is true in all the 100 situations. What is perhaps even worse is that the probability of rejecting at least one of the hypothesis is in many cases rather large. If all the tests are *independent*, the number of tests we reject follows a binomial distribution with parameters $n = 100$ and $p = 0.05$, in which case the probability of rejecting at least one hypothesis if they are all true is $1 - (1 - 0.05)^{100} = 99.4\%$.

If we carry out 100 two-sample t-tests on different[4] datasets and find that at a 5% level we rejected in 4 cases the hypothesis that the means are equal, does this support a conclusion that the means are actually different in those 4 cases? No, it does not. If we reject in 30 out of the 100 cases we are on the other hand likely to believe that for a fair part of the 30 cases the means are actually different. The binomial probability of getting more than 10 rejections is 1.1% and getting more than 20 rejections has probability 2.0×10^{-8}. But for how many and for which of the 30 cases can we conclude that there is a difference? A natural thing is to order (the absolute value of) the test statistics

$$|t_{(1)}| \leq \ldots \leq |t_{(100)}|$$

and then take them from the top and down.

Example 3.4.7. We consider the ALL microarray dataset introduced in Example 2.7.3. There are a total of 12625 genes represented on the array. We do a t-test for each gene where we test if there is a difference in the mean value between those with the BCR/ABL fusion gene and those without. To facilitate the comparison with a single t-distribution we carry out the tests under the assumption of equal variances in the two groups. The top 10 list of t-tests are as follows:

---

[4]Formally we need independent datasets for some of the subsequent quantitative computations to be justified but qualitatively the arguments hold in a broader context.
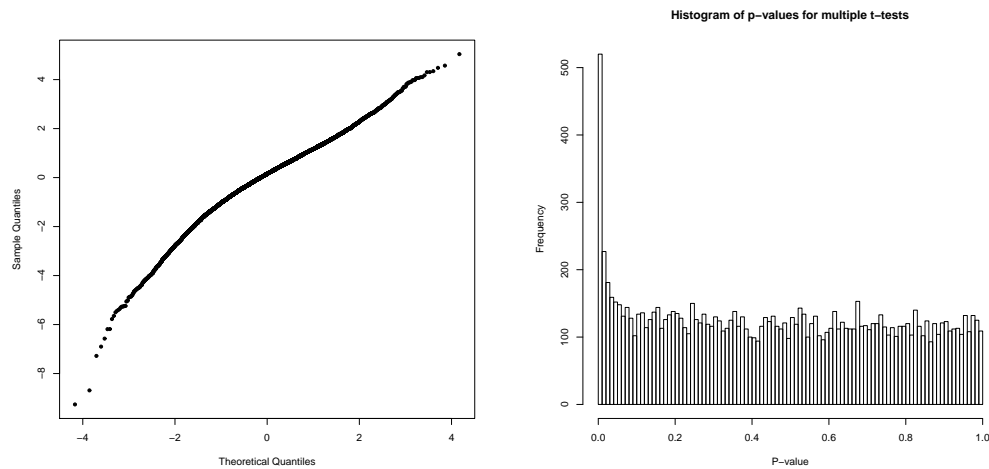
Figure 3.12: QQ-plot of the 12,625 t-test statistics for the ALL dataset (left) and histogram of the corresponding p-values.

name	t-test statistics	p-value
1636_g_at	-9.26	$3.76e - 14$
39730_at	-8.69	$4.79e - 13$
1635_at	-7.28	$2.45e - 10$
1674_at	-6.90	$1.28e - 09$
40504_at	-6.57	$5.27e - 09$
37015_at	-6.19	$2.74e - 08$
40202_at	-6.18	$2.79e - 08$
32434_at	-5.78	$1.54e - 07$
37027_at	-5.65	$2.60e - 07$
39837_s_at	-5.50	$4.74e - 07$

We find our earlier considered gene, 1635_at, as number three from the top on this list. Figure 3.12 shows that QQ-plot of the computed t-test statistics against the t-distribution with 77 degrees of freedom together with a histogram of the p-values. The QQ-plot bends in a way that indicates that there are too many large and small values in the sample. This is confirmed by the histogram of p-values, which shows that there are in the order of several hundred p-values too many in the range from 0 to 0.01. ◇

The conclusion in the example above is that there are a number of the cases where we should reject the hypothesis – even in the light of the fact that we do 12625 tests. The real question is that if we continued the list above, when should we stop? What should the threshold for the t-test statistic be in the light of the multiple tests carried out?

If we want to control the probability of including just a single wrong rejection we talk about controlling the *family wise error rate*. If all the tests are independent and each at level α the probability of not rejecting a single one if they are all true is $1 - (1 - \alpha)^n$ if we make n tests. If we want to keep this value at a 5% level, say, we solve and find that

$$\alpha = 1 - (1 - 0.05)^{1/n} = 1 - (0.95)^{1/n}.$$

With $n = 12625$ as in the example this gives an $\alpha = 4.1 \times 10^{-6}$. Thus each test has to be carried out at the level 4.1×10^{-6} to be sure not to reject a true hypothesis. This can be a very conservative procedure.

Current research suggests that for large, multiple testing problems focus should change from the family wise error rate to other quantities such as the *false discovery rate*, which is the relative number of falsely rejected hypotheses out of the total number of rejected hypotheses. The book *Multiple testing procedures with applications to genomics* by Dudoit and van der Laan (Springer, 2008) treats this and a number of other issues in relation to multiple testing problems. A very pragmatic viewpoint is that the multiple testing problem is a simple matter of choosing a suitable threshold to replace the critical value used for a single test. How to do this appropriately and how to interpret the choice correctly can be a much more subtle problem, but ordering the tests according to p-value is almost always a sensible thing to do.

Exercises

Exercise 3.4.1. Consider the setup for Exercise 3.3.4 and the null-hypothesis

$$H_0 : \beta = 0$$

Interpret the hypothesis and compute a formula for $-2 \log Q(x)$ for testing this hypothesis. What is the approximating distribution of this test statistic under the null-hypothesis?

Exercise 3.4.2. Make a simulation study to investigate the distribution of the test statistics $-2 \log Q(X)$ for the hypothesis considered in Example 3.4.5. That is, use the estimated Jukes-Cantor model to simulate new datasets, 200 say, compute the corresponding $-2 \log Q(x)$ statistics for each dataset and compare the resulting empirical distribution with the χ^2-distribution with 1 degree of freedom.

3.5   Confidence intervals

The formal statistical test answers a question about the parameter in terms of the data at hand. Is there evidence in the data for rejecting the given hypothesis about the parameter or isn't there? We can only deal with such a question in the light of the uncertainty in the data even if the hypothesis is true. The distribution of the test statistic captures this, and the test statistic needs to be sufficiently large compared to its distribution before we reject the hypothesis.

There is another, dual way of dealing with the uncertainty in the data and thus the uncertainty in the estimated parameters. If we consider a real valued parameter then instead of formulating a specific hypothesis about the parameter we report an interval, such that the values of the parameter in the interval are conceivable in the light of the given dataset. We call the intervals *confidence intervals*.

If $(P_\theta)_{\theta \in \Theta}$ is a parametrized family of probability measures on E, and if we have an observation $x \in E$, then an estimator $\hat{\theta} : E \to \Theta$ produces an estimate $\hat{\theta}(x) \in \Theta$. If the observation came to be as a realization of an experiment that was governed by one probability measure P_θ in our parametrized family (thus the *true* parameter is θ), then in most cases $\hat{\theta}(x) \neq \theta$ – but it is certainly the intention that the estimate and the true value should not be too far apart. We attempt here to quantify how far away from the estimate $\hat{\theta}(x)$ it is conceivable that the true value of the parameter is.

Example 3.5.1. Let $X_1, \ldots, X_n$ be iid Bernoulli distributed with success probability $p \in [0, 1]$. Our parameter space is $[0, 1]$ and the unknown parameter is the success probability p. Our sample space is $\{0, 1\}^n$ and the observation is an n-dimensional vector $x = (x_1, \ldots, x_n)$ of 0-1-variables. We will consider the estimator

$$\hat{p} = \frac{1}{n} \sum_{i=1}^{n} X_i,$$

which is the relative frequency of 1's (and the MLE as well). The distribution of $n\hat{p} = \sum_{i=1}^{n} X_k$ is a binomial distribution with parameters (n, p), which implicitly[5] gives the distribution of $\hat{p}$.

If we in this example take $z(p)$ and $w(p)$ to be the 0.025- and 0.975-quantiles for the binomial distribution with parameters (n, p) we know that

$$\mathbb{P}_p(z(p) \leq n\hat{p} \leq w(p)) \simeq 0.95.$$

The reason that we don't get exact equality above is that the binomial distribution is discrete, so the distribution function has jumps and we may not be able to obtain exact equality. If we now define

$$I(\hat{p}) = \{p \in [0, 1] \mid z(p) \leq n\hat{p} \leq w(p)\}$$

---

[5]The distribution of $\hat{p}$ is a distribution on $\{0, 1/n, 2/n, \ldots, 1\}$ – a set that changes with n – and the convention is to report the distribution in terms of $n\hat{p}$, which is a distribution on $\mathbb{Z}$.
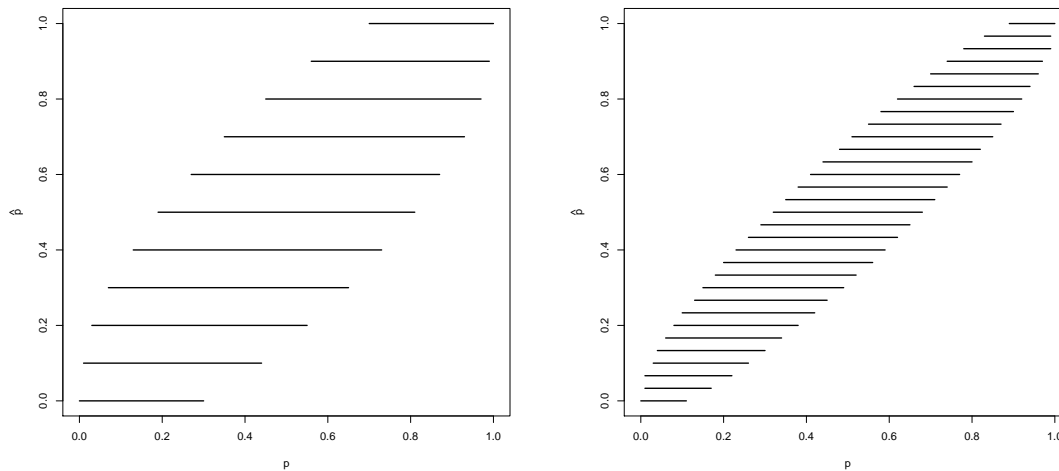
Figure 3.13: These figures show 95%-confidence intervals for the parameter p in the binomial distribution for different possible realizations of the estimator $\hat{p}$ (on the y-axis) with $n = 10$ (left) and $n = 30$ (right). For a given estimate $\hat{p}(x) = y$ we can read of which p (those on the line) that could produce such an estimate. Note the cigar shape.

we have $\mathbb{P}_p(p \in I(\hat{p})) \simeq 0.95$. We can find the interval $I(\hat{p})$ by reading it off from a figure as illustrated in Figure 3.13. Note that the probability statement is a statement about the random interval $I(\hat{p})$. It says that this random interval will contain the true parameter with probability 0.95 and we call $I(\hat{p})$ a 95%-confidence interval.

It will be shown in a later chapter that the variance of $n\hat{p}$ is $np(1 - p)$, and if we approximate the binomial distribution $B(n, p)$ with the $N(np, np(1-p))$ distribution, which will also be justified later, we arrive at the following convenient approximation

$$z(p) \simeq n\left(p - 1.96\frac{\sqrt{\hat{p}(1 - \hat{p})}}{\sqrt{n}}\right) \quad \text{and} \quad w(p) \simeq n\left(p + 1.96\frac{\sqrt{\hat{p}(1 - \hat{p})}}{\sqrt{n}}\right)$$

If we plug this approximation into the formula for the confidence interval we get

$$I(\hat{p})) = \left[\hat{p} - 1.96\frac{\sqrt{\hat{p}(1 - \hat{p})}}{\sqrt{n}}, \hat{p} + 1.96\frac{\sqrt{\hat{p}(1 - \hat{p})}}{\sqrt{n}}\right].$$

The approximation does not work well if n is too small or if the true p is too close to 0 or 1. $\diamond$

For a single parameter as in Example 3.5.1 we report an interval. How large the interval is depends upon how certain – or confident – we want to be that the true

parameter is in the interval. In the example we chose a 95%-confidence interval, which has become a de facto standard. For general parameters the following definition tells what a *confidence set* is.

Definition 3.5.2. *A confidence set for the parameter θ given the observation $x \in E$ is a subset $I(x) \subseteq \Theta$. If for each $x \in E$ we have given a confidence set $I(x)$ we say that $(I(x))_{x \in E}$ are $(1-\alpha)$-confidence sets for the unknown parameter if for all $\theta \in \Theta$*

$$\mathbb{P}_\theta(\theta \in I(X)) \geq 1 - \alpha. \tag{3.15}$$

We call $1 - \alpha$ the coverage of the confidence sets.

If $\Theta \subseteq \mathbb{R}$ and $I(x)$ is an interval we call $I(x)$ a confidence interval. We emphasize again that $\mathbb{P}_\theta(\theta \in I(X))$ is a probability statement prior to conducting the experiment about whether the random confidence set $I(X)$ will contain the parameter, and *not* whether the parameter belongs to the confidence set $I(x)$ after having observed the realization x of X. This is a very subtle point about the interpretation of confidence sets. It is the observation and therefore the confidence set that is a realization of the random experiment and not the unknown parameter. For a given realization x we simply can't tell whether $I(x)$ contains θ or not, since θ is unknown. But if we *a priori* to making the experiment decide upon a family of $(1 - \alpha)$-confidence sets that we will choose among depending on the observation, then we know that the probability that the confidence set we end up with really contains θ is at least $1 - \alpha$ no matter what θ is. If α is chosen small, $\alpha = 0.05$ say, then we are pretty confident that θ is actually in $I(x)$ and if $\alpha = 0.01$ we are even more so.

Example 3.5.3. We consider the statistical model for just group 1 in Example 3.1.4, which states that we observe

$$X_j = \mu + \sigma \varepsilon_j$$

for $j = 1, \ldots, 37$ where the ε_j's are iid $N(0,1)$. Assume for the sake of simplification that σ is known. The parameter is then just μ and the parameter space is $\mathbb{R}$, the sample space is $\mathbb{R}^{37}$ and we observe $x = (x_1, \ldots, x_{37})$.

As usual $\hat{\mu} = \frac{1}{37} \sum_{j=1}^{37} X_j$ and we introduce the statistic

$$h(X, \mu_0) = \hat{\mu} - \mu_0 = (\mu - \mu_0) + \sigma \frac{1}{37} \sum_{j=1}^{37} \varepsilon_j.$$

If $\mu = \mu_0$ the distribution of this statistic is $N(0, \frac{\sigma^2}{37})$ and if we recall that 1.96 is the 0.975-quantile for the $N(0,1)$ normal distribution

$$\mathbb{P}_{\mu_0} \left(|h(X, \mu_0)| \leq 1.96 \frac{\sigma}{\sqrt{37}} \right) = 0.95$$

We find that

$$
\begin{aligned}
I(x) &= \{\mu_0 \in \mathbb{R} \mid |h(X, \mu_0)| \le 1.96 \frac{\sigma}{\sqrt{37}}\} \\
&= \{\mu_0 \in \mathbb{R} \mid -1.96 \frac{\sigma}{\sqrt{37}} \le \hat{\mu} - \mu_0 \le 1.96 \frac{\sigma}{\sqrt{37}}\} \\
&= \left[\hat{\mu} - 1.96 \frac{\sigma}{\sqrt{37}}, \hat{\mu} + 1.96 \frac{\sigma}{\sqrt{37}}\right]
\end{aligned}
$$

If we plug in the estimated value of the standard deviation we get the 95%-confidence interval $[8.27, 8.80]$. Since the standard deviation is estimated and not known we violate the assumptions for the derivations above. Because there is an approximation involved we say that the confidence interval has *nominal* coverage 95% whereas the *actual* coverage may be lower. If n is not too small the approximation error is minor and the actual coverage is close to the nominal 95%. Below we treat a method that is formally correct – also if the variance is estimated. $\diamond$

Note the similarity of the construction above with a hypothesis test. If we formulate the simple null-hypothesis

$$H_0 : \mu = \mu_0,$$

we can introduce the test statistics $h(x, \mu_0)$ as above and reject the test if this test statistic is larger in absolute value than $1.96 \frac{\sigma}{\sqrt{37}}$. The 95%-confidence interval consists precisely of those μ_0 where the two-sided level 5% test based on the test statistic $h(x, \mu_0)$ will be accepted.

This duality between confidence intervals and statistical tests of simple hypotheses is a completely general phenomena. Consider a simple null-hypothesis

$$H_0 : \theta = \theta_0$$

then if $A(\theta_0)$ is an acceptance region for a level α test, the set

$$I(x) = \{\theta_0 \mid x \in A(\theta_0)\}$$

is a $(1 - \alpha)$-confidence set. This is because $\theta_0 \in I(x)$ if and only if $x \in A(\theta_0)$, hence

$$\mathbb{P}_{\theta_0}(X \in A(\theta_0)) = \mathbb{P}_{\theta_0}(\theta_0 \in I(X)).$$

This equality also implies that if $I(x)$ for $x \in E$ form $(1 - \alpha)$-confidence sets then the set

$$A(\theta_0) = \{x \in E \mid \theta_0 \in I(x)\}$$

forms an acceptance region for a level α test.

If $\Theta \subseteq \mathbb{R}$ we naturally ask what general procedures we have available for producing confidence intervals. We consider here intervals that are given by the test statistic

$$h(x, \theta_0) = \hat{\theta}(x) - \theta_0,$$

for any estimator $\hat{\theta}$ of the unknown parameter. With $\alpha \in (0,1)$ the fundamental procedure is to find formulas for $z_\alpha(\theta_0)$ and $w_\alpha(\theta_0)$, the $\alpha/2$ and $1-\alpha/2$ quantiles, for the distribution of $\hat{\theta}(x) - \theta_0$ under the probability measure P_{θ_0} and define the confidence set

$$I(x) = \{\theta_0 \in \Theta \mid z_\alpha(\theta_0) \le \hat{\theta}(x) - \theta_0 \le w_\alpha(\theta_0)\}.$$

This is what we did for the binomial and the normal distributions above. Unfortunately these are special cases, and even if we know the quantiles as a function of θ_0 it may not be practically possible to compute $I(x)$ above. In general the set does not even have to be an interval either! A computable alternative is obtained by plugging in the estimate of θ in the formulas for the quantiles above, which gives

$$I(x) = [\hat{\theta}(x) - w_\alpha(\hat{\theta}(x)), \hat{\theta}(x) - z_\alpha(\hat{\theta}(x))].$$

Exact distributions and thus quantiles are hard to obtain, and in most cases we have to rely on approximations. Below we present three of the most basic constructions of approximate $(1-\alpha)$-confidence intervals.

- Suppose we have a formula $\text{se}(\theta_0)$ for the standard deviation of $\hat{\theta}$ under P_{θ_0} – often referred to as the *standard error of* $\hat{\theta}$. Assume, furthermore, that the distribution of $\hat{\theta}(x) - \theta_0$ can be approximated by the $N(0, \text{se}(\theta_0)^2)$-distribution. With z_α the $1-\alpha/2$ quantile for the $N(0,1)$-distribution the general construction is

$$I(x) = \{\theta_0 \in \Theta \mid -\text{se}(\theta_0)z_\alpha \le \hat{\theta}(x) - \theta_0 \le \text{se}(\theta_0)z_\alpha\}.$$

As above, it may be practically impossible to compute $I(x)$ and it may not even be an interval. If we plug in the estimate of θ in the formula $\text{se}(\theta_0)$ we arrive at the interval

$$I(x) = [\hat{\theta}(x) - \text{se}(\hat{\theta}(x))z_\alpha, \hat{\theta}(x) + \text{se}(\hat{\theta}(x))z_\alpha].$$

- If $\hat{\theta}$ is the maximum likelihood estimator and the minus-log-likelihood function is twice differentiable then

$$\hat{i}(x) = \frac{\mathrm{d}^2 l_x}{\mathrm{d}\theta^2}(\hat{\theta}(x))$$

is known as the *observed Fisher information*, or just the observed information for short. The mean of

$$\frac{\mathrm{d}^2 l_X}{\mathrm{d}\theta^2}(\theta_0)$$

under P_{θ_0} is denoted $i(\theta_0)$ and is called the Fisher information. Under quite general conditions $1/\sqrt{i(\hat{\theta}(x))}$ or $1/\sqrt{\hat{i}(x)}$ are both valid estimates of the standard error of $\hat{\theta}$, and if we proceed as above we get using the latter estimate, say,

$$I(x) = \left[\hat{\theta}(x) - \frac{z_\alpha}{\sqrt{\hat{i}(x)}}, \hat{\theta}(x) + \frac{z_\alpha}{\sqrt{\hat{i}(x)}}\right].$$

- Estimates $\hat{z}_\alpha$ and $\hat{w}_\alpha$ of the quantiles $z_\alpha(\theta_0)$ and $w_\alpha(\theta_0)$ or an estimate, $\hat{se}$, of the standard error of $\hat{\theta}$ are found by simulations. This is known as *bootstrapping* and the technicalities will be pursued below. In any case, once the estimates have been computed one proceeds as above and computes either

$$I(x) = [\hat{\theta}(x) - \hat{w}_\alpha, \hat{\theta}(x) - \hat{z}_\alpha],$$

or

$$I(x) = [\hat{\theta}(x) - \hat{se}z_\alpha, \hat{\theta}(x) + \hat{se}z_\alpha].$$

As most of the practically usable constructions involve approximations, we say that the resulting confidence intervals have *nominal* coverage $1 - \alpha$. The actual coverage probability is the function

$$\theta \mapsto \mathbb{P}_\theta(\theta \in I(X)),$$

and the sets $I(x)$, $x \in E$, are $(1-\alpha)$-confidence sets if the actual coverage probability is larger than $1 - \alpha$ for all θ. The nominal value $1 - \alpha$ is what we aim for, the actual coverage is what we get. Hopefully, if the approximations are not too bad, the actual coverage is not much smaller than the nominal $1 - \alpha$.

Example 3.5.4. For the estimation of the probability parameter p for n iid Bernoulli random variables the MLE is

$$\hat{p} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

The minus-log-likelihood function is

$$l_x(p) = -N \log p - (n - N) \log(1 - p)$$

where $N = \sum_{i=1}^{n} X_i$. Differentiation gives

$$l_x'(p) = -\frac{N}{p} + \frac{n - N}{(1 - p)}, \qquad \hat{i}(p) = l_x''(p) = \frac{N}{p^2} + \frac{n - N}{(1 - p)^2} = \frac{N(1 - 2p) + np^2}{p^2(1 - p)^2}.$$

Since the mean of N is np the Fisher information is

$$i(p) = \frac{np(1 - 2p) + np^2}{p^2(1 - p)^2} = \frac{n((1 - 2p) + p)}{p(1 - p)^2} = \frac{n}{p(1 - p)}.$$

If we use $1/\sqrt{i(\hat{p})}$ as an estimate of the standard error of $\hat{p}$ and we proceed with the standard construction of a $(1 - \alpha)$-confidence interval we get

$$\left[\hat{p} - \frac{\sqrt{\hat{p}(1 - \hat{p})}}{\sqrt{n}} z_\alpha, \hat{p} + \frac{\sqrt{\hat{p}(1 - \hat{p})}}{\sqrt{n}} z_\alpha\right]$$

where z_α is the $(1 - \alpha/2)$-quantile for the normal distribution $N(0, 1)$. We find that this interval is identical to the approximate interval considered in Example 3.5.1. For the binomial distribution we will not use the observed Fisher information because we have a formula for the Fisher information. The observed information is useful when we don't have such a formula.   ◇
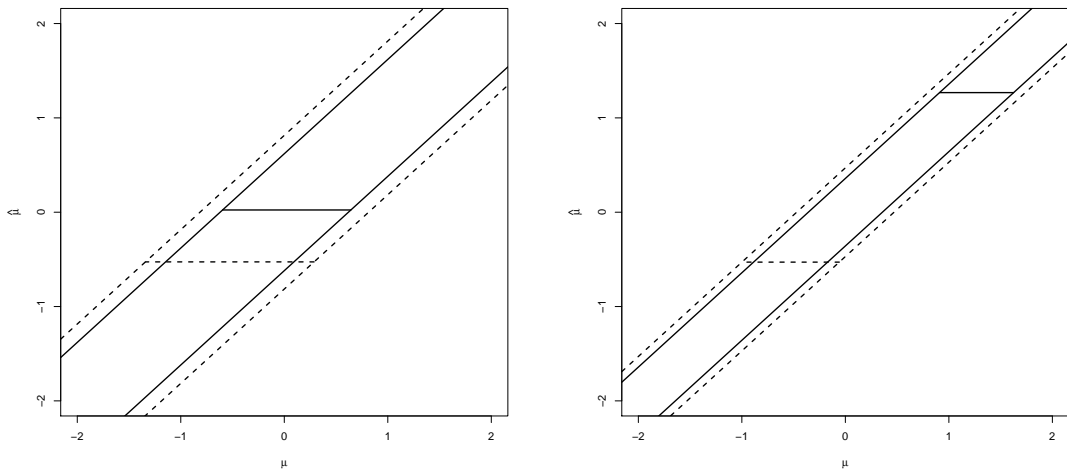
Figure 3.14: These figures show 95%- and 99%-confidence intervals for the parameter of interest μ in the normal distribution $N(\mu, \sigma^2)$ with $n = 10$ (left) or $n = 30$ (right) independent observations. On the figure $\sigma^2 = 1$. Reading the figure from the x-axis the full lines give the 0.025- and 0.975 quantiles for the normal distribution with variance $1/\sqrt{n}$ and mean value parameter μ, the dashed lines give the 0.005- and 0.995 quantiles. For a given estimate $\hat{\mu}(x) = y$ we can read the figure from the y-axis and read of which μ that could produce such an estimate. This gives the confidence intervals.

3.5.1 Parameters of interest

In the discussion above we do in reality only treat the situation with a single, univariate real parameter. In situations with more then one parameter we gave abstract definitions of confidence sets but we did not provide any practical methods. Though it is possible – also computationally and in practice – to work with confidence sets for more than a univariate parameter, such sets are notoriously difficult to relate to. When we have more than one unknown parameter we usually focus on univariate parameter transformations – the *parameters of interest*.

In general, if $\tau : \Theta \to \mathbb{R}$ is any map from the full parameter space into the real line, and if we are really interested in $\tau = \tau(\theta)$ and not so much θ, we call τ the *parameter of interest*. If $\hat{\theta}$ is an estimator of θ, then $\hat{\tau} = \tau(\hat{\theta})$ can be taken as an estimator of τ. This is the *plug-in* principle.

Example 3.5.5. We consider the statistical model of $X_1, \ldots, X_n$ where we assume that the variables are iid with the $N(\mu, \sigma^2)$-distribution. The parameter space is $\Theta = \mathbb{R} \times (0, \infty)$ but we are often mostly interested in the mean value parameter μ.

Thus we consider the parameter of interest given by

$$\tau(\mu, \sigma^2) = \mu$$

where $\tau : \mathbb{R} \times (0, \infty) \to \mathbb{R}$.

When considering any (joint) estimator $(\hat{\mu}, \hat{\sigma}^2)$ of (μ, σ^2) the plug-in estimator of μ is simply $\hat{\mu}$ – that is, nothing happens. The point here is that though the estimator does not seem to have anything to do with whether or not σ is estimated, the distribution of $\hat{\mu}$ will in general depend upon the complete, unknown parameter – in this case (μ, σ) – and not just the unknown parameter of interest itself. We have to take this into account in the construction of confidence intervals. $\diamond$

Example 3.5.6. Continuing Example 3.1.4 – and the discussion in Section 3.4.1 – we considered gene expression measures for two groups of individuals. Taking logarithms we considered the use of an additive noise model, where we assumed that the variables within each groups are iid $N(\mu_i, \sigma_i^2)$.

For the comparison of gene expression measurements – and many other assay based, quantitative measurements of concentrations – the *fold-change* between the groups is commonly regarded as an interesting parameter. Because we use the additive model on the log-measurements (base 2), the fold change is defined as

$$\tau(\mu_1, \sigma_1^2, \mu_2, \sigma_2^2) = 2^{\mu_1 - \mu_2}.$$

Thus a difference in mean of 1 corresponds to a fold-change of 2. A difference in mean of 2 corresponds to a fold-change of 4 etc. $\diamond$

Example 3.5.7. If P_θ for $\theta \in \Theta$ is a parametrized family of probability measures on E, a recurring problem is to estimate the probability of a specific event $A \subseteq E$. We define

$$\tau(\theta) = P_\theta(A).$$

Our parameter of interest is then the probability of this particular event.

If $X_1, \ldots, X_n$ are iid random variables taking values in E and with distribution P_θ for some $\theta \in \Theta$, we may estimate $\tau(\theta)$ directly as the relative frequency

$$\varepsilon_n(A) = \frac{1}{n} \sum_{i=1}^{n} 1(X_i \in A).$$

This is *not* the plug-in estimator – the plug-in estimator requires that we have an estimator, $\hat{\theta}$, of θ and then use $\tau(\hat{\theta}) = P_{\hat{\theta}}(A)$ as an estimator of $\tau(\theta)$. If necessary, $P_{\hat{\theta}}(A)$ can be computed via simulations.

$\diamond$

Example 3.5.8. The logistic regression model for $X_1, \ldots, X_n$ with $y_1, \ldots, y_n$ fixed, as considered in Example 3.3.17, is given by the point probabilities

$$\mathbb{P}(X_i = 1) = \frac{\exp(\alpha + \beta y_i)}{1 + \exp(\alpha + \beta y_i)}$$

for parameters $\alpha, \beta \in \mathbb{R}$. One interpretation of this parametrization is that the log odds are linear in y. That is, with $p(y) = \frac{\exp(\alpha+\beta y)}{1+\exp(\alpha+\beta y)}$,

$$\log \frac{p(y)}{1-p(y)} = \alpha + \beta y.$$

The log odds equal 0 precisely when $p(y) = 1/2$ and this happens when $y = -\alpha/\beta$. The value of y where $p(y) = 1/2$ is called LD_{50}, which means the *Lethal Dose* for 50% of the subjects considered. In other words, the dose that kills half the flies. We see that in terms of the parameters in our logistic regression model

$$\text{LD}_{50} = -\frac{\alpha}{\beta}.$$

It is often of greater interest to estimate LD_{50} than α or β separately. Thus the *parameter of interest* is in this case LD_{50}. For the flies in Example 3.3.17 we find that

$$\text{LD}_{50} = -1.899.$$

$\diamond$
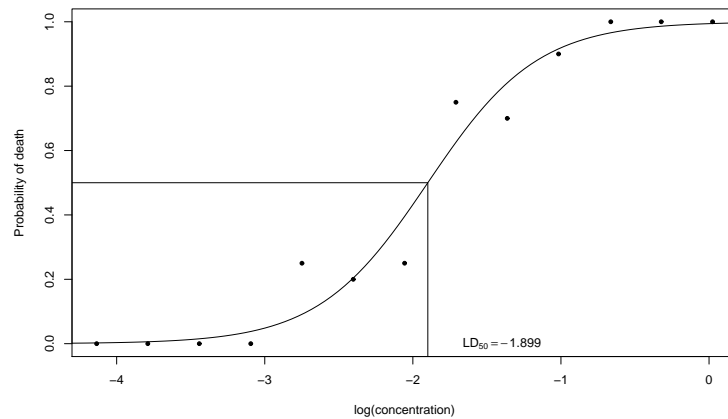
Figure 3.15: In the fly death experiment we estimate the parameter of interest, LD_{50}, to be -1.899. Thus a concentration of $0.1509 = \exp(-1.899)$ is found to be lethal for half the flies.

If we use the plug-in principle for estimation of the parameter of interest, we can proceed and try to find the distribution of the statistic

$$\hat{\tau} - \tau_0 = \tau(\hat{\theta}) - \tau(\theta_0).$$

We observe that the distribution of $\tau(\hat{\theta})$ is the transformation via τ of the distribution of $\hat{\theta}$. If we are capable of finding this distribution, we can essentially use the

procedure discussed above. However, we need a minor modification of the definition of confidence sets when we consider a parameter of interest instead of the complete parameter.

Definition 3.5.9. *A confidence set for a real valued parameter of interest $\tau = \tau(\theta)$ given the observation $x \in E$ is a subset $I(x) \subseteq \mathbb{R}$. If we for each $x \in E$ have a given confidence set $I(x)$ we say that $(I(x))_{x \in E}$ are $(1 - \alpha)$-confidence sets for the unknown parameter of interest if for all $\theta \in \Theta$*

$$\mathbb{P}_\theta(\tau(\theta) \in I(X)) \geq 1 - \alpha. \tag{3.16}$$

We call $1 - \alpha$ the coverage of the confidence sets.

For the practical construction of confidence intervals we can proceed in ways very similar to those in the previous section. We summarize the practically applicable methods below – noting that for the constructions below it is *not* in general an assumption that the estimator of τ is the plug-in estimator, but for some of the constructions we still need an estimator $\hat{\theta}$ of the full parameter.

