Introduction to Statistics

Lutz Dümbgen

May 23, 2023

Contents

Preface

These lecture notes are a translation of my textbook 'Einführung in die Statistik' published by Birhäuser 2015. They are related to a regular course for B.Sc. students in Mathematics in their second or third year. All participants are expected to be familiar with basic probability, e.g. as taught in the B.Sc. course 'Kombinatorik und Wahrscheinlichkeitsrechnung'.

The choice of topics is admittedly subjective and reflects my experiences with statistical consulting. A distinguishing feature of our field is the ability to quantify uncertainty. For that reason I focus on confidence regions and much less on point estimation. The latter topic is, in my opinion, overemphasised among academic statisticians. In the present notes numerous procedures with guaranteed properties for arbitrary finite sample sizes are introduced. One exception is the chapter on density estimation: Here the goal is to discuss an ill-posed problem and to illustrate important concepts such as bias and regularity assumptions.

The amount of material corresponds to a lecture of four hours plus two hours of exercises per week. For shorter courses as in Bern I skip various sections. More complex procedures such as regression methods or multivariate analyses are treated in advanced courses. In particular, likelihood methods are not introduced here but in later courses.

As a student and assistant at the University of Heidelberg I had the priviledge to learn a lot about probability and statistics from Hermann Rost, Dietrich W. Müller and Günter Sawitzki. My choice of topics and examples reflects partly the inspiring introductory courses by D.W. Müller. Günter Sawitzki convinced me that stochastic order is a very relevant concept. He also ignited my interest for graphical methods and numerical aspects. Richard Gill provided valuable information about the law suit of Lucia de Berk (Example 8.10 in Section 8.2).

Over the last sixteen years, numerous students and assistants have provided comments and constructive criticism which helped my to improve the material substantially. An incomplete list of names comprises Ladina Abbühl, Sofia Caprez, Mika Frei, Manuela Häfliger, Christoph Kopp, Fabio Matti, Michael Mosimann, Philipp Muri und Niki Zumbrunnen. Special thanks are due to Dominic Schuhmacher, Kaspar Stucki and Andrea Fraefel who read large parts of the manuscript.

Bern, May 2021 *Lutz Dümbgen*

Technical hint. The numerical examples and some exercises require suitable software. All calculations and graphics in this book have been produced with the open-source software R [22]. Section A.1 in the appendix contains specific hints for that. I refrained from including this into the main text. Since R is already the seventh programming environment I am working with, my confidence in the persistence of such systems is limited.

Chapter 1

Introduction

In this chapter we discuss some explicit examples of statistical analyses and procedures to illustrate several important ideas. Thereafter we introduce some basic concepts which will recur repeatedly in later chapters.

1.1 The Chocolate-Tasting Lady and Fisher's Exact Test

R. A. Fisher[1] illustrated the test carrying his name with a little randomised experiment involving a tea-tasting lady. Here we consider a similar experiment:

Example 1.1 (Chocolate-tasting lady). A lady claims to be able to distinguish chocolate out of a freshly opened package from chocolate out of a package which was opened at least one day ago by taste and smell. Since this claim is smiled at, she and her peers agree on the following *randomised experiment*: Two identical small packages, each one containing four pieces of chocolate, are placed overnight into a cupboard, one package open, the other one still closed. The next day, the second package is opened, too, and the eight pieces of chocolate are presented to the lady in random order. Her task is to identify the four pieces from the freshly opened package.

The goal of this experiment is to verify the *working hypothesis* that the lady is indeed able to distinguish between "fresh" and "old" chocolate. Easier to describe is the *null hypothesis* that she does not smell or taste any difference. Under the latter null hypothesis, the probability that the lady solves the task equals

$$1 \Big/ \binom{8}{4} \;=\; 1/70 \;\approx\; 0.0143,$$

because there are $\binom{8}{4} = 70$ possible answers she could give. If she solves the task, we may claim with confidence $69/70 \approx 0.9857$ that the working hypothesis is true. If the lady fails, we make no definitive statement.[2]

Instead of "with confidence $69/70$" one could also say "with an uncertainty of $1/70$". Both statements should be elaborated a bit. Even if the lady solves the task, we don't know for sure that the working hypothesis is true. It could happen that a few years later the lady confesses to have cheated and solved the task just by chance. The stated confidence level may be interpreted as follows: Suppose a very large number of people claim to have the same ability, and they all participate in such an experiment. If none of them has the ability, the relative proportion of people passing the test is close to $1/70$.

[1] Ronald A. Fisher (1890-1962): important British statistician and mathematical biologist.

[2] This experiment has really been carried out, and the lady solved the task flawlessly!

Next, we recall the definition of hypergeometric distributions:

Definition 1.2 (Hypergeometric distributions). A random variable X with values in $\mathbb{N}_0$ is hypergeometrically distributed with parameters $N \in \mathbb{N}$ and $\ell, n \in \{0, 1, \ldots, N\}$, if for arbitrary numbers $x \in \mathbb{N}_0$,

$$\mathbb{P}(X = x) \;=\; f_{N,\ell,n}(x) := \binom{\ell}{x}\binom{N-\ell}{n-x}\Big/\binom{N}{n}.$$

(Here we define $\binom{a}{b} := 0$ if $b < 0$ or $b > a$.) We denote this distribution by $\mathrm{Hyp}(N, \ell, n)$. The corresponding distribution function is denoted by $F_{N,\ell,n}$, i.e. $F_{N,\ell,n}(x) := \mathbb{P}(X \leq x)$ for $x \in \mathbb{R}$. Thus for $x \in \mathbb{N}_0$,

$$F_{N,\ell,n}(x) \;=\; \sum_{k=0}^{x} f_{N,\ell,n}(k).$$

These distributions may be explained with an urn model: Consider an urn with N balls from which ℓ balls are marked. Now one draws randomly and without replacement n balls from the urn. The random number X of marked balls in this sample follows $\mathrm{Hyp}(N, \ell, n)$.

In Example 1.1 one could ask whether it would be sufficient if the lady detected at least three of the four "fresh" pieces. Under the null hypothesis, the number X of correctly detected "fresh" pieces is hypergeometrically distributed with parameters 8, 4 and 4, so

$$\mathbb{P}(X \geq 3) \;=\; \binom{4}{3}\binom{4}{1}\Big/\binom{8}{4} + \binom{4}{4}\binom{4}{0}\Big/\binom{8}{4} \;=\; \frac{17}{70} \;\approx\; 0.2429.$$

Thus in case of $X \geq 3$ we could only claim with confidence $53/70 \approx 0.7571$ that the lady has a well-trained sense of smell and taste. In Exercise 1.1 we consider a different variant of this experiment where the lady gets a second chance in case of $X = 3$.

A particular feature of Example 1.1 is that all involved people knew that precisely four pieces of chocolate are "fresh" and four are "old". Now we discuss a more general version of Fisher's exact test in a different situation.

Example 1.3 (Comparison of two treatments in a randomised study). Suppose one wants to verify that a certain (medical) treatment 1 is better than a standard treatment 2 (or no treatment at all). As an explicit example consider the regular intake of vitamin C (ascorbic acid) during winter to prevent a flue (treatment 1) versus no such measure (treatment 2). To verify the benefits of treatment 1, a group of N test persons is divided randomly into two groups: The n_1 individuals in group 1 receive treatment 1, the n_2 individuals in group 2 receive treatment 2. One talks about a *blinded study* if the test persons don't know which group they belong to. With this one wants to prevent so-called placebo effects. In the explicit example with vitamin C one could give each test person small capsules to take in daily; in group 1 the capsules contain vitamin C, in group 2 they contain only a placebo.

After a certain time one determines the number of successes and failures in both groups. The results may be summarised in a two-by-two table:

	Success	Failure	
Treatment 1	H_1	$n_1 - H_1$	n_1
Treatment 2	H_2	$n_2 - H_2$	n_2
	$H_+ := H_1 + H_2$	$N - H_+$	N

Thus one observed H_+ successes in total, H_i times in group i.

The total number H_+ is random and depends on many factors and circumstances. But under the *null hypothesis* that the two treatments have precisely the same effects, the conditional distribution of H_1, given H_+, equals $\mathrm{Hyp}(N, H_+, n_1)$. For under the null hypothesis, by the end of the study there will be H_+ people with successful treatment and $N - H_+$ people with failure, regardless of the random splitting into two groups.

Under the *working hypothesis* that treatment 1 is better than treatment 2 one expects relatively large values of H_1. The question is how large H_1 should be to be convinced of the working hypothesis. To this end we fix a *test level* $\alpha \in (0, 1)$ and consider the quantiles

$$q_{1-\alpha;N,\ell,n} \;:=\; \min\bigl\{x \in \mathbb{N}_0 : F_{N,\ell,n}(x) \geq 1 - \alpha\bigr\}.$$

Under the null hypothesis we have the inequality

$$\begin{aligned}\mathbb{P}\bigl(H_1 > q_{1-\alpha;N,\ell,n_1} \,\big|\, H_+ = \ell\bigr) \;&=\; 1 - \mathbb{P}\bigl(H_1 \leq q_{1-\alpha;N,\ell,n_1} \,\big|\, H_+ = \ell\bigr)\\ &=\; 1 - F_{N,\ell,n_1}(q_{1-\alpha;N,\ell,n_1})\\ &\leq\; \alpha.\end{aligned}$$

In particular,

$$\begin{aligned}\mathbb{P}\bigl(H_1 > q_{1-\alpha;N,H_+,n_1}\bigr) \;&=\; \sum_{\ell=0}^{N} \mathbb{P}(H_+ = \ell)\, \mathbb{P}\bigl(H_1 > q_{1-\alpha;N,\ell,n_1} \,\big|\, H_+ = \ell\bigr)\\ &\leq\; \sum_{\ell=0}^{N} \mathbb{P}(H_+ = \ell)\, \alpha\\ &=\; \alpha.\end{aligned}$$

In case of $H_1 > q_{1-\alpha;N,H_+,n_1}$ we may claim with confidence $1 - \alpha$ that the null hypothesis is wrong and conclude indirectly that treatment 1 is more effective than treatment 2.

Here comes a fictitious data example: In a randomised study during November, December and January, $N = 40$ test persons took a small capsule each day. For $n_1 = 20$ persons all capsules contained a certain dose of vitamin C, for the other $n_2 = 20$ persons the capsules contained a placebo. By the end of January it turned out that in group 1, $H_1 = 15$ persons stayed healthy during the whole period while $n_1 - H_1 = 5$ persons had a flu at least once. In group 2 the numbers were $H_2 = 11$ and $n_2 - H_2 = 9$. Typically one uses the test level $\alpha = 5\%$. Here this leads to the critical value $q_{1-\alpha;N,H_+,n_1} = q_{0.95;40,26,20} = 15$, because $F_{40,26,20}(14) \approx 0.8399$ and $F_{40,26,20}(15) \approx 0.9521$. Since H_1 is not larger than 15, we cannot claim anything about the positive effect of vitamin C with confidence 95%.

1.2 Tail Regions and P-Values

Fisher's exact test and numerous other statistical procedure utilise a special transformation of test quantities into so-called p-values in the unit interval. We describe now the general underlying principle which will recur repeatedly. The starting point is a random variable X and a hypothetical probability distribution P_o of X. The question is whether X follows the distribution P_o indeed or whether the observed value of X is "suspiciously small" or "suspiciously large". An indispensable tool is the distribution function F_o of P_o. That means, $F_o(x) := P_o((-\infty, x]) = \mathbb{P}(X \leq x)$ for any real number, and $F_o(x-) := \lim_{s \to x, s<x} F_o(s) = P_o((-\infty, x)) = \mathbb{P}(X < x)$.

To judge whether X is suspiciously small, we compute the *left-sided p-value*

$$P_o((-\infty, X]) \;=\; F_o(X),$$

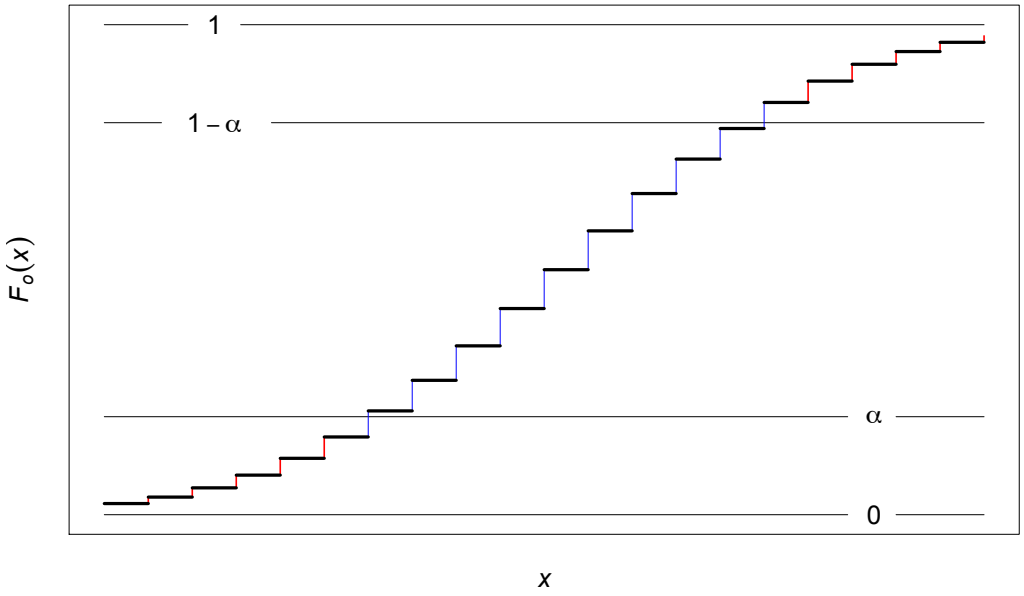

Figure 1.1: Illustration of Lemma 1.4.

and to judge whether X is suspiciously large, we compute the *right-sided p-value*

$$P_o([X,\infty)) \;=\; 1 - F_o(X-).$$

With the *two-sided p-value*

$$2 \cdot \min\big\{P_o((-\infty,X]), P_o([X,\infty))\big\} \;=\; 2 \cdot \min\{F_o(X), 1 - F_o(X-)\}$$

we may judge whether X is suspiciously small or large. In all three cases, a small p-value is evidence against the hypothesis that X follows P_o. The subsequent lemma provides a precise statement.

Lemma 1.4 (P-values). *Let X be a real-valued random variable with distribution P_o and distribution function F_o. Then*

$$\left.\begin{aligned} \mathbb{P}\big(F_o(X) \le \alpha\big) \\ \mathbb{P}\big(1 - F_o(X-) \le \alpha\big) \\ \mathbb{P}\big(2 \cdot \min\{F_o(X), 1 - F_o(X-)\} \le \alpha\big) \end{aligned}\right\} \;\le\; \alpha$$

for any $\alpha \in (0,1)$. All three inequalities are equalities if F_o is continuous.

Before proving this lemma let us consider the special case that $\mathbb{P}(X \in \mathbb{Z}) = P_o(\mathbb{Z}) = 1$. Here F_o is a step function which is constant on any interval $[x, x+1)$, $x \in \mathbb{Z}$. Figure 1.1 illustrates Lemma 1.4 in this situation. One sees the graph of F_o. The jump size of F_o at a point $x \in \mathbb{Z}$, that is the difference $F_o(x) - F_o(x-)$, equals $P_o(\{x\}) = \mathbb{P}(X = x)$. The probability that $F_o(X) \le \alpha$ is equal to the sum of all jump sizes at positions x with $F_o(x) \le \alpha$, and this sum is clearly smaller or equal to α. Analogously the probability that $1 - F_o(X-) \le \alpha$ is equal to the sum of all jump sizes at positions x with $F_o(x-) \ge 1 - \alpha$, and this sum is at most α.

Proof of Lemma 1.4. We use the well-known fact that the function F_o is non-decreasing and continuous from the right with limits $\lim_{x\to-\infty} F_o(x) = 0$ and $\lim_{x\to\infty} F_o(x) = 1$. For fixed $\alpha \in (0,1)$ the real number $x_o := \inf\{x \in \mathbb{R} : F_o(x) > \alpha\}$ is well-defined with $F_o(x) \le \alpha$ for

any $x < x_o$ and $F_o(x) > \alpha$ for any $x > x_o$. Moreover, $F_o(x_o) \geq \alpha$ because of the right-continuity of F_o. In case of $F_o(x_o) > \alpha$,

$$\mathbb{P}\big(F_o(X) \leq \alpha\big) \;=\; \mathbb{P}(X < x_o) \;=\; F_o(x_o -) \leq \alpha.$$

In case of $F_o(x_o) = \alpha$,

$$\mathbb{P}\big(F_o(X) \leq \alpha\big) \;=\; \mathbb{P}(X \leq x_o) \;=\; F_o(x_o) = \alpha.$$

If F_o is continuous, we are necessarily in the latter case.

The inequalities for $1 - F_o(X -)$ may be verified analogously or by means of a symmetry consideration: The random variable $\widetilde{X} := -X$ has distribution function $\widetilde{F}_o(x) = 1 - F_o((-x) -)$. Consequently,

$$\mathbb{P}\big(1 - F_o(X -) \leq \alpha\big) \;=\; \mathbb{P}\big(\widetilde{F}_o(\widetilde{X}) \leq \alpha\big) \;\leq\; \alpha.$$

Equality holds true if $\widetilde{F}_o$ is continuous which is equivalent to F_o being continuous.

Concerning the two-sided p-value, note that at least one of the two values $F_o(X) = P_o((-\infty, X])$ and $1 - F_o(X -) = P_o([X, \infty))$ has to be greater than or equal to $1/2$. Thus

$$\begin{aligned}\mathbb{P}\big(2 \cdot \min\{F_o(X), 1 - F_o(X -)\} \leq \alpha\big) \;&=\; \mathbb{P}\big(F_o(X) \leq \alpha/2 \;\text{ or }\; 1 - F_o(X -) \leq \alpha/2\big) \\ &=\; \mathbb{P}\big(F_o(X) \leq \alpha/2\big) + \mathbb{P}\big(1 - F_o(X -) \leq \alpha/2\big) \\ &\leq\; \alpha/2 + \alpha/2 \;=\; \alpha\end{aligned}$$

with equality in case of a continuous distribution function F_o. □

Example 1.5 (Fisher's exact test). We consider once more example 1.3. The null hypothesis is rejected if $H_1 > q_{1-\alpha;N,H_+,n_1}$. One can easily verify that the latter inequality is equivalent to the right-sided p-value $1 - F_{N,H_+,n_1}(H_1 -) = 1 - F_{N,H_+,n_1}(H_1 - 1)$ being less than or equal to α. The present setting corresponds to Lemma 1.4 with $X = H_1$ and $P_o = \mathrm{Hyp}(N, H_+, n_1)$, the conditional distribution of H_1, given H_+, under the null hypothesis.

Figure 1.2 illustrates this for the fictitious study on vitamin C. The group sizes were $n_1 = n_2 = 20$ with the success numbers $H_1 = 15$ and $H_2 = 11$. The figure shows twice a bar plot of the hypergeometric weight function $f_{40,26,20}$. In the upper part the weights $f_{40,26,20}(x)$ with $x > q_{0.95;40,26,20} = 15$ are marked dark. The sum of these weights is less than or equal to 5%, the sum of the remaining weights is at least 95%. In the lower part the weights $f_{40,26,20}(x)$ with $x \geq H_1 = 15$ are marked dark. The sum of these weights is the (right-sided) p-value $1 - F_{40,26,50}(14) \approx 0.1601$. The fact that the latter p-value is greater than the test level 5% confirmes that H_1 is not larger than the critical value $q_{0.95;N,H_+,n_1}$.

1.3 The Size of a Population

Many statistical analyses involve *samples* from a certain *population*. By means of the sample one wants to draw conclusions about the total population. Quite often one focuses on certain averages or relative proportions within the population whereas its total size is less frequently of interest. But inference about a population size is an interesting problem and useful to illustrate statistical concepts.