- We have formulas $z_\alpha(\theta_0)$ and $w_\alpha(\theta_0)$ for the $\alpha/2$ and $1 - \alpha/2$ quantiles for the distribution of $\hat{\tau} - \tau$ in which case we can compute the interval

$$I(x) = [\hat{\tau}(x) - w_\alpha(\hat{\theta}(x)), \hat{\tau}(x) - z_\alpha(\hat{\theta}(x))].$$

- We have an estimator $\hat{se}$ of the standard error of $\hat{\tau}$. With z_α the $1 - \alpha/2$ quantile for the $N(0,1)$-distribution we can compute the interval

$$I(x) = [\hat{\tau}(x) - \hat{se}z_\alpha, \hat{\tau}(x) + \hat{se}z_\alpha].$$

  If we have a formula $se(\theta_0)$ for the standard error of $\hat{\tau}$ under P_{θ_0} we can use the plug-in estimator $\hat{se} = se(\hat{\theta}(x))$. If $\hat{\theta}$ is the maximum-likelihood estimator and $\hat{\tau} = \tau(\hat{\theta})$ is the plug-in estimator, an estimator, $\hat{se}$, is obtainable in terms of the Fisher information, see Math Box 4.7.3.

- Estimates $\hat{z}_\alpha$ and $\hat{w}_\alpha$ of the quantiles $z_\alpha(\theta_0)$ and $w_\alpha(\theta_0)$ or an estimate, $\hat{se}$, of the standard error of $\hat{\tau}$ are found by bootstrapping and we compute

$$I(x) = [\hat{\tau}(x) - \hat{w}_\alpha, \hat{\tau}(x) - \hat{z}_\alpha],$$

  or

$$I(x) = [\hat{\tau}(x) - \hat{se}z_\alpha, \hat{\tau}(x) + \hat{se}z_\alpha].$$

Example 3.5.10. Just as in Example 3.5.3 we consider the statistical model specified as

$$X_j = \mu + \sigma\varepsilon_j$$

> **Math Box 3.5.1** (Multivariate information). If $\Theta \subseteq \mathbb{R}^d$ and if the minus-log-likelihood function as a function of d variables is twice differentiable, the second derivative is the $d \times d$ matrix denoted $\mathrm{D}^2 l_X(\theta)$, cf. Math Box 3.3.1.
>
> The Fisher $d \times d$ information matrix, $I(\theta)$, is the entry-by-entry mean value of $\mathrm{D}^2 l_X(\theta)$. If $\hat{\theta}$ denotes the MLE it is under quite general conditions possible to show that
> $$\hat{\theta} - \theta \overset{\text{approx}}{\sim} N\left(0, I(\theta)^{-1}\right)$$
> under P_θ where $I(\theta)^{-1}$ is the matrix-inverse of $I(\theta)$. The distribution is the multi-variate normal distribution considered in Math Box 2.13.1.
>
> An important consequence is that if $v \in \mathbb{R}^d$ then $\tau(\theta) = v^T \theta$ is a one-dimensional parameter transformation (take $v^T = (1, 0, \ldots, 0)$ to select the first coordinate of $\hat{\theta}$, say) and
> $$v^T \hat{\theta} - v^T \theta \overset{\text{approx}}{\sim} N\left(0, v^T I(\theta)^{-1} v\right).$$
>
> The i'the coordinate in $\hat{\theta}$ thus follows approximately a normal distribution with variance $(I(\theta)^{-1})_{ii}$.
>
> More generally, if $\tau : \Theta \to \mathbb{R}$ is differentiable
> $$\tau(\hat{\theta}) - \tau(\theta) \overset{\text{approx}}{\sim} N\left(0, D\tau(\theta)^T I(\theta)^{-1} D\tau(\theta)\right).$$
>
> The standard error of $\tau(\hat{\theta})$ can be estimated in two ways. First, via the plug-in method
> $$\hat{\text{se}} = \sqrt{D\tau(\hat{\theta})^T I(\hat{\theta})^{-1} D\tau(\hat{\theta})},$$
> which requires that we have a formula for $I(\theta)$. Alternatively, we can estimate the Fisher information $I(\theta)$ by the observed Fisher information $\mathrm{D}^2 l_X(\hat{\theta})$ in which case we get
> $$\hat{\text{se}} = \sqrt{D\tau(\hat{\theta})^T \mathrm{D}^2 l_X(\hat{\theta})^{-1} D\tau(\hat{\theta})}.$$

for $j = 1, \ldots, n$ where the ε_j's are iid $N(0, 1)$, but we do not assume that σ is known. The parameter of interest is μ and we seek a $(1 - \alpha)$-confidence interval. The parameter space is $\mathbb{R} \times (0, \infty)$ and we use the MLE $\hat{\mu} = \frac{1}{n} \sum_{j=1}^{n} X_j$ and the usual modification

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_{j=1}^{n} (X_j - \hat{\mu})^2$$

of the MLE as the estimator of σ^2. The computations in Example 3.5.3 combined with a plug-in of the estimate of the unknown standard deviation lead to the interval

$$\left[\hat{\mu} - \frac{\hat{\sigma}}{\sqrt{n}} z_\alpha, \hat{\mu} + \frac{\hat{\sigma}}{\sqrt{n}} z_\alpha\right]$$

with z_α the $(1 - \alpha/2)$-quantile for the $N(0, 1)$-distribution. This interval is identical to the general approximation constructions when we use $\hat{\sigma}/\sqrt{n}$ as an estimate of the

standard error of $\hat{\mu}$.

For this particular case we can come up with an exact confidence interval by considering a special statistic. It is useful to recall the relation between confidence intervals and the null-hypotheses of the form

$$H_0 : \mu = \mu_0.$$

If we introduce the t-test statistic

$$T = \frac{\sqrt{n}(\hat{\mu} - \mu_0)}{\hat{\sigma}}$$

it can be shown that under the null-hypothesis H_0 it has a t-distribution with $n - 1$ degrees of freedom – cf. also Section 3.4.1. If w_α denotes the $(1 - \alpha/2)$-quantile for the t-distribution the resulting confidence interval becomes

$$
\begin{aligned}
I(x) &= \{\mu_0 \in \mathbb{R} \mid \frac{\sqrt{n}|\hat{\mu} - \mu_0|}{\hat{\sigma}} \leq w_\alpha\} \\
&= \left[\hat{\mu} - \frac{\hat{\sigma}}{\sqrt{n}}w_\alpha, \hat{\mu} + \frac{\hat{\sigma}}{\sqrt{n}}w_\alpha\right].
\end{aligned}
$$

These intervals are exact under the assumption of iid normally distributed observations, and in this case the actual coverage is $(1 - \alpha)$.

If we return to the gene expression data that we also considered in Example 3.5.3 we found the approximate 95%-confidence interval $[8.27, 8.80]$. There are 37 observations and using the t-distribution with 36 degrees of freedom instead of the approximating normal distribution the quantile changes from 1.96 to 2.03 and the confidence interval changes to $[8.26, 8.81]$. Thus by using the more conservative t-distribution the confidence interval is increased in length by roughly 3.5%. $\diamond$

The second construction based on the t-statistic in Example 3.5.10 does not fit in among the bullet points above. When τ is the parameter of interest the general t-statistic is

$$t = t(x, \tau) = \frac{\hat{\tau}(x) - \tau}{\text{se}(\hat{\theta}(x))}$$

where $\text{se}(\theta_0)$ is the standard error of $\hat{\tau}$ under P_{θ_0} and $\hat{\theta}$ is an estimator of θ. Using this statistic requires first of all that we have a formula for the standard error. If we approximate the distribution of the t-statistic by a $N(0, 1)$-distribution, the resulting confidence intervals

$$I(x) = \{\tau \in \mathbb{R} \mid |t(x, \tau)| \leq z_\alpha\} = [\hat{\tau}(x) - \text{se}(\hat{\theta}(x))z_\alpha, \hat{\tau} + \text{se}(\hat{\theta}(x))z_\alpha]$$

where z_α is the $1 - \alpha/2$-quantile for the $N(0, 1)$-distribution are the same as we have under the second bullet point. When the parameter of interest is the mean value parameter μ for the normal distribution we found in Example 3.5.10 the formula

$$\text{se}(\mu, \sigma^2) = \frac{\sigma}{\sqrt{n}}$$

for the standard error. For this particular example – as we already discussed – it is possible theoretically to find the exact distribution of the t-statistics. As stated, it is a t-distribution with $n-1$ degrees of freedom. The consequence is that the quantile z_α from the normal distribution is replaced by a quantile w_α from the t-distribution. It happens so that $w_\alpha \geq z_\alpha$ but that the quantile for the t-distribution approaches z_α when n gets large. Using the exact t-distribution gives systematically wider confidence intervals. It is not uncommon to encounter the standard confidence interval construction as above but where the quantile from the normal distribution has been replaced by the quantile from the t-distribution, also in situations where there is no theoretical reason to believe that the t-distribution is a better approximation than the normal distribution. It is difficult to say if such a practice is reasonable, but since the resulting confidence intervals get systematically larger this way, we can regard the procedure as being conservative.

Example 3.5.11. Consider the setup in Section 3.4.1 with two groups of independent normally distributed observations where the full parameter set is

$$\Theta = \mathbb{R} \times (0, \infty) \times \mathbb{R} \times (0, \infty).$$

We consider here the parameter of interest

$$\tau(\mu_1, \sigma_1^2, \mu_2, \sigma_2^2) = \mu_1 - \mu_2$$

being the difference of the two means.

Under the hypothesis

$$H_0 : \mu_1 = \mu_2 + \delta$$

the distribution of the statistic

$$\hat{\mu}_1 - \hat{\mu}_2$$

is $N(\delta, \sigma_1^2/n + \sigma_2^2/m)$. The parameter δ is the parameter of interest and the standard, approximate $(1 - \alpha)$-confidence interval for δ is

$$\left[\hat{\mu}_1 - \hat{\mu}_2 - \sqrt{\frac{\hat{\sigma}_1^2}{n} + \frac{\hat{\sigma}_2^2}{m}} z_\alpha, \hat{\mu}_1 - \hat{\mu}_2 + \sqrt{\frac{\hat{\sigma}_1^2}{n} + \frac{\hat{\sigma}_2^2}{m}} z_\alpha \right]$$

where z_α is the $(1 - \alpha/2)$-quantile for the $N(0, 1)$-distribution.

For this particular case we can instead use the t-distribution approximation of the T-test statistic as considered in Section 3.4.1, which will lead to slightly more conservative confidence intervals as we replace the quantile z_α by the corresponding quantile for the t-distribution. For the gene expression data considered in Section 3.4.1 the 95%-confidence interval for the difference based on the normal distribution becomes $[0.874, 1.532]$. Using the t-distribution with 67.9 degrees of freedom the interval becomes $[0.868, 1.537]$. The latter interval is also reported by the `t.test` function in R – see R Box 3.4.1. If the fold-change is the parameter of interest

rather than the difference in means, we can obtain (nominal) 95%-confidence intervals as $[2^{0.874}, 2^{1.532}] = [1.83, 2.89]$ using the quantile from the normal distribution and $[2^{0.868}, 2^{1.537}] = [1.82, 2.90]$ using the t-distribution.

If we make the assumption of equal variances in the two groups, the t-distribution becomes exact. The confidence interval based on the t-distribution with $37 + 42 - 2 = 77$ degrees of freedom is $[0.874, 1.531]$. The general formula for the $(1 - \alpha/2)$-confidence interval for the difference δ under the assumption of equal variances is

$$\left[\hat{\mu}_1 - \hat{\mu}_2 - \sqrt{\frac{n + m}{nm}} \hat{\sigma} w_\alpha, \hat{\mu}_1 - \hat{\mu}_2 + \sqrt{\frac{n + m}{nm}} \hat{\sigma} w_\alpha \right]$$

where w_α is the $(1 - \alpha/2)$-quantile for the t-distribution with $n + m - 2$ degrees of freedom. Here $\hat{\sigma}^2$ is the pooled variance estimator. $\diamond$

We have in this section focused on general confidence intervals based on the statistic $\hat{\tau} - \tau$. The t-statistic is an example of another possible choice of statistic useful for confidence interval constructions. There are numerous alternatives in the literature for choosing a suitable statistic $h(x, \tau)$, but we will not pursue these alternatives here.

3.6   Regression

Regression models form the general class of models that is most important for statistical applications. It is primarily within the framework of regression models that we treat *relations* among several observables. Sometimes these relations are known up to a small number of parameters from a given scientific theory in the subject matter field. In other cases we try to establish sensible relations from the data at hand.

We specify the general regression model for a real valued observable X given y by the scale-location model

$$X = g_\beta(y) + \sigma\varepsilon$$

where ε is a random variable with mean 0 and variance 1. Here $y \in E$ for some sample space E and

$$g_\beta : E \to \mathbb{R}$$

is a function parametrized by β. The full parameter is $\theta = (\beta, \sigma) \in \Theta$, where we will assume that $\Theta \subseteq \mathbb{R}^d \times (0, \infty)$, thus the β-part of the parametrization is a d-dimensional real vector. The variable y has many names. Sometimes it is called the independent variable – as opposed to X which is then called the dependent variable – in other situations y is called a covariate. We will call y the regressor to emphasize that the observable X is regressed on the regressor y.

If ε has distribution with density f and if we observe $X_1, \ldots, X_n$ such that

$$X_i = g_\beta(y_i) + \sigma \varepsilon_i$$

where $\varepsilon_1, \ldots, \varepsilon_n$ are iid and $y_1, \ldots, y_n \in E$ then $X_1, \ldots, X_n$ become independent but in general *not* identically distributed. The value of y_i dictates through the function g_β the mean value of X_i.

We find that the general joint density for the distribution of $X_1, \ldots, X_n$ is

$$f_{\beta,\sigma}(x_1, \ldots, x_n) = \prod_{i=1}^n \frac{1}{\sigma} f\left(\frac{x_i - g_\beta(y_i)}{\sigma}\right). \tag{3.17}$$

The minus-log-likelihood function becomes

$$l_x(\beta, \sigma) = n \log \sigma - \sum_{i=1}^n \log f\left(\frac{x_i - g_\beta(y_i)}{\sigma}\right). \tag{3.18}$$

Maximum-likelihood estimation, for the regression model at hand, can thus be boiled down to a matter of minimizing the function above. We elaborate on this in the case where the distribution of ε is assumed to be the normal distribution. In that case

$$l_x(\beta, \sigma) = n \log \sigma + \frac{1}{2\sigma^2} \underbrace{\sum_{i=1}^n (x_i - g_\beta(y_i))^2}_{\text{RSS}(\beta)} + n \log \sqrt{2\pi}.$$

We can observe that for fixed σ this minus-log-likelihood is minimized by minimizing the quantity $\text{RSS}(\beta)$, and the resulting minimizer does not depend upon σ. We can summarize the conclusion of this analysis as follows. For the regression model with a normally distributed noise term the MLE exists if and only if

$$\text{RSS}(\beta) = \sum_{i=1}^n (x_i - g_\beta(y_i))^2 \tag{3.19}$$

has a unique minimizer $\hat{\beta}$ in which case

$$\tilde{\sigma}^2 = \frac{1}{n} \text{RSS}(\hat{\beta})$$

is the corresponding MLE of the variance σ^2. We call $\text{RSS}(\beta)$ the *residual sum of squares*.

It is quite common to use this MLE of β based on the normal distribution even in situations where the assumption of a normal distribution does not hold. After all, minimizing the residual sum of squares $\text{RSS}(\beta)$ is a sensible thing to do. If we estimate β by minimizing $\text{RSS}(\beta)$ we often talk about the *least squares estimator*.
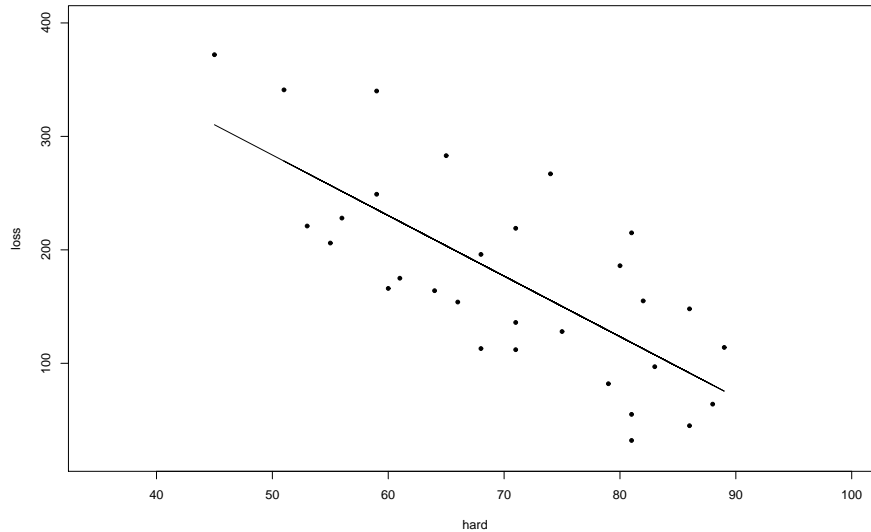
Figure 3.16: Scatter plot of the relation between hardness and abrasion loss for tires
and the straight line estimated by least squares regression.

The least squares estimator and the MLE are identical for the normal distribution
but otherwise not.

The variance is on the other hand typically *not* estimated by the MLE above – not
even if the assumption of the normal distribution holds. Instead we use the estimator

$$\hat{\sigma}^2 = \frac{1}{n-d}\text{RSS}(\hat{\beta})$$

where d is the dimension of the parameter space for β.

3.6.1   Ordinary linear regression

A linear regression model is obtained when $E = \mathbb{R}$, $d = 2$, $\beta = (\beta_0, \beta_1)$ and

$$g_\beta(y) = \beta_0 + \beta_1 y.$$

The parameter β_0 is called the *intercept* and β_1 the *slope*.

The residual sum of squares reads

$$\text{RSS}(\beta) = \sum_{i=1}^{n}(x_i - \beta_0 - \beta_1 y_i)^2.$$
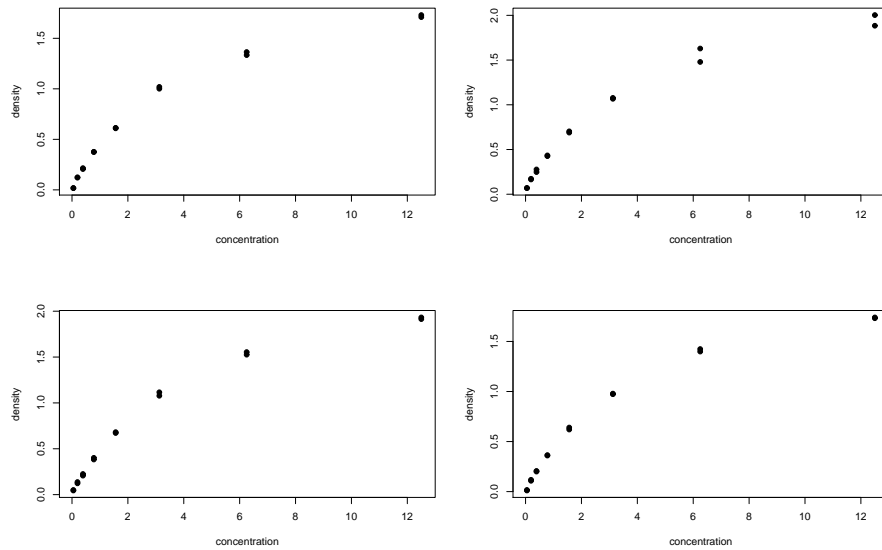
Figure 3.17: The optical density in the ELISA calibration experiment as a function of the known concentrations of DNase in four different runs of the experiment.

Under the very mild condition that the y_i's are not all equal there is a unique minimizer, and there is even an explicit, analytic formula for the minimizer, see Math Box 3.6.1.

Example 3.6.1. We consider in this example the relation between the hardness measured in Shore units of a rubber tire and the abrasion loss in gm/hr. Figure 3.16 shows the scatter plot. From the scatter plot we expect that there is a close to straight line relation; the harder the tare is the smaller is the loss. There are 30 observations in the dataset and if X_i denotes the loss and y_i the hardness for $i = 1, \ldots, 30$ we suggest the model

$$X_i = \beta_0 + \beta_1 y_i + \sigma \varepsilon_i$$

where $\varepsilon_1, \ldots, \varepsilon_{30}$ are iid $N(0, 1)$. Figure 3.16 also shows the straight line estimated by least squares estimation of the parameters (β_0, β_1). $\diamond$

Example 3.6.2 (DNase ELISA calibration)**.** In the development of an ELISA assay for the recombinant protein DNase in rat serum, a calibration experiment was carried out as follows. A known dilution series,

$$0.0488, 0.1953, 0.3906, 0.7812, 1.5625, 3.1250, 6.2500, 12.5000,$$

in ng/ml, of the DNase protein was made, and each run, out of a total of 11 runs, of the experiment consisted of measuring the optical density from the ELISA experi-
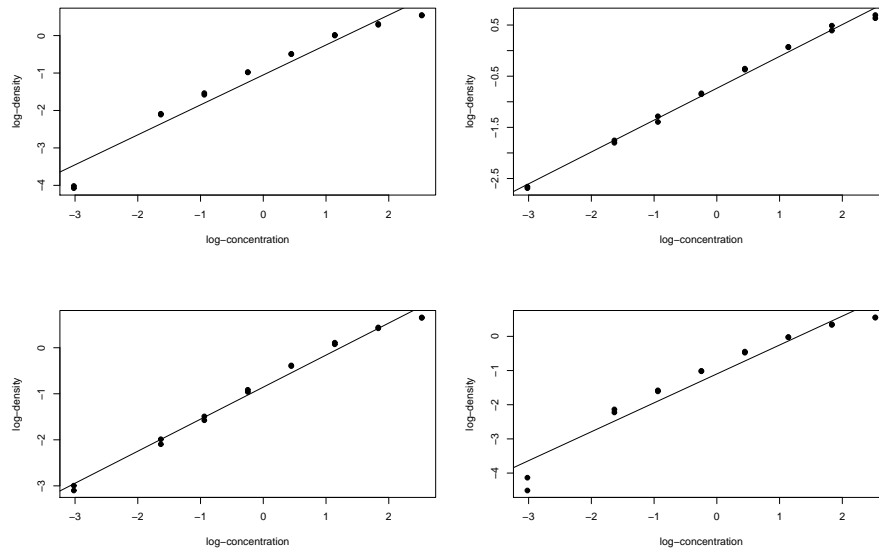
Figure 3.18: The logarithm of the optical density in the ELISA calibration experiment as a function of the known logarithm of concentrations of DNase in four different runs of the experiment. Note that the transformations make the data points lie more closely to a straight line than for the untransformed data.

ment with two measurements for each concentration. The first four runs are shown on Figure 3.17.

From that figure it is not obvious that we can use a linear regression model to capture the density as a function of concentration. However, taking a look at Figure 3.18 we see that by applying the log-transformation to both quantities we get points that approximately lie on a straight line. Thus considering only one run, the first, say, then if X_i denotes the *log-density* and y_i the *log-concentration* for $i = 1, \ldots, 16$ we can try the model

$$X_i = \beta_0 + \beta_1 y_i + \sigma \varepsilon_i.$$

The resulting estimated lines plotted on Figure 3.18 are estimated in this way – with a separate estimation for each of the four runs shown on the figure.

$\diamond$

Example 3.6.3 (Beaver body temperature)**.** As with the DNase example above it is a standard trick to be able to find a suitable transformation of the observables such that a straight line relation becomes plausible. We consider in this example the body temperature for a beaver measured every 10 minutes for a period of 16 hours and 40 minutes, that is, 100 measurements in total. We let X_i denote the i'th
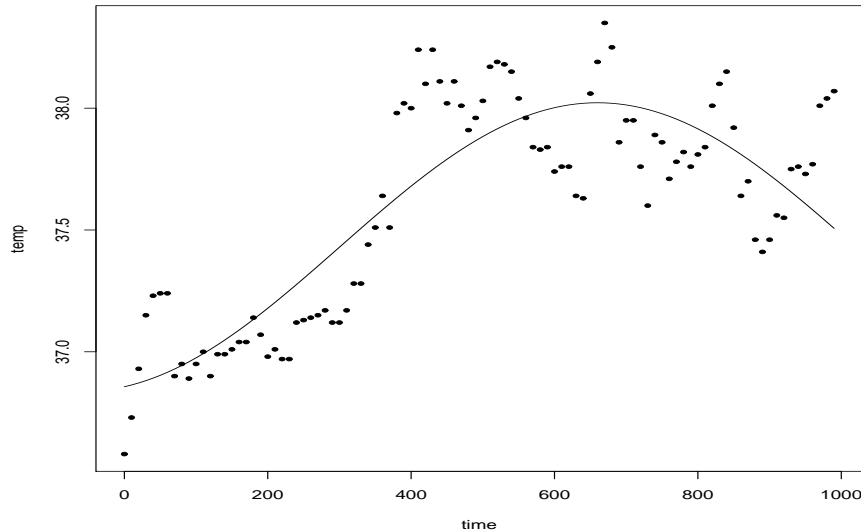
Figure 3.19: The body temperature of a beaver measured every 10 minutes together with a curve estimated by least squares regression.

temperature measurement and t_i the time in minutes since the first measurement (that is $t_1 = 0, t_2 = 10, \ldots$). We suggest that the body temperature will be periodic with a period of 24 hours (1440 minutes) and be minimal at 8.30 in the morning. As the first measurement is taken at 9.30 in the morning we introduce

$$y_i = \cos(2\pi(t_i + 60)/1440)$$

and suggest the model

$$X_i = \beta_0 + \beta_1 y_i + \sigma\varepsilon_i$$

where $\varepsilon_1, \ldots, \varepsilon_{100}$ are iid $N(0, 1)$. Figure 3.19 shows a plot of the data (x_i, t_i) together with the curve estimated by least squares linear regression of the x_i's onto the y_i's.

$\diamond$

An important technique used in the two latter examples above is that ordinary linear regression can be useful even in situations where there is no linear relation between the observable and the regressor. The question is if we can transform the regressor and/or the observable by some known transformation(s) such that linear regression is applicable for the transformed values instead.

The most interesting hypothesis to test in relation to the linear regression model is whether $\beta_1 = 0$ because this is the hypothesis that the value of the regressor does not influence the distribution of the observable. There are several formal test statistics that can be computed for testing this hypothesis. From a summary of the

R Box 3.6.1 (Linear regression). Ordinary linear least squares regression is done using the `lm` function in R. If you load the library `MASS` and read in the dataset `Rubber` you will find a data frame named `Rubber` with three columns, cf. Example 3.6.1. We can regress the loss on the hardness as follows:

```
> library(MASS)
> data(Rubber)
> rubberLm <- lm(loss~hard,data=Rubber)
```

The resulting object `rubberLm` can then be "interrogated" by several generic functions. A call of `summary(rubberLm)` gives among other things a table with the estimated parameters including their estimated standard error and a t-test for the hypothesis that the parameter is equal to zero. A call of `plot(rubberLm)` produces four diagnostic plots, `fitted(rubberLm)` returns the fitted values, `residuals(rubberLm)` returns the residuals and `rstandard(rubberLm)` gives the standardized residuals.

object returned by the `lm` function in R we can read of the value of the estimated standard error $\hat{\text{se}}$ of the estimator for β_1, the t-test statistic

$$t = \frac{\hat{\beta}_1}{\hat{\text{se}}}$$

for testing the hypothesis $H_0 : \beta_1 = 0$ together with a p-value computed based on the t-distribution with $n - 2$ degrees of freedom. The estimated standard error $\hat{\text{se}}$ can also be used for constructing confidence intervals of the form

$$[\hat{\beta}_1 - \hat{\text{se}}z_\alpha, \hat{\beta}_1 + \hat{\text{se}}z_\alpha]$$

where z_α is the $(1-\alpha/2)$-quantile for the $N(0,1)$-distribution. If the error distribution is $N(0,1)$ then there is theoretical basis for replacing the quantile for the normal distribution by the quantile for the t-distribution with $n - 2$ degrees of freedom.

With $\hat{\beta}_0$ and $\hat{\beta}_1$ the least squares estimates we introduce the *fitted values*

$$\hat{x}_i = \hat{\beta}_0 + \hat{\beta}_1 y_i$$

and the *residuals*

$$e_i = x_i - \hat{x}_i = x_i - \hat{\beta}_0 - \hat{\beta}_1 y_i.$$

It is important to investigate if the model actually fits the data. That is, we need methods that can tell if one or more of the fundamental model assumptions are violated. The residuals should resemble the noise variables $\sigma\varepsilon_1, \ldots, \sigma\varepsilon_n$, who are iid. Moreover, there is no relation between the y_i's and the $\sigma\varepsilon_i$'s, and the variance of $\sigma\varepsilon_i$

> **Math Box 3.6.1** (Linear regression). The theoretical solution to the minimization of RSS is most easily expressed geometrically. We denote by $x \in \mathbb{R}^n$ our vector of observations, by $y \in \mathbb{R}^n$ the vector of regressors and by $\mathbf{1} \in \mathbb{R}^n$ a column vector of 1's. The quantity $\mathrm{RSS}(\beta)$ is the length of the vector $x - \beta_0 \mathbf{1} - \beta_1 y$ in $\mathbb{R}^n$. This length is minimized over β_0 and β_1 by the *orthogonal projection* of x onto the space in $\mathbb{R}^n$ spanned by y and $\mathbf{1}$.
>
> One can find this projection if we know an *orthonormal basis*, and such a one is found by setting $a = \mathbf{1}/\sqrt{n}$ (so that a has length 1) and then replacing y by
>
> $$ b = \frac{y - (y^T a)a}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}. $$
>
> Both a and b have unit length and they are orthogonal as $a^T b = 0$. Note that $(ya^T)a = \bar{y}\mathbf{1}$ where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. We have to assume that at least two of the y_i's differ or the sum in the denominator above is 0. If all the y_i's are equal, the vector y and $\mathbf{1}$ are linearly dependent, they span a space of dimension one, and β_0 and β_1 are not both identifiable. Otherwise the vectors span a space of dimension two.
>
> The projection of x onto the space spanned by a and b is
>
> $$ (x^T a)a + (x^T b)b = \underbrace{\left(\bar{x} - \bar{y}\frac{x^T y - n\bar{y}\bar{x}}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \right)}_{\hat{\beta}_0} \mathbf{1} + \underbrace{\frac{x^T y - n\bar{y}\bar{x}}{\sum_{i=1}^{n}(y_i - \bar{y})^2}}_{\hat{\beta}_1} y. $$
>
> Thus the theoretical solution to the minimization problem can be written
>
> $$ \hat{\beta}_1 = \frac{x^T y - n\bar{y}\bar{x}}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \qquad \hat{\beta}_0 = \bar{x} - \bar{y}\hat{\beta}_1. $$

does not depend upon the value y_i. We typically investigate these issues by graphical methods – diagnostic plots – based on the computed residuals. There is one caveat though. The variance of the i'th residual *does* in fact depend upon the y_i's, and it can be shown to equal $\sigma^2(1 - h_{ii})$ where

$$ h_{ii} = \frac{1}{n} + \frac{(y_i - \bar{y})^2}{\sum_{j=1}^{n}(y_j - \bar{y})^2} $$

and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. The quantity h_{ii} is called the *leverage* of the i'th observation. The *standardized residuals* are defined as

$$ r_i = \frac{e_i}{\hat{\sigma}\sqrt{1 - h_{ii}}}. $$

The leverage quantifies how much the observation x_i influences the fitted value $\hat{x}_i$ relative to the other observations. A large leverage (close to 1) tells that the observation heavily influences the fitted value whereas a small value (close to $1/n$) tells that the observation has minimal influence.

Math Box 3.6.2 (Fitted values). We continue here with the setup from Math Box 3.6.1. If $X_i = \beta_0 + \beta_1 y_i + \sigma \varepsilon_i$ we find that the fitted value is

$$\hat{X}_i = (X^T a)a_i + (X^T b)b_i = \beta_0 + \beta_1 y_i + \sum_{j=1}^{n} \sigma(a_j a_i + b_j b_i)\varepsilon_j.$$

If $\varepsilon_j \sim N(0,1)$ then $\sigma(a_j a_i + b_j b_i)\varepsilon_j \sim N(0, \sigma^2(a_j a_i + b_j b_i)^2)$ and it follows from Math Box 2.10.20 that the fitted value has distribution

$$\hat{X}_i \sim N\left(\beta_0 + \beta_1 y_i, \sigma^2 \sum_{j=1}^{n}(a_j a_i + b_j b_i)^2\right).$$

The quantity usually denoted $h_{ii} = \sum_{j=1}^{n}(a_j a_i + b_j b_i)^2$ is know as the *leverage* and if we use that a and b are orthonormal we find that

$$
\begin{aligned}
h_{ii} &= \sum_{j=1}^{n}(a_j a_i + b_j b_i)^2 = a_i^2 a^T a + b_j^2 b^T b + 2a_i b_i a^T b \\
&= a_i^2 + b_i^2 = \frac{1}{n} + \frac{(y_i - \bar{y})^2}{\sum_{j=1}^{n}(y_j - \bar{y})^2}.
\end{aligned}
$$

It follows that the residual

$$
\begin{aligned}
\hat{\varepsilon}_i &= X_i - \hat{X}_i = \sigma \varepsilon_i - \sum_{j=1}^{n} \sigma(a_j a_i + b_j b_i)\varepsilon_j \\
&= \sigma(1 - \underbrace{(a_i^2 + b_i^2)}_{h_{ii}})\varepsilon_i - \sum_{j=1, j\neq i}^{n} \sigma(a_j a_i + b_j b_i)\varepsilon_j
\end{aligned}
$$

has a $N(0, \sigma^2(1 - h_{ii}))$-distribution.

Plots of the residuals or the standardized residuals against the y_i's or against the fitted values $\hat{x}_i$'s are used to check if the mean value specification $\beta_0 + \beta_1 y_i$ is adequate. We expect to see an unsystematic distribution of points around 0 at the y-axis and spread out over the range of the variable we plot against. Systematic deviations from this, such as slopes or bends, in the point cloud indicate that the model specification does not adequately capture how the mean of X_i is related to y_i. A plot of the standardized residuals, their absolute value – or the square root of their absolute value, as the choice is in R – against the fitted values can be used to diagnose if there are problems with the assumption of a constant variance. We look for systematic patterns, in particular whether the spread of the distribution of the points changes in a systematic way over the range of the fitted values. If so, there may be problems with the constant variance assumption. Finally, we can also compare the standardized residuals via a QQ-plot to the distribution of the ε_i's. A QQ-plot of the empirical distribution of the standardized residuals against the nor-
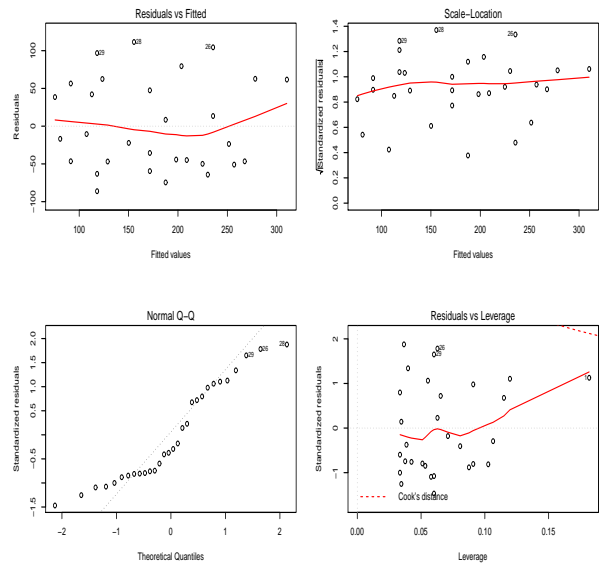
Figure 3.20: The four standard diagnostic plots produced in R for a linear regression.

mal distribution is thus a diagnostic of whether we can take the ε_i's to be normally distributed.

The value of the leverages are sometimes also investigated. A large value of the leverage for a single observation is not in itself problematic, but if the corresponding observation is an outlier, it can drastically affect the whole fit if the leverage is large. A large value of the leverage for a single observation can also indicate a mistake in the recording of the y_i-value.

Example 3.6.4. If we want to investigate the model for the tire loss, as considered in Example 3.6.1, further we can make a summary in R of the result from the call of `lm`, cf. R Box 3.6.1, by `summary(rubberLm)`.

```
Call:
lm(formula = loss ~ hard, data = Rubber)

Residuals:
   Min     1Q Median     3Q    Max
-86.15 -46.77 -19.49  54.27 111.49

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 550.4151    65.7867   8.367 4.22e-09 ***
hard         -5.3366     0.9229  -5.782 3.29e-06 ***
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 60.52 on 28 degrees of freedom
Multiple R-squared: 0.5442,      Adjusted R-squared: 0.5279
F-statistic: 33.43 on 1 and 28 DF,  p-value: 3.294e-06
```

From this we read of the estimates $\hat{\beta}_0 = 550.4$ and $\hat{\beta}_1 = -5.334$, their standard errors, the t-value for a test whether the parameter equals 0 and corresponding p-value computed using the t-distribution with 28 degrees of freedom. As a visual guidance for the eye, the stars are printed to highlight significance at different levels. The conclusion is in this case that none of the parameters can be taken equal to 0. The residual standard error is the estimate of σ.

To assess if the model is adequate we consider the four diagnostic plots shown in Figure 3.20. These are the default plots produced by a call of `plot(rubberLm)`. The plot of the residuals against the fitted values shows that the residuals are scattered randomly around 0 over the range of the fitted values. Thus the straight line seems to capture the relation between loss and hardness well. The QQ-plot shows that the distribution of the residuals is reasonably approximated by the normal distribution – the deviations from a straight line are within the limits of what is conceivable with only 30 observations. The plot of the square root of the absolute value of the standardized residuals against the fitted values shows that the variance does not seem to depend upon the mean. Moreover, there are no clear outliers. The fourth and final plot shows the standardized residuals plotted against the leverage. We should in particular be concerned with combinations of large leverage and large standardized residual as this may indicate an outlier that has considerable influence on the fitted model.

Confidence intervals for the two estimated parameters, β_0 and β_1, can be computed from the information in the summary table above. All we need is the standard error and the relevant quantile from the t-distribution or the normal distribution. For convenience, the R function `confint` can be used to carry out this computation.

```
    > confint(rubberLm)


              2.5 %      97.5 %
(Intercept) 415.657238 685.173020
hard         -7.227115  -3.445991
```

The parameters β_0 and β_1 are, however, not the only parameters of interest in the context of regression. If $y \in \mathbb{R}$ the *predicted value*

$$g_{\hat{\beta}}(y) = \hat{\beta}_0 + \hat{\beta}_1 y$$

is of interest for the prediction of a future loss for a tire of hardness y. If we have a new data frame in R, called `newdata`, say, with a column named `hard`, as our regressor
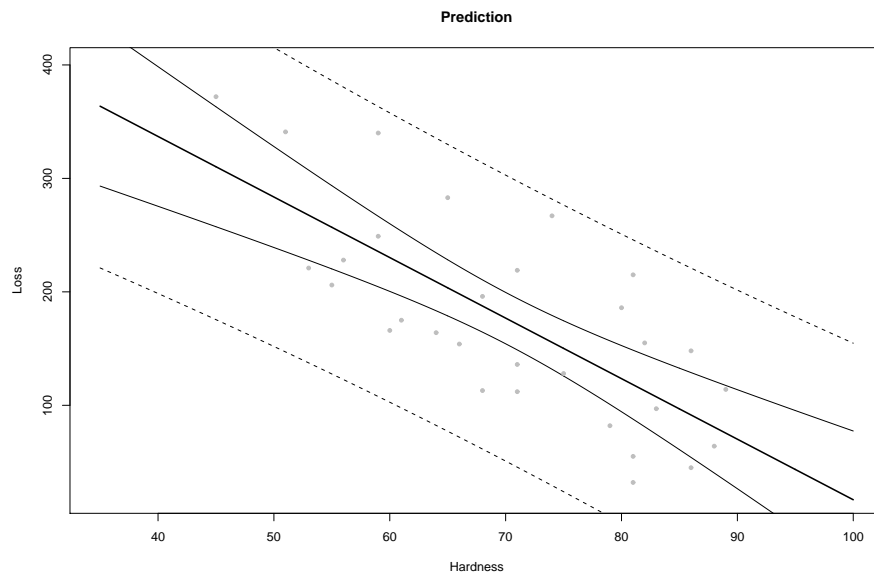
Figure 3.21: Predicted values, 95%-confidence bands and 95%-prediction bands for the linear regression model of tire abrasion loss as a function of tire hardness.

in the original dataset, we can use `predict(rubberLm,newdata)` to compute the predicted values. Moreover, 95%-confidence intervals for all predicted values are obtained by `predict(rubberLm, newdata, interval = "confidence")`. Setting the argument `level` to $1 - \alpha$ (default is 0.95) will give $(1 - \alpha)$-confidence intervals. If $\hat{\text{se}}(y)$ denotes the estimate of the standard error of $g_{\hat{\beta}}(y)$ the confidence interval computed by `predict` is

$$[g_{\hat{\beta}}(y) - \hat{\text{se}}(y)w_\alpha, g_{\hat{\beta}}(y) + \hat{\text{se}}(y)w_\alpha]$$

with w_α the $1 - \alpha/2$-quantile for the t-distribution with $n - 2$ degrees of freedom.

As a final remark we note that to predict a new value

$$X = \beta_0 + \beta_1 y + \sigma\varepsilon$$

the confidence interval computed by e.g. the `predict` function does only take into account the uncertainty in the estimate $\hat{\beta}_0 + \hat{\beta}_1 y$ of the mean value $\beta_0 + \beta_1 y$ of X and not the random component $\sigma\varepsilon$. So-called prediction or tolerance intervals are computed by `predict(rubberLm, newdata, interval = "prediction")`. They are computed as

$$\left[g_{\hat{\beta}}(y) - \sqrt{\hat{\text{se}}(y)^2 + \hat{\sigma}^2}w_\alpha, g_{\hat{\beta}}(y) + \sqrt{\hat{\text{se}}(y)^2 + \hat{\sigma}^2}w_\alpha\right].$$

Figure 3.21 shows predicted values, 95%-confidence bands and 95%-prediction bands.

$\diamond$
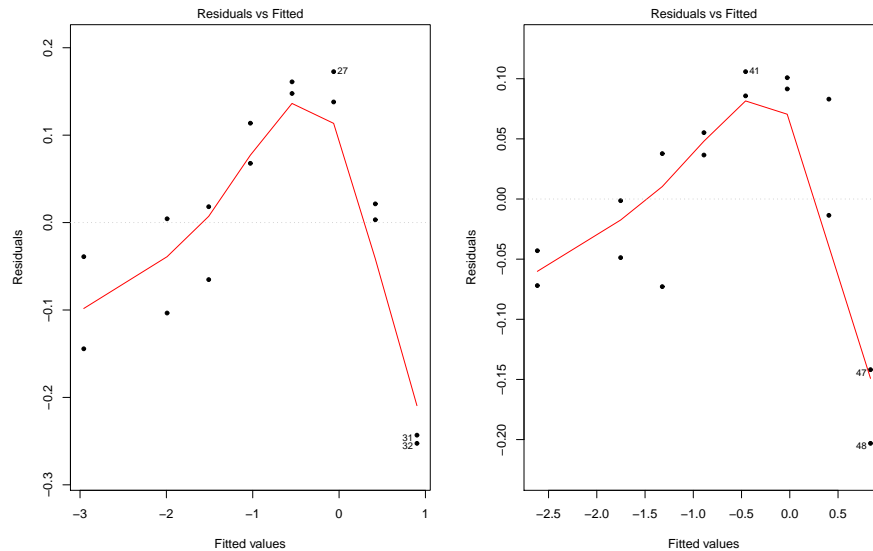
Figure 3.22: Residual plots for the linear regression model of log-density on log-concentration of run 2 and 3 for the DNase ELISA experiment.

In the example above we discussed predictions. A word of warning is appropriate. Even though the regression can be extrapolated to yield predictions outside of the range of the y-values in the dataset such extrapolations should be done with great care – if not avoided altogether. Predictions for values in the range of the observed y-values are more reliable. First of all it appears from Figure 3.21 that the confidence interval for the predicted value is most narrow in the center of the y-values, which is indeed a general phenomena. Second, and more important, we have used various model diagnostics to justify that the model is appropriate and adequate for the data at hand. This provides justification of the predictions that the model provides but only in the range where we have observations. If it cannot be justified by other means that the model extrapolates well outside of the range of the observed y-variable, for instance by subject matter knowledge, extrapolations are model dependent speculations where we have little support for the predictions.

Example 3.6.5. The summary of the `lm`-object for one of the linear regressions for the ELISA data considered in Example 3.6.2 reads

```
Call:
lm(formula = log(density) ~ log(conc), data = myDNase[myDNase[,1] == 2, ])

Residuals:
    Min      1Q   Median      3Q      Max
-0.25254 -0.07480  0.01123  0.11977  0.17263
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.85572    0.03505  -24.41 7.09e-13 ***
log(conc)    0.69582    0.02025   34.36 6.39e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1402 on 14 degrees of freedom
Multiple R-squared: 0.9883,        Adjusted R-squared: 0.9874
F-statistic:  1181 on 1 and 14 DF,  p-value: 6.391e-15
```

Figure 3.22 shows the residual plot for two of the runs. Both plots are typical for this dataset and shows a lack of fit for the model. The residuals are positive for the fitted values in the middle range and negative for the extreme fitted values. The straight line does not seem to fully capture the relation between the log-density and the log-concentration. We will return to a refined model in the next section.

The fact that the model does not fit the data does not mean that the linear regression model is useless. After all, it captures the general relation between the log-density and log-concentration up to some relatively small curvature, which on the other hand appears very clearly on the residual plot. The systematic error of the model for the mean value leads to a larger estimate of the variance than if the model was more refined. Often this results in conservative conclusions such as wider confidence intervals for parameters of interest and wider prediction intervals. Whether this is tolerable must be decided on a case-by-case basis. ⋄

Example 3.6.6. Figure 3.24 shows the residual plot and the QQ-plot for the beaver data considered in Example 3.6.3. The QQ-plot seems OK, but the residual plot shows a problem. Around the middle of the residual plot the residuals show a sudden change from being exclusively negative to being mostly positive. This indicates a lack of fit in the model and thus that the temperature variation over time cannot be ascribed to a 24 hour cycle alone.

In fact, there might be an explanation in the dataset for the change of the temperature. In addition to the time and temperature it has also been registered if the beaver is active or not. There is a 0-1 variable called `activ` in the dataset, which is 1 if the beaver is active. A refined analysis will include this variable so that we have a three parameter model of the mean with

$$\beta_0 + \beta_1 y$$

the mean temperature if the beaver is inactive and

$$\beta_0 + \beta_{\texttt{activ}} + \beta_1 y$$

if the beaver is active. Here $y = \cos(2\pi t/1440)$ where t is time in minutes since 8.30 in the morning. The summary output from this analysis is:
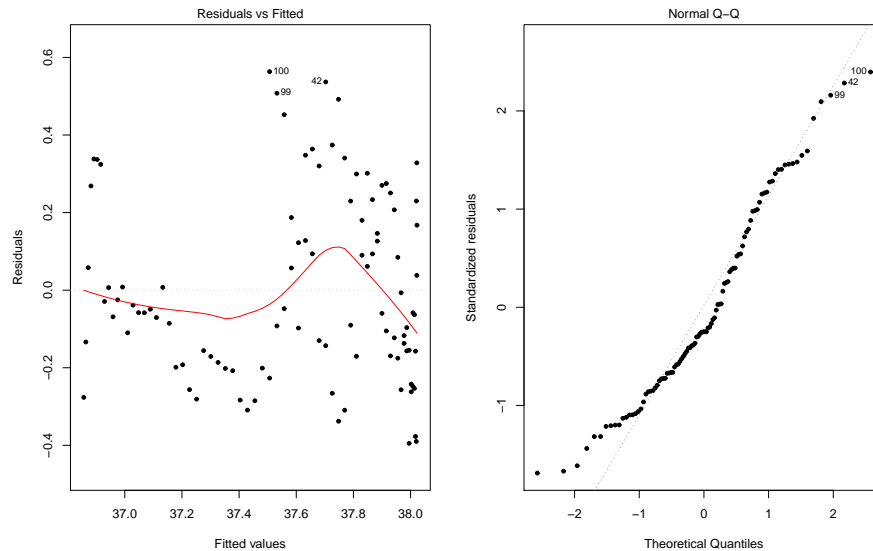
Figure 3.23: Residual plot and QQ-plot for the beaver body temperature data.

```
Call:
lm(formula = temp ~ activ + cos(2 * pi * (time + 60)/1440), data = beaver2)

Residuals:
     Min       1Q   Median       3Q      Max
-0.44970 -0.11881 -0.02208  0.17742  0.39847

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)
(Intercept)                    37.19976    0.04253 874.690  < 2e-16 ***
activ                           0.53069    0.08431   6.294 8.95e-09 ***
cos(2 * pi * (time + 60)/1440) -0.24057    0.06421  -3.747 0.000304 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2002 on 97 degrees of freedom
Multiple R-squared: 0.8034,        Adjusted R-squared: 0.7993
F-statistic: 198.1 on 2 and 97 DF,  p-value: < 2.2e-16
```

We can read from the summary that the hypothesis that $\beta_{\texttt{activ}} = 0$ is clearly rejected. The estimate $\hat{\beta}_{\texttt{activ}} = 0.531$ suggests that the beaver being active accounts for roughly half a grade of increase in the body temperature. Figure 3.24 shows the fitted mean value and the residual plot. It might be possible to refine the model even further – perhaps with a smoother transition between the inactive and active state – but we will not pursue such refinements here. Instead we will focus on an-
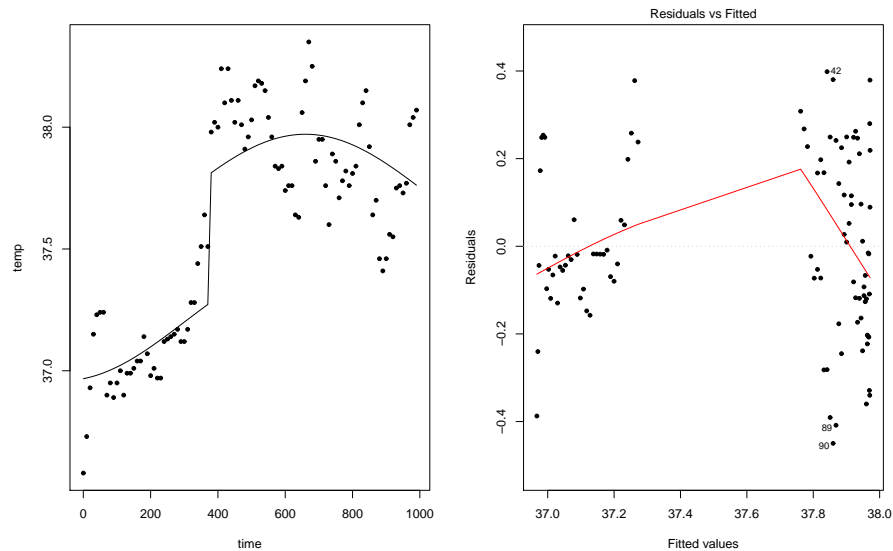
Figure 3.24: Plot of the beaver temperature data and estimated curve (left) and residual plot (right) when the additional variable `activ` is included in the analysis.

other potential problem. The observations of the temperature are taken every 10 minutes and this may give problems with the independence assumption. When we deal with observations over time, this should always be taken into consideration. If the observations are ordered according to observation time we can investigate the assumption by a lag-plot of the residuals $e_1, \ldots, e_{100}$, which is a plot of (e_{i-1}, e_i) for $i = 2, \ldots, 100$. Figure 3.25 shows that lag-plot of the residuals. This plot should look like a scatter plot of independent variables, but in this case it does not. On the contrary, it shows a clear dependence between residual e_i and the lagged residual e_{i-1}.

Though the residuals show dependence and the model assumptions of independent ε_i's are questionable, the estimated mean value function can still be used as a reasonable estimate. The problem is in general that all estimated standard errors are systematically wrong as they are based on the independence assumption, and they are typically too small, which means that we tend to draw conclusions that are too optimistic and provide confidence and prediction intervals that are too narrow. The right framework for systematically correcting for this problem is time series analysis, which is beyond the scope of these notes.   ◇
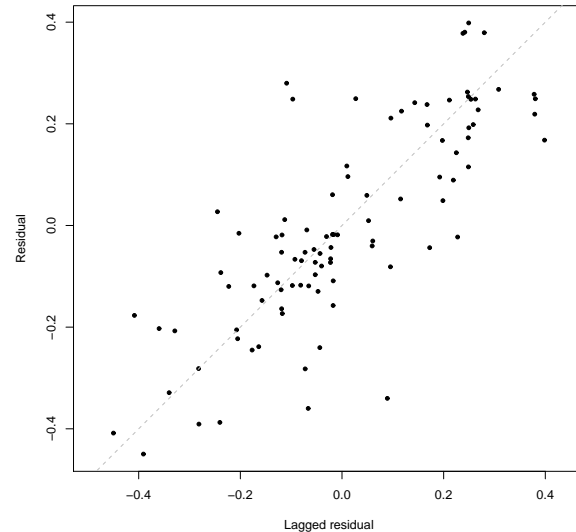
Figure 3.25: The temperature observations are ordered by time and this plot shows the residuals plotted against the lagged residuals, that is, the residuals from the previous time point.

3.6.2   Non-linear regression

Dividing regression into linear and non-linear regression is like dividing the animals in the Zoo into elephants and non-elephants. In the world of non-linear regression you can meet all sorts of beasts, and we can in this section only touch on a few simple examples that you can find outside of the world of linear regression.

In this section we will also consider estimation using the least squares method only. This is only the maximum-likelihood method if the ε_i-terms are normally distributed. In the world of regression this is often referred to as *non-linear least squares*. We can in general not expect to find closed form solutions to these minimization problems and must rely on numerical optimization. A word of warning is appropriate here. We cannot expect that the numerical optimization always goes as smoothly as desirable. To find the correct set of parameters that globally minimizes the residual sum of squares we may need to choose appropriate starting values for the algorithm to converge, and it may very easily be the case that there are multiple local minima, that we need to avoid.

For non-linear regression there are generalizations of many of the concepts from linear regression. The fitted values are defined as

$$\hat{x}_i = g_{\hat{\beta}}(y_i)$$

and the residuals are

$$e_i = x_i - \hat{x}_i = x_i - g_{\hat{\beta}}(y_i).$$

To check the model we make residual plots of the residuals against either the regressors y_i or the fitted values $\hat{x}_i$ and we look for systematic patterns that either indicate that the model of the mean via the function $g_{\hat{\beta}}$ is inadequate or that the constant variance assumption is problematic. Moreover, we can compare the empirical distribution of the residuals to the normal distribution via a QQ-plot. We don't have a simple leverage measure, nor do we have a formula for the variance of the residuals. So even though the residuals may very well have different variances it is not as easy to introduce standardized residuals that adjust for this. In the context of non-linear regression the term standardized residual often refers to $e_i/\hat{\sigma}$ where we simply divide by the estimated standard deviation.

All standard estimation procedures used will produce estimates of the standard error for the β-parameters that enter in the model. These estimates are based on the assumption that the ε_i's are normally distributed, but even under this assumption the estimates will still be approximations and can be inaccurate. The estimated standard errors can be used to construct confidence intervals for each of the coordinates in the β-vector or alternatively to test if the parameter takes a particular value. In some cases – but certainly not always – it is of particular interest to test if a parameter equals 0, because it is generally a hypothesis about whether a simplified model is adequate.

Example 3.6.7 (Michaelis-Menten)**.** Enzymes work as catalysts in the conversion of a substrate into a product. The enzyme *alcohol dehydrogenase* catalyzes the conversion of ethanol (the substrate) to acetaldehyde (the product). The data below are from Bendinskas et al. *Journal of Chemical Education, 82(7), 1068 (2005)*. The *Michaelis-Menten rate equation* states that the rate, r, for the conversion is related to the concentration y of the substrate via

$$r = \frac{\beta_1 y}{\beta_2 + y}. \tag{3.20}$$

With $\beta = (\beta_1, \beta_2)$ the two unknown parameters and measurements $r_1, \ldots, r_n$ of the conversion rate for different substrate concentrations $y_1, \ldots, y_n$ we set up a non-linear regression model

$$R = g_\beta(y) + \sigma\varepsilon$$

where $g_\beta(y) = \frac{\beta_1 y}{\beta_2 + y}$. We assume in this model that the measurement noise, $\sigma\varepsilon$, of the conversion rate enters additively.

R Box 3.6.2 (Non-linear least squares regression)**.** There are several different possibilities for fitting non-linear regression models in R. The `nls` function does non-linear least squares estimation and is quite similar in use to the `lm` function for ordinary least squares.

With `ethConv` a data frame with two columns named `rate` and `conc` we can estimate the Michaelis-Menten curve from Example 3.6.7 by

```
> nls(rate ~ beta1*conc/(beta2 + conc),
+     data=ethConv,
+     start=c(beta1=1,beta2=1))
```

Consult Example 3.6.7 on how to summarize the result of this call with the `summary` function.

A main difference from `lm` is that in the formula specification `rate ~ beta1 * conc/(beta2 + conc)` we need to explicitly include all unknown parameters and we need to explicitly give an initial guess of the parameters by setting the `start` argument. You can leave out the specification of `start` in which case `nls` automatically starts with all parameters equal to 1 – and gives a warning.

If you want to do things beyond what `nls` can do there are several solutions, some are tied up with a particular area of applications or with particular needs.

The `drc` package is developed for dose-response curve estimation but can be useful for general non-linear regression problems. The main function, `multdrc`, does non-linear least squares estimation but a particular advantage is that it allows for the simultaneous fitting of multiple dataset. Doing so it is possible to share some but not all parameters across the different curves.

Substrate concentration	Conversion rate
0.007	0.06
0.015	0.11
0.031	0.16
0.068	0.21
0.100	0.23
0.200	0.28
0.300	0.29
0.400	0.28

```
> ethConvNls <- nls(rate ~ beta1*conc/(beta2 + conc),
+                   data=ethConv,
+                   start=c(beta1=1,beta2=1))
```
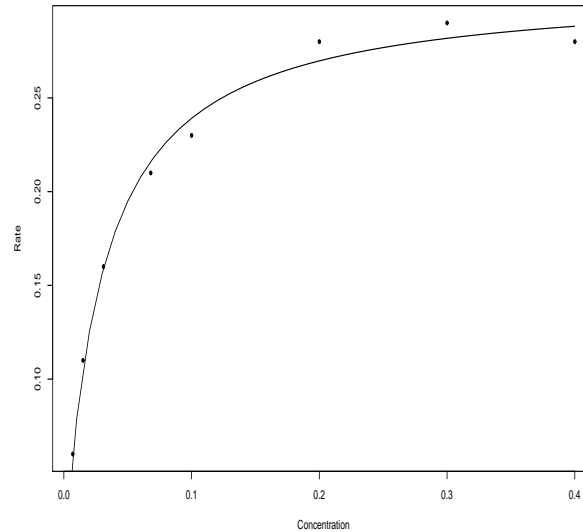
Figure 3.26: The data and estimated Michaelis-Menten rate curve from Example 3.6.7

We can then call `summary` on the resulting object to get information on the parameter estimates, estimated standard errors etc.

```
> summary(ethConvNls)
```

```
Formula: rate ~ beta1 * conc/(beta2 + conc)

Parameters:
      Estimate Std. Error t value Pr(>|t|)
beta1 0.309408   0.006420    48.19 5.35e-09 ***
beta2 0.029391   0.002503    11.74 2.30e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.00807 on 6 degrees of freedom

Number of iterations to convergence: 8
Achieved convergence tolerance: 5.363e-06
```

◇

Example 3.6.8 (Four parameter logistic model)**.** One of the flexible and popular non-linear regression models is known as the four parameter logistic model and is

given by the function

$$g_\beta(y) = \frac{\beta_2 - \beta_1}{1 + \exp(\beta_4(y - \beta_3))} + \beta_1$$

for $\beta = (\beta_1, \beta_2, \beta_3, \beta_3) \in \mathbb{R}^4$. For identifiability of the parameters we need to assume that $\beta_1 < \beta_2$, say. The parameters β_1 and β_2 then give the lower and upper asymptotes, respectively, for the graph of the function. The sign of β_4 determines if the function is decreasing or increasing and its absolute value how steep the graph is. The parameter β_3 determines the value of y where the mean response is $(\beta_2 + \beta_1)/2$, which is the mid-point between the minimal value β_1 and the maximal value β_2.

We will illustrate the use of this model on the ELISA data considered in Example 3.6.2. We use `nls` in R with initial values $\beta_1 = 0$, $\beta_2 = 2$, $\beta_3 = 1$ and $\beta_4 = -1.5$ and get the following summary of the result.

```
Formula: density ~ (beta2 - beta1)/(1 + exp(beta4 * (log(conc) - beta3))) +
    beta1

Parameters:
       Estimate Std. Error t value Pr(>|t|)
beta1 -0.007897   0.017200  -0.459    0.654
beta2  2.377241   0.109517  21.707 5.35e-11 ***
beta3  1.507405   0.102080  14.767 4.65e-09 ***
beta4 -0.941106   0.050480 -18.643 3.16e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01981 on 12 degrees of freedom

Number of iterations to convergence: 5
Achieved convergence tolerance: 5.686e-06
```

The conclusion from this summary is that the parameter β_1 can be taken equal to 0. This is sensible as we expect a 0 measurement if there is no DNase in the sample (the concentration is 0). Taking $\beta_1 = 0$ a reestimation of the model yields the new estimates

```
Parameters:
      Estimate Std. Error t value Pr(>|t|)
beta2  2.34518    0.07815   30.01 2.17e-13 ***
beta3  1.48309    0.08135   18.23 1.22e-10 ***
beta4 -0.96020    0.02975  -32.27 8.51e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01919 on 13 degrees of freedom
```
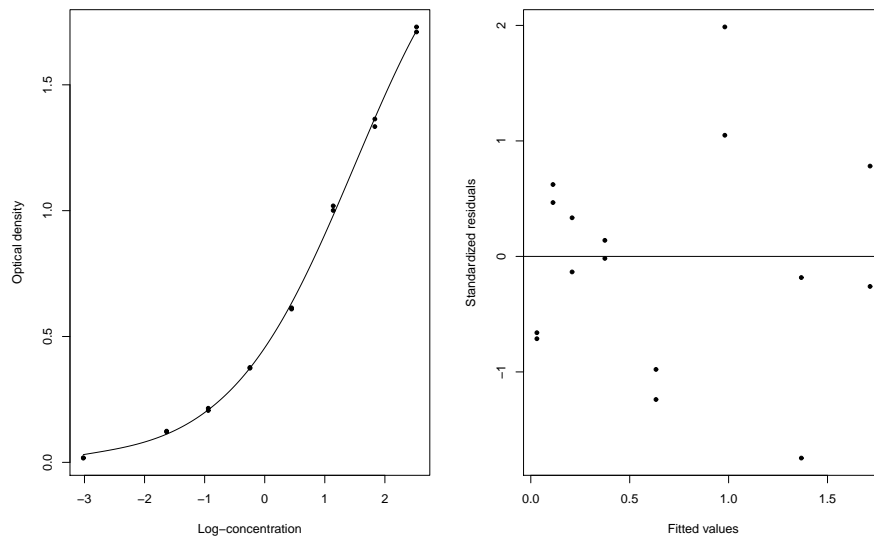
Figure 3.27: The data and estimated logistic model for the relation between optical density and log-concentration for the DNase ELISA considered in Example 3.6.8.

For this model figure 3.27 shows the data and the estimated mean value curve for run 1 in the dataset together with the residual plot, which shows that the model is adequate.

When y is the log-concentration an alternative formulation of the logistic model expressed directly in terms of the concentration is

$$g_\beta(\text{conc}) = \frac{\beta_2 - \beta_1}{1 + \exp(\beta_4(\log(\text{conc}) - \beta_2))} + \beta_1 = \frac{\beta_2 - \beta_1}{1 + \left(\frac{\text{conc}}{\exp(\beta_2)}\right)^{\beta_4}} + \beta_1.$$

$\diamond$

Example 3.6.9 (Expression calibration). It is not the rule to produce standard curves for microarray experiments, that is, dilution series with known concentrations are not routinely applied to a microarray for the purpose of estimating the relation between concentration and the actually observed light intensity. In the process of understanding and modeling the results from microarray experiments some so-called spike-in experiments have, however, been done. These are experiments where some samples of known concentration have been *spiked in* to the sample that is applied to the microarray and one can use the results from these known sample concentrations to infer the relation between concentration and light intensity. We consider here a spike-in experiment, `spikein95`, from the Bioconductor `SpikeInSubset` package.
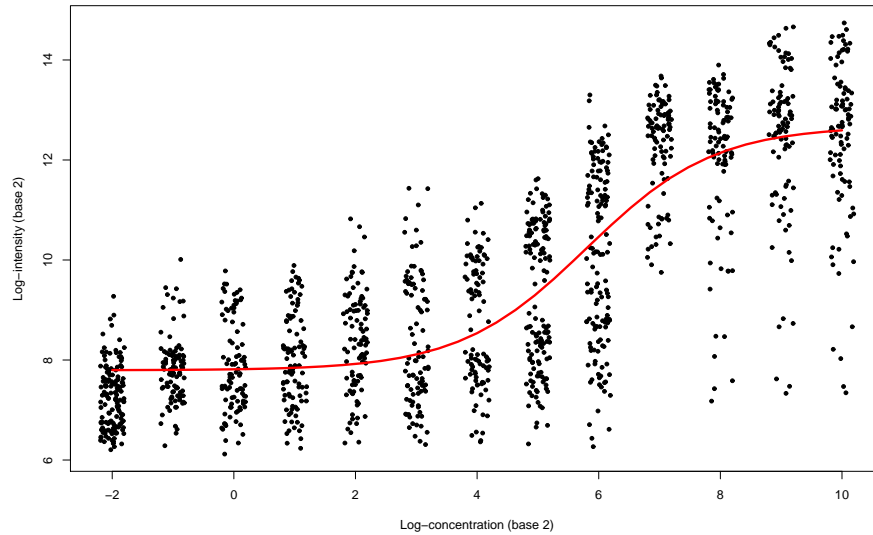
Figure 3.28: .

We will use the logistic model as considered above to capture the relation between log-concentration and log-intensity. The concentrations are in picoMolar.

Using `nls` in R the summary of the resulting object reads

```
Formula: log2(value) ~ (beta2 - beta1)/(1 + exp(beta4 * (log2(conc) -
    beta3))) + beta1

Parameters:
      Estimate Std. Error t value Pr(>|t|)
beta1  7.79500    0.06057  128.69   <2e-16 ***
beta2 12.68086    0.12176  104.15   <2e-16 ***
beta3  5.80367    0.09431   61.54   <2e-16 ***
beta4 -0.95374    0.08001  -11.92   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.293 on 1388 degrees of freedom

Number of iterations to convergence: 14
Achieved convergence tolerance: 6.411e-06
```

From the summary we see that none of the parameters can be taken equal to 0. It is most notable that β_1 is significantly larger than 0 and in fact the estimate $\hat{\beta}_1 = 7.795$ is far from 0. It suggests that there is a considerable background signal even when
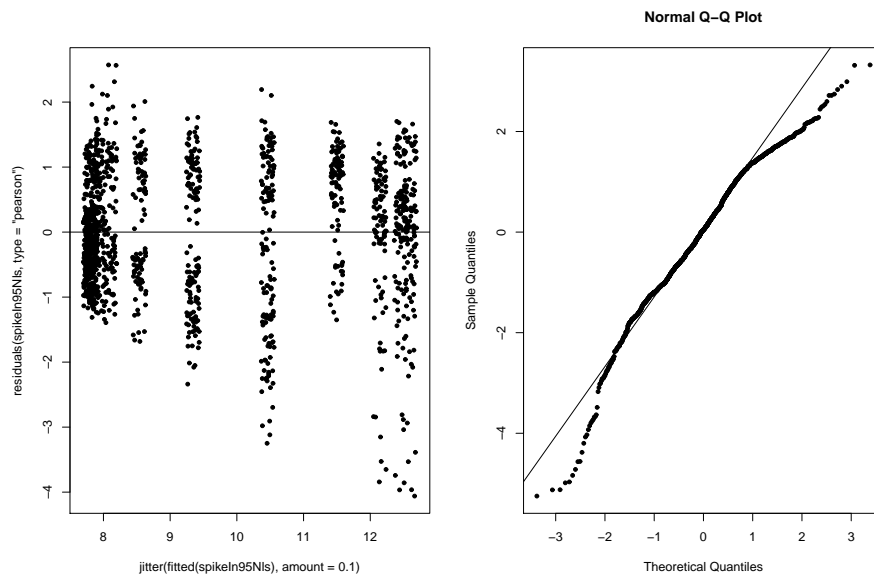
Figure 3.29: .

the concentration equals 0. Figure 3.28 shows the data and the estimated curve. There are only 13 different concentrations but a large number of replications for each concentration and the x-values have therefore been jittered to aid visualization. We see from this plot that the logistic curve captures the overall behavior rather well, but already from this plot it seems that there are problems with variance inhomogeneity and a number of observations – in particular some of the small observations for the high concentrations – seem problematic.