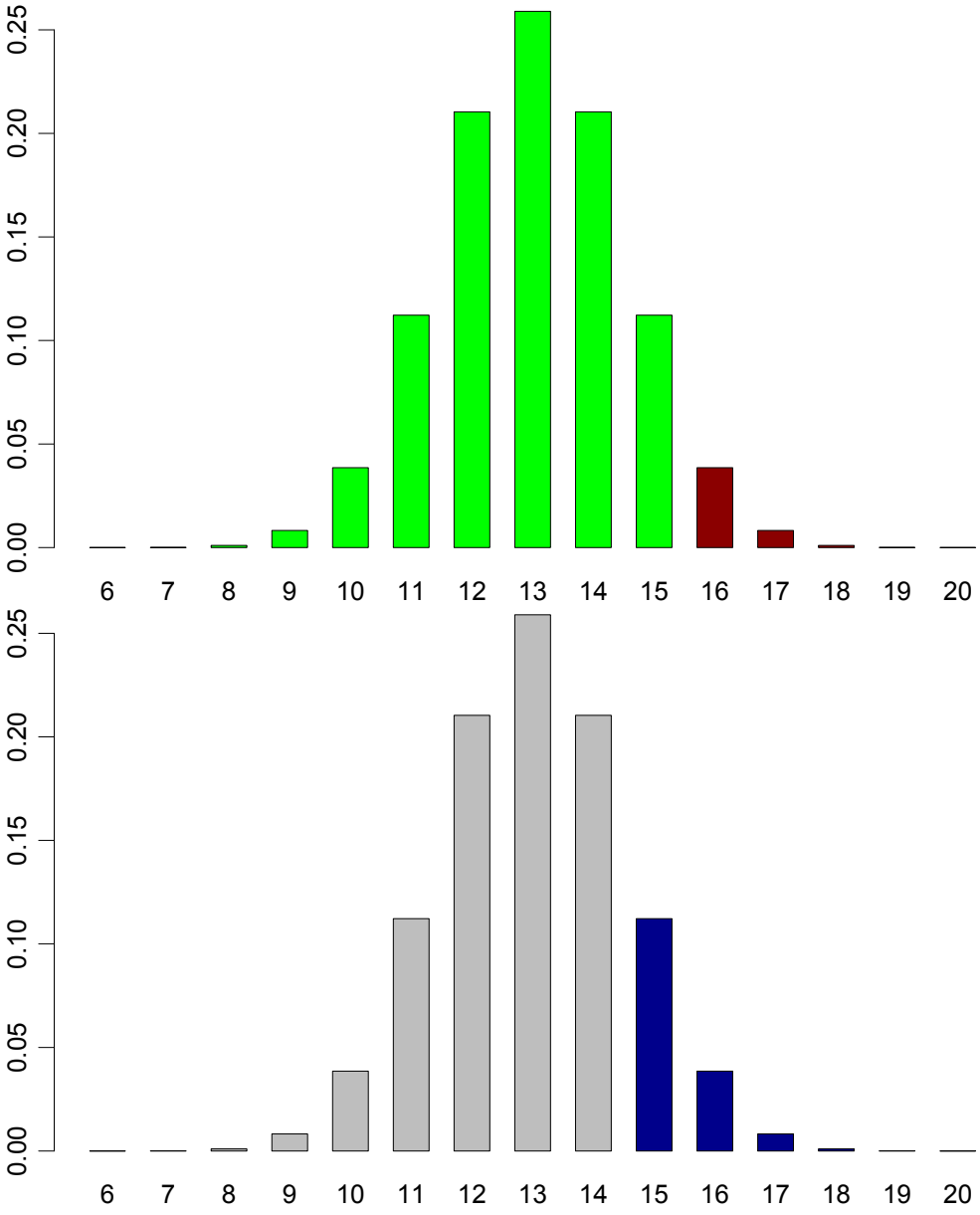

Figure 1.2: Fisher's exact test via critical value (above) or p-value (below).

Population and sample space. Let $\mathcal{M}$ be a population of "individuals" of unknown size

$$N := \#\mathcal{M}.$$

Now we draw a sample $\omega = (\omega_1, \omega_2, \ldots, \omega_n)$ of size n from that population without replacement. That means, ω is an element of the sample space

$$\big\{\omega \in \mathcal{M}^n : \omega_i \neq \omega_j \text{ whenever } i \neq j\big\}.$$

This sample space consists of

$$[N]_n := N(N-1)\cdots(N-n+1)$$

different samples. For if we want to specify an arbitrary sample ω, there are N possibilities for ω_1, then $N-1$ possibilities for ω_2, thereafter $N-2$ possibilities for ω_3 and so on. In general we write $[a]_k := \prod_{i=0}^{k-1}(a-i)$ for $a \in \mathbb{R}$ and $k \in \mathbb{N}$ while $[a]_0 := 1$.

The case of individuals with reference numbers

We assume that each individual in the population has a unique reference number between 1 and N which is easily obtained. Here we may identify the population $\mathcal{M}$ with the set $\{1, 2, \ldots, N\}$, and our sample ω corresponds to a tuple of n different natural numbers. A useful *statistic* is the number

$$X(\omega) := \max(\omega_1, \omega_2, \ldots, \omega_n) \geq n.$$

This statistic $X(\omega)$ is obviously a lower bound for the unknown population size N, and this is essentially the only conclusion we may draw with absolute certainty. The art of statistics is to draw further conclusions about N. In particular we would like to obtain an *upper bound* for N.

Example 1.6 (Matriculation numbers at the University of Bern in 2005/2006). Students receive a unique matriculation number when they sign up for a Swiss university program for the first time. This number is retained even when they switch to a different university. The eight-digit matriculation number has the form

$$J_1J_2 - Z_1Z_2Z_3 - Z_4Z_5P.$$

Here J_1 and J_2 denote the academic year of the very first registration, for instance $J_1J_2 = 05$ for students who started in the fall semester 2005 or in the spring semester 2006. The digits $Z_1, Z_2, \ldots, Z_5$ correspond to a five-digit number within a certain range depending on the university. Especially at the University of Bern, these numbers are assigned consecutively from $10'000$ through $14'999$. The last digit, P, is just a security digit to detect errors when filling out electronic forms. For instance, if a student has matriculation number $05-106-020$, this means that he was the 603-rd person to sign up at the University of Bern in the academic year 2005/2006.

Now let N be the total number of students having started their university studies in Bern in the academic year 2005/2006. In a lecture we obtained the matriculation numbers of $n = 9$ such students, resulting in a tuple $\omega = (\omega_1, \ldots, \omega_9)$ of nine different numbers in $\{1, 2, \ldots, N\}$. It turned out that $X(\omega) = 2782$.

A statistical model. To say more about the unknown quantity N, we have to make certain assumptions about our sample. For simplicity we assume that ω was chosen completely at random (even if it was obtained differently). That way the statistic X becomes a *random variable*

$$X : \Omega \to \mathbb{Z}$$

defined on the sample space

$$\Omega \;:=\; \big\{\omega \in \mathbb{N}^n : \omega_i \neq \omega_j \text{ whenever } i \neq j\big\}.$$

The latter is equipped with a probability measure $\mathbb{P}_N$ depending on the unknown *parameter* N. Precisely, $\mathbb{P}_N$ is the uniform distribution on the subset $\Omega_N = \big\{\omega \in \Omega : X(\omega) \leq N\big\}$ of Ω, i.e.

$$\mathbb{P}_N(B) \;=\; \frac{\#(B \cap \Omega_N)}{\#\Omega_N} \;=\; \frac{\#(B \cap \Omega_N)}{[N]_n}$$

for $B \subset \Omega$. Generally the dependence of various probabilities, expected values and other objects on the population size is indicated by a subscript N. In particular the following equations are true:

$$\mathbb{P}_N(X = x) \;=\; \begin{cases} \dfrac{n[x-1]_{n-1}}{[N]_n} & \text{for } x \in \{n, n+1, \ldots, N\}, \\ 0 & \text{else,} \end{cases} \tag{1.1}$$

$$F_N(x) := \mathbb{P}_N(X \leq x) \;=\; \begin{cases} 0 & \text{for } x < n, \\ \dfrac{[x]_n}{[N]_n} & \text{for } x \in \{n, n+1, \ldots, N\}, \\ 1 & \text{for } x \geq N. \end{cases} \tag{1.2}$$

As to (1.1), to generate a sample $\omega \in \Omega$ with $X(\omega) = x \in \{n, n+1, n+1, \ldots\}$, one could first specify which of the n components of ω equals x, and then one could fill the remaining $n-1$ positions with different numbers from $\{1, \ldots, x-1\}$ which amounts to $[x-1]_{n-1}$ possibilities. Formula (1.2) follows from the fact that there are $[x]_n$ samples $\omega \in \Omega$ with $X(\omega) \leq x$.

An estimator for N. By means of the sample ω we would like to compute an estimate $\widehat{N}(\omega)$ of the population size N. A first attempt would be $\widehat{N}(\omega) := X(\omega)$. But this value is systematically too small. To make this precise we compute the expected value of X.

Lemma 1.7. *For arbitrary $N \geq n$,*

$$\mathbb{E}_N(X) \;=\; \frac{n\,(N+1)}{n+1}.$$

Proof of Lemma 1.7. Equation (1.1) and the fact that $\sum_{x=n}^{N} \mathbb{P}_N(X = x) = 1$ lead to the general formula $\sum_{x=n}^{N} [x-1]_{n-1} = [N]_n/n$ for integers $1 \leq n \leq N$. In other words,

$$\sum_{j=m}^{M} [j]_m \;=\; \frac{[M+1]_{m+1}}{m+1} \quad \text{for integers } 0 \leq m \leq M. \tag{1.3}$$

Hence $\mathbb{E}_N(X)$ equals

$$\sum_{x=n}^{N} \mathbb{P}_N(X = x) \cdot x \;=\; \frac{n}{[N]_n} \sum_{x=n}^{N} [x]_n \;=\; \frac{n}{[N]_n} \frac{[N+1]_{n+1}}{n+1} \;=\; \frac{n\,(N+1)}{n+1}.$$

Exercise 1.6 provides an alternative proof of this lemma. □

Lemma 1.7 implies that

$$\widehat{N} \;:=\; \frac{n+1}{n}\, X - 1$$

is an *unbiased estimator* of N. That means, for arbitrary parameters $N \geq n$,

$$\mathbb{E}_N(\widehat{N}) \;=\; N.$$

The imprecision of an arbitrary estimator $\widehat{N}$ may be quantified by its *mean quadratic error*

$$\mathbb{E}_N\big((\widehat{N} - N)^2\big).$$

The imprecision of our specific estimator $\widehat{N} = (1 + 1/n)X - 1$ is analyzed in Exercise 1.7. It turns out that

$$\mathbb{E}_N\big((\widehat{N} - N)^2\big) \;<\; \frac{N^2}{n^2}.$$

This implies that

$$\mathbb{E}_N\left(\Big|\frac{\widehat{N}}{N} - 1\Big|\right) \;\leq\; \sqrt{\mathbb{E}_N\left(\Big(\frac{\widehat{N}}{N} - 1\Big)^2\right)} \;<\; \frac{1}{n}.$$

Here we utilised the well-known inequality $\mathbb{E}(|Y|)^2 \leq \mathbb{E}(Y^2)$ for real-valued random variables Y. Hence the relative error $|\widehat{N}/N - 1|$ is smaller than $1/n$ on average.

Example 1.8 (Matriculations 2005/2006). In Example 1.6 we observed a sample of size $n = 9$ with $X = 2782$, so

$$\widehat{N} \;=\; \frac{10}{9} \cdot 2782 - 1 \;=\; 3090.11\overline{1}.$$

Hence we guess that 3090 students enrolled in the academic year 2005/2006 at the University of Bern.

Confidence bounds for N. Instead of an estimator one can also determine bounds for N which are correct with a given *confidence*. The idea is to consider all hypothetical values of N and to check in each case whether the observed value of X is "suspiciously small" or "suspiciously large" for the corresponding distribution function F_N. To this end we use Lemma 1.4 about p-values.

In the present context it follows from Lemma 1.4 that

$$\mathbb{P}_N(F_N(X) \leq \alpha) \;\leq\; \alpha$$

for any fixed number α. In other words, with probability $1 - \alpha$ the unknown true parameter N satisfies the inequality $F_N(X) > \alpha$ which is equivalent to $[X]_n/[N]_n > \alpha$. Since $[N]_n$ is strictly increasing in $N \geq n$, these inequalities are equivalent to $N \leq b_\alpha(X)$, where

$$\begin{aligned} b_\alpha(x) \;&:=\; \max\big\{N \geq n : F_N(x) > \alpha\big\} \\ &=\; \max\big\{N \geq x : [N]_n < [x]_n/\alpha\big\} \end{aligned}$$

for integers $x \geq n$. This data-dependent number $b_\alpha(X)$ is an *upper $(1-\alpha)$-confidence bound* for N. That means,

$$\mathbb{P}_N(N \leq b_\alpha(X)) \;\geq\; 1 - \alpha$$

for any value of $N \geq n$. A simple explicit formula for $b_\alpha(x)$ is not available, but its numerical computation is straightforward. When analysing a specific sample ω, we claim *with confidence* $1 - \alpha$ that $N \leq b_\alpha(X(\omega))$. Note that we do *not* say: "With probability $1 - \alpha$, $N \leq b_\alpha(X(\omega))$." For a given data set ω, the inequality $N \leq b_\alpha(X(\omega))$ is true with probability one or zero! The formulation "with confidence $1 - \alpha$" indicates that we use a procedure with error rate at most α in the long run. See also our comments on Example 1.1.

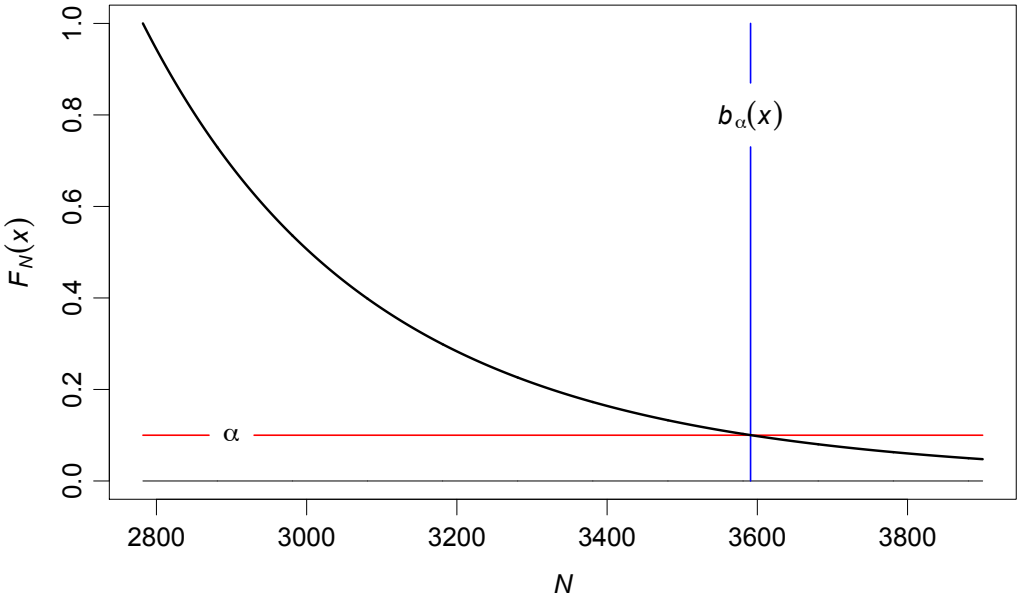

Figure 1.3: Construction of the upper confidence bound $b_{0.1}(2782)$ with $n = 9$.

x	500	1000	1500	2000	2500	3000	3500
$b_{0.5}(x)$	539	1079	1619	2159	2699	3239	3779
$b_{0.1}(x)$	644	1290	1936	2581	3227	3873	4519
$b_{0.05}(x)$	695	1393	2090	2788	3485	4183	4880
$b_{0.01}(x)$	831	1665	2499	3333	4167	5001	5835

Table 1.1: Some values of the upper confidence bound $b_\alpha(x)$ with $n = 9$.

Example 1.9 (Matriculations 2005/2006). Figure 1.3 shows for $n = 9$ and $X = 2782$ the values $F_N(X)$ as a function of $N \geq 2782$ and the resulting upper 90%-confidence bound $b_{0.1}(2782) = 3591$. Of course the latter is hard to detect by eye, but indeed $F_{3591}(X) > 0.1 > F_{3592}(X)$. Thus we claim with confidence 90% that in the academic year 2005/2006 at most 3591 students started studying in Bern.

Table 1.1 shows for $n = 9$ and $\alpha \in \{0.5, 0.1, 0.05, 0.01\}$ some values of the upper bound $b_\alpha(x)$.

Analogously one can compute a *lower* confidence bound for N. It follows from Lemma 1.4 that

$$\mathbb{P}_N(F_N(X-1) \geq 1-\alpha) \;\leq\; \alpha.$$

In other words, with probability $1-\alpha$ the unknown true parameter N satisfies the inequality $F_N(X-1) < 1-\alpha$ which is equivalent to $[X-1]_n/[N]_n < 1-\alpha$ and to $N \geq a_\alpha(X)$, where

$$\begin{aligned} a_\alpha(x) \;&:=\; \min\bigl\{N \geq n : F_N(x-1) < 1-\alpha\bigr\} \\ &=\; \min\bigl\{N \geq x : [N]_n > [x-1]_n/(1-\alpha)\bigr\} \end{aligned}$$

for integers $x \geq n$. Thus we obtain a *lower* $(1-\alpha)$*-confidence bound* $a_\alpha(X)$ for N. That means,

$$\mathbb{P}_N(N \geq a_\alpha(X)) \;\geq\; 1-\alpha$$

for any value of $N \geq n$.

Finally one can combine upper and lower bound as follows: For any value of $N \geq n$,

$$\mathbb{P}_N\bigl(a_{\alpha/2}(X) \leq N \leq b_{\alpha/2}(X)\bigr) \;\geq\; 1-\alpha.$$

This yields a $(1-\alpha)$-*confidence interval* $\big[a_{\alpha/2}(X), b_{\alpha/2}(X)\big]$ for N.

Whether a lower bound, an upper bound or an interval is of interest depends on the specific application and should be clarified before analysing the data.

Remark 1.10. The problem described in this section is known in the literature as the "taxi problem": A visitor arrives at the train station of a city and sees n taxis with licence numbers $\omega_1, \omega_2, \ldots, \omega_n$. The question is how many taxis are in use in this city. The procedures described here have been used, for instance, by the allied forces in the second world war to draw conclusions about the number of german tanks. This is described in the monograph of G. E. Noether[3] (1971).

Capture-Recapture Experiments

In ecology the size of a population is sometimes estimated by means of a capture-recapture experiment. The latter are also used in epidemiology, medicine and social sciences. In the simplest case one conducts a two-stage experiment:

Step 1 (Capture): One draws a first random sample of size ℓ (without replacement) from the population, marks these individuals and releases them.

Step 2 (Recapture): One draws a second random sample of size n (without replacement) and determines the number

$$X \;:=\; \text{number of marked individuals in the second sample.}$$

We assume tacitly that $N \geq \max(\ell, n)$. Big values of X indicate a small population size N, smaller values a larger one. A possible estimator for N is given by

$$\widehat{N} \;:=\; \frac{\ell n}{X}$$

(or $\widehat{N} := \ell n/(X+1)$, to avoid division by zero). The idea behind that estimator is as follows: After step 1, the relative fraction of marked individuals in the whole population is ℓ/N. Within the second sample the relative fraction of marked individuals is X/n. Assuming that these two fractions are similar, one may guess that N is approximately $\ell n/X$.

One can easily verify that the random variable X is hypergeometrically distributed with parameters N, ℓ and n. To compute confidence bounds for N we need a monotonicity property which is proven in Exercise 1.8: For fixed $x \in \mathbb{N}_0$, the value $F_{N,\ell,n}(x)$ is non-decreasing in N. Note that the inequality $F_{N,\ell,n}(X) > \alpha$ holds true with probability at least $1-\alpha$ according to Lemma 1.4. But this inequality is equivalent to N being greater than or equal to the *lower* $(1-\alpha)$-*confidence bound* $a_\alpha(X)$, where

$$a_\alpha(x) \;:=\; \min\big\{N \geq \max(\ell, n) : F_{N,\ell,n}(x) > \alpha\big\}$$

for $x \in \{0, 1, \ldots, \min(\ell, n)\}$.

Alternatively one could start from the inequality $F_{N,\ell,n}(X-1) < 1-\alpha$ which holds true with probability $1-\alpha$, too. This leads to the *upper* $(1-\alpha)$-*confidence bound* $b_\alpha(X)$ for N, where

$$b_\alpha(x) \;:=\; \sup\big\{N \geq \max(\ell, n) : F_{N,\ell,n}(x-1) < 1-\alpha\big\}$$

for $x \in \{0, 1, \ldots, \min(\ell, n)\}$. In case of $x = 0$ we just get $b_\alpha(x) = \infty$ because $F_{N,\ell,n}(-1) = 0$ for arbitrary $N \geq \max(n, \ell)$. In case of $x > 0$, however, $b_\alpha(x) < \infty$, see Exercise 1.9.

Again one should determine *before* the data analysis whether a lower bound, an upper bound or a combination of both is of interest.

[3] Gottfried E. Noether (1915-1991): Statistician and educational scientist; born in Germany and emigrated to the USA in 1939.

Example 1.11. Suppose that $\ell = n = 20$, and we want to determine a lower 95%-confidence bound for the population size N. Suppose that the experiment yielded the value $X = 2$. Now we have to find out for which potential parameters N the value $F_{N,20,20}(2)$ is suspiciously small, i.e. smaller than or equal to 5%. Here are some explicit numbers:

N	75	76	77	78	79	80	81	82
$F_{N,20,20}(2)$	.0417	.0455	.0495	.0537	.0580	.0625	.0671	.0719

This table shows that the lower 95%-confidence bound $a_{0.05}(2)$ equals 78. Thus one may claim with confidence 95% that $N \geq 78$. An absolutely correct lower bound would be $\ell + n - X = 38$.

1.4 Main Types of Statistical Procedures

Statistical methods fall into the following categories:

Descriptive statistics: The task is to describe or summarise the raw data quantitatively or graphically.

Inductive statistics: From empirical or experimental data one wants to draw conclusions about underlying phenomena, despite incomplete information. To this end, the data are considered as random objects and analyzed with tools from probability theory.

While many laymen associate "statistics" with big tables or colourful graphics, inductive statistics is more important and demanding. Our focus is primarily in inductive methods, although some graphical procedures will be covered as well. Our starting point are *(raw) data* $\omega \in \Omega$ which is considered as random. That means, we consider a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ with a σ-field $\mathcal{A}$ over Ω and an unknown probability distribution $\mathbb{P}$ on $\mathcal{A}$. Readers who are not familiar with measure theory should just think about a countable set Ω and a discrete probability distribution $\mathbb{P}$ on Ω.

Usually, we make certain assumptions about the probability measure $\mathbb{P}$, and often it depends on a certain unknown *parameter* θ in a given *parameter space* Θ. This is indicated by a subscript θ, writing $\mathbb{P}_\theta$ instead of $\mathbb{P}$.

The three most important types of statistical procedures are **(point) estimators**, **confidence regions** and **(statistical) tests**. Out of these, confidence regions are particularly relevant and useful. Two other types of procedures, *predictors* and *prediction regions*, are of interest in time series analyses.

(Point) Estimators

Suppose one is interested in a real or arbitrary quantity $g(\theta) \in \mathbb{G}$ of the unknown parameter θ, where $\mathbb{G}$ and $g : \Theta \to \mathbb{G}$ are given. A *(point) estimator of* $g(\theta)$ is a mapping[4]

$$\widehat{g} : \Omega \to \mathbb{G}.$$

For an arbitrary data set $\omega \in \Omega$ this defines an estimate $\widehat{g}(\omega)$ of $g(\theta)$; see Figure 1.4.

Point estimators are evaluated by their precision. The goal is to construct an estimator such that $\widehat{g}$ is "as close as possible" to the unknown value $g(\theta)$. In this context there are some concepts some of which we saw already. For simplicity we consider only the case $\mathbb{G} = \mathbb{R}$, i.e. real-valued quantities $g(\theta)$.

[4]Strictly speaking, $(\mathbb{G}, \mathcal{B})$ is a measurable space, and $\widehat{g}$ is a $\mathcal{A}$-$\mathcal{B}$-measurable mapping.

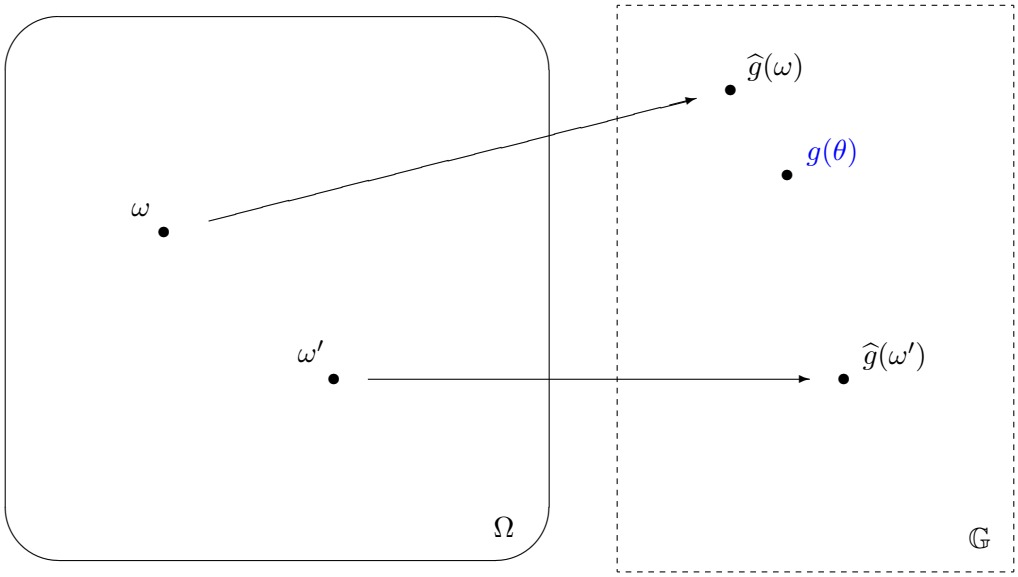

Figure 1.4: A point estimator $\widehat{g}$ of $g(\theta)$.

Bias and unbiasedness. The *bias* of an estimator is its systematic error which depends typically on θ:

$$\mathrm{Bias}_\theta(\widehat{g}) \;:=\; \mathbb{E}_\theta(\widehat{g}) - g(\theta).$$

An estimator $\widehat{g}$ of $g(\theta)$ is called *unbiased* if $\mathbb{E}_\theta(\widehat{g}) = g(\theta)$ for arbitrary parameters $\theta \in \Theta$, that means,

$$\mathrm{Bias}_\theta(\widehat{g}) \;=\; 0 \quad \text{for any } \theta \in \Theta.$$

(Recall that if $\mathbb{P}_\theta$ is a discrete distribution,

$$\mathbb{E}_\theta(\widehat{g}) \;=\; \sum_{\omega\in\Omega} P_\theta(\{\omega\})\,\widehat{g}(\omega) \;=\; \sum_{g\in\mathbb{R}} \mathbb{P}_\theta(\widehat{g}=g)\,g.$$

The second formula is true whenever the distribution of $\widehat{g}$ under $\mathbb{P}_\theta$ is discrete.)

Mean squared error. A common measure for the imprecision of a point estimator is its *mean squared error*,

$$\mathrm{MSE}_\theta(\widehat{g}) \;:=\; \mathbb{E}_\theta\big((\widehat{g} - g(\theta))^2\big),$$

or its *root mean squared error*,

$$\mathrm{RMSE}_\theta(\widehat{g}) \;:=\; \sqrt{\mathrm{MSE}_\theta(\widehat{g})}.$$

It follows from the well-known formula $\mathbb{E}(Y^2) = \mathrm{Var}(Y) + \mathbb{E}(Y)^2$ that

$$\mathrm{MSE}_\theta(\widehat{g}) \;=\; \mathrm{Var}_\theta(\widehat{g}) + \mathrm{Bias}_\theta(\widehat{g})^2.$$

Thus the mean squared error equals the sum of the variance, describing random fluctuations of $\widehat{g}$ around its mean, and the squared bias, describing the systematic error of $\widehat{g}$. For an unbiased estimator we get the simple identity $\mathrm{MSE}_\theta(\widehat{g}) = \mathrm{Var}_\theta(\widehat{g})$.

(If $\mathbb{P}_\theta$ is a discrete distribution,

$$\mathrm{MSE}_\theta(\widehat{g}) \;=\; \sum_{\omega\in\Omega} P_\theta(\{\omega\})\big(\widehat{g}(\omega) - g(\theta)\big)^2 \;=\; \sum_{g\in\mathbb{R}} \mathbb{P}_\theta(\widehat{g}=g)(g - g(\theta))^2,$$

and the latter formula is true whenever the distribution of $\widehat{g}$ under $\mathbb{P}_\theta$ is discrete.)

Example 1.12 (Estimating a population size, I). As in the first part of Section 1.3 we consider a sample $\omega = (\omega_1, \omega_2, \ldots, \omega_n)$ of n different numbers in $\{1, 2, \ldots, N\}$, where the population size N is unknown. Here $\Omega = \big\{\omega \in \mathbb{N}^n : \omega_i \neq \omega_j \text{ whenever } i \neq j\big\}$, and $\theta = N$ lies in the parameter space $\Theta = \{n, n+1, n+2, \ldots\}$. Moreover, $\mathbb{P}_N$ is the uniform distribution on the subset $\Omega_N = \big\{\omega \in \Omega : \omega_1, \ldots, \omega_n \leq N\big\}$ of Ω. This corresponds to our assumption that ω was drawn completely at random.

Now we are interested in $g(N) := N$ and consider the maximal entry $X(\omega)$ of ω. A potential point estimator of N would be X itself. But according to Lemma 1.7,

$$\mathrm{Bias}_N(X) \;=\; \mathbb{E}_N(X) - N \;=\; \frac{n\,(N+1)}{n+1} - N \;=\; \frac{n-N}{n+1}.$$

Exercise 1.7 yields the expression

$$\mathrm{Var}_N(X) \;=\; \frac{n(N+1)(N-n)}{(n+1)^2(n+2)},$$

and after a few manipulations we obtain the mean squared error

$$\mathrm{MSE}_N(X) \;=\; \mathrm{Var}_N(X) + \mathrm{Bias}_N(X)^2 \;=\; \frac{(2N-n)(N-n)}{(n+1)(n+2)}.$$

An alternative to X is the unbiased estimator $\widehat{N} := (1 + 1/n)X - 1$. This estimator satisfies the equation

$$\mathrm{MSE}_N(\widehat{N}) \;=\; \mathrm{Var}_N(\widehat{N}) \;=\; \frac{(n+1)^2}{n^2}\,\mathrm{Var}_N(X) \;=\; \frac{(N+1)(N-n)}{n(n+2)}.$$

From this one can deduce that $\mathrm{MSE}_N(\widehat{N}) < \mathrm{MSE}_N(X)$ if and only if $(n-1)N > n^2 + n + 1$. Hence the estimator $\widehat{N}$ is superior to the naive estimator X in terms of mean squared error if the sample size n is larger than 1 and the true population size N is sufficiently large.

Example 1.13 (Estimating a population size, II). Suppose the individuals of a population carry the reference numbers $a+1, a+2, \ldots, b$, where a and b are unknown integers. If we draw a random sample from this population without replacement, it corresponds to a tuple $\omega = (\omega_1, \omega_2, \ldots, \omega_n)$ in the sample space $\Omega = \big\{\omega \in \mathbb{Z}^n : \omega_i \neq \omega_j \text{ whenever } i \neq j\big\}$, and the unknown parameter $\theta = (a, b)$ lies in the parameter space $\Theta = \big\{(a, b) : a, b \in \mathbb{Z}, b - a \geq n\big\}$.

A specific example would be the matriculation numbers if one considers students having started in Bern, not knowing that the matriculation numbers $(Z_1Z_2Z_2, Z_4, Z_5)$ start at $10'000$, i.e. $a = 9'999$.

Assuming again that the sample ω has been drawn completely at random, $\mathbb{P}_{(a,b)}$ is the uniform distribution on the set $\Omega_{(a,b)} = \big\{\omega \in \Omega : a < \omega_1, \ldots, \omega_n \leq b\big\}$.

Suppose we are still interested in the parameter $N = b - a = g(a, b)$. To estimate or bound it, one could consider the statistic

$$X(\omega) \;:=\; \max(\omega_1, \ldots, \omega_n) - \min(\omega_1, \ldots, \omega_n).$$

Here we have to assume that $n \geq 2$. The distribution of X depends only on N, because it remains the same if we replace ω with $(\omega_1 - a, \omega_2 - a, \ldots, \omega_n - a)$ which is uniformly distributed on $\Omega_{(0,N)}$. With the considerations in Exercise 1.6 one can show that

$$\widehat{N} \;:=\; \frac{(n+1)X}{n-1} - 1$$

defines an unbiased estimator of N.

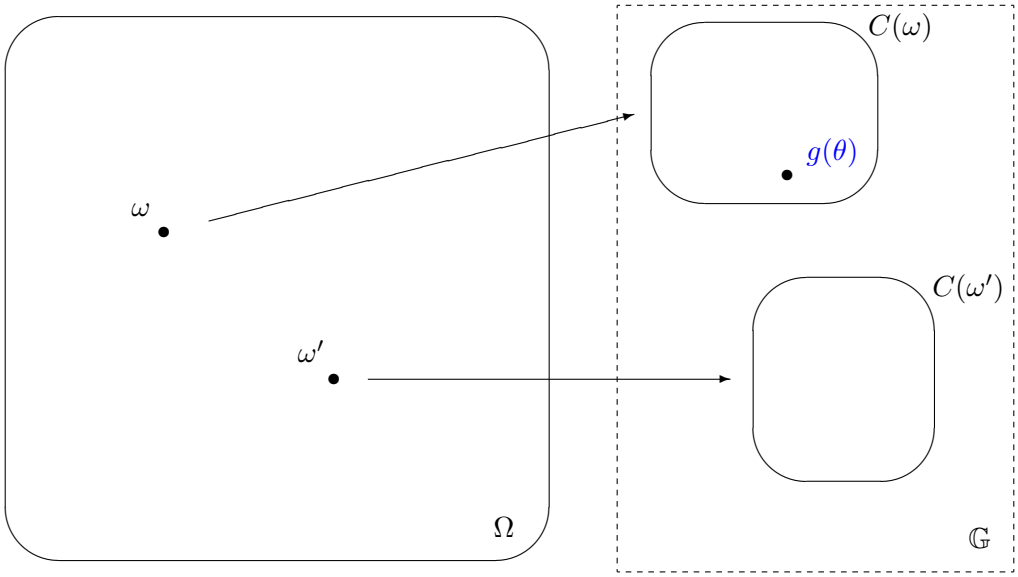

Figure 1.5: A confidence region C for $g(\theta)$.

Example 1.14 (Estimating a population size, III). In connection with capture-recapture experiments we consider a population $\mathcal{M} \subset \mathcal{M}_{\text{total}}$ of N individuals and the set Ω of all pairs $\omega = (\omega^{(1)}, \omega^{(2)})$ of two samples $\omega^{(1)} = (\omega_1^{(1)}, \dots, \omega_\ell^{(1)})$ and $\omega^{(2)} = (\omega_1^{(2)}, \dots, \omega_n^{(2)})$ from $\mathcal{M}_{\text{total}}$, each of them with pairwise different elements. Here $\mathbb{P}_{\mathcal{M}}$ is the uniform distribution on the subset of all $\omega \in \Omega$ with $\omega_i^{(1)}, \omega_j^{(2)} \in \mathcal{M}$ for all $i \le \ell$ and $j \le n$.

For instance, $\mathcal{M}$ could be the population of pigeons that are visiting a particular park in Bern regularly, and $\mathcal{M}_{\text{total}}$ is, say, the set of all pigeons worldwide.

The superpopulation $\mathcal{M}_{\text{total}}$ allows us to define a common sample space Ω, and the parameter space Θ consists of (all) subsets $\mathcal{M}$ of $\mathcal{M}_{\text{total}}$ with at least $\max(\ell, n)$ elements. Suppose we are mainly interested in the value $g(\mathcal{M}) = N = \#\mathcal{M}$. If $\omega \sim \mathbb{P}_{\mathcal{M}}$, then the random variable

$$X(\omega) := \#\big(\{\omega_1^{(1)}, \dots, \omega_\ell^{(1)}\} \cap \{\omega_1^{(2)}, \dots, \omega_n^{(2)}\}\big)$$

has distribution $\text{Hyp}(N, \ell, n)$ if $\omega \sim \mathbb{P}_{\mathcal{M}}$. In this setting, there exists *no* unbiased estimator of N. Precisely, as shown in Exercise 1.11, there exists no function $h : \{0, 1, \dots, \min(\ell, n)\} \to \mathbb{R}$ such that $\widehat{g} = h(X)$ is an unbiased estimator of N.

Confidence Regions

Instead of a point $\widehat{g}(\omega) \in \mathbb{G}$ one specifies a subset $C(\omega) \subset \mathbb{G}$, claiming or hoping that it contains the point $g(\theta)$. The corresponding mapping

$$C : \Omega \to \mathcal{P}(\mathbb{G})$$

is called a *confidence region for* $g(\theta)$; see Figure 1.5.

If one can guarantee that for a given $\alpha \in (0, 1)$,

$$\mathbb{P}_\theta\big(g(\theta) \in C\big) \ge 1 - \alpha \quad \text{for arbitrary } \theta \in \Theta,$$

then C is called a confidence region for $g(\theta)$with confidence level $1 - \alpha$ or briefly a $(1 - \alpha)$-confidence region for $g(\theta)$. The probability on the left hand side involves the fixed point $g(\theta)$ and the random set C. Written at full length, it stands for $\mathbb{P}_\theta\big(\{\omega \in \Omega : g(\theta) \in C(\omega)\}\big)$. [5]

[5] We assume tacitly that $\{\omega \in \Omega : g(\theta) \in C(\omega)\}$ belongs to the σ-field $\mathcal{A}$ for arbitrary $\theta \in \Theta$.

A special recipe. In the context of population sizes we utilised a method which will be used in later sections again. Let $X : \Omega \to \mathbb{R}$ be a real-valued statistic, and let F_θ be its distribution function, i.e. $F_\theta(x) := \mathbb{P}_\theta(X \le x)$ for $x \in \mathbb{R}$. Suppose we know F_θ for each *hypothetical* value $\theta \in \Theta$. According to Lemma 1.4, the *true* parameter θ satisfies the inequality $F_\theta(X) > \alpha$ with probability $1 - \alpha$. Consequently, if we set

$$\tilde{C}(x) \;:=\; \big\{g(\theta) : \theta \in \Theta, F_\theta(x) > \alpha\big\}$$

for $x \in \mathbb{R}$, then $\omega \mapsto \tilde{C}(X(\omega))$ defines a $(1-\alpha)$-confidence region for $g(\theta)$. In this construction we exclude all *hypothetical* parameters $\theta \in \Theta$ such that the value $X(\omega)$ is suspiciously small for F_θ.

Analogously one could exclude all hypothetical parameters in Θ such that the value $X(\omega)$ is suspiciously large. This leads to the $(1-\alpha)$-confidence region $\omega \mapsto \tilde{C}(X(\omega))$ with

$$\tilde{C}(x) \;:=\; \big\{g(\theta) : \theta \in \Theta, F_\theta(x-) < 1 - \alpha\big\}$$

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

Finally one could combine both approaches and exclude all hypothetical parameters in Θ such that the value $X(\omega)$ is suspiciously extreme. This leads to the $(1-\alpha)$-confidence region $\omega \mapsto \tilde{C}(X(\omega))$ with

$$\tilde{C}(x) \;:=\; \big\{g(\theta) : \theta \in \Theta, F_\theta(x) > \alpha/2 \text{ and } F_\theta(x-) < 1 - \alpha/2\big\}$$

for $x \in \mathbb{R}$. One could split the error probability α differently and define

$$\tilde{C}(x) \;:=\; \big\{g(\theta) : \theta \in \Theta, F_\theta(x) > \alpha_1 \text{ and } F_\theta(x-) < 1 - \alpha_2\big\}$$

with given numbers $\alpha_1, \alpha_2 > 0$ such that $\alpha_1 + \alpha_2 = \alpha$.