Figure 3.29 shows that residual plot and a QQ-plot of the residuals against the normal distribution. It is clear from the residual plot that an assumption of the same variance across the different concentrations does not hold. The normal QQ-plot shows deviations from the normal distribution in both tails. This form of deviation suggests that the distribution is left-skewed compared to the normal distribution because there are too many small residuals and too few large residuals. However, because the variance cannot be assumed constant this is in reality a mixture of different distributions with different scales.   ⋄

Exercises

For the next four exercises you are asked to consider the `Rubber` data from the `MASS` library.

Exercise 3.6.1. In the dataset there is another variable, `tens`. Plot `loss` versus `tens` and carry out a linear regression of `loss` on `tens`. This includes model diagnostics. What is the conclusion?

Exercise 3.6.2. Compute the residuals when you regress `loss` on `hard` and plot the residuals versus `tens`. Interpret the result.

Exercise 3.6.3. If we let y_i denotes the hardness and z_i the tension, consider the extended model

$$X_i = \beta_0 + \beta_1 y_i + \beta_2 z_i + \sigma \varepsilon_i$$

where $\varepsilon_1, \ldots, \varepsilon_{30}$ are iid $N(0,1)$. Use `lm` (the formula should be `loss~hard+tens`) to carry our a linear regression model including model diagnostics where you regress on hardness as well as tension.

Exercise 3.6.4. Use the function `predict` to compute 95% prediction intervals for the 30 observed values using

- the model where we only regress on hardness

- the model where we regress on hardness and tension.

Compare the prediction intervals. It can be useful to plot the prediction intervals versus the fitted values.

Exercise 3.6.5. Show that the Michaelis-Menten rate equation 3.20 can be rephrased as

$$\frac{1}{r} = \frac{\beta_2}{\beta_1}\frac{1}{y} + \frac{1}{\beta_1}.$$

Argue that this formula suggests a linear regression model for the inverse of the rate regressed on the inverse of the substrate concentration. Estimate the parameters using this linear regression model and compare with Example 3.6.7.

3.7 Bootstrapping

The basic idea in bootstrapping for constructing confidence intervals for a parameter of interest τ when we have an estimator $\hat{\tau}$ of $\tau = \tau(\theta)$ is to find an approximation of the distribution of $\hat{\tau} - \tau(\theta_0)$ – usually by doing simulations that depend upon the observed dataset $x \in E$. What we attempt is to approximate the distribution of

$$\hat{\tau} - \tau(\theta_0)$$

under P_{θ_0}, by the distribution of

$$\hat{\tau} - \hat{\tau}(x)$$

under a cleverly chosen probability measure P_x, which may depend upon x. Since the distribution is allowed to depend upon the concrete observation x the construction of the resulting confidence set – that provides information about the uncertainty of the estimate $\hat{\tau}(x)$ – depends upon the observation itself. Hence what we are going to suggest is to pull information about the uncertainty of an estimate out from the very same data used to make the estimate, and for this reason the method is known as *bootstrapping*. Supposedly one of the stories in *The Surprising Adventures*

of Baron Munchausen by Rudolf Erich Raspe (1736 - 1794) contains a passage where the Baron pulls himself out of a deep lake by pulling his own bootstraps. Such a story is, however, not to be found in the original writings by Raspe, but the stories of Baron Munchausen were borrowed and expanded by other writers, and one can find versions where the Baron indeed did something like that. To bootstrap is nowadays, with reference to the Baron Munchausen story, used to describe various seemingly paradoxical constructions or actions. To boot a computer is for instance an abbreviation of running a so-called bootstrap procedure that gets the computer up and running from scratch.

The computation of the distribution of $\hat{\tau} - \hat{\tau}(x)$ under P_x is a transformation problem that is usually solved by computer simulations. We simply use the computer to generate a large number of new datasets $x_1, \ldots, x_B$ and relevant quantities such as quantiles for the distribution of $\hat{\tau} - \hat{\tau}(x)$ or the standard deviation of $\hat{\tau}$ are estimated from the simulated data.

Algorithm 3.7.1 (The Bootstrap for Confidence Set Construction). We consider an estimator $\hat{\tau} : E \to \mathbb{R}$ for the real valued parameter of interest. For a given dataset $x \in E$ the corresponding estimate is $\hat{\tau}(x)$ and with P_x a probability measure on E, which may depend upon x, the bootstrapping algorithm for producing a nominal $(1 - \alpha)$-confidence interval proceeds as follows:

- Choose B sufficiently large and simulate B new independent, identically distributed datasets, $x_1, \ldots, x_B \in E$, each simulation being from the probability measure P_x.

- Compute, for each dataset x_i, $i = 1, \ldots, B$, new estimates $\hat{\tau}(x_i)$ using the estimator $\hat{\tau}$.

- Compute $\hat{z}_\alpha$ and $\hat{w}_\alpha$ as the $\alpha/2$ and $1 - \alpha/2$ quantiles for the empirical distribution of $\hat{\tau}(x_i) - \hat{\tau}(x)$, $i = 1, \ldots, B$.

- Define $I(x) = [\hat{\tau}(x) - \hat{w}_\alpha, \hat{\tau}(x) - \hat{z}_\alpha]$.

A minor modification of the algorithm is given by replacing the last two bullet points by

- Compute the empirical mean $\overline{\tau} = \frac{1}{B} \sum_{i=1}^{B} \hat{\tau}(x_i)$ and the empirical standard deviation
$$\hat{se} = \sqrt{\frac{1}{B-1} \sum_{i=1}^{B} (\hat{\tau}(x_i) - \overline{\tau})^2}$$

- Define $I(x) = [\hat{\tau}(x) - \hat{se} z_\alpha, \hat{\tau}(x) + \hat{se} z_\alpha]$ where z_α is the $1 - \alpha/2$ quantile for the $N(0, 1)$-distribution.

The two most important and commonly encountered choices of P_x are known as *parametric* and *non-parametric* bootstrapping respectively. We consider parametric bootstrapping here and non-parametric bootstrapping in the next section.

Definition 3.7.1. *If we have an estimator, $\hat{\theta}$, of θ then parametric bootstrapping is the bootstrapping method we get by taking*

$$P_x = P_{\hat{\theta}(x)}.$$

Example 3.7.2. As a continuation of Example 3.5.8 we compute confidence intervals for the parameters α and β as well as the LD_{50} parameter. A simple parametric bootstrap with $B = 1000$ yields the following 95% confidence intervals:

	Parameters		
	α	β	LD_{50}
$\hat{\mathrm{se}}$	0.713	0.358	0.080
$[\hat{\tau}(x) - 1.96\hat{\mathrm{se}}, \hat{\tau}(x) + 1.96\hat{\mathrm{se}}]$	$[3.74, 6.53]$	$[2.00, 3.41]$	$[-2.06, -1.74]$
$[\hat{\tau}(x) - \hat{w}_{0.05}, \hat{\tau}(x) - \hat{z}_{0.05}]$	$[3.47, 6.14]$	$[1.81, 3.19]$	$[-2.06, -1.74]$

Note that the confidence interval(s) for LD_{50} are quite narrow and the estimated standard deviation is likewise small – at least compared to the parameters α and β. The parameter LD_{50} is simply much better determined than the two original parameters. Note also that the two different types of intervals differ for α and β but not for LD_{50}. This can be explained by a *right*-skewed distribution of the estimators for the two former parameters, which turn into a *left* translation of the second confidence intervals as compared to the symmetric intervals based on $\hat{\mathrm{se}}$. The distribution of the estimator for LD_{50} is much more symmetric around $\hat{\mathrm{LD}}_{50}$. $\diamond$

3.7.1   The empirical measure and non-parametric bootstrapping

Definition 3.7.3. *Given a dataset $x_1, \ldots, x_n$ we define the empirical probability measure, or simply the empirical measure, ε_n, on E by*

$$\varepsilon_n(A) = \frac{1}{n} \sum_{k=1}^{n} 1(x_k \in A) \tag{3.21}$$

for all events $A \subseteq E$.

The empirical measure is the collection of relative frequencies, $\varepsilon_n(A)$, for all events $A \subseteq E$, which we encountered when discussing the frequency interpretation in Section 2.3. It is also the frequency interpretation that provides the rationale for considering the empirical measure.

To define non-parametric bootstrapping we need to assume that the dataset considered consists of realizations of iid random variables. Thus we assume that $E = E_0^n$ and that the dataset $x = (x_1, \ldots, x_n)$ is a realization of n iid random variables $X_1, \ldots, X_n$. The *empirical measure* based on $x_1, \ldots, x_n$ on E_0 is denoted ε_n.

Definition 3.7.4. *Non-parametric bootstrapping is defined by letting*

$$P_x = \varepsilon_n \otimes \ldots \otimes \varepsilon_n.$$

That is, random variables $X_1, \ldots, X_n$ with distribution P_x are iid each having the empirical distribution ε_n.

Whereas the procedure for doing parametric bootstrapping is straight forward, the definition of non-parametric bootstrapping seems a little more difficult. In fact, the definition of non-parametric bootstrapping is somewhat complicated. In practice it is really the other way around. How to do parametric bootstrap simulations from $P_{\hat{\theta}(x)}$ relies upon the concrete model considered, and sometimes it can be difficult to actually simulate from $P_{\hat{\theta}(x)}$. It does as a minimum require a simulation algorithm that is model dependent. To do non-parametric bootstrap simulations using the empirical measure is on the contrary easy and completely model independent. Simulating from P_x is a matter of simulating (independently) from ε_n, which in turn is a matter of sampling with replacement from the dataset $x_1, \ldots, x_n$. Theorem 3.7.5 below is an adaption of the general Algorithm 2.11.1, which is suitable for simulating from any empirical measure. The recommended approach is to sample *indices* from the dataset uniformly. This approach is efficient and completely generic. The implementation requires no knowledge about the original sample space whatsoever.

R Box 3.7.1 (Simulation from the empirical measure). If `x` is a vector of length n containing the dataset $x_1, \ldots, x_n$ we can obtain a sample of size $B = 1000$ from the empirical measure by

```
> y <- sample(x,1000,replace=TRUE)
```

The vector `y` then contains 1000 simulations from the empirical measure. Note the parameter `replace` which by default is `FALSE`.

Result 3.7.5. *Let $x_1, \ldots, x_n$ be a dataset with values in the sample space E and corresponding empirical measure. If U is uniformly distributed on $\{1, \ldots, n\}$ then the distribution of*

$$X = x_U$$

is the empirical measure ε_n.

Proof: With $E_n = \{z \in E \mid x_i = z \text{ for some } i = 1, \ldots, n\}$ we find that for $z \in E_n$ and with $I_z = \{i \in \{1, \ldots, n\} \mid x_i = z\}$ then

$$\mathbb{P}(X = z) = \mathbb{P}(x_U = z) = \mathbb{P}(U \in I_z) = \frac{|I_z|}{n} = \frac{1}{n} \sum_{i=1}^{n} 1(x_i = z) = \varepsilon_n(z).$$

$\square$

An alternative formulation of Theorem 3.7.5, making the transformation explicit, is via the $x_1, \ldots, x_n$-dependent map

$$h_{x_1, \ldots, x_n} : \{1, \ldots, n\} \to E$$

defined by

$$h_{x_1, \ldots, x_n}(i) = x_i.$$

Then if U has the uniform distribution on $\{1, \ldots, n\}$ the theorem states that the transformed variable $h_{x_1, \ldots, x_n}(U)$ has the empirical distribution.

Remark 3.7.6. It follows from the theorem that if $U_1, \ldots, U_B$ are B iid uniformly distributed random variables taking values in $\{1, \ldots, n\}$ then $X_1, \ldots, X_n$ defined by

$$X_i = x_{U_i}$$

for $i = 1, \ldots, n$ are iid with distribution ε_n. Taking $U_1, \ldots, U_n$ to be iid uniformly from $\{1, \ldots, n\}$ is known as *sampling with replacement* n times from $\{1, \ldots, n\}$. The random variables $X_1, \ldots, X_n$ can therefore be regarded as n samples with replacement from the set $\{x_1, \ldots, x_n\}$. How we choose to perform the simulation of $U_1, \ldots, U_n$ is another issue. It could be done by Algorithm 2.11.1, but the details are not important. The implementation of that simulation can be done once and for all and optimized sufficiently.

Non-parametric bootstrapping makes a priori only sense if our observables are iid. In a regression setup we cannot use non-parametric bootstrapping directly. However, we can non-parametrically bootstrap the residuals, which gives a bootstrapping algorithm where

$$X_i = g_{\hat{\beta}}(y_i) + e_{U_i}$$

with $U_1, \ldots, U_n$ are iid uniformly distributed on $\{1, \ldots, n\}$. Thus we use the parametric estimate of the mean value relation between the observable and the regressor and then we use the non-parametric, empirical distribution of the residuals to sample new error variables.

3.7.2   The percentile method

For the construction of confidence intervals it is tempting to proceed in the following way. If $\hat{\tau}(x)$ is the estimate then with z_α and w_α the $\alpha/2$ and $1 - \alpha/2$ quantiles of the distribution of $\hat{\tau}$ under the probability measure $P_{\hat{\theta}(x)}$ we simply take the interval $[z_\alpha, w_\alpha]$ as a $(1 - \alpha)$-"confidence interval". This corresponds to reading a figure like Figure 3.13 vertically compared to the usual horizontal reading. By this we mean that instead of locating the estimate on the vertical axis and read of horizontally the values of p in the case of Figure 3.13 that could produce the estimate with high probability, we simply read of the distribution of $\hat{p}$ under $P_{\hat{p}(x)}$ vertically – and locate relevant quantiles.

> **Math Box 3.7.1** (Percentile interval). Assume that there exists a strictly increasing function $\varphi : \mathbb{R} \to \mathbb{R}$ such that for all $\theta \in \Theta$ the distribution of $\varphi(\hat{\tau})$ is symmetric around $\varphi(\tau)$ under P_θ. This means that the distribution of $\varphi(\hat{\tau}) - \varphi(\tau)$ equals the distribution of $\varphi(\tau) - \varphi(\hat{\tau})$, and in particular it follows that
>
> $$\varphi(\tau) - z'_\alpha(\theta) = w'_\alpha(\theta) - \varphi(\tau),$$
>
> where $z'_\alpha(\tau)$ and $w'_\alpha(\tau)$ are the $\alpha/2$ and $1 - \alpha/2$ quantiles for the distribution of $\varphi(\hat{\tau})$ under P_θ. The construction of standard $(1 - \alpha)$-confidence intervals for $\varphi(\tau)$ gives
>
> $$[2\varphi(\hat{\tau}(x)) - w'_\alpha(\hat{\tau}(x)), 2\varphi(\hat{\tau}(x)) - z'_\alpha(\hat{\tau}(x))] = [z'_\alpha, w'_\alpha].$$
>
> Using that φ is strictly increasing allows us to take the inverse and find that the corresponding quantiles for the distribution of $\hat{\tau}$ are $z_\alpha = \varphi^{-1}(z'_\alpha)$ and $w_\alpha = \varphi^{-1}(w'_\alpha)$. The confidence interval obtained in this way for τ – by transforming back and forth using φ – is therefore precisely the percentile interval $[z_\alpha, w_\alpha]$.
>
> This argument in favor of the percentile method relies on the existence of the function φ, which introduces symmetry and justifies the interchange of the quantiles. For any practical computation it is completely irrelevant to actually know φ. What is really the crux of the matter is whether there exists such a φ that also makes the distribution of $\varphi(\hat{\tau}) - \varphi(\tau)$ largely independent of θ.

We may discard this construction by arguing that the $[z_\alpha, w_\alpha]$ is simply a misunderstanding of the idea in confidence intervals. Confidence intervals are intervals of parameters for which the observation is reasonably likely. The interval $[z_\alpha, w_\alpha]$ is an interval where the estimator will take its value with high probability if $\theta = \hat{\theta}(x)$. This is in principle something entirely different. There are arguments, though, to justify the procedure. They are based on the existence of an implicit parameter transformation combined with a symmetry consideration; see Math Box 3.7.1 The interval $[z_\alpha, w_\alpha]$ – and its refinements – is known in the literature as the *percentile confidence interval* or *percentile interval* for short.

Symmetry, or approximate symmetry, of the distribution of the estimators around τ makes the percentile interval and the classical interval very similar in many practical cases. But rigorous arguments that justify the use of the percentile intervals are even more subtle; see Math Box 3.7.1. If the distribution of the estimator is skewed, the percentile intervals may suffer from having actual coverage which is too small. There are various ways to remedy this, known as the bias corrected percentile method and the accelerated bias corrected percentile method, but we will not pursue these matters here.

There is one argument in favor of the percentile method, and that is invariance under monotone parameter transformations. This means that if $\varphi : \mathbb{R} \to \mathbb{R}$ is an increasing function and $\tau' = \varphi(\tau)$ is our new parameter of interest, then if $[z_\alpha, w_\alpha]$ is a $(1 - \alpha)$-percentile confidence interval for τ, then $[\varphi(z_\alpha), \varphi(w_\alpha)]$ is a $(1 - \alpha)$-percentile confidence interval for $\tau' = \varphi(\tau)$.

Exercises

Exercise 3.7.1. Consider the Poisson model from Exercise 3.3.4. Implement parametric bootstrapping for computing confidence intervals for the two parameters α and β. Compute for the Ceriodaphnia data 95% confidence intervals for the parameters. The Poisson model can be estimated using `glm` in R:

```
glm(organisms~concentration,family=poisson,data=Ceriodaphnia)
```

use `summary` on the result to compute estimates of the standard error and compare the bootstrapped confidence intervals with standard confidence intervals based on the estimated standard error.

Exercise 3.7.2. Consider the four parameter logistic model for the ELISA data from Example 3.6.8. Implement a bootstrapping algorithm (parametric or non-parametric) for the computation of confidence intervals for the predicted value $g_{\hat{\beta}}(x)$. Use the algorithm for computing 95%-confidence intervals for predicted values using the ELISA data from Example 3.6.8 for different choices of x – make a plot of the confidence bands. Compare with corresponding confidence bands computed using a linear regression on the log-log-transformed data as in Example 3.6.2.

4

Mean and Variance

We have previously introduced the mean and the variance for probability measures on $\mathbb{R}$ given by a density or given by point probabilities. For the further development it is beneficial to put these definitions into a more general context. In this chapter we deal with expectations of real valued random variables in general. We get several convenient results about how to compute expectations (means) and variances, and we get access to a better understanding of how the empirical versions approximate the theoretical mean and variance. We also touch upon higher order moments and we discuss the multivariate concept of covariance. As an illustration of how some of these methods can be applied, we discuss Monte Carlo simulations as a general method based on random simulation for computing integrals numerically and we discuss aspects of asymptotic theory. The chapter ends with a brief treatment of entropy.

4.1 Expectations

For the general development and computations of means and variances of real valued random variables it is useful to introduce some notation. With reference to Sections 2.4 and 2.6 we define the *expectation* of a real valued random variable by one of the two following definitions.

Definition 4.1.1. *If X is a real valued random variable with density f then*

$$\mathbb{E}X = \int_{-\infty}^{\infty} x f(x) \mathrm{d}x$$

denotes its expectation provided that

$$\int_{-\infty}^{\infty} |x| f(x) \mathrm{d}x < \infty,$$

in which case we say that X has finite expectation.

Definition 4.1.2. *If X is a discrete random variable taking values in $E \subseteq \mathbb{R}$ with point probabilities $(p(x))_{x \in E}$ then*

$$\mathbb{E}X = \sum_{x \in E} xp(x)$$

denotes its expectation provided that

$$\sum_{x \in E} |x|p(x) < \infty.$$

in which case we say that X has finite expectation.

If X_1 and X_2 are two real valued random variables with a joint distribution having density $f(x_1, x_2)$ Result 2.13.5 shows that their marginal densities are

$$f_1(x_1) = \int f(x_1, x_2)\mathrm{d}x_1 \quad f_2(x_2) = \int f(x_1, x_2)\mathrm{d}x_1.$$

Provided that the integrals make sense we can compute the expectation of X_1 as

$$\mathbb{E}X_1 = \int x_1 f_1(x_1)\mathrm{d}x_1 = \int \int x_1 f(x_1, x_2)\mathrm{d}x_2\mathrm{d}x_1$$

and similarly for the expectation of X_2. What about the expectation of $X_1 + X_2$? In principle we have to compute the distribution of $X_1 + X_2$ first and then compute the mean value for that distribution. The computation of the distribution of the transformed variable $X_1 + X_2$ can, however, be bypassed using the following result, which we state without a derivation.

Result 4.1.3 (Transformations). *If $h : \mathbb{R}^n \to \mathbb{R}$ is a real valued function, if the distribution of X has density $f : \mathbb{R}^n \to [0, \infty)$ and if $h(X)$ has finite expectation then*

$$\begin{aligned} \mathbb{E}h(X) &= \int h(x)f(x)\mathrm{d}x \\ &= \underbrace{\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty}}_{n} h(x_1, \ldots, x_n)f(x_1, \ldots, x_n)\mathrm{d}x_1 \cdots \mathrm{d}x_n. \end{aligned}$$

Remark 4.1.4. Whether h has finite expectation can be checked by computing

$$\int |h(x)|f(x)\mathrm{d}x = \underbrace{\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty}}_{n} |h(x_1, \ldots, x_n)|f(x_1, \ldots, x_n)\mathrm{d}x_1 \cdots \mathrm{d}x_n.$$

The result in Result 4.1.3 is very useful since we may not be able to find an explicit analytic expression for the density of the distribution of $h(X)$ – the distribution may not even have a density – but often the distribution of X is specified in terms of the density f. The computation of the iterated integrals can, however, be a horrendous task.

Taking $h(x_1, x_2) = x_1 + x_2$ we find that

$$
\begin{aligned}
\mathbb{E}(X_1 + X_2) &= \int\int (x_1 + x_2)f(x_1, x_2)\mathrm{d}x_1\mathrm{d}x_2 \\
&= \int\int x_1 f(x_1, x_2)\mathrm{d}x_2\mathrm{d}x_1 + \int\int x_2 f(x_1, x_2)\mathrm{d}x_1\mathrm{d}x_2 \\
&= \mathbb{E}X_1 + \mathbb{E}X_2.
\end{aligned}
$$

The computations are sensible if $X_1 + X_2$ has finite expectation, but using the remark above and noting that $|x_1 + x_2| \leq |x_1| + |x_2|$ it follows that $X_1 + X_2$ indeed has finite expectation if X_1 and X_2 have finite expectations. In conclusion, the expectation of the sum is the sum of the expectations.

A result similar to Result 4.1.3 but for a discrete distribution can also be derived. In fact, we will do that here. If X is a random variable with values in a discrete set E, the random variable $h(X)$ takes values in the discrete subset $E' \subseteq \mathbb{R}$ given by

$$E' = \{h(x) \mid x \in E\}.$$

For each $z \in E'$ we let $A_z = \{x \in E \mid h(x) = z\}$ denote the set of all x's in E, which h maps to z. Note that each $x \in E$ belongs to exactly one set A_z. We say that the sets A_z, $z \in E'$, form a *disjoint partition* of E. The distribution of $h(X)$ has point probabilities $(q(z))_{z \in E'}$ given by

$$q(z) = P(A_z) = \sum_{x \in A_z} p(x)$$

by Definition 2.9.3, and using Result 4.1.2 the expectation of $h(X)$ can be written as

$$\mathbb{E}h(X) = \sum_{z \in E'} zp(z) = \sum_{z \in E'} z \sum_{x \in A_z} q(x).$$

Now the function h is constantly equal to z on A_z so we get

$$\mathbb{E}h(X) = \sum_{z \in E'} \sum_{x \in A_z} h(x)p(x).$$

Since the sets A_z, $z \in E'$, form a disjoint partition of the sample space E the sum on the r.h.s. above is precisely a sum over all elements in E, hence

$$\mathbb{E}h(X) = \sum_{x \in E} h(x)p(x).$$

We have derived the following result.

Result 4.1.5 (Transformations). *Let X be a random variable taking values in a discrete set E with distribution P given by the point probabilities $(p(x))_{x \in E}$. If $h : E \to \mathbb{R}$ is any real valued function then if $h(X)$ has finite expectation*

$$\mathbb{E}h(X) = \sum_{x \in E} h(x)p(x).$$

Remark 4.1.6. Similar to the continuous case, $h(X)$ has finite expectation if and only if

$$\sum_{x \in E} |h(x)|p(x) < \infty.$$

Remark 4.1.7. If X is a Bernoulli variable with success probability p we find that

$$\mathbb{E}X = 1 \times \mathbb{P}(X = 1) + 0 \times \mathbb{P}(X = 0) = p. \tag{4.1}$$

In Section 4.2 we will develop the theoretical details for the assignment of an expectation to a real valued random variable. In summary, a *positive* real valued random variable X can be assigned an expectation $\mathbb{E}X \in [0, \infty]$ – but it may be equal to ∞. A real valued random variable X with $\mathbb{E}|X| < \infty$ – the expectation of the positive real valued random variable $|X|$ is finite – can be assigned an expectation $\mathbb{E}X \in \mathbb{R}$. If $\mathbb{E}|X| < \infty$ we say that the random variable X has finite expectation. The main conclusion from Section 4.2 can be stated as the following result.

Result 4.1.8. *If X and Y are two real valued random variables with finite expectation then $X + Y$ has finite expectation and*

$$\mathbb{E}(X + Y) = \mathbb{E}X + \mathbb{E}Y.$$

Furthermore, if $c \in \mathbb{R}$ is a real valued constant then cX has finite expectation and

$$\mathbb{E}(cX) = c\mathbb{E}X.$$

Moreover, if X and Y are independent real valued random variables with finite expectation then

$$\mathbb{E}(XY) = \mathbb{E}X\,\mathbb{E}Y.$$

Example 4.1.9. If $X_1, \ldots, X_n$ are iid Bernoulli variables with success probability p then

$$X = X_1 + \ldots + X_n \sim \text{Bin}(n, p).$$

We can find the expectation for the binomially distributed random variable X by using Result 4.1.2

$$\mathbb{E}X = \sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k},$$

but it requires a little work to compute this sum. It is much easier to use Result 4.1.8 together with (4.1) to obtain that

$$\mathbb{E}X = \mathbb{E}X_1 + \ldots + \mathbb{E}X_n = p + \ldots + p = np.$$

<div align="right">◇</div>

Example 4.1.10. Let $X_1, \ldots, X_n$ be iid random variables with values in the four letter alphabet $E = \{A, C, G, T\}$. Let $w = w_1 w_2 \ldots w_m$ denote a *word* from this alphabet with m letters. Assume that $m \leq n$ (we may think of $m \ll n$) and define

$$N = \sum_{i=1}^{n-m+1} 1(X_i X_{i+1} \ldots X_{i+m-1} = w).$$

Thus N is the number of times word w occurs in the sequence. It follows using (4.1) and Result 4.1.8 that the expectation of N is

$$
\begin{aligned}
\mathbb{E}N &= \sum_{i=1}^{n-m+1} \mathbb{E}1(X_i X_{i+1} \ldots X_{i+m-1} = w) \\
&= \sum_{i=1}^{n-m+1} \mathbb{P}(X_i X_{i+1} \ldots X_{i+m-1} = w)
\end{aligned}
$$

Since the X-variables are independent we have that

$$
\begin{aligned}
\mathbb{P}(X_i X_{i+1} \ldots X_{i+m-1} = w) &= \mathbb{P}(X_i = w_1)\mathbb{P}(X_{i+1} = w_2) \ldots \mathbb{P}(X_{i+m-1} = w_m) \\
&= p(w_1)p(w_2) \ldots p(w_m) \\
&= p(A)^{n_w(A)} p(C)^{n_w(C)} p(G)^{n_w(G)} p(T)^{n_w(T)}
\end{aligned}
$$

where $p(A), p(C), p(G)$ and $p(T)$ are the point probabilities for the distribution of the X-variables and $n_w(A), n_w(C), n_w(G)$ and $n_w(T)$ are the number of A's, C's, G's and T's in w. Thus

$$\mathbb{E}N = (n - m + 1)p(A)^{n_w(A)} p(C)^{n_w(C)} p(G)^{n_w(G)} p(T)^{n_w(T)}. \tag{4.2}$$

$\diamond$

If X is a real valued random variable and if we consider the family of transformations $h_k(x) = x^k$ for $k \in \mathbb{N}$ the corresponding family of expectations of the transformed variables $h_k(X) = X^k$ for $k \in \mathbb{N}$ have special names. These expectations are known as the *moments* of X, and when the distribution of X is given by either a density or by point probabilities on a discrete subset of $\mathbb{R}$, Results 4.1.3 and 4.1.5 provide methods for computing these moments.

Definition 4.1.11. *If X is a real valued random variable with $\mathbb{E}|X|^k < \infty$ we say that X has finite k'th moment and call*

$$\mathbb{E}X^k$$

the k'th moment of X. The central k'th moment of X is

$$\mathbb{E}(X - \mathbb{E}X)^k.$$

For $k = 2$ the *central second moment* is the variance, and will be treated in greater detail below.

4.1.1   The empirical mean

The average or empirical mean of $x_1, \ldots, x_n \in \mathbb{R}$ is defined as

$$\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Occasionally we may also use the notation $\overline{x}$ to denote the average of n real numbers.

Though the average is always a well defined quantity, we can only really interpret this quantity if we regard the x's as a realization of identically distributed random variables. If those variables have distribution P and if X is a random variable with distribution P, then we regard $\hat{\mu}_n$ as an estimate of $\mu = \mathbb{E}X$. In Section 4.5 we derive some results that say more precisely how $\hat{\mu}_n$ behaves as an estimate of $\mu = \mathbb{E}X$.

In this section we will make one observation – namely that the average is also the expectation of a random variable, whose distribution is *the empirical distribution*, ε_n, given by $x_1, \ldots, x_n$. The empirical distribution on $\mathbb{R}$, which is defined in terms of $x_1, \ldots, x_n$, can be seen as a transformation of the uniform distribution on $\{1, \ldots, n\}$ via

$$h_{x_1, \ldots, x_n} : \{1, \ldots, n\} \to \mathbb{R},$$

which is defined by

$$h_{x_1, \ldots, x_n}(i) = x_i.$$

By Result 4.1.5 the expectation of a random variable with distribution ε_n is therefore given as

$$\sum_{i=1}^{n} h_{x_1, \ldots, x_n}(i)p(i) = \sum_{i=1}^{n} x_i \frac{1}{n} = \hat{\mu}_n$$

where $p(i) = 1/n$ are the point probabilities for the uniform distribution on $\{1, \ldots, n\}$.

Notationally it may be useful to express what the distribution of X is when we compute its expectation – if that is not clear from the context. Therefore we write $\mathbb{E}_P X$ for the expectation, $\mathbb{E}X$, if the distribution of X is P. This allows us for instance to write

$$\hat{\mu}_n = \mathbb{E}_{\varepsilon_n} X,$$

indicating that if X has the empirical distribution ε_n given by $x_1, \ldots, x_n$, then its expectation is the average of $x_1, \ldots, x_n$.

If $X_1, \ldots, X_n$ are iid with distribution P and (finite) expectation $\mu = \mathbb{E}_P X_1$, and if $x_1, \ldots, x_n$ are realizations of $X_1, \ldots, X_n$, then since we regard the empirical distribution ε_n to be an approximation of P due to the frequency interpretation we may regard $\hat{\mu}_n$ as an approximation of μ. This can be regarded as an application of the plug-in principle.

Exercises

> **Exercise 4.1.1.** Consider the setup from Example 4.1.10 with the word $w =$ `TATAAA`. Compute the expectation of N when $n = 10000$ and the random variables have the uniform distribution on the alphabet.

4.2   More on expectations

This section contains a rather technical development of the results for computing expectations based on some fundamental results from *measure and integration theory* that is beyond the scope of these notes. The section can be skipped in a first reading.

Result 4.2.1 (Fact). *There exists an expectation operator $\mathbb{E}$ that assigns to any positive random variable X (i.e. X takes values in $[0, \infty)$) a number*

$$\mathbb{E}X \in [0, \infty]$$

such that:

1. *If X is a Bernoulli random variable then*

$$\mathbb{E}X = \mathbb{P}(X = 1). \tag{4.3}$$

2. *If $c \geq 0$ is a positive constant and X is a positive random variable then*

$$\mathbb{E}(cX) = c\mathbb{E}X. \tag{4.4}$$

3. *If X and Y are two positive random variables then*

$$\mathbb{E}(X + Y) = \mathbb{E}X + \mathbb{E}Y. \tag{4.5}$$

The number $\mathbb{E}X$ is called the expectation or mean of X.

One should note two things. First, the expectation operator assigns an expectation to *any* random variable provided it is real valued and positive. Second, that the expectation may be $+\infty$. The theorem presented as a fact above gives little information about how to *compute* the expectation of a random variable. For this, the reader is referred to the previous section.

One of the consequences that we can observe right away is, that if X and Y are two positive random variables such that $X \leq Y$ (meaning $\mathbb{P}(X \leq Y) = 1$), then $Y - X \geq 0$ is a positive random variable with $\mathbb{E}(Y - X) \geq 0$. Consequently, if $X \leq Y$, then $X + Y - X = Y$ and

$$\mathbb{E}X \leq \mathbb{E}X + \mathbb{E}(Y - X) = \mathbb{E}(X + X - Y) = \mathbb{E}Y. \tag{4.6}$$

Math Box 4.2.1 (The expectation operator). The expectation operator of a positive random variable X can be defined in the following way. First, for $0 = s_0 < s_1 < \ldots < s_n$ some positive real numbers we form a subdivision of the positive half line $[0, \infty)$ into $n + 1$ disjoint intervals $I_0, I_2, \ldots, I_n$ given as

$$
\begin{array}{ccccccccc}
[0, \infty) = & I_0 & \cup & I_1 & \cup \ldots \cup & I_{n-1} & \cup & I_n \\
& \| & & \| & \cdots & \| & & \| \\
= & [0, s_1] & \cup & (s_1, s_2] & \cup \ldots \cup & (s_{n-1}, s_n] & \cup & (s_n, \infty)
\end{array}
$$

Then we can compute the average of the s_i's *weighted* by the probabilities that $X \in I_i$:

$$
\xi_n(X) := \sum_{i=0}^{n} s_i \mathbb{P}(X \in I_i).
$$

If the size of each of the intervals shrinks towards zero as $n \to \infty$ and $s_n \to \infty$ then it is possible to show that $\xi_n(X)$ always converges to something. Either a positive real number or $+\infty$. This limit is called the expectation of X and we may write

$$
\mathbb{E}X = \lim_{n \to \infty} \xi_n(X).
$$

Note that $\xi_n(X)$ is defined entirely in terms of the distribution of X and so is the limit.

Moreover, we can also regard

$$
X_n = \sum_{i=0}^{n} s_i 1(X \in I_i)
$$

as an approximation to the variable X. We round to the largest s_i smaller than X. Then if we want the three properties of $\mathbb{E}$ in Result 4.2.1 to be fulfilled, it follows that

$$
\mathbb{E}X_n = \sum_{i=1}^{n} s_i \mathbb{E}1(X \in I_i) = \sum_{i=1}^{n} s_i P(X \in I_i) = \xi_n(X)
$$

using for instance that $1(X \in I_i)$ is a Bernoulli variable. It is definitely beyond the scope of these notes to show that $\xi_n(X)$ converges let alone that Result 4.2.1 holds for the limit. The mathematically inclined reader is referred to the literature on measure and integration theory.

Moreover, note that if X is any random variable and A an event then $1(X \in A)$ is a Bernoulli random variable and by (4.3)

$$
\mathbb{E}1(X \in A) = \mathbb{P}(X \in A).
$$

Example 4.1.10 in the previous section actually showed how the expectation of a *positive* random variable could be computed from the few elementary rules in Result 4.2.1. One question arises. Why does Result 4.2.1 require that X is positive and not

just real valued? The answer lies in the following consideration: If X is a real valued random variable we can define two positive real valued random variables by

$$X^+ = \max\{X, 0\} \quad \text{and} \quad X^- = \max\{-X, 0\},$$

which are called the positive and negative part of X respectively. They are both transformations of X, and one can get X back from these two variables by

$$X = X^+ - X^-.$$

This holds because in general for any $x \in \mathbb{R}$ we have that

$$x = \max\{x, 0\} + \min\{x, 0\} = \max\{x, 0\} - \max\{-x, 0\}.$$

A natural and desirable extension of the additivity property, as given in (4.5), is that

$$\mathbb{E}X = \mathbb{E}(X^+ - X^-) = \mathbb{E}X^+ - \mathbb{E}X^-. \tag{4.7}$$

However, for the right hand side to make sense, we can not have that both the terms are equal to ∞. There is simply no way to make sense out of subtracting ∞ from ∞. On the other hand, by Result 4.2.1 above both terms $\mathbb{E}X^+$ and $\mathbb{E}X^-$ are well defined, and *if* they are both finite, we can subtract them. Moreover, in analogy with the reconstruction of X from X^+ and X^-, one can also construct the transformed variable $|X|$ – the absolute value of X – from X^+ and X^-. Indeed, for $x \in \mathbb{R}$ we have that

$$|x| = \max\{x, 0\} + \max\{-x, 0\},$$

hence

$$|X| = X^+ + X^-.$$

Since $|X|$ is a positive random variable $\mathbb{E}|X|$ exists by Result 4.2.1, and by (4.5) we have that

$$\mathbb{E}|X| = \mathbb{E}X^+ + \mathbb{E}X^-.$$

The sum is finite if and only if both terms are finite. This leads us to the definition:

Definition 4.2.2. *If X is a real valued random variable we say that it has finite expectation if*

$$\mathbb{E}|X| < \infty.$$

In this case the expectation of X, $\mathbb{E}X$, is well defined by (4.7), that is

$$\mathbb{E}X = \mathbb{E}X^+ - \mathbb{E}X^-.$$

Without too much trouble, one can now show that the properties listed above as (4.4) and (4.5) for the expectation of a positive random variable carries over in an appropriate form for the expectation of a general real valued variable. This provides a proof of the fundamental Result 4.1.8.

The derivation of Result 4.1.8 from the properties of $\mathbb{E}$ for positive random variables is mostly a matter of bookkeeping. First we note that

$$|X + Y| \leq |X| + |Y|,$$

so if $\mathbb{E}|X| < \infty$ and $\mathbb{E}|Y| < \infty$ then by (4.6) $\mathbb{E}|X + Y| < \infty$. So $X + Y$ does have finite expectation if X and Y do.

To show the additivity we need a little trick. We have that

$$(X + Y)^+ - (X + Y)^- = X + Y = X^+ - X^- + Y^+ - Y^-,$$

which can be rearranged to give

$$(X + Y)^+ + X^- + Y^- = X^+ + Y^+ + (X + Y)^-,$$

where all terms on both sides are *positive* random variables. Then we can apply Result 4.2.1, (4.5), to obtain

$$\mathbb{E}(X + Y)^+ + \mathbb{E}X^- + \mathbb{E}Y^- = \mathbb{E}(X + Y)^- + \mathbb{E}X^+ + \mathbb{E}Y^+$$

and rearranging this equality back yields

$$\mathbb{E}(X + Y) = \mathbb{E}(X + Y)^+ - \mathbb{E}(X + Y)^- = \mathbb{E}X^+ - \mathbb{E}X^- + \mathbb{E}Y^+ - \mathbb{E}Y^- = \mathbb{E}X + \mathbb{E}Y.$$

If $c \in \mathbb{R}$ and X is a real valued random variable $|cX| = |c||X|$, hence

$$\mathbb{E}|cX| = \mathbb{E}|c||X| = |c|\mathbb{E}|X|,$$

and we see that cX has finite expectation if and only if $|X|$ has finite expectation. Moreover, if $c \geq 0$

$$(cX)^+ = cX^+ \quad \text{and} \quad (cX)^- = cX^-,$$

thus

$$\mathbb{E}(cX) = \mathbb{E}(cX^+) - \mathbb{E}(cX^-) = c\mathbb{E}X^+ - c\mathbb{E}X^- = c(\mathbb{E}X^+ - \mathbb{E}X^-) = c\mathbb{E}X.$$

If $c < 0$

$$(cX)^+ = -cX^- \quad (cX)^- = -cX^+$$

and

$$\mathbb{E}(cX) = \mathbb{E}(-cX^-) - \mathbb{E}(-cX^+) = -c(\mathbb{E}X^- - \mathbb{E}X^+) = -c(-\mathbb{E}X) = c\mathbb{E}X.$$

The proof of $\mathbb{E}(XY) = \mathbb{E}X\,\mathbb{E}Y$ for *independent* variables is skipped since it requires more knowledge of the underlying definition of the expectation. Note, however, that if X and Y are *independent* Bernoulli random variables then XY is a Bernoulli random variable and

$$\mathbb{E}(XY) = \mathbb{P}(XY = 1) = \mathbb{P}(X = 1, Y = 1) = \mathbb{P}(X = 1)\mathbb{P}(Y = 1) = \mathbb{E}X\,\mathbb{E}Y$$

where the third equality follows form independence of X and Y. Without independence the conclusion is wrong. This is the starting point for a general proof – relying on the construction discussed in Math Box 4.2.1.

At this point in the abstract development it is natural to note that the Definitions 4.1.1 and 4.1.2 of the expectation of course coincide with the abstract definition. In fact, Definitions 4.1.1 and 4.1.2 can be viewed as computational techniques for actually computing the expectation. As we showed in Example 4.1.10 we may be able to derive the mean of a random variable without getting even close to understanding the actual distribution of the random variable. The integer variable N introduced in Example 4.1.10 has a far more complicated distribution than we can handle, but its mean can be derived based on the simple computational rules for $\mathbb{E}$. Returning to Definitions 4.1.1 and 4.1.2 for computing the expectation should be regarded as a last resort.

Example 4.2.3 (Mixtures)**.** Most of the standard distributions like the normal distribution are not suited for models of *multimodal data*. By this we mean data that cluster around several different points in the sample space. This will show up on a histogram as multiple modes (peaks with valleys in between). If we for instance consider gene expressions for a single gene in a group of patients with a given disease, multiple modes may occur as a result of, yet unknown, subdivisions of the disease on the gene expression level. In general, we want to capture this phenomena that there is a subdivision of the observed variable according to an *unobserved* variable. This is captured by a triple of independent variables (Y, Z, W) where Y and Z are real valued random variables and W is a Bernoulli variable with $\mathbb{P}(W = 1) = p$. Then we define

$$X = YW + Z(1 - W).$$

The interpretation is that the distribution of X is given as a *mixture* of the distribution of Y and Z in the sense that either $W = 1$ (with probability p) in which case $X = Y$ or else $W = 0$ (with probability $1 - p$) in which case $X = Z$. Since $|X| \leq |Y|W + |Z|(1 - W) \leq |Y| + |Z|$ it follows by Result 4.1.8 that

$$\mathbb{E}|X| \leq \mathbb{E}|Y| + \mathbb{E}|Z|.$$

Thus X has finite expectation if Y and Z have finite expectation. Moreover, Result 4.1.8 implies, since Y and W are independent and Z and W are independent, that

$$\mathbb{E}X = \mathbb{E}(YW) + \mathbb{E}(Z(1 - W)) = \mathbb{E}Y \, \mathbb{E}W + \mathbb{E}Z \, (1 - \mathbb{E}(W)) = p\mathbb{E}Y + (1 - p)\mathbb{E}Z.$$

The probabilities p and $1 - p$ of $(W = 1)$ and $(W = 0)$ respectively are called the *mixture proportions*, and the equality above says, quite intuitively, that the expectation of X is a weighted average of the expectation of Y and the expectation of Z – weighted according to the mixture proportions.

If $Y \sim N(\xi_1, \sigma_1^2)$ and $Z \sim N(\xi_2, \sigma_2^2)$ then

$$\mathbb{E}X = p\xi_1 + (1 - p)\xi_2,$$

Math Box 4.2.2. If X is a positive real valued random variable with distribution function F then it has finite expectation if and only if $\int_0^\infty 1 - F(x)\mathrm{d}x < \infty$ in which case

$$\mathbb{E}X = \int_0^\infty 1 - F(x)\mathrm{d}x.$$

This makes us in principle capable of computing the expectation of any real valued random variable X. Both X^+ and X^- are positive random variables, and with F^+ and F^- denoting their respective distribution functions we get that if

$$\int_0^\infty 1 - F^+(x)\mathrm{d}x < \infty \quad \text{and} \quad \int_0^\infty 1 - F^-(x)\mathrm{d}x < \infty$$

then X has finite expectation and

$$\begin{aligned} \mathbb{E}X &= \mathbb{E}X^+ - \mathbb{E}X^- \\ &= \int_0^\infty 1 - F^+(x)\mathrm{d}x - \int_0^\infty 1 - F^-(x)\mathrm{d}x \\ &= \int_0^\infty F^-(x) - F^+(x)\mathrm{d}x. \end{aligned}$$

but note that the distribution of X is far from being a normal distribution. $\diamond$

4.3   Variance

Definition 4.3.1. *If X is a real valued random variable with expectation $\mathbb{E}X$, then if X has finite second moment, that is, if X^2 has finite expectation, we define the variance of X as*

$$\mathbb{V}X = \mathbb{E}(X - \mathbb{E}X)^2 \tag{4.8}$$

and the standard deviation is defined as $\sqrt{\mathbb{V}X}$.

The variance is the expectation of the squared difference between X and its expectation $\mathbb{E}X$. This is a natural way of measuring how variable X is.

Remark 4.3.2. Writing out $(X - \mathbb{E}X)^2 = X^2 - 2X\mathbb{E}X + (\mathbb{E}X)^2$ and using Result 4.1.8 we obtain

$$\mathbb{V}X = \mathbb{E}X^2 - 2\mathbb{E}X\mathbb{E}X + (\mathbb{E}X)^2 = \mathbb{E}X^2 - (\mathbb{E}X)^2, \tag{4.9}$$

which is a useful alternative way of computing the variance. The expectation of X^2, $\mathbb{E}X^2$, is called the *second moment* of the distribution of X.

Remark 4.3.3. For any $\mu \in \mathbb{R}$ we can write

$$\begin{aligned} (X - \mu)^2 &= (X - \mathbb{E}X + \mathbb{E}X - \mu)^2 \\ &= (X - \mathbb{E}X)^2 + 2(X - \mathbb{E}X)(\mathbb{E}X - \mu) + (\mathbb{E}X - \mu)^2, \end{aligned}$$

from which

$$
\begin{aligned}
\mathbb{E}(X - \mu)^2 &= \mathbb{E}(X - \mathbb{E}X)^2 + 2(\mathbb{E}X - \mathbb{E}X)(\mathbb{E}X - \mu) + (\mathbb{E}X - \mu)^2 \\
&= \mathbb{V}X + (\mathbb{E}X - \mu)^2 \geq \mathbb{V}X
\end{aligned}
\tag{4.10}
$$

with equality if and only if $\mathbb{E}X = \mu$. The number $\mathbb{E}(X - \mu)^2$ is the expected squared difference between μ and X, and as such a measure of how much the outcome deviates from μ on average. We see that the expectation $\mathbb{E}X$ is the unique value of μ that minimizes this measure of deviation. The expectation is therefore in this sense the best constant approximation to any outcome of our experiment.

Example 4.3.4. If X is a Bernoulli random variable with success probability p we know that $\mathbb{E}X = p$. We find using Result 4.1.5 that

$$
\begin{aligned}
\mathbb{V}X &= \mathbb{E}(X - p)^2 = (1 - p)^2 \mathbb{P}(X = 1) + p^2 \mathbb{P}(X = 0) \\
&= (1 - p)^2 p + p^2 (1 - p) = (1 - p)p(1 - p + p) = (1 - p)p.
\end{aligned}
$$

$\diamond$

Example 4.3.5. If X is a random variable with mean 0 and variance 1, if $\mu \in \mathbb{R}$ and $\sigma > 0$, then

$$
\mathbb{E}(\sigma X + \mu) = \sigma \mathbb{E}(X) + \mu = \mu
$$

and

$$
\mathbb{V}(\sigma X + \mu) = \mathbb{E}(\sigma X + \mu - \mu)^2 = \mathbb{E}(\sigma^2 X^2) = \sigma^2 \mathbb{V}X = \sigma^2.
$$

This shows that the location scale-transformation, as considered in Example 2.9.9, of X has mean μ and variance σ^2. This was shown in Example 2.9.9 for the case where the distribution had a density. In the other direction we find that if $\mathbb{E}X = \mu$ and $\mathbb{V}X = \sigma^2$ then

$$
\mathbb{E}\left(\frac{X - \mu}{\sigma}\right) = 0 \quad \text{and} \quad \mathbb{V}\left(\frac{X - \mu}{\sigma}\right) = 1.
$$

We refer to $\frac{X - \mu}{\sigma}$ as the normalization of X, which has mean 0 and standard deviation 1. $\diamond$

If $x_1, \ldots, x_n$ are realizations of n identically distributed random variables with finite second moment and whose variance is σ^2, the sample variance or empirical variance of $x_1, \ldots, x_n$ is defined as

$$
\tilde{\sigma}_n^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu}_n)^2
\tag{4.11}
$$

where $\hat{\mu}_n$ is the empirical mean. The empirical variance is an estimate of the variance σ^2. Like the empirical mean, the empirical variance is the variance of a random

variable having distribution ε_n. With the notation as in Section 4.1.1 and using Result 4.1.5 the variance of X having distribution ε_n is

$$
\begin{aligned}
\mathbb{V}_{\varepsilon_n} X &= \mathbb{E}_{\varepsilon_n}(X - \mathbb{E}_{\varepsilon_n}X)^2 \\
&= \sum_{i=1}^{n}(h_{x_1,\ldots,x_n}(i) - \hat{\mu}_n)^2 \frac{1}{n} \\
&= \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu}_n)^2 = \tilde{\sigma}_n^2.
\end{aligned}
$$

As for the expectation we use the subscript notation, $\mathbb{V}_P X$, to denote the variance, $\mathbb{V}X$, of X if the distribution of X is P. The square root of the empirical variance,

$$
\tilde{\sigma}_n = \sqrt{\tilde{\sigma}_n^2},
$$

is called the *sample standard deviation* and is an estimate of the standard deviation, σ, of the random variables.

Since $\tilde{\sigma}_n^2$ is the variance of a random variable X having distribution ε_n we can use computational rules for variances. For instance, (4.9) can be used to obtain the alternative formula

$$
\tilde{\sigma}_n^2 = \mathbb{V}_{\varepsilon_n} X = \mathbb{E}_{\varepsilon_n} X^2 - (\mathbb{E}_{\varepsilon_n}X)^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - \hat{\mu}_n^2. \tag{4.12}
$$

It should be remarked that whereas (4.9) can be quite useful for theoretical computations it may *not* be suitable for numerical computations. This is because both $\frac{1}{n}\sum_{i=1}^{n}x_i^2$ and $\hat{\mu}_n^2$ can attain very large numerical values, and subtracting numerically large numbers can lead to a serious loss of precision.

Example 4.3.6 (Empirical normalization)**.** We consider a dataset $x_1,\ldots,x_n \in \mathbb{R}$ and let X be a random variable with distribution ε_n (the empirical distribution). Then by definition

$$
\mathbb{E}_{\varepsilon_n} X = \hat{\mu}_n \quad \text{and} \quad \mathbb{V}_{\varepsilon_n} X = \tilde{\sigma}_n^2.
$$

The normalized dataset, $\{x_1', \ldots, x_n'\}$, is defined by

$$
x_i' = \frac{x_i - \hat{\mu}_n}{\tilde{\sigma}_n}
$$

and the normalized empirical distribution, ε_n', is given by the normalized dataset. If X has distribution ε_n the normalized random variable

$$
X' = \frac{X - \hat{\mu}_n}{\tilde{\sigma}_n^2}
$$

has distribution ε_n', and we find, referring to Example 4.3.5 that

$$
\mathbb{E}_{\varepsilon_n'} X' = \mathbb{E}_{\varepsilon_n}\left(\frac{X - \hat{\mu}_n}{\tilde{\sigma}_n^2}\right) = 0
$$

and

$$\mathbb{V}_{\varepsilon'_n} X' = \mathbb{V}_{\varepsilon_n} \left(\frac{X - \hat{\mu}_n}{\widetilde{\sigma}_n^2} \right) = 1.$$

In other words, if we normalize the dataset using the empirical mean and sample standard deviation, the resulting normalized dataset has empirical mean 0 and sample standard deviation 1. ◇

Example 4.3.7. In Example 3.1.4 we considered the two model paradigms – additive noise and multiplicative noise. The additive noise model for X was formulated as

$$X = \mu + \sigma\varepsilon \tag{4.13}$$

where $\mu \in \mathbb{R}$, $\sigma > 0$ and ε is a real valued random variable with $\mathbb{E}\varepsilon = 0$ and $\mathbb{V}\varepsilon = 1$. Using the rules for $\mathbb{E}$ developed in this chapter we find that

$$\mathbb{E}X = \mu + \sigma\mathbb{E}\varepsilon = \mu,$$

and

$$\mathbb{V}X = \sigma^2 \mathbb{V}\varepsilon = \sigma^2.$$

The multiplicative noise model was given as

$$X = \mu\varepsilon \tag{4.14}$$

with ε a positive random variable, $\mu > 0$. In this case

$$\mathbb{E}X = \mu\mathbb{E}\varepsilon.$$

and

$$\mathbb{V}X = \mu^2 \mathbb{V}\varepsilon.$$

The notable property of the multiplicative noise model is that the standard deviation, $\sqrt{\mathbb{V}X} = \mu\sigma$, scales with the mean. That is, the larger the expected value of X is the larger is the noise. For the additive noise model the size of the noise is unrelated to the mean value.

As discussed in Example 3.1.4 one often transforms the multiplicative noise model via the logarithm to get

$$\log X = \log \mu + \log \varepsilon,$$

which is an additive noise model. Logarithms and expectations are not interchangeable, though, and it actually holds that

$$\mathbb{E} \log \varepsilon < \log \mathbb{E}\varepsilon,$$

hence if we want $\mathbb{E} \log \varepsilon = 0$ to assure that $\mathbb{E} \log X = \log \mu$ the expectation of X is always $> \mu$. If we, in addition, assume that $\log \varepsilon_i \sim N(0, \sigma^2)$ as is often done for the additive noise model the distribution of ε is called the *log-normal distribution*. In other words, $\varepsilon = \exp(Z)$ where $Z \sim N(0, \sigma^2)$. ◇

4.4 Multivariate Distributions

If we consider two real valued random variables X and Y, the bundled variable (X, Y) takes values in $\mathbb{R}^2$. The mean and variance of each of the variables X and Y rely exclusively on the marginal distributions of X and Y. Thus they tell us nothing about the joint distribution of X and Y. We introduce the covariance as a measure of dependency between X and Y.

Definition 4.4.1. *If XY has finite expectation the covariance of the random variables X and Y is defined as*

$$\mathbb{V}(X, Y) = \mathbb{E}\left((X - \mathbb{E}X)(Y - \mathbb{E}Y)\right) \tag{4.15}$$

and the correlation is defined by

$$\text{corr}(X, Y) = \frac{\mathbb{V}(X, Y)}{\sqrt{\mathbb{V}X\mathbb{V}Y}}. \tag{4.16}$$

The covariance is a measure of the covariation, that is, the dependency between the two random variables X and Y. The correlation is a standardization of the covariance by the variances of the coordinates X and Y.

We should note that the covariance is symmetric in X and Y:

$$\mathbb{V}(X, Y) = \mathbb{V}(Y, X).$$

Furthermore, if $X = Y$ then

$$\mathbb{V}(X, X) = \mathbb{E}(X - \mathbb{E}X)^2 = \mathbb{V}X,$$

that is, the covariance of X with X is simply the variance of X. Moreover,

$$
\begin{aligned}
\mathbb{V}(X, Y) &= \mathbb{E}\left((X - \mathbb{E}X)(Y - \mathbb{E}Y)\right) \\
&= \mathbb{E}(XY - X\mathbb{E}Y - Y\mathbb{E}X + \mathbb{E}X\mathbb{E}Y) \\
&= \mathbb{E}(XY) - \mathbb{E}X\mathbb{E}Y - \mathbb{E}X\mathbb{E}Y + \mathbb{E}X\mathbb{E}Y \\
&= \mathbb{E}(XY) - \mathbb{E}X\mathbb{E}Y, \tag{4.17}
\end{aligned}
$$

which gives an alternative formula for computing the covariance. Using this last formula, it follows from Result 4.1.8 that if X and Y are *independent* then

$$\mathbb{V}(X, Y) = 0.$$

On the other hand it is important to know that the covariance being equal to zero does *not* imply independence. We also obtain the generally valid formula

$$\mathbb{E}(XY) = \mathbb{E}X\mathbb{E}Y + \mathbb{V}(X, Y)$$

for the expectation of the product of two random variables.

If we have two random variables, X and Y, with finite variance, the we use (4.9) to compute the variance of $X + Y$:

$$
\begin{aligned}
\mathbb{V}(X + Y) &= \mathbb{E}(X + Y)^2 - (\mathbb{E}(X + Y))^2 \\
&= \mathbb{E}(X^2 + Y^2 + 2XY) - ((\mathbb{E}X)^2 + (\mathbb{E}Y)^2 + 2\mathbb{E}X\mathbb{E}Y) \\
&= \mathbb{E}X^2 - (\mathbb{E}X)^2 + \mathbb{E}Y^2 - (\mathbb{E}Y)^2 + 2(\mathbb{E}XY - \mathbb{E}X\mathbb{E}Y) \\
&= \mathbb{V}X + \mathbb{V}Y + 2\mathbb{V}(X, Y)
\end{aligned}
$$

using (4.9) again together with (4.17) for the last equality.

We also find that

$$
\mathbb{V}(cX) = \mathbb{E}(cX - c\mathbb{E}X)^2 = c^2\mathbb{E}(X - \mathbb{E}X)^2 = c^2\mathbb{V}X.
$$

We summaries these derivations as follows.

Result 4.4.2. *If X and Y are two random variables with finite variance then the sum $X + Y$ has finite variance and*

$$
\mathbb{V}(X + Y) = \mathbb{V}X + \mathbb{V}Y + 2\mathbb{V}(X, Y). \tag{4.18}
$$

If X is a random variable with finite variance and $c \in \mathbb{R}$ is a constant then cX has finite variance and

$$
\mathbb{V}(cX) = c^2\mathbb{V}X. \tag{4.19}
$$

Remark 4.4.3. We observe that the formula

$$
\mathbb{V}(X + Y) = \mathbb{V}X + \mathbb{V}Y \tag{4.20}
$$

holds if and only if $\mathbb{V}(X, Y) = 0$, which in particular is the case if X and Y are independent. Note also that it follows from the theorem that

$$
\begin{aligned}
\mathbb{V}(X - Y) &= \mathbb{V}(X + (-1)Y) = \mathbb{V}X + \mathbb{V}((-1)Y) + 2\mathbb{V}(X, -Y) \\
&= \mathbb{V}X + (-1)^2\mathbb{V}Y + 2\mathbb{V}(X, -Y) \\
&= \mathbb{V}X + \mathbb{V}Y - 2\mathbb{V}(X, Y).
\end{aligned}
$$

Result 4.4.4 (Cauchy-Schwarz)**.** *The covariance fulfills the Cauchy-Schwarz inequality. If X and Y are real valued random variables with finite variance*

$$
|\mathbb{V}(X, Y)| \leq \sqrt{\mathbb{V}X}\sqrt{\mathbb{V}Y}.
$$

The inequality is a classical mathematical result, but the derivation is, in fact, quite elementary given the tools we already have at our disposal, so we provide a derivation here for completeness.

Proof: A computation shows that for all $t \in \mathbb{R}$

$$
\begin{aligned}
0 &\leq \mathbb{E}((X - \mathbb{E}X) + t(Y - \mathbb{E}Y))^2 \\
&= \mathbb{E}(X - \mathbb{E}X)^2 + 2t\mathbb{E}\left((X - \mathbb{E}(X)(Y - \mathbb{E}Y)\right) + t^2\mathbb{E}(Y - \mathbb{E}(Y)^2 \\
&= \mathbb{V}X + 2t\mathbb{V}(X,Y) + t^2\mathbb{V}Y.
\end{aligned}
$$

Defining $g(t) = \mathbb{V}X + 2t\mathbb{V}(X,Y) + t^2\mathbb{V}Y$ as a second order polynomial in t we know that the polynomial can only be positive if the *discriminant* of the polynomial is negative. That is,

$$4\mathbb{V}(X,Y)^2 - 4\mathbb{V}X\mathbb{V}Y \leq 0,$$

which implies the Cauchy-Schwarz inequality. $\square$

The Cauchy-Schwarz inequality implies that the correlation is always a number between -1 and 1, that is,

$$-1 \leq \mathrm{corr}(X,Y) \leq 1.$$

If we consider not just two random variables but an n-dimensional vector $X = (X_1, \ldots, X_n)$ of real valued random variables with finite variance we can compute the n^2 covariances for each pair of variables X_i and X_j. One usually arranges the covariances in a $n \times n$ matrix Σ given by

$$\Sigma_{ij} = \mathbb{V}(X_i, Y_j).$$

That is

$$
\Sigma = \left\{
\begin{array}{cccc}
\mathbb{V}X_1 & \mathbb{V}(X_1, X_2) & \cdots & \mathbb{V}(X_1, X_n) \\
\mathbb{V}(X_2, X_1) & \mathbb{V}(X_2) & \cdots & \mathbb{V}(X_2, X_n) \\
\vdots & \vdots & \ddots & \vdots \\
\mathbb{V}(X_n, X_1) & \mathbb{V}(X_n, X_2) & \cdots & \mathbb{V}X_n
\end{array}
\right\}.
$$

Note that due to the symmetry of the covariance we have that the covariance matrix Σ is symmetric:

$$\Sigma_{ij} = \Sigma_{ji}.$$

As a direct consequence of Result 4.4.2 we have the following result about the variance of the sum of n real valued random variables.

Result 4.4.5. *If $X_1, \ldots, X_n$ are n real valued random variables with finite variance and covariance matrix Σ then*

$$
\begin{aligned}
\mathbb{V}\left(\sum_{i=1}^{n} X_i\right) &= \sum_{i=1}^{n}\sum_{j=1}^{n} \Sigma_{ij} \\
&= \sum_{i=1}^{n} \mathbb{V}X_i + 2\sum_{i<j} \mathbb{V}(X_i, X_j).
\end{aligned}
$$

Example 4.4.6. Continuing Example 4.1.10 we want to compute the variance of the counting variable N. This is a quite complicated task, and to this end it is useful to introduce some auxiliary variables. Recall that $X_1, \ldots, X_n$ are iid random variables with values in $E = \{A, C, G, T\}$ and $w = w_1 w_2 \ldots w_m$ is an m-letter word. Define for $i = 1, \ldots, n - m + 1$

$$Y_i = 1(X_i X_{i+1} \ldots X_{i+m-1} = w)$$

to be the Bernoulli random variable that indicates whether the word w occurs with starting position i in the sequence of random variables. Then

$$N = \sum_{i=1}^{n-m+1} Y_i.$$

To compute the variance of N we compute first the covariance matrix for the variables $Y_1, \ldots, Y_{n-m+1}$. If $i + m \leq j$ then Y_i and Y_j are independent because the two vectors $(X_i, X_{i+1}, \ldots, X_{i+m-1})$ and $(X_j, X_{j+1}, \ldots, X_{j+m-1})$ are independent. By symmetry

$$\Sigma_{ij} = \mathbb{V}(Y_i, Y_j) = 0$$

if $|i - j| \geq m$. If $|i - j| < m$ the situation is more complicated, because then the variables are actually dependent. We may observe that since the X-variables are iid, then for all $i \leq j$ with fixed $j - i = k < m$ the variables

$$(X_i, X_{i+1}, \ldots, X_{i+m-1}, X_j, X_{j+1}, \ldots, X_{j+m-1})$$

have the same distribution. Thus if we define

$$\rho(k) = \mathbb{V}(Y_1, Y_k)$$

for $k = 0, \ldots, m - 1$ then, using symmetry again,

$$\Sigma_{ij} = \mathbb{V}(Y_i, Y_j) = \rho(|i - j|)$$

for $|i - j| < m$. Thus for $2m - 1 \leq n$ the covariance matrix Σ has the structure

$$\Sigma = \left\{ \begin{array}{ccccccccccc}
\rho(0) & \rho(1) & \rho(2) & \ldots & \rho(m{-}1) & 0 & 0 & \ldots & 0 & 0 & 0 \\
\rho(1) & \rho(0) & \rho(1) & \ldots & \rho(m{-}2) & \rho(m{-}1) & 0 & \ldots & 0 & 0 & 0 \\
\rho(2) & \rho(1) & \rho(0) & \ldots & \rho(m{-}3) & \rho(m{-}2) & \rho(m{-}1) & \ldots & 0 & 0 & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \\
\rho(m{-}1) & \rho(m{-}2) & \rho(m{-}3) & \ldots & \rho(0) & \rho(1) & \rho(2) & \ldots & 0 & 0 & 0 \\
0 & \rho(m{-}1) & \rho(m{-}2) & \ldots & \rho(1) & \rho(0) & \rho(1) & \ldots & 0 & 0 & 0 \\
0 & 0 & \rho(m{-}1) & \ldots & \rho(2) & \rho(1) & \rho(0) & \ldots & 0 & 0 & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \\
0 & 0 & 0 & \ldots & 0 & 0 & 0 & \ldots & \rho(0) & \rho(1) & \rho(2) \\
0 & 0 & 0 & \ldots & 0 & 0 & 0 & \ldots & \rho(1) & \rho(0) & \rho(1) \\
0 & 0 & 0 & \ldots & 0 & 0 & 0 & \ldots & \rho(2) & \rho(1) & \rho(0)
\end{array} \right\},$$

which is non-zero in a diagonal band. Using Result 4.4.5 we see that there are $n - m + 1$ terms $\rho(0)$ (in the diagonal) and $2(n - m + 1 - k)$ terms $\rho(k)$ for $k = 1, \ldots, m - 1$, and therefore

$$\mathbb{V}(N) = (n - m + 1)\rho(0) + \sum_{k=1}^{m-1} 2(n - m + 1 - k)\rho(k).$$

We can also find

$$\rho(0) = \mathbb{V}(Y_1) = p_w(1 - p_w)$$

where $p_w = p(w_1)p(w_2)\dots p(w_m)$ is the "word probability". The value of $\rho(k)$ for $k = 1, \dots, m-1$ depends, however, crucially and in a complicated way on the structure of *self-overlap* in the word w.