In all three cases we reduce the raw data ω to the value $X(\omega)$ and then determine the set of all hypothetical parameters $\theta \in \Theta$ which are plausible for $X(\omega)$. Whether the resulting confidence regions are really useful and which shape they have depends on the specific situation.

(Statistical) Tests

By means of the data $\omega \in \Omega$ one wants to verify a certain "effect" (*working hypothesis*, *alternative hypothesis*). For this purpose one formulates a *null hypothesis*. That means, one describes the distribution of the data under the assumption that the effect in question is *not* present. Then one specifies for which data one rejects this null hypothesis. That means, one divides Ω into an *acceptance region* Ω_o and a *rejection region* (also called *critical region*) $\Omega_1 = \Omega \setminus \Omega_o$.[6] In case of $\omega \in \Omega_o$, no conclusion is drawn; the null hypothesis is considered as possibly true. In case of $\omega \in \Omega_1$ one claims that the null hypothesis is wrong and that the working hypothesis is true or at least plausible.

This procedure is a so-called statistical test. When applying it one has to be aware of two types of potential errors:

Error of the first type: The null hypothesis is true, but we reject it because $\omega \in \Omega_1$.

Error of the second type: The working hypothesis is true, but we don't reject the null hypothesis because $\omega \in \Omega_o$.

Since these two types of errors cannot be avoided simultaneously, we focus on the probability of an error of the first type. Indeed, often the null hypothesis is easier to describe and handle than the alternative hypothesis. If we can guarantee that for a given test level $\alpha \in (0,1)$,

$$\mathbb{P}(\Omega_1) \;\le\; \alpha \quad \text{under the null hypothesis,}$$

[6] The sets Ω_o, Ω_1 should belong to the σ-field $\mathcal{A}$.

then our test is called a *test with level* α. Under this condition, if $\omega \in \Omega_1$ one may claim with *confidence* $1-\alpha$ that the null hypothesis is wrong. In other terms, in case of $\omega \in \Omega_1$ we reject the null hypothesis at test level α.

Example 1.15 (Fisher's exact test for randomised studies). In Example 1.3 the goal was to verify, if possible, the working hypothesis that treatment 1 is more effective than treatment 2. The null hypothesis is that there is no difference between the two treatments. Now let Ω be the set of all two-by-two tables which could result from a randomised study:

h_1	$n_1 - h_1$	n_1
h_2	$n_2 - h_2$	n_2
$h_+ = h_1 + h_2$	$N - h_+$	N

The distribution $\mathbb{P}$ takes into account the recruitment of test persons, their assignment to the two treatment groups, and all factors having an impact on the success or non-success of the treatments. Typically we don't know $\mathbb{P}$ completely. But we assume that under the null hypothesis the conditional distribution of the upper left entry h_1, given the group sizes and the total number of successes, equals $\mathrm{Hyp}(N, h_+, n_1)$.

The critical region Ω_1 consists of all two-by-two tables in which h_1 is suspiciously large in the sense that it exceeds the critical value $q_{1-\alpha;N,h_+,n_1}$. This is equivalent to the right-sided p-value $1 - F_{N,h_+,n_1}(h_1 - 1)$ being less than or equal to α.

Before describing another example of a statistical test, let us recall the definition of binomial distributions:

Definition 1.16 (Binomial distributions). A random variable X is binomially distributed with parameters $n \in \mathbb{N}$ and $p \in [0,1]$, if for arbitrary $x \in \{0, 1, \ldots, n\}$,

$$\mathbb{P}(X = x) \;=\; f_{n,p}(x) := \binom{n}{x} p^x (1-p)^{n-x}.$$

We denote this distribution with $\mathrm{Bin}(n,p)$. The corresponding distribution function is denoted by $F_{n,p}$, i.e. $F_{n,p}(x) = \sum_{k=0}^{x} f_{n,p}(k)$ for $x \in \{0, 1, \ldots, n\}$.

The binomial distribution $\mathrm{Bin}(n,p)$ describes the distribution of a random sum $X = \sum_{i=1}^{n} X_i$, where the summands $X_1, X_2, \ldots, X_n$ are stochastically independent with $\mathbb{P}(X_i = 1) = p$ and $\mathbb{P}(X_i = 0) = 1 - p$.

Example 1.17 (A binomial test of randomness). Before carrying on with reading, the reader should write down a "completely random" sequence of 50 digits in $\{0, 1\}$.

If one asks people to write down a random sequence $\omega = (\omega_1, \omega_2, \ldots, \omega_n)$ of n digits $\omega_i \in \{0, 1\}$, they often produce sequences with too many changes. To quantify this effect we define the test statistic

$$X(\omega) \;:=\; \#\{i < n : \omega_i \neq \omega_{i+1}\}.$$

Under the null hypothesis the sequence has been chosen completely at random from $\{0,1\}^n$. One can easily verify that under this null hypothesis, X has distribution $\mathrm{Bin}(n-1, 0.5)$. To judge whether the observed value of X is suspiciously large, we compute the right-sided p-value $1 - F_{n-1,0.5}(X-1)$. It follows from Lemma 1.4 that

$$\mathbb{P}\big(1 - F_{n-1,0.5}(X-1) \leq \alpha\big) \;\leq\; \alpha \quad \text{under the null hypothesis.}$$

Hence if this p-value is less than or equal to α, we may claim with confidence $1-\alpha$ that the sequence hasn't been chosen completely at random.

Here the sample space equals $\Omega = \{0,1\}^n$, and the critical region is given by

$$\begin{aligned}\Omega_1 &= \big\{\omega \in \{0,1\}^n : 1 - F_{n-1,0.5}(X(\omega) - 1) \le \alpha\big\} \\ &= \big\{\omega \in \{0,1\}^n : X(\omega) > q_{1-\alpha;n-1,0.5}\big\}\end{aligned}$$

with the critical value

$$q_{1-\alpha;n-1,0.5} := \min\big\{x \in \{0,1,\ldots,n-1\} : F_{n-1,0.5}(x) \ge 1-\alpha\big\}.$$

Numerical example. If $n = 50$ and $\alpha = 0.05$, the critical value is $q_{1-\alpha;n-1,0.5} = q_{0.95;49,0.5} = 30$ because $F_{49,0.5}(29) \approx 0.9238$ and $F_{49,0.5}(30) \approx 0.9573$. Hence if a sequence ω has more than 30 changes, we claim with confidence 95% that it hasn't been chosen completely at random.

1.5 Data Sets and Variables

In the previous sections we saw already some important procedures and ideas. In the subsequent chapters we shall introduce and discuss numerous additional procedures and concepts. This material will be arranged in terms of data and variable types.

Data sets. A *data set (sample)* consists of *observations (cases)*. For each observation there are values of one or several *variables (features)*. The number of observations is called *sample size*.

Example 1.18 (A poll among students). In the first-year course "Introduction to statistics for economics and social sciences (Bern 2003/2004)", 263 students filled out a form. Each student corresponds to one observation. The students were asked to provide values of the following variables:

(1) Gender: `female` oder `male`
(2) Age: in years
(3) Month of birth: a number from $\{1,2,\ldots,12\}$
(4) Origin: Canton or country of birth
(5-6) Body height and weight : in `cm` and `kg`, respectively
(7) Monthly rent: Net rent in `CHF`
(8) Smoking: never = `0`, occasionally = `1`, regularly = `2`
(9) Random digit: a digit in {0,1,...,9}, chosen "completely at random"
(10) Number of siblings: a number in $\{0,1,2,\ldots\}$
(11) Estimated body height of lecturer: in `cm`

One distinguishes two or three main types of variables:

Categorical (qualitative) variables. These variables take values in an finite set.

In Example 1.18 the following variables are categorical: 'Gender', 'Month of birth', 'Origin', 'Smoking', 'Random digit'.

If a variable has precisely two possible values, e.g. 'Gender', it is called *dichotomous* or *binary*.

Numerical (quantitative) variables. These variables contain numbers with an objective (physical, economical, ...) meaning.

In Example 1.18 the following variables are numerical: 'Age', 'Body height' and 'Body weight', 'Monthly rent', 'Number of siblings', 'Estimated body height of lecturer'. The variables 'Month of birth' and 'Smoking' are also coded by numbers, but these numbers have been chosen somewhat arbitrary. (In the author's opinion, the variable 'Random digit' is categorical.)

Ordinal variables. These are categorical variables whose potential values stand in a natural order with a smallest and a largest value.

Such variables are quite common in medicine, psychology and social sciences. One example are questions about satisfaction with something with possible answers 'dissatisfied', 'partly dissatisfied', 'mostly satisfied' and 'completely satisfied'. Also grade in exams may be viewed as ordinal variables. Sometimes ordinal variables are deduced from numerical variables by "binning", i.e. by partitioning the range of a numerical variable into finitely many intervals and just recording the interval the raw variable falls into.

In Example 1.18 the variable 'Smoking' is ordinal, 0 (never) $\leq$ 1 (occasionally) $\leq$ 2 (regularly). At first glance one might think that 'Month of birth' and 'Random digit' are ordinal, too. But after the month of December follows January. also the random digits may be imagined as sectors on a small Roulette wheel. Hence these two variables are 'cyclical' rather than ordinal.

Data matrices. Typically a data set is stored as a data matrix, that is a table with rows corresponding to observations and columns corresponding to variables. Often there is an additional first row containing the variables' names.

1.6 Exercises

Exercise 1.1. In Example 1.1 one could carry out a multiple phase experiment. The base experiment with eight pieces of chocolate is repeated until for the first time $X \leq 2$ or $X = 4$. Precisely, let X_i be the result of the i-th round. In case of $X_i = 3$ the base experiment is repeated and yields a new value X_{i+1}. All in all we get a random number J of rounds, where $X_i = 3$ for $1 \leq i < J$ and $X_J \neq 3$. In case of $X_J = 4$ we would claim that the working hypothesis is true; in case of $X_J \leq 2$ we would draw no conclusion.

Suppose that the null hypothesis is true. What is the probability that we arrive at the wrong conclusion, i.e. claim that the working hypothesis is true because $X_J = 4$? Further, how many pieces of chocolate may (or need to be) tasted on average?

Exercise 1.2 (Wine-tasting gentlemen). Mr. Perfect and Mr. Good, two gourmets of wine, claim that they are able to distinguish between two vintages A and B of a particular wine. Precisely, Mr. Perfect claims that if presented with a glass of this wine, he can tell whether it is vintage A or vintage B. Mr. Good claims that if presented with two glasses of this wine, he can tell whether they are from the same vintage or not.

Describe a randomized experiment for each gentleman in which he is presented with n glasses of the given wine and has to solve a certain task. What is our confidence that his claim is correct if the task is solved? How large should n be if you aim for 95% confidence?

Exercise 1.3 (An experiment in social sciences). During a special training, 48 future managers participated in an experiment without knowing its true purpose. Each of them got the resume of a fictitious employee, and each manager had to decide whether this employee is promoted or not to a higher job level. the 48 resumes were identical except for the candidate's name: In 24 cases the name was Mr. Miller, in 24 cases it was Mrs. Miller. The 48 resumes have been distributed among the 48 managers completely at random. Here is a two-by-two table of the managers' decisions:

	promotion	no promotion	
Mr. Miller	21	3	24
Mrs. Miller	14	10	24
	35	13	48

The question is whether these data confirm the general impression of men being being promoted more easily than women. Analyze the data as in Example 1.3 with test level $\alpha = 5\%$ aus. To this end you may use the following table of the hypergeometric distribution $\mathrm{Hyp}_{48,35,24}$. It contains the weights $f_{48,35,24}(x)$ (rounded to four decimal digits):

x	11	12	13	14	15	16	17
$f_{48,35,24}(x)$	0.0000	0.0003	0.0036	0.0206	0.0720	0.1620	0.2415
x	18	19	20	21	22	23	24
$f_{48,35,24}(x)$	0.2415	0.1620	0.0720	0.0206	0.0036	0.0003	0.0000

Here one should think carefully about the null hypothesis being tested. One could consider the 48 future managers as a random sample from some population, and one wants to draw a conclusion about the latter. But this is possibly far-fetched. Alternatively one could focus on these specific 48 persons and consider the null hypothesis that all of them judged the fictitious person regardless of gender. The working hypothesis would be that among these 48 persons some people would prefer men more easily than women.

Exercise 1.4. Let X be a random variable with the following distribution:

x	-1	0	1	2	3	4
$\mathbb{P}(X = x)$	0.05	0.10	0.20	0.25	0.25	0.15

Draw

(a) the distribution function F_o of X, i.e. $F_o(x) := \mathbb{P}(X \le x)$ for $x \in \mathbb{R}$.

(b) the three functions

$$\alpha \;\mapsto\; \begin{cases} \mathbb{P}\big(F_o(X) \le \alpha\big) \\ \mathbb{P}\big(1 - F_o(X-) \le \alpha\big) \\ \mathbb{P}\big(2 \cdot \min\{F_o(X), 1 - F_o(X-)\} \le \alpha\big). \end{cases}$$

for $\alpha \in [-0.1, 1.4]$.

Exercise 1.5. Suppose that X is a random variable with values in a countable set $\mathcal{X}$, and let P_o be a probability distribution on $\mathcal{X}$ with probability mass function f_o, that is, $f_o(x) = P_o(\{x\})$. Show that

$$\pi(X) \;:=\; \sum_{x \in \mathcal{X}:\; f_o(x) \le f_o(X)} f_o(x)$$

defines a p-value for the null hypothesis that $X \sim P_o$. That means, if $X \sim P_o$, then

$$\mathbb{P}(\pi(X) \le \alpha) \;\le\; \alpha \quad \text{for any } \alpha \in (0,1).$$

Exercise 1.6. In this exercise we prove Lemma 1.7 by means of a symmetry consideration. We consider the uniform distribution $\mathbb{P}_N$ on the set $\Omega_N = \big\{\omega \in \Omega : X(\omega) \le N\big\}$. For a tuple $\omega \in \Omega_N$ let $1 \le \omega_{(1)} < \omega_{(2)} < \cdots < \omega_{(n)} \le N$ be its ordered components; in particular, $\omega_{(n)} = X(\omega)$. With $\omega_{(0)} := 0$ and $\omega_{(n+1)} := N + 1$ we define the random vector $\boldsymbol{Z} = (Z_i)_{i=1}^{n+1}$ with entries $Z_i(\omega) := \omega_{(i)} - \omega_{(i-1)}$. That means, we partition the numbers from 1 through $N+1$ into $n+1$ random intervals:

$$0, \underbrace{\ldots\ldots, \omega_{(1)}}_{Z_1(\omega) \text{ elem.}}, \underbrace{\ldots\ldots, \omega_{(2)}}_{Z_2(\omega) \text{ elem.}}, \;\ldots\; \ldots, \omega_{(n-1)}, \underbrace{\ldots\ldots, \omega_{(n)}}_{Z_n(\omega) \text{ elem.}}, \underbrace{\ldots\ldots, N+1}_{Z_{n+1}(\omega) \text{ elem.}}.$$

(a) Show that $\boldsymbol{Z}$ is uniformly distributed on the set

$$\mathcal{Z}_N \;:=\; \Big\{\boldsymbol{z} \in \mathbb{N}^{n+1} : \sum_{i=1}^{n+1} z_i = N+1\Big\}.$$

(One has to show that for each $\boldsymbol{z} \in \mathcal{Z}_N$, the set $\big\{\omega \in \Omega_N : \boldsymbol{Z}(\omega) = \boldsymbol{z}\big\}$ has the same number of elements.)

(b) Show that the random variables $Z_1, Z_2, \ldots, Z_{n+1}$ are identically distributed. (For this purpose one could consider, for instance, the mapping

$$(z_1, z_2, \ldots, z_{n+1}) \;\mapsto\; (z_{n+1}, z_1, z_1, \ldots, z_n)$$

from $\mathcal{Z}_N$ into $\mathcal{Z}_N$.)

(c) Determine by means of part (b) the expected values $\mathbb{E}_N(Z_i)$ and $\mathbb{E}_N(X)$.

Exercise 1.7. Consider the proof of Lemma 1.7. Starting from the general formula (1.3) it was shown that $\mathbb{E}_N(X) = (N+1)n/(n+1)$.

(a) Now determine $\mathbb{E}(X(X+1))$, $\mathbb{E}_N(X^2)$ and $\mathrm{Var}_N(X)$.

(b) Show that the standard deviation of the estimator $\widehat{N} := (n+1)X/n - 1$ satisfies the inequality

$$\mathrm{Std}(\widehat{N}) \;<\; \frac{N}{n}.$$

Exercise 1.8 (Monotonicity of $\mathrm{Hyp}(N, \ell, n)$ in N). Show that for any fixed $x \in \mathbb{N}_0$, the distribution function $F_{N,\ell,n}(x)$ of $\mathrm{Hyp}(N, \ell, n)$ is non-decreasing in N. One can even show that

$$F_{N+1,\ell,n}(x) \;=\; F_{N,\ell,n}(x) + \frac{x+1}{N+1}\, f_{N,\ell,n}(x+1).$$

Hint: These claims may be verified with tedious calculations. More elegant is a coupling argument: Design a random experiment with two random variables X and $\tilde{X}$ such that $X \sim \mathrm{Hyp}(N, \ell, n)$, $\tilde{X} \sim \mathrm{Hyp}(N+1, \ell, n)$ and $\tilde{X} \le X$. For instance, imagine an urn with ℓ blue, $N - \ell$ white and one black ball. From this urn one draws $n+1$ balls without replacement ...

Exercise 1.9 (Capture-recapture method).

(a) An absolutely certain lower bound for N is given by $\ell + n - X$, because in the first step one has marked ℓ individuals, and in the second step one has seen $n - X$ new individuals. Verify that the lower confidence bound $a_\alpha(\cdot)$ satisfies the inequality $a_\alpha(x) \ge \ell + n - x$ for arbitrary $x \in \{0, 1, \ldots, \min(\ell, n)\}$.

(b) Show that $b_\alpha(x) < \infty$ whenever $x \ge 1$.

Exercise 1.10. An ecologist is concerned that the population of bugs of a certain kind in a certain region has grown too much recently. To verify this, he performs a capture-recapture experiment with $\ell = n = 40$ bugs.

(a) Should he compute a lower or an upper confidence bound for the total number N of bugs?

(b) Suppose he finds in the second step $X = 3$ animals he has seen in the first step already. Determine the resulting 90%-confidence bound for N by means of the following table with values of $F_N(x) = F_{N,40,40}(x)$ for various values of N and $x = 2, 3$:

N	256	257	258	259	260	261	262	263	264
$F_N(3)$	.0902	.0920	.0939	.0957	.0976	.0994	.1013	.1032	.1052
N	1416	1417	1418	1419	1420	1421	1422	1423	1424
$F_N(2)$	.8996	.8998	.8999	.9001	.9002	.9004	.9006	.9007	.9009

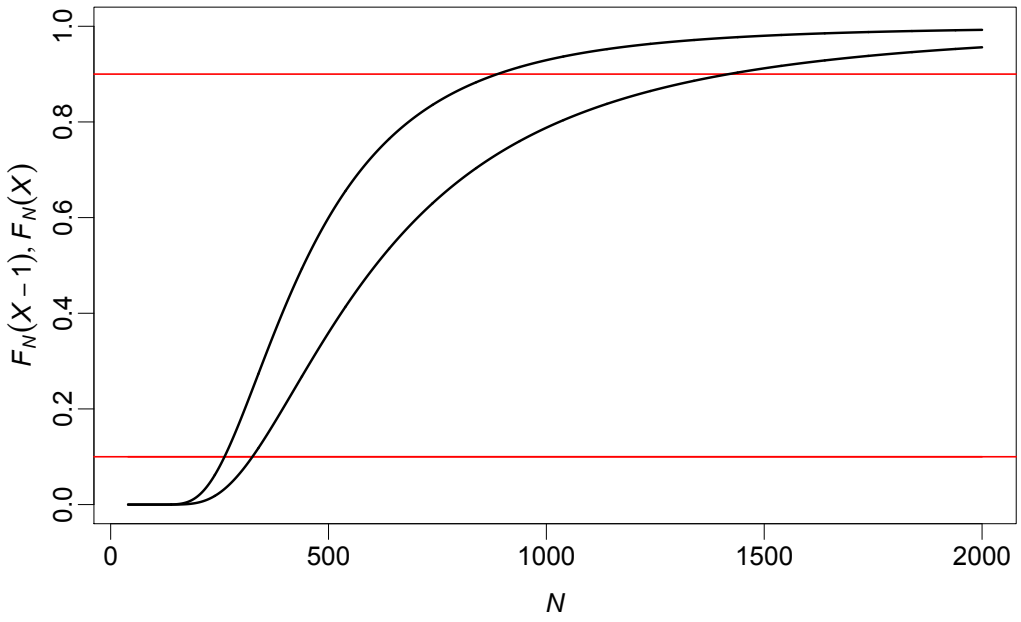

Figure 1.6: Example for capture-recapture method.