If w does *not* self-overlap, that is, no suffix of the word equals a prefix of the same word, then $\mathbb{E}(Y_1 Y_k) = \mathbb{P}(Y_1 Y_k = 1) = 0$ for $k = 1, \dots, m-1$ and we see that

$$\rho(k) = \mathbb{V}(Y_1, Y_k) = \mathbb{E}(Y_1 Y_k) - \mathbb{E}Y_1 \mathbb{E}Y_k = -p_w^2.$$

Therefore, if w does not self-overlap we find the formula

$$
\begin{aligned}
\mathbb{V}(N) &= (n - m + 1)p_w(1 - p_w) - 2p_w^2 \sum_{k=1}^{m-1}(n - m + 1 - k) \\
&= (n - m + 1)p_w(1 - p_w) - p_w^2(m - 1)(2n + 2 - 3m)
\end{aligned}
$$

for the variance of the number of occurrences N of the word w in a sequence of n iid random variables.

We say that w has a k-shift overlap if

$$w_{1+k}w_{2+k}\dots w_m = w_1 w_2 \dots w_{m-k},$$

which means that the $m - k$ prefix of the word equals the $m - k$ suffix of the word. In that case

$$
\begin{aligned}
\mathbb{E}(Y_1 Y_k) &= \mathbb{P}(Y_1 Y_k = 1) = \mathbb{P}(X_1 \dots X_m = w, X_{m+1} \dots X_{m+k} = w_{m-k+1} \dots w_m) \\
&= p_w p(w_{m-k+1}) \dots p(w_m),
\end{aligned}
$$

and therefore

$$\rho(k) = p_w p(w_{m-k+1}) \dots p(w_m) - p_w^2$$

if w has a k-shift overlap.   ◇

With $x_1, \dots, x_n \in E$ a dataset where $x_l = (x_{1l}, \dots, x_{kl}) \in \mathbb{R}^k$ the computation of the covariance under the empirical measure gives

$$
\begin{aligned}
\tilde{\sigma}_{ij,n}^2 &= \mathbb{V}_{\varepsilon_n}(X_i, X_j) = \mathbb{E}_{\varepsilon_n}((X_i - \mathbb{E}_{\varepsilon_n} X_i)(X_j - \mathbb{E}_{\varepsilon_n} X_j)) \\
&= \frac{1}{n} \sum_{l=1}^{n}(x_{il} - \hat{\mu}_{i,n})(x_{jl} - \hat{\mu}_{j,n})
\end{aligned}
$$

where

$$\hat{\mu}_{i,n} = \frac{1}{n} \sum_{l=1}^{n} x_{il}.$$

> **Math Box 4.4.1** (Multidimensional confidence sets). If $\Theta \subseteq \mathbb{R}^d$ the θ parameter is a d-dimensional (column) vector and if $\Sigma(\theta)$ denotes the covariance matrix of $\hat{\theta}$ under P_θ a possible test statistic for testing the null-hypothesis $H_0 : \theta = \theta_0$ is
>
> $$h(x, \theta_0) = (\hat{\theta}(x) - \theta_0)^t \Sigma(\hat{\theta}(x))^{-1} (\hat{\theta}(x) - \theta_0). \tag{4.22}$$
>
> Large values are critical. The corresponding multivariate confidence set becomes
>
> $$\begin{aligned} I(x) &= \{\theta_0 \in \Theta \mid h(x, \theta_0) \leq z\} \\ &= \{\theta_0 \in \Theta \mid (\hat{\theta}(x) - \theta_0)^T \Sigma(\hat{\theta}(x))^{-1} (\hat{\theta}(x) - \theta_0) \leq z\}. \end{aligned}$$
>
> Such a set is known as an *ellipsoid* in $\mathbb{R}^d$, and if $d = 2$ the set is an ellipse.

Using (4.17) instead we obtain that

$$\widetilde{\sigma}_{ij,n}^2 = \frac{1}{n} \sum_{l=1}^n x_{il} x_{jl} - \widehat{\mu}_{i,n} \widehat{\mu}_{j,n}. \tag{4.21}$$

As for the variance this is not a recommended formula to use for the actual computation of the empirical covariance.

The empirical covariance matrix $\widetilde{\Sigma}_n$ is given by

$$\widetilde{\Sigma}_{ij,n} = \widetilde{\sigma}_{ij,n}^2$$

The empirical correlation becomes

$$\widetilde{\mathrm{corr}}_{ij,n} = \frac{\widetilde{\sigma}_{ij,n}^2}{\widetilde{\sigma}_{i,n} \widetilde{\sigma}_{j,n}} = \frac{\sum_{l=1}^n (x_{il} - \widehat{\mu}_{i,n})(x_{jl} - \widehat{\mu}_{j,n})}{\sqrt{\sum_{l=1}^n (x_{il} - \hat{\mu}_{i,n})^2 \sum_{l=1}^n (x_{jl} - \hat{\mu}_{j,n})^2}}.$$

If $x_1, \ldots, x_n$ are realizations of n identically distributed random variables with values in $\mathbb{R}^k$ the empirical covariances and correlations are estimates of the theoretical covariances and correlations.

4.5 Properties of the Empirical Approximations

We consider $X_1, \ldots, X_n$ independent and identically distributed with distribution P on E and ε_n is the corresponding empirical probability measure,

$$\varepsilon_n(A) = \frac{1}{n} \sum_{k=1}^n 1(X_k \in A), \quad A \subseteq E,$$

Let A be any event, then since $1(X_i \in A)$ is a Bernoulli variable we can use Result 4.2.1 to find that

$$\mathbb{E}1(X_i \in A) = P(A)$$

so

$$\begin{aligned}
\mathbb{E}\varepsilon_n(A) &= \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}1(X_i \in A)\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}1(X_i \in A) \\
&= \frac{1}{n}\sum_{i=1}^{n}P(A) = P(A).
\end{aligned}$$

Example 4.3.4 gives that

$$\mathbb{V}1(X_i \in A) = P(A)(1 - P(A)).$$

Hence by independence of the random variables $X_1, \ldots, X_n$ and Result 4.4.2

$$\begin{aligned}
\mathbb{V}\varepsilon_n(A) &= \mathbb{V}\left(\frac{1}{n}\sum_{i=1}^{n}1(X_i \in A)\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}1(X_i \in A) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}P(A)(1 - P(A)) = \frac{1}{n}P(A)(1 - P(A)).
\end{aligned}$$

We have derived the following result about the empirical probability measure.

Result 4.5.1. *With ε_n the empirical probability measure, and A any event it holds that*

$$\mathbb{E}\varepsilon_n(A) = P(A) \tag{4.23}$$

and

$$\mathbb{V}\varepsilon_n(A) = \frac{1}{n}P(A)(1 - P(A)). \tag{4.24}$$

As for all other probability measures the collection of numbers $\varepsilon_n(A)$ for all events $A \subseteq E$ is enormous even for a small, finite set E. If E is finite we will therefore prefer the smaller collection of frequencies

$$\varepsilon_n(z) = \frac{1}{n}\sum_{i=1}^{n}1(x_i = z)$$

for $z \in E$ – which is also sufficient for completely determining the empirical measure just like for any other probability measure on a discrete set. If P is given by the point probabilities $(p(z))_{z \in E}$ Result 4.5.1 tells us that

$$\mathbb{E}\varepsilon_n(z) = p(z) \tag{4.25}$$

and

$$\mathbb{V}\varepsilon_n(z) = \frac{1}{n}p(z)(1 - p(z)). \tag{4.26}$$

> **R Box 4.5.1** (Mean and variance). If `x` is a numeric vector one can compute the (empirical) mean of `x` simply by
>
> > ```
> > > mean(x)
> > ```
>
> Likewise, the (empirical) variance can be computed by
>
> > ```
> > > var(x)
> > ```
>
> Using `var` results in $\hat{\sigma}_n^2$ where we divide by $n-1$.
>
> The normalized dataset where we subtract the mean and divide by the variance can be efficiently computed by
>
> > ```
> > > y <- scale(x)
> > ```

We recall that the empirical measure can be regarded as a realization of a random variable. Thus with a different realization we would get a different empirical mean $\hat{\mu}_n$ and a different empirical variance $\tilde{\sigma}_n^2$. To evaluate the performance of these empirical quantities as approximations of the expectation and variance respectively we can study their distributions when regarded as random variables, that is, when regarded as estimators. In particular we can compute the expectation and variance of $\hat{\mu}_n$ and $\tilde{\sigma}_n^2$.

The computation is an exercise in using the properties of the expectation operator and independence of the random variables $X_1, \ldots, X_n$. First

$$\mathbb{E}\hat{\mu}_n = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n X_i\right) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}X_i = \mathbb{E}X,$$

then using independence

$$\mathbb{V}\hat{\mu}_n = \mathbb{V}\left(\frac{1}{n}\sum_{i=1}^n X_i\right) = \frac{1}{n^2}\sum_{i=1}^n \mathbb{V}X_i = \frac{1}{n}\mathbb{V}X,$$

and finally

$$\mathbb{E}\tilde{\sigma}_n^2 = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n X_i^2 - \hat{\mu}_n^2\right) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}X_i^2 - \mathbb{E}\hat{\mu}_n^2$$

$$= \mathbb{E}X^2 - \mathbb{V}\hat{\mu}_n - (\mathbb{E}\hat{\mu}_n)^2 = \mathbb{E}X^2 - (\mathbb{E}X)^2 - \frac{1}{n}\mathbb{V}X = \frac{n-1}{n}\mathbb{V}X.$$

Thus we have the following result.

Result 4.5.2. *Considering the empirical mean $\hat{\mu}_n$ and the empirical variance $\tilde{\sigma}_n^2$ as estimators of the mean and variance respectively we have*

$$\mathbb{E}\hat{\mu}_n = \mathbb{E}X \quad and \quad \mathbb{V}\hat{\mu}_n = \frac{1}{n}\mathbb{V}X \tag{4.27}$$

together with

$$\mathbb{E}\tilde{\sigma}_n^2 = \frac{n-1}{n}\mathbb{V}X. \tag{4.28}$$

The theorem shows that the expected value of $\hat{\mu}_n$ equals the true expectation $\mathbb{E}X$ and that the variance of $\hat{\mu}_n$ decreases as $1/n$. Thus for large n the variance of $\hat{\mu}_n$ becomes negligible and $\hat{\mu}_n$ will always be a very close approximation to $\mathbb{E}X$. How large n should be depends on the size of $\mathbb{V}X$. Regarding the empirical variance its expectation does not equal the true variance $\mathbb{V}X$. The expected value is always smaller than $\mathbb{V}X$. The relative deviation is

$$\frac{\mathbb{V}X - \mathbb{E}\tilde{\sigma}_n^2}{\mathbb{V}X} = \frac{1}{n},$$

which becomes negligible when n becomes large. However, for $n = 5$, say, the empirical variance undershoots the true variance by 20% on average. For this reason the empirical variance is *not* the preferred estimator of the variance. Instead the standard estimator is

$$\hat{\sigma}_n^2 = \frac{n}{n-1}\tilde{\sigma}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{\mu}_n)^2. \tag{4.29}$$

It follows from Result 4.5.2 and linearity of the expectation operator that

$$\mathbb{E}\hat{\sigma}_n^2 = \mathbb{V}(X).$$

The square root $\hat{\sigma}_n = \sqrt{\hat{\sigma}_n^2}$ naturally becomes the corresponding estimator of the standard deviation. Note, however, that the expectation argument doesn't carry over to the standard deviations. In fact, it is possible to show that

$$\mathbb{E}\hat{\sigma}_n < \sqrt{\mathbb{V}(X)}$$

so $\hat{\sigma}_n$ is still expected to undershoot the standard deviation.

The standard deviation of $\hat{\mu}_n$ is $\sqrt{\frac{1}{n}\mathbb{V}X}$, which is usually called the *standard error of the mean*, and we call $\frac{1}{\sqrt{n}}\hat{\sigma}_n$ (or alternatively $\frac{1}{\sqrt{n}}\tilde{\sigma}_n$) the *sample standard error of the mean*.

It is also possible to compute the variance of $\tilde{\sigma}_n^2$ but the derivation is long and tedious so we will skip it. The result is

$$\mathbb{V}\tilde{\sigma}_n^2 = \frac{n-1}{n^3}\left((n-1)\mathbb{E}(X - \mathbb{E}X)^4 - (n-3)(\mathbb{V}X)^2\right), \tag{4.30}$$

which is not a particularly nice formula either. One can observe though that the variance decreases approximately as $1/n$, which shows that also the empirical variance becomes a good approximation of the true variance when n becomes large. But regardless of whether we can compute the variance of the empirical variance, we can compare the variance of $\tilde{\sigma}_n^2$ with the variance of $\hat{\sigma}_n^2$ and find that

$$\mathbb{V}(\hat{\sigma}_n^2) = \left(\frac{n}{n-1}\right)^2 \mathbb{V}(\tilde{\sigma}_n^2).$$

Hence the variance of $\hat{\sigma}_n^2$ is *larger* than the variance of the empirical variance $\tilde{\sigma}_n^2$. This is not necessarily problematic, but it should be noticed that what we gain by correcting the empirical variance so that the expectation becomes right is (partly) lost by the increased variance.

If we consider an n-dimensional random variable $X = (X_1, \ldots, X_n)$ we can also derive a result about the expectation of the empirical covariance.

Using (4.21) yields

$$
\begin{aligned}
\mathbb{E}\tilde{\sigma}_{ij,n} &= \frac{1}{n}\sum_{l=1}^{n}\mathbb{E}X_{il}X_{jl} - \mathbb{E}\widehat{\mu}_{i,n}\widehat{\mu}_{j,n} \\
&= \mathbb{E}X_iX_j - \frac{1}{n^2}\sum_{l=1}^{n}\sum_{m=1}^{n}\mathbb{E}X_{il}X_{jm}
\end{aligned}
$$

Observing then that due to independence of X_{il} and X_{jm} when $m \neq l$

$$\mathbb{E}X_{il}X_{jm} = \mathbb{E}X_i\mathbb{E}X_j.$$

There are $n(n-1)$ such terms in the last sum above. There are n terms equaling $\mathbb{E}X_iX_j$. This gives that

$$
\begin{aligned}
\mathbb{E}\tilde{\sigma}_{ij,n} &= \mathbb{E}X_iX_j - \frac{1}{n}\mathbb{E}X_iX_j - \frac{n-1}{n}\mathbb{E}X_i\mathbb{E}X_j \\
&= \frac{n-1}{n}\left(\mathbb{E}X_iX_j - \mathbb{E}X_i\mathbb{E}X_j\right) = \frac{n-1}{n}\mathbb{V}(X_i, X_j).
\end{aligned}
$$

Thus we have the result.

Result 4.5.3. *Considering the empirical covariance as an estimator of the covariance, its expectation is*

$$\mathbb{E}\tilde{\sigma}_{ij,n} = \frac{n-1}{n}\mathbb{V}(X_i, X_j).$$

As we can see the empirical covariance also generally undershoots the true covariance leading to the alternative estimate

$$\hat{\sigma}_{ij,n} = \frac{1}{n-1}\sum_{l=1}^{n}(x_{il} - \widehat{\mu}_{i,n})(x_{jl} - \widehat{\mu}_{j,n}) \tag{4.31}$$

of the true covariance with $\mathbb{E}\hat{\sigma}_{ij,n} = \mathbb{V}(X_i, X_j)$.

To conclude this section we present a simple yet fundamental result that states in a rather precise way how the empirical mean approximates the true mean when the number of observations grows to infinity.

Consider the random variable $(X - \mu)^2 1(|X - \mu| > \varepsilon)$, which equals $(X - \mu)^2$ if $|X - \mu| > \varepsilon$ and 0 otherwise. Clearly

$$\varepsilon^2 1(|X - \mu| > \varepsilon) \le (X - \mu)^2 1(|X - \mu| > \varepsilon) \le (X - \mu)^2$$

so taking the expectation on both sides yields

$$\varepsilon^2 \mathbb{P}(|X - \mu| > \varepsilon) \le \mathbb{E}(X - \mu)^2 = \mathbb{V}X,$$

which, slightly rewritten yields the *Chebychev inequality*.

Result 4.5.4 (Chebychevs inequality). *If X is a random variable with finite variance and expectation $\mu = \mathbb{E}X$ then for all $\varepsilon > 0$*

$$\mathbb{P}(|X - \mu| > \varepsilon) \le \frac{\mathbb{V}X}{\varepsilon^2}.$$

A most important derivation based on Chebychevs inequality is the Law of Large Numbers. Recall that $\mathbb{E}(\hat{\mu}_n) = \mu$ and $\mathbb{V}(\hat{\mu}_n) = \frac{\sigma^2}{n}$, thus if we apply the Chebychev inequality to $\hat{\mu}_n$ we find that

$$\mathbb{P}(|\hat{\mu}_n - \mu| > \varepsilon) \le \frac{\mathbb{V}(\hat{\mu}_n)}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2}.$$

This gives us the Law of Large Numbers:

Result 4.5.5. *If $X_1, \ldots, X_n$ are n iid real valued random variables with finite variance σ^2 and expectation μ, then with*

$$\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

we have for all $\varepsilon > 0$ that

$$\mathbb{P}(|\hat{\mu}_n - \mu| > \varepsilon) \le \frac{\sigma^2}{n\varepsilon^2} \to 0$$

for $n \to \infty$.

Exercises

Exercise 4.5.1. Compute $\mathbb{V}X$ where X is exponentially distributed with intensity parameter $\lambda > 0$. Recall that the expectation or mean is $\mu = \mathbb{E}X = \lambda^{-1}$ and the MLE of the mean based on iid exponentially distributed random variables $X_1, \ldots, X_n$ is

$$\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Compute $\mathbb{V}\hat{\mu}_n$ and the standard error of $\hat{\mu}_n$.

4.6   Monte Carlo Integration

Expectations have a special and classical role to play in probability theory and statistics. There are quite a few examples of distributions on $\mathbb{R}$ where we can analytically compute the expectation. However, there are also many many situations where we have no chance of computing the expectation analytically. Remember that one of the analytic tools we have at our disposal is the formula from Definition 4.1.1 or more generally from Result 4.1.3

$$\mathbb{E}h(X) = \int_{-\infty}^{\infty} h(x)f(x)\mathrm{d}x.$$

when X is a real valued random variable, whose distribution has density f and $h : \mathbb{R} \to \mathbb{R}$ such that $h(X)$ has finite expectation. Thus the success of an analytic computation of an expectation depends heavily on our ability to compute integrals.

The process can be reversed. Instead of computing the integral analytically we can rely on the empirical mean as an approximation of the expectation and thus of the integral. As discussed in Section 4.5 the more (independent) observations we have the more precise is the approximation, so if we can get our hands on a large number of iid random variables $X_1, \ldots, X_n$, whose distribution is given by the density f, we can approximate the integral by the (random) empirical mean

$$\frac{1}{n} \sum_{i=1}^{n} h(X_i).$$

According to Result 4.5.2 the variance of this random quantity decays like $1/n$ for n tending to ∞, and thus for sufficiently large n the empirical mean is essentially not random anymore. The empirical mean becomes a numerical approximation of the theoretical integral, and with modern computers and simulation techniques it is often easy to generate the large number of random variables needed. This technique is called *Monte Carlo integration* referring to the fact that we use randomness to do numerical integration.

Example 4.6.1. We know from Example 2.6.15 that the density for the Gumbel distribution is

$$f(x) = \exp(-x)\exp(-\exp(-x)) = \exp(-x - \exp(-x)).$$

We should note that the latter formula is more suitable than the former for numerical computations. The mean value for the Gumbel distribution can then be written as

$$\int_{-\infty}^{\infty} x \exp(-x - \exp(-x))\mathrm{d}x.$$

There is no easy way to compute this integral, but it can be computed numerically in R using the `integrate` function. It can also be computed using Monte-Carlo
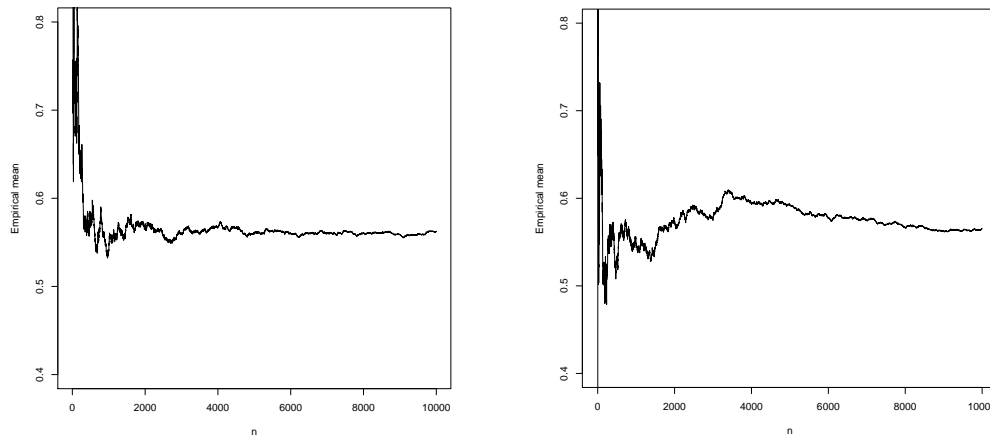
Figure 4.1: The average for n simulated Gumbel variables (left) as a function of n and the average of $-\log y_1, \ldots, -\log y_n$ for n simulated exponential variables (right) as a function of n.

integration by simulating iid Gumbel distributed random variables and computing the average. Figure 4.1 shows how the average looks as a function of n. We use a figure like this to investigate how large an n we should choose.

One attempt to compute the integral above analytically would be to make the substitution $y = \exp(-x)$ so that $\mathrm{d}y = \exp(-x)\mathrm{d}x$ and the integrals equals

$$ -\int_0^\infty \log(y)\exp(-y)\mathrm{d}y. $$

We recognize this integral as the expectation of $-\log Y$ where Y has the exponential distribution. This integral is of course as difficult as the former to compute explicitly, but we can use Monte-Carlo integration where we take $f(x) = \exp(-x)$ the density for the exponential distribution and $h(x) = -\log x$.

Most likely the two Monte-Carlo algorithms turn out to be identical when it comes to implementations. $\diamond$

Elaborating a little on the idea, we may start with an integral $\int_{-\infty}^\infty g(x)\mathrm{d}x$ that we would like to compute. Initially we have no reference to a probability measure, but we may take f to be any density, which we for technical reasons assume to be strictly positive everywhere, that is $f(x) > 0$ for all $x \in \mathbb{R}$. Then

$$ \int_{-\infty}^\infty g(x)\mathrm{d}x = \int_{-\infty}^\infty \frac{g(x)}{f(x)} f(x)\mathrm{d}x, $$

so if we define $h(x) = g(x)/f(x)$, which is well defined as $f(x) > 0$ for all $x \in \mathbb{R}$, we

find that

$$\int_{-\infty}^{\infty} g(x)\mathrm{d}x = \mathbb{E}h(X) = \mathbb{E}\left(\frac{g(X)}{f(X)}\right)$$

where the distribution of X has density f. Simulating $X_1, \ldots, X_n$ as independent random variables, whose distribution has density f, we get the empirical approximation

$$\frac{1}{n}\sum_{i=1}^{n}\frac{g(X_i)}{f(X_i)} \simeq \int_{-\infty}^{\infty} g(x)\mathrm{d}x,$$

which is valid for large n. In this particular setup the Monte Carlo integration is also known as *importance sampling*. This is basically because we are free here to choose f, and a good choice is in general one such that $f(x)$ is large when $g(x)$ is large. The x's where $g(x)$ is large (and g is a reasonably nice function, not oscillating too rapidly) contribute the most to the integral, and by taking $f(x)$ large when $g(x)$ is large means that we put a *large weight* $f(x)$ on the *important points* x where we get the largest contribution to the integral. This is a heuristic, not a precise mathematical result.

Example 4.6.2. Take $g(x) = 1_{[a,\infty)}(x)f_0(x)$ where $f_0(x) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)$ is the density for the normal distribution with mean 0 and variance 1 and $a > 0$. We think of a as quite large, and thus if Y is a random variable with distribution having density f_0 we see that

$$\mathbb{P}(Y \geq a) = \int_a^{\infty} f_0(x)\mathrm{d}x = \int_{-\infty}^{\infty} 1_{[a,\infty)}(x)f_0(x)\mathrm{d}x = \int_{-\infty}^{\infty} g(x)\mathrm{d}x.$$

It is possible to compute the probability $\mathbb{P}(Y \geq a)$ by Monte Carlo integration where we simulate $Y_1, \ldots, Y_n$ being iid with the $N(0,1)$-distribution and $h(x) = 1_{[a,\infty)}(x)$. However, the empirical mean becomes

$$\frac{1}{n}\sum_{i=1}^{n}1(Y_i \geq a)$$

and if a is large, we may very well risk that no or only a few of the Y's are $\geq a$. This is a central problem in computing small probabilities by Monte Carlo integration. Even though the absolute error will be small almost by definition (a small probability is close to zero, so even if we get an empirical mean being 0 with high probability it is in absolute values close to the true probability) the relative error will be very large. Using importance sampling, taking

$$f(x) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(x-a)^2}{2}\right) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2 + a^2 - 2xa}{2}\right),$$

which is the density for the normal distribution with mean a, we find that simulating

iid $X_1, \ldots, X_n$ with the $N(a,1)$-distribution the empirical mean becomes

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\frac{g(X_i)}{f(X_i)} &= \frac{1}{n}\sum_{i=1}^{n}1(X_i \geq a)\frac{f_0(X_i)}{f(X_i)} \\
&= \frac{1}{n}\sum_{i=1}^{n}1(X_i \geq a)\exp\left(2X_i a - \frac{a^2}{2}\right).
\end{aligned}$$

$\diamond$

The previous example, though representing a problem of real interest, does not do justice to the general applicability of Monte Carlo integration and the impact it has had due to the rapid development of computer technology. One-dimensional numerical integration of a known f does not in general pose a real challenge. The real challenge is to do high-dimensional numerical integration. Using the full version of Result 4.1.3 we have that

$$\mathbb{E}h(X) = \int h(x)f(x)\mathrm{d}x$$

where X is a random variable with values in $\mathbb{R}^n$, whose distribution has density $f : \mathbb{R}^n \to [0,\infty)$, and where $h : \mathbb{R}^n \to \mathbb{R}$ such that $h(X)$ has finite expectation. Note that the integral above represents n successive integrals. If n is large, say in the hundreds or even thousands, it becomes very difficult to write a successful numerical integration algorithm without relying on (some form of) Monte Carlo integration.

Monte Carlo integration is used in physics, mathematical finance, structural biology and many other areas to compute quantities such as probabilities or expectations of random variables that arise from an awfully complicated underlying setup. In statistics we have encountered bootstrapping as a simulation based method. When we use bootstrapping for computing an estimate of the standard error the actual computation is a Monte Carlo integration. In fact, most computations of a numerical quantity based on simulations can be regarded as a Monte-Carlo integration. It seems, however, that the most influential application of Monte Carlo integration in statistics is to Bayesian statistical analysis. A particular flavor of Monte Carlo integration is used here, which is called *Markov Chain Monte Carlo* or *MCMC* for short. The objective is to compute the so-called *posterior distribution*, or at least expectations w.r.t. the posterior distribution, but it is often not possible to simulate directly from the posterior distribution. This is a problem that physicists also had for some of their models in statistical physics. It turns out that much like with importance sampling, we can simulate in a different way and rely on our ability to compute certain fractions – and by a miracle both in the problems from physics and in Bayesian statistics we can actually carry this out in practice. The technique of Markov Chain Monte Carlo integration has certainly provided the Bayesian statisticians with a tool that is indispensable for a proper practical data analysis according

to Bayesian principles. Prior to MCMC practical Bayesian data analysis was often obstructed by the difficulties of computing the posterior distribution.

Even with a strictly frequentistic interpretation of probabilities the Bayesian methodology for high-dimensional parameter estimation has turned out to be useful. In Chapter 3 the primary approach to estimation was through the minimization of the minus-log-likelihood function. The density for the posterior distribution is in effect the likelihood function multiplied by a penalization factor, and the resulting minus-log becomes a penalized minus-log-likelihood function.

A Bayesian estimator for the parameter suggests itself. Instead of minimizing the penalized minus-log-likelihood function we can compute the expectation of the posterior as an estimator of the parameter. In principle we get rid of the difficult practical problem of minimizing a high-dimensional function and replace it with a much simpler problem of computing an average of some simulated random variables, but there is a caveat. Problems with local minima of the penalized minus-log-likelihood function can lead to poor simulations and just as for the minimization problem one has to pay careful attention to whether the simulation algorithm in reality got caught in an area of the parameter space that is located around a local minimum.

Exercises

Exercise 4.6.1. Let

$$f_\lambda(x) = \frac{1}{c(\lambda)(1 + \frac{x^2}{2\lambda})^{\lambda + \frac{1}{2}}}$$

denote the density for the t-distribution, cf. Exercise 2.6.4. Random variables from the t-distribution can be simulated using `rt` in R where degrees of freedom (`df`) equals 2λ. Compute using Monte-Carlo integration the mean

$$\int x f_\lambda(x) \mathrm{d}x$$

for $\lambda = 2, 4, 10, 100$. Plot the average computed as a function of n. Try also to compute the mean by Monte-Carlo integration for $\lambda = \frac{1}{2}, 1$ – what happens?

Exercise 4.6.2. Consider the following integral

$$\int_{\mathbb{R}^{100}} \frac{1}{(2\pi)^{50}} e^{-\sum_{i=1}^{100} \frac{x_i^2}{2} - \rho \sum_{i=1}^{99} x_i x_{i+1}} \mathrm{d}x$$

Compute it using Monte Carlo integration with $\rho = 0.1$. Provide an estimate for the variance of the result. Can you do it for $\rho = 0.2$ also? What about $\rho = 0.6$?

Hint: You should recognize that this can be seen as an expectation of a function of $d = 100$ iid $N(0, 1)$-distributed random variables. It may also be useful to plot the running mean as a function of the number of simulations n.

4.7 Asymptotic Theory

Asymptotic theory is an important topic in theoretical statistics and has had a great influence on how practical statisticians actually carry out their data analysis. In asymptotic theory we consider how the distribution of transformations of n random variables behaves when n becomes large. The phenomena that occurs, which makes this idea of studying the large n sample behavior so important, is that while we can rarely tell what the actual distribution of the transformation is, we can often tell very precisely what the limiting distribution is. The limit is, moreover, largely independent of many of the fine details in the distribution of the variables considered. In addition, the limit distributions are often useful approximation even for moderately large n, and this makes asymptotic theory useful in practice.

First we may recall from Result 4.5.5 (the law of large numbers) that

$$\mathbb{P}(|\hat{\mu}_n - \mu| > \varepsilon) \to 0$$

for $n \to \infty$ for all $\varepsilon > 0$. The theorem defines a notion of convergence, which we call *convergence in probability* . We write

$$\hat{\mu}_n \xrightarrow{P} \mu$$

if $\mathbb{P}(|\hat{\mu}_n - \mu| > \varepsilon) \to 0$ for $n \to \infty$ for all $\varepsilon > 0$. The much stronger result that we will discuss in this section also gives us asymptotically the distribution of $\hat{\mu}_n$ and is known as the *central limit theorem* or CLT for short.

Result 4.7.1 (CLT). *If $X_1, \ldots, X_n$ are n iid real valued random variables with finite variance σ^2 and expectation μ then for all $x \in \mathbb{R}$*

$$\mathbb{P}\left(\frac{\sqrt{n}(\hat{\mu}_n - \mu)}{\sigma} \le x\right) \to \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{x^2}{2}\right) \mathrm{d}x$$

for $n \to \infty$. We write

$$\hat{\mu}_n \overset{as}{\sim} N\left(\mu, \frac{1}{n}\sigma^2\right)$$

and say that $\hat{\mu}_n$ asymptotically follows a normal distribution.

First note that what we are considering here is the distribution function of the normalization of the random variable $\hat{\mu}_n$. We say that $\frac{\sqrt{n}(\hat{\mu}_n - \mu)}{\sigma}$ *converges in distribution* to the standard normal distribution.

Example 4.7.2. Consider as in Example 4.1.10 n iid random variables, $X_1, \ldots, X_n$, with values in $E = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$. Then

$$\hat{p}_n(\mathtt{A}) = \frac{1}{n} \sum_{i=1}^{n} 1(X_i = \mathtt{A})$$

is the relative frequency of A-occurrences – the empirical mean of the iid Bernoulli variables $1(X_1 = \text{A}), \ldots, 1(X_n = \text{A})$. Since

$$\mathbb{E}1(X_1 = \text{A}) = \mathbb{P}(X_1 = \text{A}) = p(\text{A}) \quad \text{and} \quad \mathbb{V}1(X_1 = \text{A}) = p(\text{A})(1 - p(\text{A}))$$

we have from Result 4.7.1 that

$$\hat{p}_n(\text{A}) \overset{\text{as}}{\sim} N\left(p(\text{A}), \frac{1}{n}p(\text{A})(1 - p(\text{A}))\right).$$

Likewise, the same result for $\hat{p}_n(\text{C})$, $\hat{p}_n(\text{G})$ and $\hat{p}_n(\text{T})$ holds. ◇

Monte-Carlo integration is one situation where the use of the central limit theorem is almost certainly justified because we want to run so many simulations that the empirical average is very close to the theoretical average. This will typically ensure that n is so large that the distribution of the average is extremely well approximated by the normal distribution. It means that if $X_1, \ldots, X_n$ are iid having density f and we consider the empirical average

$$\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} h(X_i) \simeq \mu = \int h(x)f(x)\mathrm{d}x$$

then $\hat{\mu}_n \overset{\text{as}}{\sim} N\left(\mu, \frac{1}{n}\sigma^2\right)$ where

$$\sigma^2 = \mathbb{V}h(X_1) = \int (h(x) - \mu)^2 f(x)\mathrm{d}x.$$

We should regard the Monte-Carlo integration as a computer experiment for estimation of the unknown parameter μ. As such, the estimator $\hat{\mu}_n$ has an approximating normal distribution, which we can use to give a confidence interval for the parameter, and the standard $(1 - \alpha)$-confidence interval is

$$\left[\hat{\mu}_n - z_\alpha \frac{\sigma}{\sqrt{n}}, \hat{\mu}_n + z_\alpha \frac{\sigma}{\sqrt{n}}\right]$$

where z_α is the $(1 - \alpha/2)$-quantile for the normal distribution. If we cannot compute the mean μ analytically we can probably not compute the variance σ^2 analytically either, hence we need to estimate the standard error using the estimator

$$\hat{\sigma}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n} (X_i - \hat{\mu}_n)^2,$$

and we plug this estimate into the formula for the confidence interval above.