(c) Since the ecologist is not a statistician, he asks you to formulate a short and concise conclusion.

(d) Where are the lower and upper 90%-confidence bound in Figure 1.6?

Exercise 1.11 (Point estimation for capture-recapture experiments).

(a) Show that there exists an unbiased estimator of the quantity $g(N) := 1/N$. Determine also $\mathrm{RMSE}_N(\widehat{g})$.

(b) Consider an estimator of N of the form $\widehat{N} = h(X)$ with a real-valued function h on $\{0, 1, \ldots, \min(\ell, n)\}$. Explain why this estimator can never be unbiased.

(c) Determine the bias of the estimator $\widehat{N} := (\ell+1)(n+1)/(X+1)$ of N. Show that $\mathrm{Bias}_N(\widehat{N}) \leq 1$ for any $N \geq \max(\ell, n)$.

Exercise 1.12. Determine for Example 1.13 explicit formulae for $\mathbb{P}_{(0,N)}(X = x)$ and $F_N(x) := \mathbb{P}_{(0,N)}(X \leq x)$, where $N \geq n \geq 2$ and $x \in \mathbb{N}_0$.

Exercise 1.13 (First confidence bounds for a probability). A player of 'Eile mit Weile' is convinced that a certain dice returns the number 1 too rarely. To verify this, he throws the dice several times and determines the number X of throws until the number 1 appears for the first time.

(a) How could one construct confidence bounds for the unknown probability p of the number 1?
(a.1) Determine the distribution function F_p of X, that means, $F_p(x) = \mathbb{P}_p(X \leq x)$, and its monotonicity with respect to p,
(a.2) Determine explicit formulae for a lower and an upper $(1-\alpha)$-confidence bound $a_\alpha(X)$ and $b_\alpha(X)$, respectively, utilizing our general recipe.

(b) Which of the two confidence bounds is relevant for the player? For which values of X could he claim with confidence 90% that $p < 1/6$?

(c) What happens in (a.1) if one considers the number X of throws until the number 1 appears for the *second* time?

Exercise 1.14 (A biological experiment). A group of biologists wanted to verify that a special type of ants from Central America who is dwelling in Acacia trees has strong preferences concerning the choice of its home tree.

In a certain region all but 28 Acacia trees have been removed. Out of these trees, 15 were of type A and 13 of type B; none of them hosted a colony of ants so far, and all trees stood approximately in a circle. Now 16 colonies of ants who had previously dwelled in trees of type A were released at the center of that circle. After a certain time period each colony had found a new home tree:

	with ants	without ants	
Type A	13	2	15
Type B	3	10	13
	16	12	28

Formulate an appropriate working and null hypothesis. Test your null hypothesis at level $\alpha = 0.01$.

Exercise 1.15. Daniel Düsentrieb has developed a brand-new random number generator and wants to convince you of its quality. To this end he presents to you a "random" sequence $\omega \in \{0,1\}^{100}$ (to be read row by row):

1	1	0	0	1	0	0	0	0	1	0	0	0	1	0	1	0	1	1	0
1	0	1	1	1	1	0	1	1	1	1	0	0	0	0	0	0	0	1	1
1	1	1	0	0	1	1	0	0	0	1	0	0	0	0	0	0	0	0	1
0	0	1	0	0	0	0	0	1	0	0	1	0	0	0	0	1	1	0	1
0	1	1	1	1	1	1	1	1	1	0	1	1	1	0	0	0	0	0	0

Compute a two-sided p-value for the test statistic $X(\omega)$ (= number of changes in ω) and the null hypothesis that this sequence has been chosen completely at random. Test this null hypothesis at level $\alpha = 5\%$. You may use the following table of the binomial distribution function $F_{99,0.5}$ (rounded to four digits):

x	35	36	37	38	39	40	41	42
$F_{99,0.5}(x)$	0.0023	0.0043	0.0077	0.0133	0.0219	0.0350	0.0537	0.0795

Investigate also the following sequence ω:

1	0	0	1	1	1	1	0	1	1	0	1	1	0	0	1	0	1	1	0
0	1	1	0	0	1	0	1	1	0	1	0	1	1	0	0	1	0	1	0
1	1	1	0	1	0	0	1	0	1	1	1	0	1	1	0	1	0	1	0
1	1	1	1	1	0	0	1	1	0	0	0	0	1	1	1	0	1	0	0
1	0	0	1	0	0	1	0	0	0	0	1	0	1	1	0	1	0	1	1

Chapter 2

Categorical Variables

In this chapter we consider a categorical variable with $K \geq 2$ potential values $x_1, x_2, \ldots, x_K$. The values in the sample are denoted with $X_1, X_2, \ldots, X_n$. These are viewed as stochastically independent random variables where

$$\mathbb{P}(X_i = x_k) \;=\; p_k \quad \text{for } 1 \leq k \leq K$$

with certain parameters $p_1, p_2, \ldots, p_K \geq 0$. In particular, $\sum_{k=1}^{K} p_k = 1$.

Examples

- Consider the variable 'Smoking' in Example 1.18 with potential values $x_1 =$ 'never', $x_2 =$ 'occasionally' and $x_3 =$ 'regularly'. If we consider the 263 persons as a sample from the population of all people living in Switzerland and about 18–30 years old, we may assume the model above. Then p_1, p_2, p_3 are the relative proportions of non-smokers, occasional smokers and regular smokers in the whole population.
- We stay with Example 1.18 but now consider the variable 'Random digit'. Let p_k be the probability, that a randomly chosen person from the population would choose the digit $k - 1 \in \{0, 1, \ldots, 9\}$.
- Prior to an election of parliament, n potential voters are asked which of the available parties $x_1, x_2, \ldots, x_K$ they would choose. If the poll size n is substantially smaller than the total number of potential voters, we may assume the model above with p_k being the current proportion of supporters of party x_k.
- Consider a technical device fulfilling a specified task for a certain time period. It could work flawlessly (x_1) or one of $K - 1$ potential problems could occur ($x_2, \ldots, x_K$). Now n such devices are tested under these conditions, and p_k is the probability that for a single device we observe the outcome x_k.

2.1 Point Estimators and Graphical Representations

For each of the K possible outcomes we determine its (absolute) frequency

$$H_k \;:=\; \#\{i \leq n : X_i = x_k\}$$

and its relative frequency

$$\widehat{p}_k \;:=\; \frac{H_k}{n}$$

within the sample. As indicated by this notation, $\widehat{p}_k$ may be viewed as a point estimator for p_k. The next lemma summarises some properties of these random variables H_k and $\widehat{p}_k$.

Lemma 2.1 (Multinomial distribution). *The tuple $\boldsymbol{H} = (H_k)_{k=1}^K$ follows a multinomial distribution with parameters n and $\boldsymbol{p} = (p_k)_{k=1}^K$. That means, for any tuple $\boldsymbol{h} = (h_k)_{k=1}^K \in \mathbb{N}_0^K$,*

$$\mathbb{P}(\boldsymbol{H} = \boldsymbol{h}) \;=\; f_{n,\boldsymbol{p}}(\boldsymbol{h}) := \binom{n}{h_1, h_2, \ldots, h_K} \prod_{k=1}^K p_k^{h_k}$$

with the multinomial coefficient

$$\binom{n}{h_1, h_2, \ldots, h_K} \;:=\; \begin{cases} \dfrac{n!}{h_1!\, h_2! \cdots h_K!} & \text{if } h_1 + h_2 + \cdots + h_K = n, \\ 0 & \text{else.} \end{cases}$$

This distribution is denoted with $\mathrm{Mult}(n, \boldsymbol{p})$.

For $k = 1, 2, \ldots, K$, the absolute frequency H_k has distribution $\mathrm{Bin}(n, p_k)$*, and the estimators $\widehat{p}_k$ fulfill*

$$\begin{aligned} \mathbb{E}(\widehat{p}_k) \;&=\; p_k, \\ \mathrm{Var}(\widehat{p}_k) \;&=\; \frac{p_k(1-p_k)}{n} \;\le\; \frac{1}{4n}, \\ \mathrm{Cov}(\widehat{p}_k, \widehat{p}_\ell) \;&=\; \frac{-p_k p_\ell}{n} \quad \text{for } \ell \neq k. \end{aligned}$$

This lemma shows that $\widehat{p}_k$ is an unbiased estimator for p_k with estimation error of order $O(n^{-1/2})$. More precisely,

$$\mathbb{E}\big|\widehat{p}_k - p_k\big| \;\le\; \mathrm{Std}(\widehat{p}_k) \;\le\; \frac{1}{2\sqrt{n}}.$$

Proof of Lemma 2.1. We write $\boldsymbol{H} = \boldsymbol{H}(\boldsymbol{X})$ with the observation vector $\boldsymbol{X} = (X_i)_{i=1}^n$ and $\mathcal{X} := \{x_1, x_2, \ldots, x_K\}$. Then $\mathbb{P}(\boldsymbol{H} = \boldsymbol{h})$ equals

$$\begin{aligned} \sum_{\tilde{\boldsymbol{x}} \in \mathcal{X}^n \,:\, \boldsymbol{H}(\tilde{\boldsymbol{x}}) = \boldsymbol{h}} \mathbb{P}(\boldsymbol{X} = \tilde{\boldsymbol{x}}) \;&=\; \sum_{\tilde{\boldsymbol{x}} \in \mathcal{X}^n \,:\, \boldsymbol{H}(\tilde{\boldsymbol{x}}) = \boldsymbol{h}} \prod_{i=1}^n p_{\tilde{x}_i} \\ &=\; \#\big\{\tilde{\boldsymbol{x}} \in \mathcal{X}^n : \boldsymbol{H}(\tilde{\boldsymbol{x}}) = \boldsymbol{h}\big\} \prod_{k=1}^K p_k^{h_k}. \end{aligned}$$

Now the question is how many tuples $\tilde{\boldsymbol{x}} \in \mathcal{X}^n$ with $\boldsymbol{H}(\tilde{\boldsymbol{x}}) = \boldsymbol{h}$ exist. To generate such a tuple, one could first specify the h_1 positions with value x_1. This amounts to $\binom{n}{h_1}$ possibilities. Therafter we have $\binom{n-h_1}{h_2}$ possibilities for placing x_2, then $\binom{n-h_1-h_2}{h_3}$ possibilities for x_3, and so on. All in all the number of possibilities is

$$\binom{n}{h_1}\binom{n-h_1}{h_2}\binom{n-h_1-h_2}{h_3} \cdots \binom{n-h_1-\cdots h_{K-1}}{h_K},$$

and elementary calculations reveal that this product equals the multinomial coefficient $\binom{n}{h_1,\ldots,h_K}$. Analogously one can show that $H_k \sim \mathrm{Bin}(n, p_k)$. Now we write $\widehat{p}_k = n^{-1} \sum_{i=1}^n 1_{[X_i = x_k]}$. Here we use the notation

$$1_{[A]} \;:=\; \begin{cases} 1 & \text{if } A \text{ is correct} \\ 0 & \text{else} \end{cases}$$

for an arbitrary statement A. This representation shows that $\mathbb{E}(\widehat{p}_k) = n^{-1}\sum_{i=1}^n \mathbb{P}(X_i = x_k) = p_k$. Moreover it follows from stochastic independence of the random variables X_i that

$$\begin{aligned}\mathrm{Cov}(\widehat{p}_k, \widehat{p}_\ell) &= n^{-2}\sum_{i=1}^n \mathrm{Cov}\big(1_{[X_i=x_k]}, 1_{[X_i=x_\ell]}\big) \\ &= n^{-1}\big(1_{[k=\ell]}p_k - p_k p_\ell\big).\end{aligned}$$

In case of $k = \ell$ this leads to the formula $\mathrm{Var}(\widehat{p}_k) = n^{-1}p_k(1-p_k)$, and $p_k(1-p_k)$ equals $1/4 - (p_k - 1/2)^2 \le 1/4$. □

Graphical representation. The absolute or relative frequencies H_k or $\widehat{p}_k$ may be visualised with a *bar chart* or a *pie chart*. For a bar chart, the potential values x_k are listed horizontally, and at each x_k we draw a vertical bar of height H_k or $\widehat{p}_k$, respectively.

For a pie chart, a disc is divided into K sectors ('slices of pie'). Each sector corresponds to a value x_k, and its area is proportional to H_k. In other words, the angle of the sector for x_k equals $2\pi \cdot \widehat{p}_k$.

Example 2.2 ('Random digit'). In Example 1.18, $n = 262$ students specified a 'random digit'. The resulting absolute and relative frequencies are provided in Table 2.1. Figure 2.1 shows the corresponding bar chart and pie chart. Although pie charts are quite popular, bar charts are typically easier to read and interpret.

x_k	0	1	2	3	4	5	6	7	8	9
H_k	8	6	12	32	25	23	28	70	41	17
$\widehat{p}_k$	.0305	.0229	.0458	.1221	.0954	.0878	.1069	.2672	.1565	.0649

Table 2.1: Absolute and relative frequencies for the variable 'random digit' in Example 2.2.

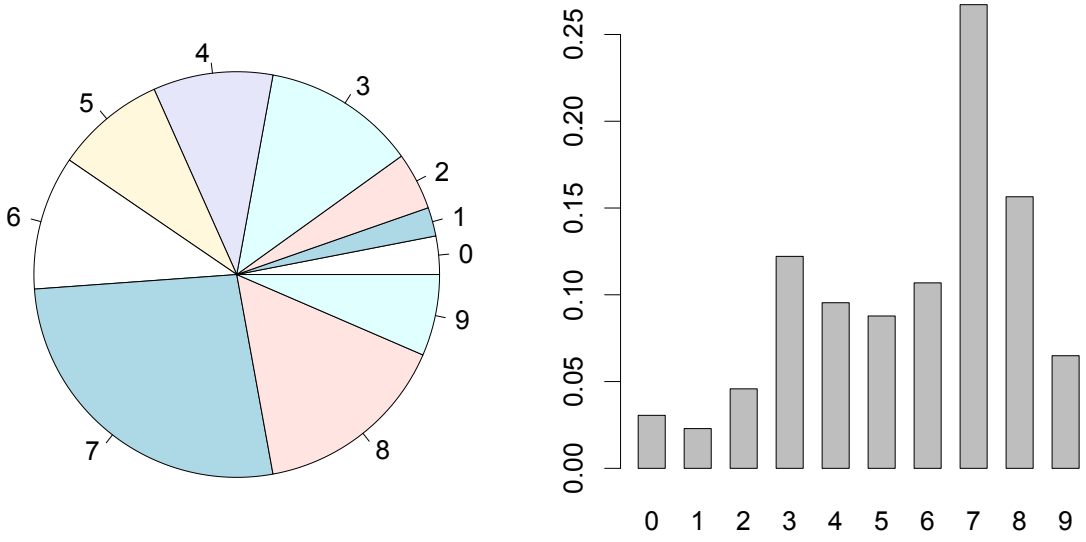

Figure 2.1: Pie and bar chart for the variable 'random digit' in Example 2.2.

2.2 Confidence Bounds for a Binomial Parameter

Now we focus on one potential value x_k and the corresponding quantities $p = p_k$, $H = H_k$ and $\widehat{p} = \widehat{p}_k$. As mentioned already, H is binomially distributed with parameters n and p. At this point we recommend Exercises 2.1 and 2.2.

Exact Confidence Bounds for p

We use our recipe from Chapter 1, this time with the binomial distribution functions $F_{n,p}$, $p \in [0,1]$. That means, $F_{n,p}(x) = \mathbb{P}_p(H \le x) = \sum_{k=0}^{x} \binom{n}{k} p^k (1-p)^{n-k}$ for $x = 0, 1, \ldots, n$. At first we have to clarify whether $F_{n,p}(x)$ is monotone in p.

Lemma 2.3. *For arbitrary $x \in \{0, 1, \ldots, n-1\}$,*

$$p \;\mapsto\; F_{n,p}(x)$$

is continuous and strictly decreasing on $[0,1]$ with boundary values $F_{n,0}(x) = 1$ and $F_{n,1}(x) = 0$. More precisely,

$$F_{n,p}(x) \;=\; n\binom{n-1}{x} \int_p^1 u^x (1-u)^{n-1-x}\, du.$$

The explicit representation of $F_{n,p}(x)$ by an integral will be used in a later chapter.

Proof of Lemma 2.3. The function $p \mapsto F_{n,p}(x)$ is a polynomial and thus continuously differentiable. The equations $F_{n,0}(x) = 1$ and $F_{n,1}(x) = 0$ are easily verified. Moreover, elementary calculations reveal that

$$\frac{d}{dp} F_{n,p}(x) \;=\; -n\binom{n-1}{x} p^x (1-p)^{n-1-x} \;<\; 0 \quad \text{for } 0 < p < 1.$$

This proves strict monotonicity of $p \mapsto F_{n,p}(x)$, and

$$F_{n,p}(x) \;=\; F_{n,p}(x) - F_{n,1}(x) \;=\; n\binom{n-1}{x} \int_p^1 u^x (1-u)^{n-1-x}\, du.$$

□

Figure 2.2 illustrates the monotonicity property stated in Lemma 2.3. This property implies the following three procedures:

(I) With probability at least $1 - \alpha$, the true parameter p and the random variable H satisfy $F_{n,p}(H) > \alpha$. The latter inequality is equivalent to

$$p \begin{cases} < b_\alpha(H) & \text{if } H < n, \\ \le 1 & \text{if } H = n. \end{cases}$$

Here we set

$$b_\alpha(h) \;:=\; \begin{cases} \text{unique solution } p \text{ of } F_{n,p}(h) = \alpha & \text{for } h = 0, 1, \ldots, n-1, \\ 1 & \text{for } h = n. \end{cases}$$

Thus we obtain an *upper $(1-\alpha)$-confidence bound* $b_\alpha(H)$ for p. That means,

$$\mathbb{P}_p(p \le b_\alpha(H)) \;\ge\; 1 - \alpha \quad \text{for arbitrary } p \in [0,1].$$

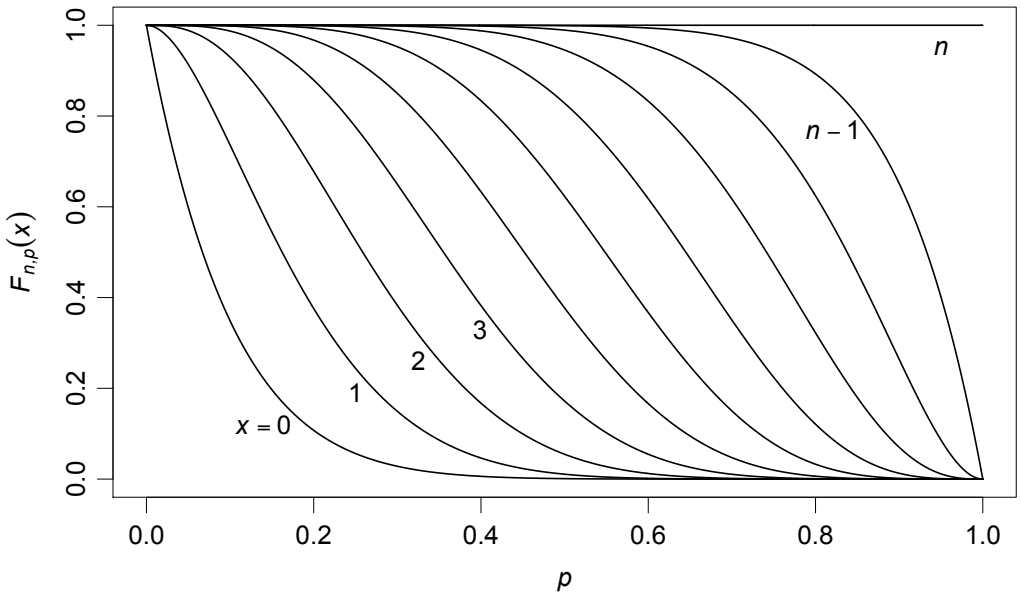

Figure 2.2: The functions $p \mapsto F_{n,p}(x)$ for $n = 10$ and $x = 0, 1, \ldots, n$.

(II) With probability at least $1 - \alpha$, the true parameter p and the random variable H satisfy $F_{n,p}(H - 1) < 1 - \alpha$, which is equivalent to

$$p \begin{cases} \geq 0 & \text{if } H = 0, \\ > a_\alpha(H) & \text{if } H > 0. \end{cases}$$

Here we set

$$a_\alpha(h) := \begin{cases} 0 & \text{for } h = 0, \\ \text{unique solution } p \text{ of } F_{n,p}(h-1) = 1 - \alpha & \text{for } h = 1, 2, \ldots, n. \end{cases}$$

This leads to a *lower* $(1 - \alpha)$*-confidence bound* $a_\alpha(H)$ for p, that means,

$$\mathbb{P}_p(p \geq a_\alpha(H)) \;\geq\; 1 - \alpha \quad \text{for arbitrary } p \in [0, 1].$$

(III) If we intend to bound the unknown parameter p from below and from above, we may compute the $(1 - \alpha)$*-confidence interval* $\big[a_{\alpha/2}(H), b_{\alpha/2}(H)\big]$ for p. This is the method of C. Clopper and E. S. Pearson[1] (1934). Other methods yield somewhat smaller intervals but are more difficult to justify and compute.

Remark 2.4. The equation $F_{n,p}(x) = \gamma$ may be solved for $x = 0$ and $x = n - 1$ explicitly. Otherwise one needs numerical procedures, for instance a bisection algorithm; see Exercise 2.3.

Example 2.5 (Quality control). The manufacturer of a certain device is convinced that the probability p of a failure under certain standard conditions is close to zero. To support this claim he tests n such devices and determines the number H of failures in this series. From his point of view an upper confidence bound $b_\alpha(H)$ is desirable.

Suppose he observes $H = 0$ failures. Then $\widehat{p} = 0$, and the upper confidence bound $b_\alpha(0)$ is the solution p of the equation $F_{n,p}(0) = (1 - p)^n = \alpha$. Hence the manufacturer may claim with confidence $1 - \alpha$ that p is no larger than

$$b_\alpha(0) \;=\; 1 - \alpha^{1/n}\,.$$

[1] Karl Pearson (1857-1936) and Egon S. Pearson (1885-1980): Father and son, influential british statisticians.

In case of $n = 50$ devices and $\alpha = 0.05$ one obtains the upper 95%-confidence bound $b_{0.05}(0) \approx 0.0582$.

Suppose the manufacturer tests $n = 50$ devices and precisely one of them fails. Then $\widehat{p} = 0.02$, and the upper confidence bound $b_{0.05}(1)$ is the unique solution p of the equation $(1-p)^{50} + 50p(1-p)^{49} = 0.05$. Here one can verify numerically that $0.0913 \leq b_{0.05}(1) \leq 0.0914$.

Example 2.6 (Opinion poll). The members of a special interest group want to persuade their city government that a majority of citizens is in favour of keeping a certain tram line. To this end $n = 100$ citizens are interviewed, and $H = 67$ persons turn out to be on the interest group's side. This yields the estimate $\widehat{p} = 0.67$ for the unknown proportion p of supporters of the tram line. To take into account the uncertainty of this small poll, the special interest group calculates a lower confidence bound $a_\alpha(67)$ for p. This is the unique solution p of the equation $F_{n,p}(66) = 1 - \alpha$. Particularly for $\alpha = 0.05$, numerical calculations reveal that $0.5845 \leq a_{0.05}(67) \leq 0.5846$; see also Figure 2.3. Hence one may claim with confidence 95% that p is no smaller than 0.5845.

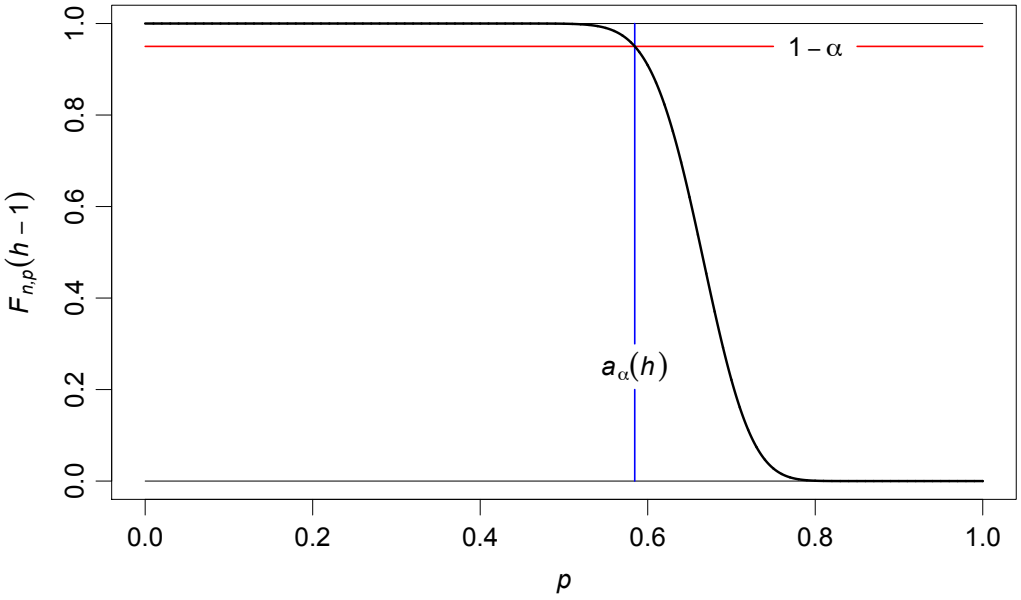

Figure 2.3: The lower confidence bounds $a_{0.05}(67)$ for p with $n = 100$.

A generalisation. The first part of Lemma 2.3 is a special case of a more general statement about monotonicity of distribution functions which will be needed later:

Lemma 2.7 (Monotonicity in distribution families). *Let $w_0, w_1, w_2, \ldots$ be nonnegative weights such that $0 < \sum_{k\geq 0} w_k\theta^k < \infty$ for arbitrary $\theta > 0$. For an arbitrary parameter $\theta \in (0, \infty)$ we define probability weights*

$$f_\theta(x) \;:=\; w_x\theta^x \Big/ \sum_{k\geq 0} w_k\theta^k, \quad x \in \mathbb{N}_0,$$

and a distribution function F_θ with

$$F_\theta(x) \;:=\; \sum_{k=0}^{x} f_\theta(k), \quad x \in \mathbb{N}_0.$$

In case of $\min\{k : w_k > 0\} \leq x < \sup\{k : w_k > 0\}$, the value $F_\theta(x)$ is continuous and strictly decreasing in $\theta > 0$, where $\lim_{\theta\to 0} F_\theta(x) = 1$ and $\lim_{\theta\to\infty} F_\theta(x) = 0$.

α	0.1	0.05	0.025	0.01	0.005
$1-\alpha$	0.9	0.95	0.975	0.99	0.995
$\Phi^{-1}(1-\alpha)$	1.2816	1.6449	1.9600	2.3264	2.5759

Table 2.2: Some values of Φ^{-1}.

Example 2.8. Here are two examples of such distribution families:

- Poisson distributions $\mathrm{Poiss}(\theta)$, $\theta > 0$: $w_k = 1/k!$;
- Binomial distributions $\mathrm{Bin}(n,p)$, $0 < p < 1$: $\theta = p/(1-p)$ and $w_k = \binom{n}{k}$.

In connection with 'odds ratios' we'll encounter a further distribution family of this type.

Approximate Confidence Bounds for p

Numerous text books provide approximate confidence bounds. This is okay if one wants to compute quickly some preliminary bounds. But in view of the powerful computing devices available today, the computation of exact bounds is no problem. Nevertheless we describe now two variants of approximate confidence bounds. But first we recall the definition of Gaussian distributions.

Definition 2.9 (Normal distribution). A real-valued random variable X follows the *normal (or Gaussian) distribution with mean* $\mu \in \mathbb{R}$ *and standard deviation* $\sigma > 0$, if its distribution is given by the density function $\phi_{\mu,\sigma}$; here

$$\phi_{\mu,\sigma}(x) \;:=\; \frac{1}{\sigma}\,\phi\Big(\frac{x-\mu}{\sigma}\Big) \quad \text{with} \quad \phi(z) \;:=\; (2\pi)^{-1/2}\exp(-z^2/2).$$

This is equivalent to saying that $\mathbb{P}(X \le x) = \Phi((x-\mu)/\sigma)$ for arbitrary $x \in \mathbb{R}$, where

$$\Phi(x) \;:=\; \int_{-\infty}^{x} \phi(z)\,dz.$$

As a symbol for this distribution we use $\mathcal{N}(\mu,\sigma^2)$. In the special case of $\mu = 0$ and $\sigma = 1$ we say that X follows the *standard normal (or Gaussian) distribution* $\mathcal{N}(0,1)$.

That a random variable X has distribution $\mathcal{N}(\mu,\sigma^2)$ with $\sigma > 0$ is equivalent to saying that $Z := (X-\mu)/\sigma$ has a standard normal distribution, see also Section A.2. In other words, X may be written as $X = \mu + \sigma Z$ with a standard gaussian random variable Z. It follows from Exercise 2.5 that indeed $\mathbb{E}(X) = \mu$ and $\mathrm{Std}(X) = \sigma$.

The distribution function $\Phi : \mathbb{R} \to (0,1)$ of the standard normal distribution is bijective with limits $\Phi(-\infty) = 0$ and $\Phi(\infty) = 1$. Its inverse function is denoted with Φ^{-1}. The symmetry of $\mathcal{N}(0,1)$ around 0 implies that

$$\Phi(-x) \;=\; 1 - \Phi(x) \quad \text{for } x \in \mathbb{R}$$

and

$$\Phi^{-1}(\gamma) \;=\; -\Phi^{-1}(1-\gamma) \quad \text{for } \gamma \in (0,1).$$

Table 2.2 contains some values of Φ^{-1}, rounded up to four digits.

Wilson's Method. The Central Limit Theorem (see appendix) implies that for arbitrary numbers $-\infty \le r < s \le \infty$,

$$\mathbb{P}_p\Big(\frac{\widehat{p}-p}{\sqrt{p(1-p)/n}} \in [r,s]\Big) \;\to\; \Phi(s) - \Phi(r) \quad \text{if } np(1-p) \to \infty. \tag{2.1}$$

For large values of $np(1-p) = \mathrm{Var}(H)$ each of the following inequalities is satisfied with probability approximately $1-\alpha$, respectively:

$$\begin{aligned} \widehat{p} &\le p + c_{\alpha,n}\sqrt{p(1-p)}, \\ \widehat{p} &\ge p - c_{\alpha,n}\sqrt{p(1-p)}, \\ |\widehat{p} - p| &\le c_{\alpha/2,n}\sqrt{p(1-p)}, \end{aligned}$$

where

$$c_{\alpha,n} := \Phi^{-1}(1-\alpha)/\sqrt{n}.$$

The preceding inequalities may be solved for p; see Exercise 2.6. They are equivalent to

$$\begin{aligned} p &\ge \frac{\widehat{p} + c^2/2 - c\sqrt{\widehat{p}(1-\widehat{p}) + c^2/4}}{1+c^2} \quad \text{with } c = c_{\alpha,n}, \\ p &\le \frac{\widehat{p} + c^2/2 + c\sqrt{\widehat{p}(1-\widehat{p}) + c^2/4}}{1+c^2} \quad \text{with } c = c_{\alpha,n}, \\ p &\in \left[\frac{\widehat{p} + c^2/2 \pm c\sqrt{\widehat{p}(1-\widehat{p}) + c^2/4}}{1+c^2}\right] \quad \text{with } c = c_{\alpha/2,n}. \end{aligned} \tag{2.2}$$

Thus we obtain approximate $(1-\alpha)$-confidence regions for p. These have been developed by E. B. Wilson[2].

Example 2.10. Figure 2.4 shows for $n = 30$ and $\alpha = 0.05$ the curves $p \mapsto p \pm c\sqrt{p(1-p)}$ with $c = c_{\alpha/2,n}$ who form together an ellipse. For three different numbers $p \in (0,1)$ the intervals $\left[p \pm c\sqrt{p(1-p)}\right]$ are drawn as vertical lines. In addition one sees for three different estimates $\widehat{p} \in (0,1)$ the corresponding confidence intervals (2.2) as horizontal lines.

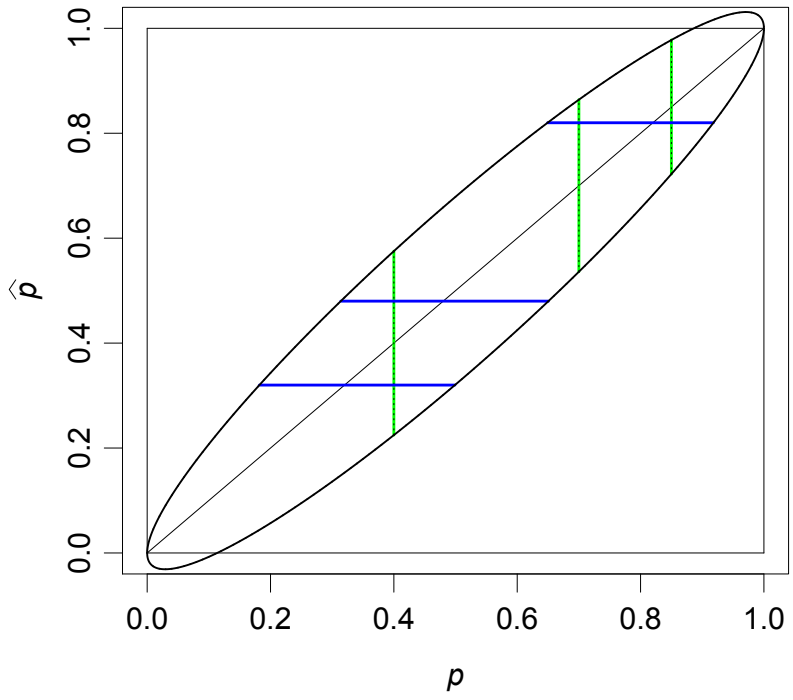

Figure 2.4: Wilson's method.

[2]Edwin B. Wilson (1879-1964): US-american mathematician with various fields of interest.

For practitioners the question is in which situations one may use Wilson's method. A simple answer would be 'never!', because nowadays the computation of the honest confidence bounds derived before is not a problem at all. Experience shows that the honest bounds and Wilson's bounds are very similar if $n\widehat{p}(1-\widehat{p}) \geq 5$.

Wald's Method. Now we describe a widely used and rather simple method which is a special case of a much more general recipe due to A. Wald[3]. In addition to the Central Limit Theorem which led to statement (2.1) we also have the following inequality for $\widehat{p}$:

$$\mathbb{E}\Big|\frac{\widehat{p}(1-\widehat{p})}{p(1-p)} - 1\Big| \;\leq\; \frac{\mathbb{E}\,|\widehat{p}-p|}{p(1-p)} \;\leq\; \frac{1}{\sqrt{np(1-p)}}.$$

Both facts together imply that in (2.1) the term $\sqrt{p(1-p)/n}$ may be replaced with $\sqrt{\widehat{p}(1-\widehat{p})/n}$; see also Exercise 4.27 (b). Thus each of the following inequalities is satisfied with probability about $1-\alpha$, respectively, provided that $np(1-p)$ is sufficiently large:

$$\begin{aligned} p \;&\geq\; \widehat{p} - c_{\alpha,n}\sqrt{\widehat{p}(1-\widehat{p})}, \\ p \;&\leq\; \widehat{p} + c_{\alpha,n}\sqrt{\widehat{p}(1-\widehat{p})}, \\ p \;&\in\; \big[\widehat{p} \pm c_{\alpha/2,n}\sqrt{\widehat{p}(1-\widehat{p})}\big]. \end{aligned}$$

The confidence bounds on the right hand side would also follow from Wilson's bounds if one replaced all terms c^2 with 0.

Wald's bounds are substantially easier to compute, but the true confidence level may be drastically smaller than the intended value of $1-\alpha$ if p is close to 0 or 1. We consider the true coverage probabilities $\mathbb{P}_p(p \in C(H))$ as a function of $p \in (0,1)$. Here $C(H)$ stands for the confidence interval $C_{\text{Wilson}}(H)$ via Wilson's method or $C_{\text{Wald}}(H)$ via Wald's method. In both cases the function

$$(0,1) \ni p \;\mapsto\; \mathbb{P}_p(p \in C(H))$$

is symmetric around 0.5. Thus we show in Figure 2.5 for $n = 100$ and $\alpha = 0.05$ the function $p \mapsto \mathbb{P}_p(p \in C_{\text{Wilson}}(H))$ on $(0, 0.5]$ and the function $p \mapsto \mathbb{P}_p(p \in C_{\text{Wald}}(H))$ on $[0.5, 1)$. Note that on the vertical axis only the range $[0.7, 1]$ is shown. In fact, $\mathbb{P}_p(p \in C_{\text{Wald}}(H))$ converges to 0 as $p \to 1$.

Upper Confidence Bounds for $|p - p_o|$

By means of our $(1-\alpha)$-confidence interval $\big[a_{\alpha/2}(H), b_{\alpha/2}(H)\big]$ for p one can possibly claim with confidence $1-\alpha$ that p is different from a given value p_o. For if the confidence interval does not contain p_o, then we may deduce with confidence $1-\alpha$ the sign of $p - p_o$ and a lower bound for $|p - p_o|$.

In some applications, however, one wants to show that the unknown parameter p is *close* to the special value p_o, even if possibly $p \neq p_o$. The confidence interval mentioned above implies the following statement: With confidence $1-\alpha$, the distance $|p - p_o|$ is not larger than

$$\begin{aligned} &\max\big\{|p' - p_o| : a_{\alpha/2}(H) \leq p' \leq b_{\alpha/2}(H)\big\} \\ &\qquad = \;\max\big\{b_{\alpha/2}(H) - p_o,\, p_o - a_{\alpha/2}(H)\big\}. \end{aligned}$$

[3] Abraham Wald (1902-1950): Romanian and US-American mathematician who developed, among many things, sequential statistical procedures, i.e. procedures with data-driven sample size.

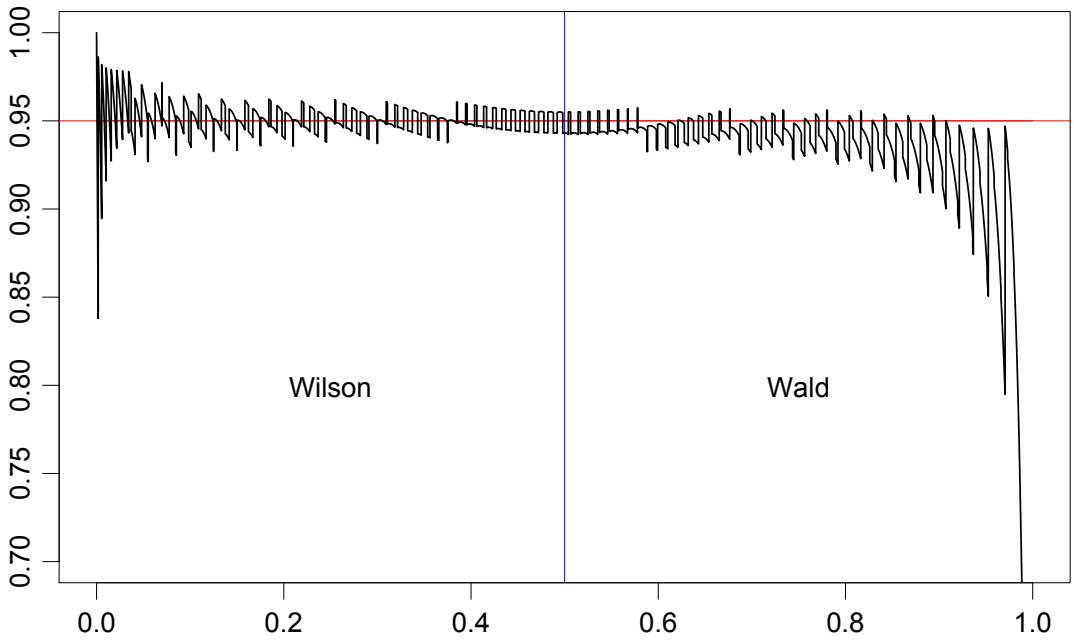

Figure 2.5: Coverage probabilities of the Wilson and Wald confidence interval in case of $n = 100$ and $\alpha = 0.05$.