Example 4.7.3. If we return to the Monte-Carlo integration of the mean for the Gumbel distribution as considered in Example 4.6.1 we find that the estimated mean and standard deviation is

$$\hat{\mu}_{10000} = 0.562 \qquad \hat{\sigma}_{10000} = 1.28$$
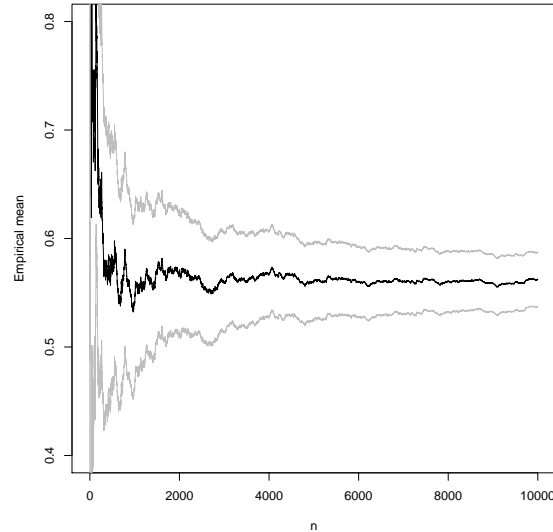
Figure 4.2: The average for n simulated Gumbel variables as a function of n including a 95%-confidence band based on the asymptotic normal distribution of the average for large n.

Here we base the estimates on 10000 simulations. Figure 4.2 shows the same plot of the average as a function of n as Figure 4.1 but this time with 95%-confidence bands. The resulting confidence interval is $[0.537, 0.587]$. $\diamond$

Example 4.7.4. As in the Example above, but this time considering an m-length word w as in Example 4.1.10, the relative frequency of w is

$$\hat{p}_w = \frac{1}{n}N = \frac{1}{n}\sum_{i=1}^{n-m+1} 1(X_i \dots X_{i+m-1} = w).$$

Note that $\mathbb{E}\hat{p}_w = p_w\frac{n-m+1}{n}$ where p_w is the word probability, see Example 4.1.10, and one may choose to divide by $n - m + 1$ instead of n to get expectation p_w. For $m \ll n$ and especially for $n \to \infty$ this does not really make a difference. Since $\hat{p}_w$ is the empirical mean of *dependent* random variables, Result 4.7.1 does not apply as it stands. There are versions of the central limit theorem that can cope with dependent variables, and the result is that indeed, $\hat{p}_w$ also asymptotically has a normal distribution. In this case

$$\hat{p}_w \overset{\text{as}}{\sim} N\left(p_w, \frac{1}{n}p_w(1 - p_w) + 2\sum_{k=1}^{m-1} \rho(k)\right)$$

where $\rho(k)$ is defined in Example 4.4.6. If w is not self-overlapping we know that

$\rho(k) = -p_w^2$ for $k = 1, \ldots, m-1$, hence for non self-overlapping words

$$\hat{p}_w \overset{as}{\sim} N\left(p_w, \frac{1}{n}p_w(1-(2m-1)p_w)\right)$$

since

$$
\begin{aligned}
p_w(1-p_w) + 2\sum_{k=1}^{m-1}\rho(k) &= p_w(1-p_w) - 2(m+1)p_w^2 \\
&= p_w(1-p_w - 2(m-1)p_w) = p_w(1-(2m-1)p_w).
\end{aligned}
$$

Perhaps a little surprisingly, this asymptotic variance, $p_w(1-(2m-1)p_w)$, is actually always positive. It must be, and a formal derivation of the asymptotic normality of $\hat{p}_w$ will actually prove this. It is a consequence of the non self-overlapping assumption that puts some restrictions on the word probability p_w – no matter what the probabilities for the single letters are! Positivity of $p_w(1-(2m-1)p_w)$ is equivalent to

$$p_w < \frac{1}{2m-1},$$

which therefore gives us a direct upper bound on the probability of a non self-overlapping word of length m. ◇

Result 4.7.5 (The Δ-method). *If $Z_n \overset{as}{\sim} N\left(\mu, \frac{1}{n}\sigma^2\right)$ and if $h : \mathbb{R} \to \mathbb{R}$ is differentiable in μ then*

$$h(Z_n) \overset{as}{\sim} N\left(h(\mu), \frac{1}{n}h'(\mu)^2\sigma^2\right).$$

The Δ-method provides first of all a means for computing approximately the variance of a non-linear transformation of a real valued random variable – provided that the transformation is differentiable and that the random variable is asymptotically normally distributed. If Z_n is a real valued random variable with density f_n, if $h : \mathbb{R} \to \mathbb{R}$, if $h(Z)$ has finite variance, and if $\mu_h = \mathbb{E}h(Z_n)$ we know that

$$\mathbb{V}(h(Z_n)) = \int_{-\infty}^{\infty}(h(x)-\mu_h)^2 f_n(x)\mathrm{d}x.$$

If Z_n also has finite variance and $\mathbb{V}Z_n = \sigma^2/n$, this variance is small for large n and Z_n does not fluctuate very much around its mean $\mu = \mathbb{E}Z_n$. Therefore we can make the approximation $\mu_h \simeq h(\mu)$. Moreover, if h is differentiable we can make a first order Taylor expansion of h around μ to get

$$h(x) \simeq h(\mu) + h'(\mu)(x-\mu),$$

from which we find

$$
\begin{aligned}
\mathbb{V}h(Z_n) &= \int_{-\infty}^{\infty}(h(x)-\mu_h)^2 f_n(x)\mathrm{d}x \simeq \int_{-\infty}^{\infty}(h(\mu)+h'(\mu)(x-\mu)-h(\mu))^2 f_n(x)\mathrm{d}x \\
&= h'(\mu)^2\int_{-\infty}^{\infty}(x-\mu)^2 f_n(x)\mathrm{d}x = h'(\mu)^2\mathbb{V}Z_n = \frac{1}{n}h'(\mu)^2\sigma^2.
\end{aligned}
$$

Math Box 4.7.1 (Multivariate CLT). The notion of convergence in distribution can be generalized to several dimensions. If $Z_n \in \mathbb{R}^d$ we say that

$$Z_n \overset{\text{as}}{\sim} N\left(\xi, \frac{1}{n}\Sigma\right)$$

where $\xi \in \mathbb{R}^d$ and Σ is a positive definite $d \times d$ matrix if for all $v \in \mathbb{R}^d$ we have

$$v^t Z_n \overset{\text{as}}{\sim} N\left(v^t\xi, \frac{1}{n}v^t\Sigma v\right).$$

If $X_1, \ldots, X_n$ are iid real valued random variables with finite *fourth* order moment, and if the X-variables have mean μ and variance σ^2 then

$$\begin{pmatrix} \hat{\mu}_n \\ \hat{\sigma}_n \end{pmatrix} \overset{\text{as}}{\sim} N\left(\begin{pmatrix} \mu \\ \sigma \end{pmatrix}, \frac{1}{n}\Sigma\right)$$

so the bivariate random variable consisting of the empirical mean and variance converges in distribution to a bivariate normal distribution. The asymptotic covariance matrix is

$$\Sigma = \sigma^2 \left\{ \begin{matrix} 1 & \frac{\mu_3}{2} \\ \frac{\mu_3}{2} & \frac{\mu_4 - 1}{4} \end{matrix} \right\}$$

where μ_3 and μ_4 are the third and fourth moments of the normalized random variable

$$Y = \frac{X_1 - \mu}{\sigma}.$$

Thus Y has mean 0 and variance 1 and

$$\mu_3 = \mathbb{E}Y^3 \quad \text{and} \quad \mu_4 = \mathbb{E}Y^4.$$

Note the the values of μ_3 and μ_4 do not depend upon μ or σ^2. If the distribution of the X-variables is $N(\mu, \sigma^2)$ it holds that $\mu_3 = 0$ and $\mu_4 = 3$.

Computing the true variance analytically is in most cases impossible, but differentiating h and computing the approximation is most likely doable. But the Δ-method actually provides much more information than just a variance approximation. It also tells that the differentiable transformation preserves the asymptotic normality, so that you actually have a way to approximate the entire distribution of $h(Z_n)$ if Z_n asymptotically has a normal distribution.

Example 4.7.6. Continuing Examples 4.7.2 and 4.7.4 we consider the function $h : (0,1) \to (0,\infty)$ given by

$$h(x) = \frac{x}{1-x},$$

then for any event, B, we see that $h(P(B)) = \xi(B)$ is the odds for that event. We

Math Box 4.7.2 (Multivariate Δ-method)**.** If $h : \mathbb{R}^d \to \mathbb{R}^k$ is differentiable in ξ with $(d \times k$ matrix of) derivatives

$$
Dh(\xi) = \begin{pmatrix} \frac{\partial h_1}{\partial x_1}(\xi) & \cdots & \frac{\partial h_k}{\partial x_1}(\xi) \\ \vdots & \ddots & \vdots \\ \frac{\partial h_1}{\partial x_d}(\xi) & \cdots & \frac{\partial h_k}{\partial x_d}(\xi) \end{pmatrix}
$$

and if $Z_n \overset{\text{as}}{\sim} N\left(\xi, \frac{1}{n}\Sigma\right)$ then

$$
h(Z_n) \overset{\text{as}}{\sim} N\left(f(\xi), \frac{1}{n}Dh(\xi)^t \Sigma Dh(\xi)\right).
$$

The function h needs only to be defined in a neighborhood of ξ for this result to be true.

As an example take $h : \mathbb{R} \times (0, \infty) \to \mathbb{R}$ to be

$$
h(x, y) = \frac{y}{x},
$$

Then

$$
Dh(x, y) = \left\{ \begin{array}{c} \frac{-y}{x^2} \\ \frac{1}{x} \end{array} \right\}.
$$

We see that $h(\mu, \sigma)$ is the coefficient of variation, and using the results in Math Box 4.7.1 and the multivariate Δ-method above we find that

$$
\widehat{\mathrm{CV}}_n = \frac{\hat{\sigma}_n}{\hat{\mu}_n} \overset{\text{as}}{\sim} N\left(\frac{\sigma}{\mu}, \frac{1}{n}\left(\frac{\sigma^4}{\mu^4} - \frac{\sigma^3 \mu_3}{\mu^3} + \frac{\sigma^2(\mu_4 - 1)}{4\mu^2}\right)\right)
$$

$$
= N\left(\mathrm{CV}, \frac{1}{n}\left(\mathrm{CV}^4 - \mathrm{CV}^3 \mu_3 + \mathrm{CV}^2 \frac{\mu_4 - 1}{4}\right)\right)
$$

since

$$
Dh(\mu, \sigma)^t \Sigma Dh(\mu, \sigma) = \sigma^2 \left\{ \begin{array}{cc} \frac{-\sigma}{\mu^2} & \frac{1}{\mu} \end{array} \right\} \left\{ \begin{array}{cc} 1 & \frac{\mu_3}{2} \\ \frac{\mu_3}{2} & \frac{\mu_4 - 1}{4} \end{array} \right\} \left\{ \begin{array}{c} \frac{-\sigma}{\mu^2} \\ \frac{1}{\mu} \end{array} \right\}
$$

$$
= \frac{\sigma^4}{\mu^4} - \frac{\sigma^3 \mu_3}{\mu^3} + \frac{\sigma^2(\mu_4 - 1)}{4\mu^2}
$$

find the derivative to be

$$
h'(x) = \frac{(1 - x) + x}{(1 - x)^2} = \frac{1}{(1 - x)^2},
$$

If we consider the iid Bernoulli random variables $1(X_1 = \mathtt{A}), \ldots, 1(X_n = \mathtt{A})$ with $p(\mathtt{A}) \in (0, 1)$ we find, using Example 4.7.2 and the Δ-method in Result 4.7.5, that the empirical odds for the nucleotide $\mathtt{A}$, $\hat{\xi}(\mathtt{A}) = \hat{p}(\mathtt{A})/(1 - \hat{p}(\mathtt{A}))$, is asymptotically

normal,

$$\hat{\xi}(\mathtt{A}) = \frac{\hat{p}(\mathtt{A})}{1 - \hat{p}(\mathtt{A})} \overset{\text{as}}{\sim} N\left(\frac{p(\mathtt{A})}{1 - p(\mathtt{A})}, \frac{1}{n}\frac{p(\mathtt{A})}{1 - p(\mathtt{A})}\right) = N\left(\xi(\mathtt{A}), \frac{1}{n}\xi(\mathtt{A})\right).$$

If w is a length m word that is not self overlapping and if we consider the (dependent) Bernoulli random variables $1(X_1 \ldots X_m = w), \ldots, 1(X_{n-m+1} \ldots X_m = w)$ we find instead, using Example 4.7.4 and the Δ-method, that the empirical odds for the word w, $\hat{\xi}(w) = \hat{p}_w/(1 - \hat{p}_w)$, is asymptotically normal,

$$\hat{\xi}(w) = \frac{\hat{p}_w}{1 - \hat{p}_w} \overset{\text{as}}{\sim} N\left(\frac{p_w}{1 - p_w}, \frac{1}{n}\frac{p_w(1 - (2m-1)p_w)}{(1 - p_w)^2}\right).$$

$\diamond$

Exercises

Exercise 4.7.1. Consider the Monte-Carlo integration from Exercise 4.6.1 for the computation of the mean in the t-distribution. Continue this exercise by computing estimates for the variance and plot the estimates as a function of n. Try the range of λ's $\frac{1}{2}, 1, 2, 10$. What happens and how can you interpret the result? Choose n and compute 95%-confidence intervals for the value of the mean for different choices of λ.

4.7.1   MLE and Asymptotic Theory

It is the exception, not the rule, that we analytically can find the distribution of an estimator. On the contrary, it is often the case that estimators, like most maximum likelihood estimators, do not even have an explicit analytic expression but are given as solutions to equations or maximizers of a function. An alternative to knowing the actual distribution of an estimator is to know a good and useful approximation.

For an estimator, $\hat{\theta}$, of a one-dimensional real parameter θ, we may be able to compute the expectation, $\xi(\theta) = \mathbb{E}_\theta\hat{\theta}$, and the variance, $\sigma^2(\theta) = \mathbb{V}_\theta\hat{\theta}$. A possible approximation of the distribution, which turns out to be quite well founded, is $N(\xi(\theta), \sigma^2(\theta))$ – the normal distribution with mean $\xi(\theta)$ and variance $\sigma^2(\theta)$. We present this as the following pseudo definition.

Definition 4.7.7. *We say that an estimator $\hat{\theta}$ of a one-dimensional parameter θ approximately follows a normal distribution with parameters $\xi(\theta)$ and $\sigma^2(\theta)$ if*

$$\mathbb{P}\left(\frac{\hat{\theta} - \xi(\theta)}{\sigma(\theta)} \leq x\right) \simeq \Phi(x)$$

for $x \in \mathbb{R}$. We write

$$\hat{\theta} \overset{approx}{\sim} N(\xi(\theta), \sigma^2(\theta))$$

It is a pseudo definition because we do not try to quantify what we mean by $\simeq$ above. The idea of this section is that it is possible to make mathematical sense out of the approximation, but it is a technical business just to formulate the results with rigorous mathematical assumptions let alone to give correct and complete derivations.

Quite a lot of estimators follow approximately a normal distribution. Next to the actual quantification of the quality of the approximation, the real obstacle is to compute the approximating parameters. In reality it is the approximating variance that presents a problem, as most reasonable estimators follow approximately a normal distribution with $\xi(\theta) = \theta$, in which case we say that the estimator is approximately unbiased – the estimator centers approximately around the true value of θ.

Result 4.7.8. *Consider a statistical model $(P_\theta)_{\theta \in \Theta}$ with $\Theta \subseteq \mathbb{R}$ a nice parameter set, and with a minus-log-likelihood function l_X, which is twice differentiable, then in a wide range of cases the MLE approximately follows a normal distribution with $\xi(\theta) = \theta$ and $\sigma^2(\theta) = i(\theta)^{-1}$ where*

$$i(\theta) = \mathbb{E}_\theta \left(\frac{\mathrm{d}^2 l_X}{\mathrm{d}\theta^2}(\theta) \right) \in (0, \infty).$$

That is

$$\hat{\theta} \overset{approx}{\sim} N(\theta, i(\theta)^{-1}).$$

We use the subscript X instead of x when writing the minus-log-likelihood function l_X and its derivatives in θ to emphasize that in this section we will consider them, for fixed θ, as transformations of X and will not be so interested in them as functions of θ for a given observation x.

Remark 4.7.9. To somehow justify the theorem we make the following informal computations. A first order Taylor expansion of the derivative of the minus-log-likelihood function around θ results in

$$\frac{\mathrm{d}l_X}{\mathrm{d}\theta}(\hat{\theta}) = \frac{\mathrm{d}l_X}{\mathrm{d}\theta}(\theta + (\hat{\theta} - \theta)) = \frac{\mathrm{d}l_X}{\mathrm{d}\theta}(\theta) + \frac{\mathrm{d}^2 l_X}{\mathrm{d}\theta^2}(\theta)(\hat{\theta} - \theta) + \text{residual term}.$$

Since $\hat{\theta}$ is a maximizer of the likelihood function (the minimizer of the minus-log-likelihood) the derivative is zero (if $\hat{\theta}$ is an interior point in Θ, i.e. not a point on the boundary). Moreover, ignoring the residual term we then rearrange the equation to yield

$$\hat{\theta} \simeq \theta - \left(\frac{\mathrm{d}^2 l_X}{\mathrm{d}\theta^2}(\theta) \right)^{-1} \frac{\mathrm{d}l_X}{\mathrm{d}\theta}(\theta).$$

One major part of the technical difficulties in a formal mathematical derivation has to do with justifying this approximation. Another major point, which has much more to do with the statistical model considered and the probabilistic theory, is to verify that the following two results hold:

$$\frac{\mathrm{d}^2 l_X}{\mathrm{d}\theta^2}(\theta) \simeq i(\theta) \quad \text{and} \quad \frac{\mathrm{d}l_X}{\mathrm{d}\theta}(\theta) \overset{\text{approx}}{\sim} N(0, i(\theta)).$$

If we freely can interchange integration w.r.t. x and differentiation w.r.t. θ in the following, we find, since $\int_{-\infty}^{\infty} f_\theta(x)\mathrm{d}x = 1$, that by differentiation w.r.t. θ

$$
\begin{aligned}
0 &= \int \frac{\mathrm{d}f_\theta(x)}{\mathrm{d}\theta}\mathrm{d}x \\
&= \int \frac{\mathrm{d}\log f_\theta(x)}{\mathrm{d}\theta} f_\theta(x)\mathrm{d}x \\
&= -\mathbb{E}_\theta\left(\frac{\mathrm{d}l_X(\theta)}{\mathrm{d}\theta}\right).
\end{aligned}
$$

So from the fact that the integral of the density is 1 we obtain the identity that the expectation of the minus-log-likelihood function is 0. A second differentiation yields

$$
\begin{aligned}
0 &= \int \frac{\mathrm{d}^2\log f_\theta(x)}{\mathrm{d}\theta^2} f_\theta(x) + \left(\frac{\mathrm{d}\log f_\theta(x)}{\mathrm{d}\theta}\right)^2 f_\theta(x)\mathrm{d}x \\
&= \mathbb{E}_\theta\left(\frac{\mathrm{d}^2\log f_\theta(x)}{\mathrm{d}\theta^2}\right) + \mathbb{V}_\theta\left(\frac{\mathrm{d}l_X(\theta)}{\mathrm{d}\theta}\right).
\end{aligned}
$$

Thus

$$
i(\theta) = -\mathbb{E}_\theta\left(\frac{\mathrm{d}^2\log f_\theta(x)}{\mathrm{d}\theta^2}\right) = \mathbb{V}_\theta\left(\frac{\mathrm{d}l_X(\theta)}{\mathrm{d}\theta}\right).
$$

These computations show that the parameters $\xi(\theta) = \theta$ and $\sigma^2(\theta) = i(\theta)$ are correct for the approximating normal distribution above. They also show that $i(\theta) \geq 0$, since the number equals a variance. We show below for a specialized model setup with iid observations that one can obtain the two approximations above from the law of large numbers and the central limit theorem respectively.

In a final step we argue that since

$$
\hat{\theta} \simeq \theta - i(\theta)^{-1}Z
$$

with the right hand side a scale-location transformation of the random variable $Z \sim N(0, i(\theta))$ then $\hat{\theta} \overset{\mathrm{approx}}{\sim} N(\theta, i(\theta)^{-1})$.

Since the inverse of $i(\theta)$ appears as a variance in the approximating normal distribution, we can see that $i(\theta)$ quantifies how precisely θ is estimated using $\hat{\theta}$. The larger $i(\theta)$ is the more precise is the estimator. This justifies the following definition.

Definition 4.7.10. *The quantity $i(\theta)$ is called the Fisher information, the expected information or just the information – after the English statistician R. A. Fisher. The quantity*

$$
\frac{\mathrm{d}^2 l_X}{\mathrm{d}\theta^2}(\theta)
$$

is called the observed information.

The computations in Remark 4.7.9 suffered from two deficiencies. One was the ability to formalize a number of *analytic* approximations, which in fact boils down to control of the error in the Taylor expansion used and thus detailed knowledge of the "niceness" of the likelihood function. The other claim was that

$$\frac{\mathrm{d}^2 l_X}{\mathrm{d}\theta^2}(\theta) \simeq i(\theta) \quad \text{and} \quad \frac{\mathrm{d}l_X}{\mathrm{d}\theta}(\theta) \overset{\text{approx}}{\sim} N(0, i(\theta)).$$

This claim is not completely innocent, and does not always hold. But it will hold in situations where we have sufficient replications – either in an iid setup or a regression setup. To show why, lets consider the situation where $X = (X_1, \ldots, X_n)$ and $X_1, \ldots, X_n$ are iid such that the minus-log-likelihood function is

$$l_X(\theta) = -\sum_{i=1}^{n} \log f_\theta(X_i)$$

then $i(\theta) = n i_0(\theta)$ where

$$i_0(\theta) = -\mathbb{E}_\theta \left(\frac{\mathrm{d}^2 \log f_\theta(X_1)}{\mathrm{d}\theta^2}(\theta) \right) = \mathbb{V}_\theta \left(\frac{\mathrm{d} \log f_\theta(X_1)}{\mathrm{d}\theta}(\theta) \right).$$

The second derivative of the minus-log-likelihood is a sum of iid random variables $-\frac{\mathrm{d}^2 \log f_\theta(X_1)}{\mathrm{d}\theta^2}, \ldots, -\frac{\mathrm{d}^2 \log f_\theta(X_n)}{\mathrm{d}\theta^2}$, and the law of large numbers, Result 4.5.5, gives that

$$\frac{1}{n}\frac{\mathrm{d}^2 l_X}{\mathrm{d}\theta^2}(\theta) = \frac{1}{n}\sum_{i=1}^{n} -\frac{\mathrm{d}^2 \log f_\theta(X_i)}{\mathrm{d}\theta^2} \overset{P}{\to} -\mathbb{E}_\theta \left(\frac{\mathrm{d}^2 \log f_\theta(X_1)}{\mathrm{d}\theta^2}(\theta) \right) = i_0(\theta).$$

Likewise the first derivative of the minus-log-likelihood is a sum of iid random variables $-\frac{\mathrm{d} \log f_\theta(X_1)}{\mathrm{d}\theta}, \ldots, -\frac{\mathrm{d} \log f_\theta(X_n)}{\mathrm{d}\theta}$, whose mean is 0 and whose variance is $i_0(\theta)$, and the central limit theorem, Result 4.7.1, gives that

$$\frac{1}{n}\frac{\mathrm{d}l_X}{\mathrm{d}\theta}(\theta) = \frac{1}{n}\sum_{i=1}^{n} -\frac{\mathrm{d} \log f_\theta(X_i)}{\mathrm{d}\theta} \overset{as}{\sim} N\left(0, \frac{1}{n}i_0(\theta) \right).$$

Result 4.7.11. *In the setup above, if $i_0(\theta) < \infty$ it holds that*

$$\frac{1}{n}\frac{\mathrm{d}^2 l_X}{\mathrm{d}\theta^2}(\theta) \overset{P}{\to} i_0(\theta)$$

for $n \to \infty$ and

$$\frac{1}{n}\frac{\mathrm{d}l_X}{\mathrm{d}\theta}(\theta) \overset{as}{\sim} N\left(0, \frac{1}{n}i_0(\theta) \right).$$

Remark 4.7.12. If $\hat{\theta}_n$ denotes the MLE in the previous theorem the conclusion is that

$$\hat{\theta}_n \overset{as}{\sim} N\left(\theta, \frac{1}{n i_0(\theta)} \right).$$

> **Math Box 4.7.3** (Multivariate information). If $\Theta \subseteq \mathbb{R}^d$ and if the minus-log-likelihood function as a function of d variables is twice differentiable, the second derivative is the matrix we denote $\mathrm{D}^2 l_X(\theta)$, cf. Math Box 3.3.1.
>
> The Fisher information is
> $$I(\theta) = \mathbb{E}_\theta(\mathrm{D}^2 l_X(\theta)).$$
>
> Like in the one-dimensional setup a similar interchange of integration and differentiation shows that the Fisher information is also the covariance matrix of $\nabla l_X(\theta)$. If $\hat\theta$ denotes the MLE the result then reads that
> $$\hat\theta \overset{\mathrm{approx}}{\sim} N\left(0, I(\theta)^{-1}\right)$$
>
> where $I(\theta)^{-1}$ is the matrix-inverse of $I(\theta)$.
>
> An important consequence is that if $v \in \mathbb{R}^d$ then $v^t\hat\theta$ is a one-dimensional parameter transformation (take $v = (1, 0, \ldots, 0)$ to select the first coordinate of $\hat\theta$, say) and
> $$v^t\hat\theta \overset{\mathrm{approx}}{\sim} N\left(0, v^t I(\theta)^{-1} v\right).$$
>
> The i'th coordinate in $\hat\theta$ thus follows approximately a normal distribution with variance $(I(\theta)^{-1})_{ii}$. This number differs from $(I(\theta)_{ii})^{-1}$. The latter is in fact the asymptotic variance parameter if we estimate θ_i using MLE *while keeping the other parameters fixed*, since the Fisher information in that case is $i(\theta_i) = I(\theta)_{ii}$. The point is that we need the entire Fisher information matrix $I(\theta)$ to compute the asymptotic variance parameter for each coordinate $\hat\theta_i$.

To use this distribution in practice we can use the plug-in estimate $i_0(\hat\vartheta_n)$ as a substitute for the unknown $i_0(\theta)$. This requires that we know a formula for $i_0(\theta)$. As an alternative one can take the observed information evaluated in the MLE

$$\frac{1}{n}\frac{\mathrm{d}^2 l_X}{\mathrm{d}\theta^2}(\hat\vartheta_n)$$

as an approximation to $i_0(\theta)$.

Clearly considering a one-dimensional parameter setup only is not sufficient for most practical applications. The multivariate version of the Fisher information is a matrix and the relevant approximations are described in Math Box 4.7.3.

4.8   Entropy

We consider in this section mostly probability measures or random variables on a discrete sample space E. The definitions and results presented can all be formulated for continuous distributions given in terms of a density instead of in terms of the point probabilities as used here.

Definition 4.8.1 (Entropy)**.** *For a probability measure P on E with point probabilities $(p(x))_{x \in E}$ we define the entropy of P as*

$$H(P) = -\sum_{x \in E} p(x) \log p(x).$$

If X is a random variable with distribution P the entropy of X is defined as the entropy of P, i.e.

$$H(X) = H(P).$$

We use the convention that $0 \log 0 = 0$ in the definition. One may also note from the definition that $\log p(x) \leq 0$, hence $H(P) \geq 0$, but the sum can be divergent if E is infinite in which case $H(P) = \infty$. We may note that $-\log p(X)$ is a positive random variable and that

$$H(X) = \mathbb{E}(-\log p(X)).$$

The interpretation of $H(X)$ is as a measure of the uncertainty about the outcome of the random variable X. This is best illustrated by a couple of examples.
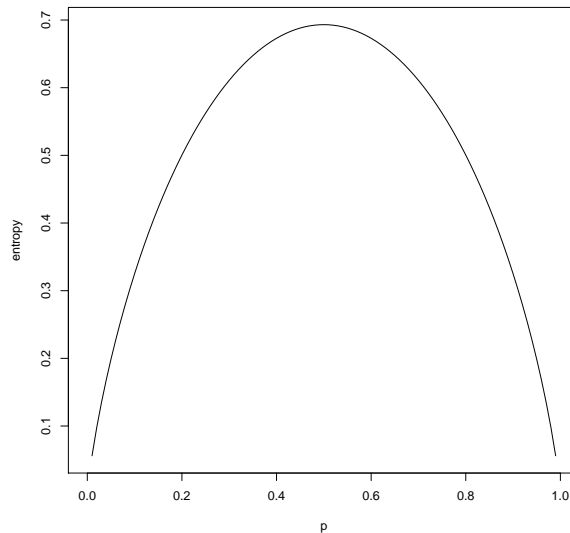
Figure 4.3: The entropy of a Bernoulli variable as a function of the success probability p.

Example 4.8.2 (Bernoulli variables)**.** If $E = \{0, 1\}$ and X is a Bernoulli variable with $\mathbb{P}(X = 1) = p$ then

$$H(X) = -p \log p - (1 - p) \log(1 - p).$$

As a function of the probability p, $H(X)$ takes the value 0 for $p = 0$ or $p = 1$. If we differentiate w.r.t. p for $p \in (0, 1)$ we find that

$$\frac{\mathrm{d}H(X)}{\mathrm{d}p} = -\log p - 1 + \log(1 - p) + 1 = \log(1 - p) - \log p = \log \frac{1 - p}{p}.$$

The derivative is > 0 for $p \in (0, 1/2)$, $= 0$ for $p = 1/2$, and < 0 for $p \in (1/2, 1)$. Thus $H(X)$ is, as a function of p, monotonely increasing in the interval from 0 to $1/2$ where it reaches its maximum. From $1/2$ to 1 it decreases monotonely again. This fits very well with the interpretation of $H(X)$ as a measure of uncertainty. If p is close to 0, X will quite certainly take the value 0, and likewise if p is close to 1, X will quite certainly take the value 1. If $p = 1/2$ we have the greatest trouble to tell what value X will take, as the two values are equally probable. $\diamond$

If E is a *finite* sample space and X takes values in E it holds that $H(X) = 0$ if and only if $\mathbb{P}(X = x) = 1$ for some $x \in E$. Moreover, if $|E| = n$ is the number of elements in E then $H(X)$ is maximal if and only if X has the uniform distribution, i.e. $P(X = x) = 1/n$ for all $x \in E$. These two cases represent the *extreme* cases with minimal and maximal uncertainty, respectively.
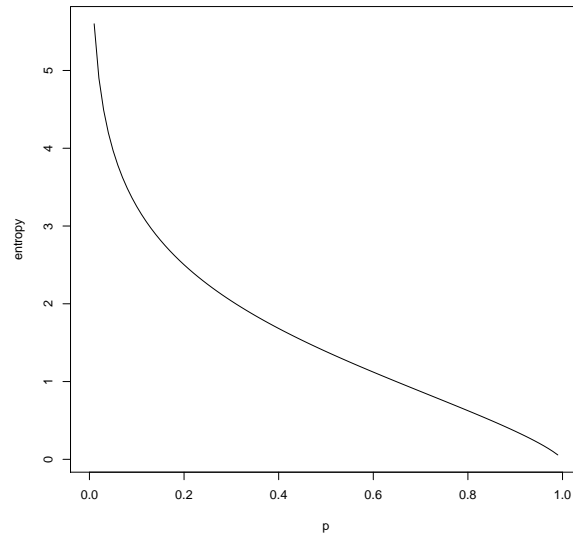
Figure 4.4: The entropy of a geometrically distributed random variable as a function of the success probability p.

Example 4.8.3 (Geometric distribution)**.** If $E = \mathbb{N}_0$ and X follows a geometric

distribution with success probability p, then

$$
\begin{aligned}
H(X) &= -\sum_{n=0}^{\infty} p(1-p)^n \log(p(1-p)^n) \\
&= -\log p \sum_{n=0}^{\infty} p(1-p)^n - \log(1-p) \sum_{n=0}^{\infty} np(1-p)^n \\
&= -\log p - \frac{(1-p)}{p} \log(1-p) \\
&= -\frac{p \log p + (1-p) \log(1-p)}{p}.
\end{aligned}
$$

For $p \in (0,1)$ this function decreases monotonely, which concurs with the interpretation of the entropy. The closer p is to 1, the more certain we are that X will take a small value close to or equal to 0. For small values of p the distribution of X becomes much more spread out, and we are more uncertain what the value of X will be. ⋄

Definition 4.8.4 (Conditional entropy). *If X and Y are two random variables taking values in E_1 and E_2 respectively we define*

$$
H(X \,|\, Y = y) = -\sum_{x \in E_1} p(x|y) \log p(x|y)
$$

with $(p(x|y))_{x \in E_1}$ being the point probabilities for the conditional distribution of X given $Y = y$, and we define the conditional entropy of X given $Y = y$ to be

$$
H(X \,|\, Y) = \sum_{y \in E_2} p(y) H(X|Y = y)
$$

where $(p(y))_{y \in E_2}$ are the point probabilities for the marginal distribution of Y.