But this bound is too conservative. A better one results if we compute the $(1 - \alpha)$-confidence interval

$$\big[\min\big(a_\alpha(H), p_o\big), \max\big(b_\alpha(H), p_o\big)\big]$$

for p. Here one combines the lower and upper $(1 - \alpha)$-confidence bound for p without replacing α with $\alpha/2$, but one forces the confidence interval to contain p_o. Behind this construction is a more general principle treated in Exercise 2.12. For the distance $|p - p_o|$ this yields the upper $(1 - \alpha)$-confidence bound

$$\max\big\{b_\alpha(H) - p_o, p_o - a_\alpha(H)\big\}.$$

2.3 The Chi-Squared Goodness-of-Fit Test and Alternatives

In various applications one is wondering whether the vector $\boldsymbol{p} = (p_k)_{k=1}^K$ coincides with a given vector $\boldsymbol{p}^o = (p_k^o)_{k=1}^K$ (null hypothesis).

Examples.

- A manufacturer of toys is producing dice. Now he wants to verify that with a newly produced dice all six numbers (faces) have the same probability of showing up. Here $K = 6$, $x_k = k$ and $p_k^o = 1/6$ for all k. From the manufacturer's viewpoint it is desirable that all probabilities p_k are close to the ideal value p_k^o.
- The roulette wheel of a casino is to be tested. The question is whether all 37 potential outcomes 0, 1, ..., 36 have the same probability $p_k^o = 1/37$. Both the owner and controllers of the casino want to detect potential deviations of the p_k from the ideal value p_k^o.
- When interviewing students (Example 1.18) they were asked to choose a 'random digit' in $\{0, 1, \ldots, 9\}$. The question is whether some and which probabilities p_k deviate significantly from $p_k^o = 1/10$.

• During a different lecture the students have been asked to produce a 'random sequence' with ten entries in $\{0, 1\}$. As our categorical variable we consider for each of these n sequences the number X of changes, so $X \in \{0, 1, \ldots, 9\}$; see also Example 1.17. Under the null hypothesis that the random sequences have been chosen completely at random, the probability p_k of $k-1$ changes is equal to

$$p_k^o \;:=\; \binom{9}{k-1} 2^{-9}.$$

The Chi-Squared Test

We want to devise a test of the null hypothesis that $\boldsymbol{p} = \boldsymbol{p}^o$. That means, we want to verify the working hypothesis that $\boldsymbol{p} \neq \boldsymbol{p}^o$ with a certain confidence if appropriate.

A test statistic. To test the former null hypothesis we need a test statistic $T = T(\boldsymbol{H})$ quantifying the empirical deviation from the null hypothesis: Each value $\widehat{p}_k$ is compared with its hypothetical value p_k^o and we form the sum

$$T \;:=\; n \sum_{k=1}^{K} \frac{(\widehat{p}_k - p_k^o)^2}{p_k^o} \;=\; \sum_{k=1}^{K} \frac{(H_k - n p_k^o)^2}{n p_k^o}.$$

This is the *chi-squared test statistic* of Karl Pearson. We'll see later why it makes sense to use the special weights $1/p_k^o$. One can easily deduce from Lemma 2.1 that

$$\mathbb{E}(T) \;=\; K - 1 \quad \text{if } \boldsymbol{p} = \boldsymbol{p}^o.$$

The exact test. Under the null hypothesis, the test statistic T has a well-defined distribution function G_o, namely

$$G_o(x) \;=\; \sum_{\boldsymbol{h} \in \mathbb{N}_0^K} 1_{[T(\boldsymbol{h}) \leq x]}\, f_{n,\boldsymbol{p}^o}(\boldsymbol{h})$$

for $x \in \mathbb{R}$; see Lemma 2.1. If the null hypothesis is violated, T tends to attain larger values. Thus we want to reject the null hypothesis if T is 'suspiciously large'. Hence if the *(right-sided) p-value*

$$1 - G_o(T-)$$

is less than or equal to α, we reject the null hypothesis at level α; that means, we claim with confidence $1 - \alpha$ that $\boldsymbol{p} \neq \boldsymbol{p}^o$. In case of this p-value being larger than α we don't make a definitive statement. This procedure is justified by Lemma 1.4 in Chapter 1.

Monte Carlo tests. The explicit computation of the p-value $1 - G_o(T-)$ is often too involved. An alternative to the exact p-value may be produced as follows: One simulates with the computer m stochastically independent random vectors $\boldsymbol{H}^{(1)}$, $\boldsymbol{H}^{(2)}$, …, $\boldsymbol{H}^{(m)}$ with distribution $\mathrm{Mult}(n, \boldsymbol{p}^o)$. For each of these one computes the test statistic $T_s = T(\boldsymbol{H}^{(s)})$ and then the Monte Carlo p-value

$$\frac{\#\big\{s \in \{1, \ldots, m\} : T_s \geq T\big\} + 1}{m + 1}.$$

If this p-value is less than or equal to α, we may claim with confidence $1-\alpha$ that the null hypothesis is violated. This is justified by the following 'Monte Carlo version' of Lemma 1.4.

Lemma 2.11. *Let $T_0, T_1, \ldots, T_m$ be real-valued random variables with the following property: For any fixed permutation σ of $\{0, 1, \ldots, m\}$ the random tuples $(T_{\sigma(0)}, T_{\sigma(1)}, \ldots, T_{\sigma(m)})$ and $(T_0, T_1, \ldots, T_m)$ are identically distributed. Then the random variable*

$$\widehat{\pi} \;:=\; \frac{\#\big\{s \in \{0, 1, \ldots, m\} : T_s \geq T_0\big\}}{m+1},$$

satisfies the inequalities

$$\mathbb{P}(\widehat{\pi} \leq \alpha) \;\leq\; \frac{\lfloor (m+1)\alpha \rfloor}{m+1} \;\leq\; \alpha.$$

for each $\alpha \in (0, 1)$. The second last inequality is an equality if the $m+1$ values $T_0, T_1, \ldots, T_m$ are almost surely different.

The property of a random tuple that its distribution remains invariant under arbitrary permutations of its components will play an important role in later chapters, too. It is satisfied, for instance, if the random variables $T_0, T_1, \ldots, T_m$ are stochastically independent and identically distributed.

Proof of Lemma 2.11. The assumption on the random variables T_0, T_1, …, T_m implies that the $m+1$ random variables $\widehat{\pi}_0$, $\widehat{\pi}_1$, …, $\widehat{\pi}_m$ with

$$\widehat{\pi}_j \;:=\; \frac{\#\big\{s \in \{0, \ldots, m\} : T_s \geq T_j\big\}}{m+1}$$

are identically distributed. Thus $\mathbb{P}(\widehat{\pi} \leq \alpha) = \mathbb{P}(\widehat{\pi}_0 \leq \alpha)$ is equal to

$$\frac{1}{m+1} \sum_{j=0}^{m} \mathbb{P}(\widehat{\pi}_j \leq \alpha) \;=\; \frac{1}{m+1} \sum_{j=0}^{m} \mathbb{E}\big(1_{[\widehat{\pi}_j \leq \alpha]}\big) \;=\; \frac{1}{m+1}\, \mathbb{E}\Big(\sum_{j=0}^{m} 1_{[\widehat{\pi}_j \leq \alpha]}\Big).$$

Now it suffices to show that

$$\sum_{j=0}^{m} 1_{[\widehat{\pi}_j \leq \alpha]} \;\leq\; \lfloor (m+1)\alpha \rfloor$$

with equality in case of the $m+1$ numbers $T_0, T_1, \ldots T_m$ being different. To this end let $t_0 \leq t_1 \leq \cdots \leq t_m$ be the values $T_0, T_1, \ldots, T_m$ in ascending order. Then $\sum_{j=0}^{m} 1_{[\widehat{\pi}_j \leq \alpha]}$ is equal to

$$\begin{aligned}
&\#\Big\{j \in \{0, \ldots, m\} \,:\, \underbrace{\#\big\{s \in \{0, \ldots, m\} : t_s \geq t_j\big\}}_{\geq m+1-j} \leq (m+1)\alpha\Big\} \\
&\quad\leq\; \#\big\{j \in \{0, \ldots, m\} : m+1-j \leq (m+1)\alpha\big\} \\
&\quad=\; \#\big\{k \in \{1, \ldots, m+1\} : k \leq (m+1)\alpha\big\} \\
&\quad=\; \lfloor (m+1)\alpha \rfloor.
\end{aligned}$$

The preceding inequalities are equalities if $t_0 < t_1 < \cdots < t_m$. □

Monte Carlo tests are very easy to implement. But not all practitioners appreciate them, because the resulting p-values do not only depend on the data but also on the (pseudo)random simulations of the tuples $\boldsymbol{H}^{(s)}$. On the other hand one can easily show that the exact p-value and its Monte Carlo version $\widehat{\pi}$ are essentially identical if m is large, see Exercise 2.15.

Chi-squared distributions and an approximate test. From a historical viewpoint, the test described now was the first procedure because in the old days of statistics the computation of exact or Monte Carlo p-values was out of reach. Let us first define a family of distributions which appear in many statistical contexts:

Definition 2.12 (Chi-squared distributions). The *chi-squared distribution with* $\ell \in \mathbb{N}$ *degrees of freedom* is defined as the distribution of $\sum_{j=1}^{\ell} Z_j^2$. Here $Z_1, Z_2, \ldots, Z_\ell$ are stochastically independent and standard Gaussian random variables. A symbol for this distribution is χ^2_ℓ.

In our specific testing problem the chi-squared distribution provides an approximation for the true distribution function G_o of T under the null hypothesis:

Theorem 2.13 (Chi-squared approximation). *Let F_{K-1} be the (continuous) distribution function of χ^2_{K-1}. Then*

$$\sup_{c\geq 0} \big|G_o(c) - F_{K-1}(c)\big| \;\to\; 0 \quad \text{as} \quad \min_{k=1,\ldots,K} np_k^o \to \infty.$$

Note that the number $K-1$ of degrees of freedom is equal to the number of possible values minus one. For our testing problem Theorem 2.13 yields the *approximate p-value*

$$1 - F_{K-1}(T).$$

Here is a rule of thumb proposed in various text books: If $\min_{k=1,\ldots,K} np_k^o \geq 5$, then the approximation above is sufficiently accurate.

Illustration of the approximation. In Figure 2.6 we illustrate the approximation of G_o by F_{K-1} in two special cases with $K = 10$. The two upper plots show the distribution functions G_o (step function) and F_9 (smooth function) in case of $p_k^o = 1/10$ for $k = 1, 2, \ldots, 10$ and $n = 20$ (left) and $n = 50$ (right). The quantity $\min_k np_k^o$ equals $n/10$, and indeed the approximation is quite good for $n = 50$. For the two lower plots we used $p_k^o = 2^{-9}\binom{9}{k-1}$ and $n = 20$ (left) and $n = 100$ (right). Here $\min_k np_k^o = n/512$, and indeed the difference between G_o and F_9 is clearly visible, even for sample size $n = 100$.

Example 2.14 ('Random digits'). For the data in Example 2.2 we want to test the null hypothesis that all p_k are equal to 0.1 at level $\alpha = 0.01$. The χ^2-test statistic equals

$$T \;=\; 262 \sum_{k=1}^{10} \frac{(\widehat{p}_j - 0.1)^2}{0.1} \;\approx\; 122.580.$$

Since $\min_k np_k^o \;=\; 26.2$ we trust in the approximation of G_o by F_9; see also Figure 2.6. The approximate p-value equals $1 - F_9(122.580) < 10^{-4}$, and the Monte Carlo method yielded extremely small p-values, too. Thus we may claim with confidence 99% that the 'random digits' are *not* uniformly distributed on the ten possible digits.

Justification of Theorem 2.13. The χ^2-test statistic T is equal to $\|\boldsymbol{Y}\|^2$ with the random vector

$$\boldsymbol{Y} \;:=\; \sqrt{n}\Big(\frac{\widehat{p}_k - p_k^o}{\sqrt{p_k^o}}\Big)_{k=1}^{K}.$$

This random vector lies within the $(K-1)$-dimensional vector space

$$\mathbb{H} \;:=\; \Big\{\boldsymbol{y} \in \mathbb{R}^K : \sum_{k=1}^{K} y_k \sqrt{p_k^o} = 0\Big\}.$$

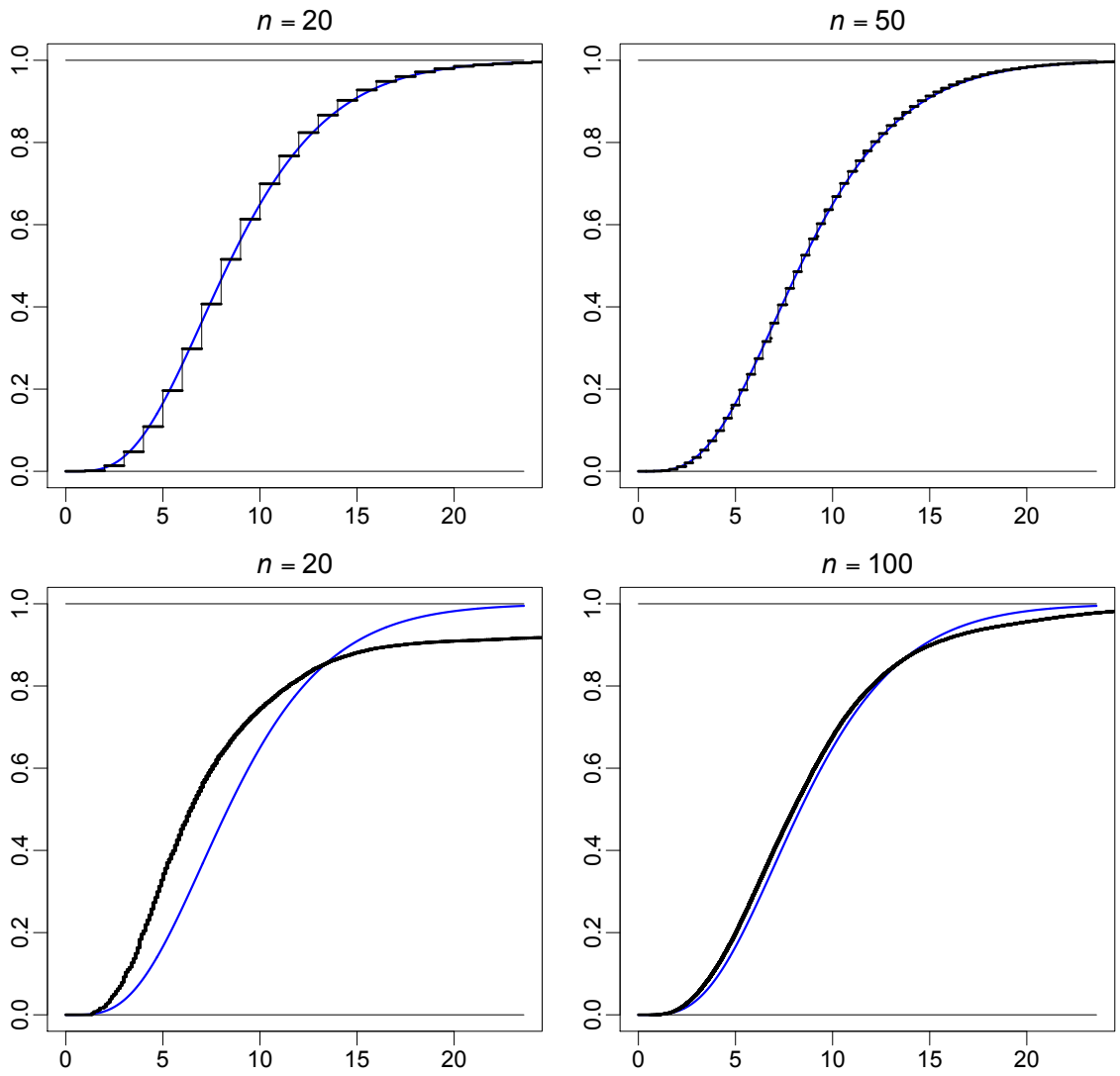

Figure 2.6: Approximation of the exact chi-squared test.

It follows from the Multivariate Central Limit Theorem that the random vector is approximatly standard normally distributed on $\mathbb{H}$, provided that $\boldsymbol{p} = \boldsymbol{p}^o$ and $\min_k np_k^o \to \infty$. That means, $\boldsymbol{Y}$ is approximately distributed like $\sum_{j=1}^{K-1} Z_j \boldsymbol{b}_j$ with stochastically independent standard Gaussian random variables $Z_1, Z_2, \ldots, Z_{K-1}$ and an orthonormal basis $\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_{K-1}$ of $\mathbb{H}$. But this means, that $T = \|\boldsymbol{Y}\|^2$ is approximately distributed like

$$\left\| \sum_{j=1}^{K-1} Z_j \boldsymbol{b}_j \right\|^2 = \sum_{j=1}^{K-1} Z_j^2 \sim \chi^2_{K-1}.$$

□

An Alternative Procedure

The chi-squared test just described has two weak points. Even if we reject the null hypothesis that $\boldsymbol{p} = \boldsymbol{p}^o$, we have no information about which components p_k are deviate from p_k^o and in which direction. In other situations it may be our goal to show that $\boldsymbol{p}$ is 'rather close' to $\boldsymbol{p}^o$.

A possible alternative to statistical tests is the computation of a confidence interval $[\tilde{a}_k, \tilde{b}_k]$ for p_k, *simultaneously for all* $k = 1, \ldots, K$. Precisely, with the given data we want to compute

confidence bounds $\tilde{a}_k = \tilde{a}_k(\boldsymbol{H})$ and $\tilde{b}_k = \tilde{b}_k(\boldsymbol{H})$ such that for a given test level $\alpha \in (0,1)$,

$$\mathbb{P}\big(p_k \in \big[\tilde{a}_k, \tilde{b}_k\big] \text{ for } k = 1, \ldots, K\big) \;\geq\; 1 - \alpha.$$

In other words, we want to compute a *confidence rectangle*

$$C(\boldsymbol{H}) \;=\; \big[\tilde{a}_1, \tilde{b}_1\big] \times \big[\tilde{a}_2, \tilde{b}_2\big] \times \cdots \times \big[\tilde{a}_K, \tilde{b}_K\big]$$

for the parameter vector $\boldsymbol{p}$ such that

$$\mathbb{P}_{\boldsymbol{p}}\big(\boldsymbol{p} \in C(\boldsymbol{H})\big) \;\geq\; 1 - \alpha \quad \text{for arbitrary } \boldsymbol{p}.$$

Then one may claim with confidence $1 - \alpha$ that *each* parameter p_k lies in the corresponding interval $[\tilde{a}_k, \tilde{b}_k]$. In particular, one may then check whether each hypothetical parameter p_k^o lies in the interval $\big[\tilde{a}_k, \tilde{b}_k\big]$.

This overall confidence can be achieved by means of a Bonferroni[4] adjustment: For each single parameter p_k one computes a $(1 - \alpha/K)$-confidence interval $\big[\tilde{a}_k, \tilde{b}_k\big]$,that means, one replaces α with α/K. Then

$$\begin{aligned}
&\mathbb{P}\big(p_k \in \big[\tilde{a}_k, \tilde{b}_k\big] \text{ for } k = 1, \ldots, K\big) \\
&\quad= \; 1 - \mathbb{P}\big(p_k \notin \big[\tilde{a}_k, \tilde{b}_k\big] \text{ for at least one } k \in \{1, \ldots, K\}\big) \\
&\quad\geq \; 1 - \sum_{k=1}^{K} \mathbb{P}\big(p_k \notin \big[\tilde{a}_k, \tilde{b}_k\big]\big) \\
&\quad\geq \; 1 - \sum_{k=1}^{K} \alpha/K \\
&\quad= \; 1 - \alpha.
\end{aligned}$$

The advantage of this method is that one can possibly identify components p_k which are with high confidence strictly smaller or strictly larger than p_k^o. However there are also situations in which the chi-squared test rejects the null hypothesis although $p_k^o \in [\tilde{a}_k, \tilde{b}_k]$ for all $k = 1, \ldots, K$.