The conditional entropy tells us about the average uncertainty about X given that we know the value of Y. The gain in information about X, that is, the loss in uncertainty as measured by entropy, that we get by observing Y tells us something about how strong the dependency is between the variables X and Y. This leads to the following definition.

Definition 4.8.5 (Mutual information). *Let X and Y be two random variables taking values in E_1 and E_2 respectively. The mutual information is defined as*

$$
I(X, Y) = H(X) - H(X|Y).
$$

A value of $I(X, Y)$ close to 0 tells us that the variables are close to being independent – not much knowledge about X is gained by observing Y. A value of $I(X, Y)$ close to $H(X)$ tells that the variables are strongly dependent. We therefore regard $I(X, Y)$ as a quantification of the dependence between X and Y.

Another way to view mutual information is through the relative entropy measure between two probability measures.

Definition 4.8.6 (Relative Entropy). *If P and Q are two probability measures on E with point probabilities $(p(x))_{x \in E}$ and $(q(x))_{x \in E}$ respectively then provided that $q(x) = 0$ implies $p(x) = 0$, we define the relative entropy of P w.r.t. Q as*

$$D(P \,|\, Q) = \sum_{x \in E} p(x) \log \frac{p(x)}{q(x)}.$$

If $q(x) = 0$ for some $x \in E$ with $p(x) > 0$ then $D(P \,|\, Q) = \infty$. If X and Y are two random variables with distribution P and Q respectively we define

$$D(X \,|\, Y) = D(P \,|\, Q)$$

as the relative entropy from X to Y.

From the definition it follows that

$$
\begin{aligned}
D(X \,|\, Y) &= \sum_{x \in E} p(x) \log \frac{p(x)}{q(x)} \\
&= \sum_{x \in E} p(x) \log p(x) - \sum_{x \in E} p(x) \log q(x) \\
&= -H(X) - \sum_{x \in E} p(x) \log q(x),
\end{aligned}
$$

or alternatively that

$$H(X) + D(X \,|\, Y) = - \sum_{x \in E} p(x) \log q(x). \tag{4.32}$$

Note that the relative entropy is non-symmetric in X and Y.

The concepts of mutual information and relative entropy are related. Let $r(x, y)$ denote the point probabilities for the probability measure R, then the point probabilities for P and Q are

$$p(x) = \sum_y r(x, y) \quad \text{and} \quad q(y) = \sum_x r(x, y)$$

respectively. Writing out how $D(R \,|\, P \otimes Q)$ is defined we find that

$$
\begin{aligned}
D(R \,|\, P \otimes Q) &= \sum_{x,y} r(x, y) \log \frac{r(x, y)}{p(x) q(y)} \\
&= -\sum_{x,y} r(x, y) \log p(x) + \sum_{x,y} r(x, y) \log \frac{r(x, y)}{q(y)} \\
&= -\sum_x \left(\sum_y r(x, y) \right) \log p(x) + \sum_{x,y} r(x|y) q(y) \log r(x|y) \\
&= -\sum_x p(x) \log p(x) + \sum_y q(y) \sum_x r(x|y) \log r(x|y) \\
&= H(X) - H(X|Y) = I(X, Y).
\end{aligned}
$$

Thus we have the result:

Result 4.8.7. *If X and Y are two random variables with joint distribution R, if X has (marginal) distribution P and Y has (marginal) distribution Q then*

$$I(X,Y) = D(R \,|\, P \otimes Q).$$

We may observe that

$$D(P|Q) = \mathbb{E}\left(\log \frac{p(X)}{q(X)}\right)$$

is the expectation of the log-likelihood ratio $\log \frac{p(X)}{q(X)}$. It can be shown that $D(P|Q) \geq 0$ with equality if and only if $P = Q$.

If Q_θ for $\theta \in \Theta$ denotes a parametrized statistical model and if ε_n is the empirical probability measured based on $x_1, \ldots, x_n$ we find that

$$D(\varepsilon_n|Q_\theta) = -H(\varepsilon_n) - \frac{1}{n}\sum_{i=1}^{n} \log q_\theta(x_i) = -H(\varepsilon_n) + l_x(\theta)$$

where $l_x(\theta)$ is the minus-log-likelihood function. Thus minimizing the minus-log-likelihood function is equivalent to minimizing the relative entropy of the empirical measure ε_n w.r.t. Q_θ over θ. The relative entropy $D(P|Q_\theta)$ is often regarded as a (asymmetric) distance measure from P to Q_θ. The maximum likelihood estimator can thus be seen as the estimator that takes the probability measure $Q_{\hat{\theta}}$ in the model that is closest to the empirical measure ε_n as measured by the relative entropy.

A

R

The program R is "GNU S", a freely available environment for statistical computing and graphics that runs on a variety of platforms including Mac OS X, Linux and Windows. It is an implementation of the S language developed by John Chambers and colleagues at the Bell Laboratories.

R consists of the base program with a number of standard packages and a large and ever growing set of additional packages. The base program offers a Command Line Interface (CLI), where you can interactively type in commands (R expressions). One will (or should) quickly learn that the proper use of R is as a high level programming language for writing R-scripts, R-functions and R-programs. One may eventually want to extend the functionality of R by writing an entire R package and/or implement various time-consuming algorithms in a lower level language like C with an R-interface. Indeed, many of the base functions are implemented in C or Fortran. Such advanced use of R is far beyond the scope of this appendix.

This appendix deals with a few fundamental questions that inevitably arise early on when one wants to use R. Questions like how to obtain R, how to run R, what is R all about, how to handle graphics, how to load packages and data, how to run scripts, and similar problems. We also give directions for locating more information on using and running R. The appendix can not stand alone, and you will for instance need the manual *An Introduction to R* – see Section A.2.

A.1 Obtaining and running R

The R program and all additional packages are available for download at the Comprehensive R Archive Network (CRAN) at `http://cran.r-project.org/`. The Danish mirror is `http://mirrors.dotsrc.org/cran/`. You can download binaries for Linux, Windows, and Mac OS X, or you can download the source code and compile

it yourself if you want. R is available as Free Software under the terms of the Free Software Foundation's GNU General Public License.

You can also download the packages from CRAN, but once you have installed R, it is quite easy to download and install packages from within R.

When R is properly installed[1] on your computer, you can run it. How you do that and the appearance of R differ a little from Linux to Windows. On a Linux machine, you simply type R in a terminal, the program starts and you are now running R. Whatever you type in next is interpreted by R. Graphics will appear in separate windows. If you start R in Windows (by locating it from e.g. the Start Menu), R runs in a new window called the RGui, with a window inside called the R console containing a command prompt. Don't be fooled – the fact that it says RGui doesn't mean that the Windows version provides a graphical user interface for using R, just that a few things like running R-scripts, loading and installing packages, and saving graphics can be done from the menu. It is recommended that you learn how to do that from the command line anyway.

R runs in a *working directory*. All interaction with the file system like reading or writing data files, running scripts or saving graphics will take place relative to the working directory unless you specify a complete alternative path instead of just a file-name. You get the current working directory by the command `getwd()`. You change the working directory to be `path` by `setwd("path")`. In the RGui on Windows, this can be done from the File menu as well.

You quit R by typing `quit()` or simply `q()`. On a Windows machine you can also close the program in the File menu (Exit). When you quit, you are always asked if you want to save the *workspace*. Doing so, all objects and the command history are stored in the working directory. When starting R it automatically loads the workspace most recently stored in the default working directory.

A.2   Manuals, FAQs and online help

In general, we refer to the manuals that come with R for detailed information on R and how to use R. The manuals in PDF-format are located in the subdirectory doc/manual. For the Windows version you can also find them through the Help menu, and you can always find the most recent version on the R homepage: `http://www.r-project.org/`. The most important manual for ordinary use of R is *An Introduction to R*.

You can also access the manuals in HTML-format by issuing the `help.start()` command in R. The HTML-page that opens contains links to the manuals, links to help pages grouped according to the package they belong to, as well as a search

---

[1]How one installs the program can be machine and platform dependent. Installation of the Windows binary should be completely straight forward.

engine for searching the help pages. Links to some additional material like FAQs are also given. When running R on Windows you can find the HTML-help page in the Help menu together with direct links to FAQs, the search engine, etc.

You can access the help pages from within R by the command `help`. For instance, `help(plot)` will give you the help page for the `plot` function. A shortcut is `?plot`. You can also search for help pages containing the word "plot" by `help.search("plot")`. Note the quotation marks when using `help.search`. For a few functions like the binary operator plus (`+`), you will need quotation marks when you call the `help` function, i.e. `help("+")`.

A.3   The R language, functions and scripts

You may find this section a little difficult, if you are not that familiar with computer programming and programming languages. However, this may explain a little about why things are the way they are.

A.3.1   Functions, expression evaluation, and objects

Formally the R language belongs to the family of functional programming languages. Every command issued is a function call, or *expression evaluation*, with or without arguments, and the result of the evaluation is returned. The functions work on *objects*. The objects store the data that the functions work on, and there are a number of data structures for doing this. The objects provide *interfaces* between the computer memory, which can not be accessed directly, and the functions. It depends on the data structure how we access the data stored in the object. For instance, an object `x` of *type* integer is a *vector* containing integers, i.e. the data structure is an integer vector. The elements in the vector can be accessed by subscripting, e.g. `x[10]` is the 10'th integer in the vector `x`. Another object `y` of type *list* is a list with each element in the list being some data structure. The first element in the list is accessed by `y[[1]]`, which could be an integer vector – or even another list. You can list all existing objects by `objects()` or alternatively `ls()`.

A thing to remember about the language is that any command typed in is actually a function evaluation. The function takes some (maybe zero) arguments, and when evaluated it returns something. Syntactically this means that whenever you want to execute a command with no arguments, for instance the `help.start()` as discussed above, you will still need to provide the parentheses `()`. If you just type `help.start` the R code for the function is printed. Some functions are evaluated because they have *side effects* like producing a graph or starting up the HTML-help pages, in which case the function returns `NULL`.

R and the standard packages that ships with R provide a large number of commands – or functions. Many more than a low level language like C. This means that you can

get R to do rather complicated things by evaluating a few functions, and it is often much more efficient to use already existing functions than to write you own R code based on more elementary R functions. The downside is that it can be difficult for a new user to figure out what functions are available. That is, there is a quite large body of knowledge that one needs to obtain. The present set of notes introduces many functions in an appropriate context.

R is an object oriented programming language. This implies that certain functions are *generic* meaning that the behaviour of the function depends on the object you give as argument. A generic function calls the function associated with the data structure of the object you give as argument. A good example is the function `plot`. A large number of objects have their own way to be plotted by default, and the generic function `plot` simply calls the correct plotting function for the object. If `x` is a numeric vector then `plot(x)` plots the values in `x` against the index. If `x` is the fitted model returned by the `lm` function (linear model), then `plot(x)` will plot the residuals.

One consequence of the object oriented line of thought is that when you fit a model to data the result is an object that (hopefully) stores all relevant information. You don't want to print all the information on screen. Instead, you can subsequently "interrogate" the fitted model object by a number of generic functions. The interrogation can in principle proceed the same way no matter what kind of model we fit, as the resulting fitted model object is able to provide the information we ask for by the generic function. You can then extract and format the results in a way that suits you.

A.3.2   Writing functions and scripts

In principle, we can use the functions provided by R and the packages to perform the computations we want or to produce the graphics we want by typing in function calls one after the other. If we want, we can even define an entirely new function and then use it – doing everything from the command line interface. Defining a function, `sq`, that squares its argument can be done as follows

```
> sq <-  function(x) x^2
```

One will, however, very quickly find it to be tedious, boring and inconvenient to type in one function call after the other and in particular to define functions using the command line editor – even though the command line editor keeps a history of your commands.

To avoid working with the command line interface you can use R together with any text editor for writing a file containing R function calls, which can then be loaded into R for sequential evaluation. You use `source("foo.r")` to load the file `foo.r`.

Note that you may need to include either the full path or the relative path (relative to the working directory, cf. Section A.1) if `foo.r` is not in the working directory.

The usage of `source` ranges from writing simple scripts that basically collect a number of function calls over implementations of new functions to entire R programs that perform a number of different tasks. It is good practice to get used to working with R this way, i.e. to write R-scripts – then you can always experiment by copy-pasting to the command line editor, if you are not sure that the whole script will run, or that you only need to run some parts of the script.

You can in principle use any text editor. There is a quite extensive environment called ESS (Emacs Speaks Statistics) for the family of Emacs editors. There is more information on the homepage `http://www.sciviews.org/_rgui/projects/Editors.html`, which gives links to a number of editors that support R script editing with features such a syntax highlighting and indentation. The RGui for Windows provides in the File menu its own simple editor for editing scripts, and you can also execute the `source` command from the File menu.

A.4 Graphics

R provides a versatile and customisable environment for producing graphics. You can produce graphics on the screen, print it, and/or save it in a number of formats. You can do this interactively or from within a script.

In R you *define* the graphics you want, and a *device driver* is used to actually produce the graphics of the appropriate format on the appropriate device. Often you want the graphics to be displayed on the screen in a window in which case the screen device (the default device) is needed. A new screen device can be opened by the `dev.new()`, command. The functions can take a number of parameters specifying the size, position, etc. of the window. There are other device drivers like the `postscript` and `pdf` device driver for producing postscript or pdf files, respectively.

If you just use a high-level plotting command like `plot`, you will not have to worry about devices in the first place. Just issue the `plot` command appropriately and you will see the result plotted in a window. However, the moment that you want plot in multiple windows, print out the graphics, or save the graphics, you will need to handle devices. Each time you issue the `dev.new()` command, you will open a new device, and the most recently opened will be the active device to which all subsequent graphics commands are directed. You can get a list with names and numbers of the open devices by `dev.list()`, you can get the current active device by `dev.cur()`, and you can set the active device to be device number `n` by `dev.set(n)`. Finally, you can close down all devices by `graphics.off()`.

Having generated a nice piece of graphics on the screen you want to save or print it. This is done by copying the content of the active device to a new device, e.g. a

postscript device. The command `dev.copy(device = postscript, file = "my-graph.ps")` will copy the content of the currently active device to the file `my-graph.ps` in postscript format. You can also produce e.g. pdf, jpeg, bmp, and bitmap formats by copying to the device with the corresponding name. Use `help(Devices)` for more information. For printing, use `dev.print()` to print the currently active device.

Saving graphics can, like executing scripts and changing the working directory, be done from the File menu when running the RGui on Windows. The R Graphics window, whose content you want to save, needs to be the active window. Then simply choose Save in the File menu and choose the appropriate format.

When you use a typical high-level plotting command like `plot`, you will often provide labels for the axis and a title for the plot. Labels are given by specifying the additional parameters `xlab` and `ylab` to the `plot` command and the title by specifying the `main` parameter or subsequently by the `title` command.

Every time you call a high-level plotting function the content of the current device will be erased. If you want to plot several graphs in the same window, you can do this by calling `points` or `lines` instead of `plot`. Some high level plotting functions can also take a parameter, `add`, which by default equals `FALSE`. If you set `add=TRUE` the graphics will be added to the existing plot. When producing multiple graphs in the same window, you may often need to explicitly control the range of the x- and y-axis. You do this by setting the `xlim` and `ylim` parameters for the `plot` function or another high-level plotting function.

The basic functions for producing graphics in R have by now been extended considerably by packages providing a range of more advanced plotting tools. Two packages should be mentioned. The `lattice` package and the `ggplot2` package both provide comparable functionality, and are in particular useful for producing multilayered or stratified plots.

A.5   Packages

Packages form a very important part of R. A large number of researchers in statistics choose to implement their work as an R package. At the time of writing the number of packages available from the CRAN archieve is of the order of thousands. The biggest problem is actually to figure out which package implement the methods that one wants to apply. If you don't know, the CRAN archive provides the possibility of browsing the packages by topic.

If we want to install and use the `ggplot2` package, we proceed as follows:

```
> install.packages("ggplot2")
```

installs the package and installs in addition all packages that `ggplot2` depends upon.

```
> t <- seq(0,10,length=101)
> #Compute sinus of the time points in t
> x <- sin(t)
> #Using plot with type="l" produces lines between the points
> plot(t,x,type="l",main="Sinus")
> #Add normally distributed "noise" to the x values
> y <- x + rnorm(101,0,0.2)
> #Plot the noisy values (no annotation) and the mean
> windows()
> plot(t,y,xlim=c(0,10),ylim=c(-2,2),ann=F,pch=20)
> points(t,x,xlim=c(0,10),ylim=c(-2,2),type="l",col="red")
> #Add a title to the plot
> title("Normally distributed variables with oscillating mean")
> #Add a legend to the plot
> legend(0,2,
+ legend=c("Observations","Mean"),col=c("black","red"),
+ lwd=1,lty=c(0,1),pch=c(20,NA))
```

Table A.1: This example illustrates some uses of R for making graphics. The symbol
produces comments in R.

```
> library("ggplot2")
```

loads the package. Note that you need an Internet connection to install the packages.
One can also load a package by the command require. Using require returns a
logical, which is TRUE if the package is available. This is useful in e.g. scripts for
checking that needed packages have actually been loaded.

A.5.1   Bioconductor

There is an entire software project called Bioconductor based primarily on R, which
provides a number of packages for the analysis of genomic data – in particular for
treating data from microarray chips. Information can be found on

http://www.bioconductor.org/

With

```
> source("http://www.bioconductor.org/biocLite.R")
> biocLite()
```

you install a fundamental subset of the Bioconductor libraries. See also

http://www.bioconductor.org/install

To install a specific package from the Bioconductor repository use biocLite("package
name").

A.6   Literature

How you are actually going to get R to do anything interesting is a longer story. The present lecture notes contains information embedded in the text via R boxes that describe functions that are useful in the context they are presented. These boxes can not entirely stand alone, but must be regarded as directions for further study. An indispensable reference is the manual *An Introduction to R* as mentioned above and the online help pages, whether you prefer the HTML-interface or the `help` function.

The homepage `http://www.r-project.org/` contains a list of, at the time of writing, 94 books related to R. This author is particularly familiar with three of the books. An introduction to statistics in R is given in

> *Peter Dalgaard. Introductory Statistics with R.*
> *Springer, 2002. ISBN 0-387-95475-9,*

which treats statistics more thoroughly than the manual. This book combined with the manual *An Introduction to R* provides a great starting point for using R to do statistics. When it comes to using R and S-Plus for more advanced statistical tasks the bible is

> *William N. Venables and Brian D. Ripley. Modern Applied Statistics with S.*
> *Fourth Edition. Springer, 2002. ISBN 0-387-95457-0.*

A more in-depth book on the fundamentals of the S language is

> *William N. Venables and Brian D. Ripley. S Programming.*
> *Springer, 2000. ISBN 0-387-98966-8..*

There is also a book on using R and Bioconductor for Bioinformatics:

> *Gentleman, R.; Carey, V.; Huber, W.; Irizarry, R.; Dudoit, S. (Eds.)*
> *Bioinformatics and Computational Biology Solutions Using R and Bioconductor.*
> *Springer, 2005. ISBN: 0-387-25146-4.*

A.7   Other resources

The R user community is growing at an increasing rate. The language R has for some time been far more than an academic language for statistical experiments. Today, R is just as much a workhorse in practical data analysis and statistics – in business and in science. The expanding user community is also what drives R forward and users contribute with packages at many different levels and there is a growing number of R-related blogs and web-sites. When you want to find an answer to your question, Google is often your friend. In many cases questions have been asked on one of the R emailing lists or treated somewhere else, and you will find it if you search. Being new to R it may be difficult to know what to search for. Two recommended places to look for information is the R wiki

`http://rwiki.sciviews.org/doku.php`

and the list of contributed documents

`http://cran.r-project.org/other-docs.html`

The latter is a mixed collection of documents on specialized R topics and some beginners guides that may be more friendly than the manual.

B

Mathematics

The mathematical prerequisites for reading this introduction to probability theory and statistics is an elementary understanding of set theory and a few concepts from calculus such as integration and differentiation. You will also need to understand a few things about limits and infinite sums. This appendix discusses briefly the most important mathematical concepts and results needed.

B.1   Sets

A set E is (informally) a collection of elements. If an element x is contained in or belongs to a the set E we write $x \in E$. If A is a collection of elements all belonging to E we say that A is a subset of E and write $A \subseteq E$. Thus A is in itself a set, which is included in the larger set E. The complement of A, denoted A^c, *within* E is the set of elements in E that *do not* belong to A. We write

$$A^c = \{x \in E \mid x \notin A\}.$$

If $A, B \subseteq E$ are two subsets of E we define the *union*

$$A \cup B = \{x \in E \mid x \in A \text{ or } x \in B\}$$

and the *intersection*

$$A \cap B = \{x \in E \mid x \in A \text{ and } x \in B\}.$$

We also define

$$A \backslash B = A \cap B^c,$$

which is the set of elements in A that *do not* belong to B.

The integers $\mathbb{Z}$, the non-negative integers $\mathbb{N}_0$, the positive integers $\mathbb{N}$ (also called the natural numbers), the rational numbers $\mathbb{Q}$, and the real numbers $\mathbb{R}$ are all examples of sets of numbers. We have the following chain of inclusions

$$\mathbb{N} \subseteq \mathbb{N}_0 \subseteq \mathbb{Z} \subseteq \mathbb{Q} \subseteq \mathbb{R}.$$

There is also the even larger set of complex numbers $\mathbb{C}$. We find for instance that

$$\mathbb{N}_0 \backslash \mathbb{N} = \{0\},$$

and that $\mathbb{Z} \backslash \mathbb{N}_0$ is the set of negative integers. Note that this is the complement of $\mathbb{N}_0$ within $\mathbb{Z}$. The complement of $\mathbb{N}_0$ within $\mathbb{R}$ is a larger set. The set $\mathbb{R} \backslash \mathbb{Q}$ (which is the complement of the rational numbers within $\mathbb{R}$) is often called the set of irrational numbers.

B.2   Combinatorics

In the derivation of the point probabilities for the binomial distribution, Example 3.2.1, we encountered the combinatorial quantity $\binom{n}{k}$. This number is the number of ways we can pick k out of n elements *disregarding the order*, since this corresponds to the number of ways we can pick out k x_i's to be equal to 1 and the remaining $n - k$ x_i's to equal 0 such that the sum $x_1 + \ldots + x_n = k$. If we take the order into account there are n possibilities for picking out the first element, $n - 1$ for the second, $n - 2$ for the third and so on, hence there are $n(n-1)(n-2)\cdots(n-k+1)$ ways of picking out k elements. We use the notation

$$n^{(k)} = n(n-1)(n-2)\cdots(n-k+1).$$

With $k = n$ this argument reveals that there are $k^{(k)} = k!$ orderings of a set of k elements. In particular, if we pick k elements in order there are $k!$ ways of reordering the set hence

$$\binom{n}{k} = \frac{n^{(k)}}{k!} = \frac{n!}{k!(n-k)!}.$$

The numbers $\binom{n}{k}$ are known as *binomial coefficients*. They are often encountered in combinatorial problems. One useful formula that relies on binomial coefficients is the following: For $x, y \in \mathbb{R}$ and $n \in \mathbb{N}$

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}, \tag{B.1}$$

which is simply known as the binomial formula. Letting $x = p$ and $y = 1 - p$ we see that

$$\sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} = (p + (1-p))^n = 1,$$

which shows that the point probabilities for the binomial distribution indeed sum to one (as they necessarily must).

A simple continuation of the argument also gives a formula for the *multinomial coefficients*

$$\binom{n}{k_1 \ldots k_m}$$

encountered in the multinomial distribution in Example 3.2.2. As we argued above there are $n!$ orderings of the n elements. If we assign labels from the set $\{1, \ldots, m\}$ by choosing one of the $n!$ orderings and then assign a 1 to the k_1 first elements, a 2 to the following k_2 elements and so on and so forth we get $n!$ ways of assigning labels to the elements. However, for any ordering we can reorder within each group and get the same labels. For a given ordering there are $k_1! k_2! \cdots k_m!$ other orderings that result in the same labels. Hence

$$\binom{n}{k_1 \ldots k_m} = \frac{n!}{k_1! k_2! \cdots k_m!}.$$

B.3   Limits and infinite sums

A sequence of real numbers, $x_1, x_2, x_3, \ldots$, often written as $(x_n)_{n \in \mathbb{N}}$, can have a *limit*, which is a value that x_n is close to for n large. We say that x_n *converges* to x if we, for all $\varepsilon > 0$ can find $N \geq 1$ such that

$$|x_n - x| \leq \varepsilon$$

for $n \geq N$. If x_n converges to x we write

$$x_n \to x, \text{ for } n \to \infty \quad \text{or} \quad \lim_{n \to \infty} x_n = x.$$

A sequence $(x_n)_{n \in \mathbb{N}}$ is *increasing* if

$$x_1 \leq x_2 \leq x_3 \ldots.$$

An increasing sequence is either upper bounded, in which case there is a least upper bound, and the sequence will approach this least upper bound, or the sequence is unbounded, in which case the sequence grows towards $+\infty$. An increasing sequence is therefore always convergent if we allow the limit to be $+\infty$. Likewise, a sequence is decreasing if

$$x_1 \geq x_2 \geq x_3 \ldots,$$

and a decreasing sequence is always convergent if we allow the limit to be $-\infty$.

Let $(x_n)_{n \in \mathbb{N}}$ be a sequence of *non-negative* reals, i.e. $x_n \geq 0$, and define

$$s_n = \sum_{k=1}^{n} x_k = x_1 + x_2 + \ldots + x_n,$$

then, since the x's are non-negative, the sequence $(s_n)_{n \in \mathbb{N}}$ is increasing, and it has a limit, which we denote

$$\sum_{n=1}^{\infty} x_n = \lim_{n \to \infty} s_n.$$

It may be $+\infty$. We write

$$\sum_{n=1}^{\infty} x_n < \infty$$

if the limit is not ∞.

If $(x_n)_{n \in \mathbb{N}}$ is any sequence of reals we define

$$x_n^+ = \max\{x_n, 0\} \quad \text{and} \quad x_n^- = \max\{-x_n, 0\}.$$

Then $x_n = x_n^+ - x_n^-$ and the sequences $(x_n^+)_{n \in \mathbb{N}}$ and $(x_n^-)_{n \in \mathbb{N}}$ are sequences of positive numbers. They are known as the positive respectively the negative part of the sequence $(x_n)_{n \in \mathbb{N}}$. If

$$s^+ = \sum_{n=1}^{\infty} x_n^+ < \infty$$

and

$$s^- = \sum_{n=1}^{\infty} x_n^- < \infty$$

then we define the infinite sum

$$\sum_{n=1}^{\infty} x_n = s^+ - s^- = \sum_{n=1}^{\infty} x_n^+ - \sum_{n=1}^{\infty} x_n^-$$

and we say that the sum is convergent. If one of the sums, s^+ or s^-, is $+\infty$ we say that the sum is divergent. We may also observe that

$$|x_n| = x_n^+ + x_n^-$$

and we conclude that the sum is convergent if and only if

$$\sum_{n=1}^{\infty} |x_n| < \infty.$$

A classical infinite sum is the geometric series with $x_n = \rho^{n-1}$ for $\rho \in (-1, 1)$, then

$$\sum_{n=1}^{\infty} \rho^{n-1} \left(= \sum_{n=0}^{\infty} \rho^n \right) = \frac{1}{1 - \rho}. \tag{B.2}$$

Another example is the infinite series representation of the exponential function,

$$\exp(\lambda) = \sum_{n=0}^{\infty} \frac{\lambda^n}{n!} \tag{B.3}$$

valid for $\lambda \in \mathbb{R}$.

B.4   Integration

Integration as introduced in elementary calculus courses is a way to compute the area under the graph of a continuous function. Thus if $f : \mathbb{R} \to \mathbb{R}$ is a continuous function we can introduce a number,

$$\int_a^b f(x)\mathrm{d}x,$$

which is the area under the graph of f from a to b. This area being computed with a sign. In some cases it makes sense to let a or b tend to $-\infty$ or ∞ respectively. In that case we get that

$$\int_{-\infty}^{\infty} f(x)\mathrm{d}x$$

is the entire area under the graph from $-\infty$ to ∞. If f is a positive function, this "area" always makes sense, though it may be infinite.

If $f(x) \geq 0$ for all $x \in \mathbb{R}$, the sequence of numbers

$$I_n = \int_{-n}^{n} f(x)\mathrm{d}x$$

is an increasing sequence, hence it has a limit, which may equal $+\infty$, for $n \to \infty$. We write

$$\int_{-\infty}^{\infty} f(x)\mathrm{d}x = \lim_{n\to\infty} I_n = \lim_{n\to\infty} \int_{-n}^{n} f(x)\mathrm{d}x.$$

We say that the function f is *integrable* over $\mathbb{R}$ if this number is finite.

As for the infinite sums above, if f is any continuous function one can define the two positive functions

$$f^+(x) = \max\{f(x), 0\} \quad \text{and} \quad f^-(x) = \max\{-f(x), 0\}$$

such that $f(x) = f^+(x) - f^-(x)$ and $|f(x)| = f^+(x) + f^-(x)$. We say that f is integrable over $\mathbb{R}$ if the two positive functions f^+ and f^- are integrable or equivalently if the positive function $|f|$ is integrable. In this case we also have that

$$\int_{-\infty}^{\infty} f(x)\mathrm{d}x = \lim_{n\to\infty} \int_{-n}^{n} f(x)\mathrm{d}x.$$

Prior to using this identity we have to check if

$$\lim_{n\to\infty} \int_{-n}^{n} |f(x)|\mathrm{d}x < +\infty.$$

B.4.1   Gamma and beta integrals

The integral

$$\int_0^\infty x^{\lambda-1}\exp(-x)\mathrm{d}x$$

is finite for $\lambda > 0$. It is known as the Γ-integral (gamma integral), and we define the Γ-function by

$$\Gamma(\lambda) = \int_0^\infty x^{\lambda-1}\exp(-x)\mathrm{d}x. \tag{B.4}$$

The Γ-function is a classical and much studied function. It holds that

$$\Gamma(\lambda+1) = \lambda\Gamma(\lambda) \tag{B.5}$$

for $\lambda > 0$, which together with $\Gamma(1) = 1$ implies that

$$\Gamma(n+1) = n!$$

for $n \in \mathbb{N}_0$. For non-integer λ the Γ-function takes more special values. One of the peculiar results about the Γ-function that can give more insight into the values of $\Gamma(\lambda)$ for non-integer λ is the *reflection formula*, which states that for $\lambda \in (0,1)$

$$\Gamma(\lambda)\Gamma(1-\lambda) = \frac{\pi}{\sin(\pi\lambda)}.$$

For $\lambda = 1/2$ we find that $\Gamma(1/2)^2 = \pi$, thus

$$\Gamma(1/2) = \int_0^\infty \frac{1}{\sqrt{x}}\exp(-x)\mathrm{d}x = \sqrt{\pi}.$$

This can together with (B.5) be used to compute $\Gamma(1/2 + n)$ for all $n \in \mathbb{N}_0$. For instance,

$$\Gamma(3/2) = \Gamma(1/2+1) = \frac{\Gamma(1/2)}{2} = \frac{\sqrt{\pi}}{2}.$$

The B-function (B is a capital β – quite indistinguishable from the capital b – and the pronunciation is thus beta-function) is defined by

$$B(\lambda_1,\lambda_2) = \frac{\Gamma(\lambda_1)\Gamma(\lambda_2)}{\Gamma(\lambda_1+\lambda_2)} \tag{B.6}$$

for $\lambda_1, \lambda_2 > 0$. The B-function has an integral representation as

$$B(\lambda_1,\lambda_2) = \int_0^1 x^{\lambda_1-1}(1-x)^{\lambda_2-1}\mathrm{d}x. \tag{B.7}$$

B.4.2   Multiple integrals

If $f : \mathbb{R}^2 \to \mathbb{R}$ is a continuous function then for any $a < b$, $a, b \in \mathbb{R}$, function

$$x \mapsto \int_a^b f(x, y) \mathrm{d}y$$

is also a continuous function from $\mathbb{R}$ to $\mathbb{R}$. We can integrate this function over the interval $[c, d]$ for $c < d$, $c, d \in \mathbb{R}$ and get the multiple integral

$$\int_c^d \int_a^b f(x, y) \mathrm{d}y \mathrm{d}x.$$

We can interpret the value of this integral as the volume under the function f over the rectangle $[c, d] \times [a, b]$ in the plane.

It is possible to interchange the order of integration so that

$$\int_c^d \int_a^b f(x, y) \mathrm{d}y \mathrm{d}x = \int_a^b \int_c^d f(x, y) \mathrm{d}x \mathrm{d}y.$$

In general, if $f : \mathbb{R}^k \to \mathbb{R}$ is a continuous function of k variables the k-times iterated integral

$$\int_{a_1}^{b_1} \cdots \int_{a_k}^{b_k} f(x_1, \ldots, x_k) \mathrm{d}x_k \ldots \mathrm{d}x_1$$

is a sensible number, which we can interpret as an $k + 1$-dimensional volume under the graph of f over the k-dimensional box

$$[a_1, b_1] \times \ldots \times [a_k, b_k].$$

It is notable that the order of the integrations above do not matter.

As for the univariate case we can for positive, continuous functions $f : \mathbb{R}^k \to [0, \infty)$ always define

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, \ldots, x_k) \mathrm{d}x_k \ldots \mathrm{d}x_1 = \lim_{n \to \infty} \int_{-n}^{n} \cdots \int_{-n}^{n} f(x_1, \ldots, x_k) \mathrm{d}x_k \ldots \mathrm{d}x_1$$

but the limit may be equal to $+\infty$. For any continuous function $f : \mathbb{R}^k \to \mathbb{R}$ we have that if

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} |f(x_1, \ldots, x_k)| \mathrm{d}x_k \ldots \mathrm{d}x_1 < \infty$$

then

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, \ldots, x_k) \mathrm{d}x_k \ldots \mathrm{d}x_1 = \lim_{n \to \infty} \int_{-n}^{n} \cdots \int_{-n}^{n} f(x_1, \ldots, x_k) \mathrm{d}x_k \ldots \mathrm{d}x_1.$$

Index