Example 2.15 ('Random digits'). For the data in Example 2.2 we compute confidence intervals for the 10 parameters p_k, each with confidence level $(1 - \alpha/10) = 0.995$, $\alpha = 5\%$. Precisely, for each p_k we compute the honest $(1 - \alpha/20)$-confidence bounds $\tilde{a}_k = a_{\alpha/20}(H_k)$ and $\tilde{b}_k = b_{\alpha/20}(H_k)$:

x_k	0	1	2	3	4	5	6	7	8	9
$\tilde{a}_k$	.009	.005	.017	.072	.052	.046	.060	.194	.099	.030
$\tilde{b}_k$	.074	.063	.095	.189	.157	.148	.171	.350	.229	.119

In particular one may claim with overall confidence 95% that the probabilities of the digits 0, 1, 2 are strictly smaller than 0.1 while the digit 7 is chosen with probability strictly larger than 0.1.

If the goal is to verify that $\boldsymbol{p}$ is close to $\boldsymbol{p}^o$, one may construct the confidence intervals $[\tilde{a}_k, \tilde{b}_k]$ as follows: Let $\tilde{a}_k^* = \tilde{a}_k^*(\boldsymbol{H})$ and $\tilde{b}_k^* = \tilde{b}_k^*(\boldsymbol{H})$ be a lower and an upper $(1 - \alpha/K)$-confidence bound for p_k, respectively. Then

$$\big[\tilde{a}_k, \tilde{b}_k\big] \;:=\; \big[\min(\tilde{a}_k^*, p_k^o), \max(\tilde{b}_k^*, p_k^o)\big]$$

defines a $(1 - \alpha/K)$-confidence interval for p_k containing the value p_k^o by construction.

[4] Carlo E. Bonferroni (1892-1960): Italian mathematician who used probability inequalities in actuarial mathematics and statistics.

Example 2.16 (Mendel's law). In a cross fertilisation experiment someone wants to illustrate Mendel's law. Starting from two parental plants $n = 400$ 'daughters' are produced which may have genotype 'AA', 'AB' oder 'BB' with respect to a particular gene. If both parents are of type 'AB', Mendel's law predicts that the genotype of a 'daughter' follows

$$\big(p^o_{\mathrm{AA}}, p^o_{\mathrm{AB}}, p^o_{\mathrm{BB}}\big) \;=\; (1/4, 1/2, 1/4).$$

Suppose that the experiment yielded

$$\big(H_{\mathrm{AA}}, H_{\mathrm{AB}}, H_{\mathrm{BB}}\big) \;=\; (106, 178, 116),$$

so

$$\big(\widehat{p}_{\mathrm{AA}}, \widehat{p}_{\mathrm{AB}}, \widehat{p}_{\mathrm{BB}}\big) \;=\; (0.265, 0.445, 0.290).$$

Now we compute for each of the three parameters p_{AA}, p_{AB} and p_{BB} a lower and an upper $(1 - \alpha/3)$-confidence bound, where $\alpha = 0.05$:

type	AA	AB	BB
lower bound	0.2190	0.3915	0.2424
upper bound	0.3151	0.4994	0.3412

Thus we may claim with confidence $1 - \alpha = 95\%$ that

$$\big(p_{\mathrm{AA}}, p_{\mathrm{AB}}, p_{\mathrm{BB}}\big) \;\in\; [0.2190,\, 0.3151] \times [0.3915,\, \mathbf{0.5}] \times [0.2424,\, 0.3412].$$

In particular we may claim with confidence 95% that the maximal deviation between the true probabilities from Mendel's values is at most 0.1085.

A Variation of the Chi-Squared Test

Usually the chi-squared goodness-of-fit test serves the purpose of showing that $\boldsymbol{p} \neq \boldsymbol{p}^o$. But one can use it also to detect manipulated data. That means, one could check whether the vector $\widehat{\boldsymbol{p}}$ is 'suspiciously close' to $\boldsymbol{p}^o$. To this end one should compute the *left-sided* p-value

$$G_o(T)$$

or its Monte Carlo approximation

$$\frac{\#\big\{s \in \{1, \ldots, m\} : T_s \leq T\big\} + 1}{m + 1}$$

or the approximation

$$F_{K-1}(T)$$

with the distribution function F_{K-1} of χ^2_{K-1}. If this p-value is less than or equal to α, one may claim with confidence $1 - \alpha$ that the observed frequency vector $\boldsymbol{h} = (H_k)_{k=1}^K$ is *not* a realisation of a random vector with distribution $\mathrm{Mult}(n, \boldsymbol{p}^o)$.

Example 2.17. We consider once more the preceding example with Mendel's law. Suppose a scientist claims to have observed $(H_{\mathrm{AA}}, H_{\mathrm{AB}}, H_{\mathrm{BB}}) = (102, 199, 99)$. This would fit Mendel's law remarkably well. Indeed, here $T = 0.055$, and the approximate left-sided p-value equals $F_2(0.055) \approx 0.0271$. (We use the χ^2 approximation since $\min_k np_k = 100$.) Thus the results look too good to be true. It could be that he or she manipulated the data or picked one particular experiment out of several without mentioning the latter.

2.4 Exercises

Exercise 2.1 (Point estimation of p). Let H be a random variable with distribution $\mathrm{Bin}(n,p)$, where $n \in \mathbb{N}$ is given while $p \in [0,1]$ is unknown. Consider for $c \geq 0$ the estimator

$$\widehat{p}_c \;:=\; \frac{H + c/2}{n + c}.$$

For $c = 0$ this is the standard estimator $\widehat{p} = H/n$, and for $c > 0$ the latter is shifted toward the value $1/2$.

(a) Determine bias, variance and mean squared error of $\widehat{p}_c$. You should realise that $\mathrm{MSE}_p(\widehat{p}_c)$ is a function of n, c and $|p - 1/2|$.

(b) Draw the function $p \mapsto \mathrm{MSE}_p(\widehat{p}_c)$ for $n = 25$ and $c = 0, 1, 2, \ldots, 7$.

(c) Determine a value $c = c(n)$ such that the maximal mean quadratic error,

$$\max_{0 \leq p \leq 1} \mathrm{MSE}_p(\widehat{p}_c),$$

is as small as possible.

Exercise 2.2 (Unbiased estimation of $g(p)$). Let H, n and p be as in Exercise 2.1, and let $g : [0,1] \to \mathbb{R}$ be an arbitrary function. For $g(p)$ we now consider all estimators of the form $\widehat{g} = s(H)$ with some mapping $s : \{0, 1, \ldots, n\} \to \mathbb{R}$.

(a) Suppose that the estimator $\widehat{g} = s(H)$ is unbiased for $g(p)$. Show that $p \mapsto g(p)$ has to be a polynomial of degree at most n.

(b) Suppose that $p \mapsto g(p)$ is a polynomial of degree at most n. Show that there exists a unique unbiased estimator $\widehat{g} = s(H)$ of $g(p)$.

Hint: Determine for $k = 0, 1, \ldots, n$ the expected value of $[H]_k$.

(c) The preceding considerations show that unbiasedness is a seemingly nice but possibly very restrictive property. Compare under this aspect the unbiased estimator of $g(p) := (1-p)^n$ with the naive estimator $(1 - H/n)^n$.

Exercise 2.3 (Implementing exact confidence bounds for p). To compute exact confidence bounds for a binomial parameter one has to solve equations of the form

$$F_{n,p}(x) \;=\; \gamma$$

for given $n \in \mathbb{N}$, $x \in \{0, 1, \ldots, n-1\}$ and $\gamma \in (0,1)$. The algorithm described in Table 2.3 solves this problem with prescribed accuracy $\delta > 0$. The result are two numbers $p_1, p_2 \in [0,1]$ such that $0 < p_2 - p_1 \leq \delta$, $F_{n,p_1}(x) \geq \gamma \geq F_{n,p_2}(x)$ and $F_{n,p_1}(x) - F_{n,p_2}(x) \leq \delta$.

Implement this algorithm. Check your program by means of Example 2.6.

Exercise 2.4. Prove Lemma 2.7. Then describe how you would compute exact confidence bounds for an unknown parameter $\theta > 0$, if you observed a random variable X with distribution function F_θ. How could one adapt the algorithm in Table 2.3 for the present setting?

Exercise 2.5 (Moments of the standard Gaussian distribution). Let Z be a standard Gaussian random variable. Show by means of a symmetry consideration and via partial integration that $\mathbb{E}(Z^{2m-1}) = 0$ and

$$\mathbb{E}(Z^{2m}) \;=\; \prod_{i=1}^{m} (2i-1) \quad \text{for } m \in \mathbb{N}.$$

An alternative derivation will be given in Exercise 4.13.

Algorithmus $(p_1, p_2) \leftarrow$ **BinoCB**(x, n, γ, δ)
$p_1 \leftarrow 0,\ F_1 \leftarrow 1$
$p_2 \leftarrow 1,\ F_2 \leftarrow 0$
while $p_2 - p_1 > \delta$ **or** $F_1 - F_2 > \delta$ **do**
  $p_m \leftarrow (p_1 + p_2)/2,\ F_m \leftarrow F_{n,p_m}(x)$
  if $F_m \geq \gamma$ **then**
    $p_1 \leftarrow p_m,\ F_1 \leftarrow F_m$
  else
    $p_2 \leftarrow p_m,\ F_2 \leftarrow F_m$
  end if
end while

Table 2.3: Computation of exact confidence bounds for p.

Exercise 2.6 (Inequalities for Wilson's and Wald's method). Show that for arbitrary numbers $p, \widehat{p} \in [0,1]$ and $c > 0$,

$$\widehat{p} \ \leq_{(\geq)} \ p +_{(-)} c\sqrt{p(1-p)}$$

if and only if

$$p \ \geq_{(\leq)} \ \frac{\widehat{p} + c^2/2 \ -_{(+)} \ c\sqrt{\widehat{p}(1-\widehat{p}) + c^2/4}}{1 + c^2}.$$

For which values $\widehat{p} \in [0,1]$ is Wald's interval

$$\left[\widehat{p} \pm c\sqrt{\widehat{p}(1-\widehat{p})}\right]$$

shorter or longer than Wilson's interval

$$\left[\frac{\widehat{p} + c^2/2 \ \pm \ c\sqrt{\widehat{p}(1-\widehat{p}) + c^2/4}}{1 + c^2}\right] ?$$

Exercise 2.7 (Examples of confidence regions for a binomial parameter p). Define for the subsequent applications a suitable probability parameter and determine whether a lower confidence bound, an upper confidence bound or a confidence interval would be appropriate. Then compute this confidence region with $\alpha = 5\%$. You may use either (i) exact bounds or (ii) Wilson's method.

(a) How common is airplane anxiety? After a spectacular attempt of a person to escape an airplane just before take-off, 335 swiss people answered the question whether they suffer from airplane anxiety. Result: 70 people answered 'yes'.

(b) Does a majority of voters prefer online-elections? 29 persons have been asked whether they would prefer showing up at an election office, sending in their ballot by regular mail or voting online. Result: 22 people would prefer voting online.

(c) A provider of a WLAN router wants to demonstrate that the majority of his customers can handle the device with the standard installation software and brochure. To this end he investigates via his call center how many of 2500 new customers needed additional customer support via phone. Result: 42 customers needed extra support.

(d) Citizens in a town want to convince their mayor that a certain quarter is problematic in terms of security. To this end, 250 citizens are interviewed and asked whether they would walk through this quarter at night time alone. Result: 139 people answered 'no'.

Exercise 2.8 (Comparison of two Poisson parameters). In some applications one observes two independent random variables with Poisson distributions, $Y_1 \sim \text{Poiss}(\lambda_1)$ and $Y_2 \sim \text{Poiss}(\lambda_2)$

with unknown parameters $\lambda_1, \lambda_2 > 0$. The question is whether λ_1 and λ_2 are different. Specific examples are (i) the comparison of cell concentrations of two cell cultures in biological or medical experiments or (ii) the comparison of two radioactive probes or (iii) the comparison of two event rates in actuarial applications.

(a) Show that the conditional distribution of Y_1, given $Y_1 + Y_2 = s$, is a binomial distribution with parameters s and $p := \lambda_1/(\lambda_1 + \lambda_2)$. That means,

$$\mathbb{P}(Y_1 = k \,|\, Y_1 + Y_2 = s) \;=\; \binom{s}{k} p^k (1-p)^{s-k} \quad \text{for } k = 0, \ldots, s.$$

(b) Describe confidence bounds for the ratio λ_1/λ_2 by means of part (a). What would be your result if $Y_1 = 14$, $Y_2 = 21$ and $\alpha = 5\%$?

Exercise 2.9 (Wilson's method for Poisson parameters). Let Y be a random variable with distribution $\mathrm{Poiss}(\lambda)$ with unknown parameter $\lambda \geq 0$. For λ one could compute exact confidence bounds, but now we want to imitate Wilson's method (for binomial parameters). The Central Limit Theorem implies that for arbitrary numbers $-\infty \leq r < s \leq \infty$,

$$P\Big(\frac{Y - \lambda}{\sqrt{\lambda}} \in [r, s]\Big) \;\to\; \Phi(s) - \Phi(r) \quad \text{as } \lambda \to \infty.$$

Now construct approximate $(1-\alpha)$-confidence bounds and intervals for λ.

Exercise 2.10 (Sample sizes for estimating a binomial parameter). So far we considered the sample size n as given. Sometimes one can determine an appropriate sample size prior to the experiment or study. We illustrate this with Wilson's $(1-\alpha)$-confidence interval for binomial parameter p based on an observation $H \sim \mathrm{Bin}(n, p)$.

(a) How large should the sample size be to guarantee an interval of length at most $\delta > 0$? What is your result if $\alpha = 0.05$ and $\delta = 0.1$?

(b) For given values $0 < p_1 < p_2 < 1$ the confidence interval should never contain both p_1 and p_2. How large should n be such that this cannot happen? Hint: Exercise 2.6.

Numerical example: For the german FDP (liberal party) a percentage of $p_1 = 5\%$ in an election would be a desaster, a percentage of $p_2 = 15\%$ or more a good reason to party. How large should the sample size be such that at least one of these scenarios may be excluded with confidence 99% ?

Exercise 2.11 (McNemar's test). Let $\boldsymbol{H}$ have distribution $\mathrm{Mult}(n, \boldsymbol{p})$ with unknown probability vector $\boldsymbol{p} = (p_j)_{j=1}^K$. The question is whether $p_1 \leq p_2$ (null hypothesis) or $p_1 > p_2$ (working hypothesis). Instead of a statistical test we construct now a suitable confidence bound for p_1/p_2:

(a) Show that the conditional distribution of H_1, given $H_1 + H_2$, is a binomial distribution with parameters $H_1 + H_2$ and $\rho := p_1/(p_1 + p_2)$. That means, for arbitrary numbers $m \in \{0, 1, \ldots, n\}$ and $x \in \{0, 1, \ldots, m\}$,

$$\mathbb{P}(H_1 = x \,|\, H_1 + H_2 = m) \;=\; \binom{m}{x} \rho^x (1-\rho)^{m-x}.$$

(b) Describe explicit confidence bounds for p_1/p_2 by means of exact confidence bounds for binomial parameters.

(c) Now analyze the following fictitious data: For the diagnosis of a certain disease two different medical tests A and B are available. The working hypothesis is that test A is more sensitive than test B. That means, for a person suffering from that disease, $\mathbb{P}(\text{test A is positive})$ exceeds $\mathbb{P}(\text{test B is positive})$. Now both tests are applied to $n = 60$ diseased persons. Test A was positive for 57 people, test B was positive for 50 people; for 48 people both tests turned out positive. Do

these numbers support the working hypothesis?
Hint: For each person four different outcomes are possible. Identify these four outcomes and formulate the working hypothesis in terms of the corresponding probabilities. Then apply one of the confidence bounds of part (b).

Exercise 2.12 (Confidence bounds for proving small differences). So far we constructed $(1-\alpha)$-confidence intervals for a real valued quantity $g(\theta)$ by combining a lower $(1-\alpha/2)$-confidence bound and an upper $(1-\alpha)$-confidence bound for $g(\theta)$. If the main goal is to show that $g(\theta)$ is close to a given value g_o, one can modify this approach as follows:

Let $a_\alpha = a_\alpha(\text{data})$ and $b_\alpha = b_\alpha(\text{data})$ be a lower and an upper confidenc bound for $g(\theta)$, respectively. That means, for arbitrary parameters θ,

$$\left.\begin{array}{l} \mathbb{P}_\theta\big(g(\theta) \geq a_\alpha\big) \\ \mathbb{P}_\theta\big(g(\theta) \leq b_\alpha\big) \end{array}\right\} \;\geq\; 1-\alpha.$$

Show that

$$\big[\min\left(a_\alpha, g_o\right), \max(b_\alpha, g_o)\big]$$

defines a $(1-\alpha)$-confidence interval for $g(\theta)$.

Exercise 2.13. To clarify whether the probability p of a newborn child being male differs significantly from 0.5, the data of $n = 429'440$ newborns have been analyzed. It turned out that $H = 221'023$ of these newborns were boys.

(a) Compute a 99%-confidence interval for p by means of Wilson's method. How would you answer the question about p above?

(b) Compute an upper 99%-confidence bound for $|p - 0.5|$.

Exercise 2.14 (Geometric interpretation of the chi-squared test statistic). For a probability vector $\boldsymbol{p}$ we consider $\sqrt{\boldsymbol{p}} := \big(\sqrt{p_k}\big)_{k=1}^K$. This defines a mapping $\boldsymbol{p} \mapsto \sqrt{\boldsymbol{p}}$ from the unit simplex to a subset of the unit sphere in $\mathbb{R}^K$, see Figure 2.7 for the case $K = 3$. Now we define for two probability vectors $\boldsymbol{p}, \boldsymbol{q}$ the following quantities:

$$T(\boldsymbol{p},\boldsymbol{q}) \;:=\; \sum_{k=1}^K \frac{(q_k - p_k)^2}{p_k}, \qquad \tilde{T}(\boldsymbol{p},\boldsymbol{q}) \;:=\; 4\big\|\sqrt{\boldsymbol{q}} - \sqrt{\boldsymbol{p}}\big\|^2$$

and

$$\delta(\boldsymbol{p},\boldsymbol{q}) \;:=\; \max_{k=1,\ldots,K}\Big|\frac{q_k}{p_k} - 1\Big|.$$

(a) Show that in case of $\delta(\boldsymbol{p},\boldsymbol{q}) > 0$,

$$1 - \frac{3\delta(\boldsymbol{p},\boldsymbol{q})}{4} \;\leq\; \frac{T(\boldsymbol{p},\boldsymbol{q})}{\tilde{T}(\boldsymbol{p},\boldsymbol{q})} \;\leq\; 1 + \frac{\delta(\boldsymbol{p},\boldsymbol{q})}{2}.$$

(b) Suppose that $\widehat{\boldsymbol{p}} = n^{-1}\boldsymbol{H}$ with $\boldsymbol{H} \sim \mathrm{Mult}(n, \boldsymbol{p}^o)$. Show that

$$\mathbb{E}(\delta(\boldsymbol{p}^o, \widehat{\boldsymbol{p}})^2) \;\leq\; \frac{K-1}{\min_{k=1,\ldots,K} np_k^o}.$$

Exercise 2.15. For a test statistic $T = T(\text{data})$ we consider the p-value

$$\pi \;:=\; 1 - G_o(T-)$$

with a given distribution function G_o and the Monte Carlo p-value

$$\widehat{\pi} \;:=\; \frac{\#\big\{s \in \{1,\ldots,m\} : T_s \geq T\big\} + 1}{m+1}.$$

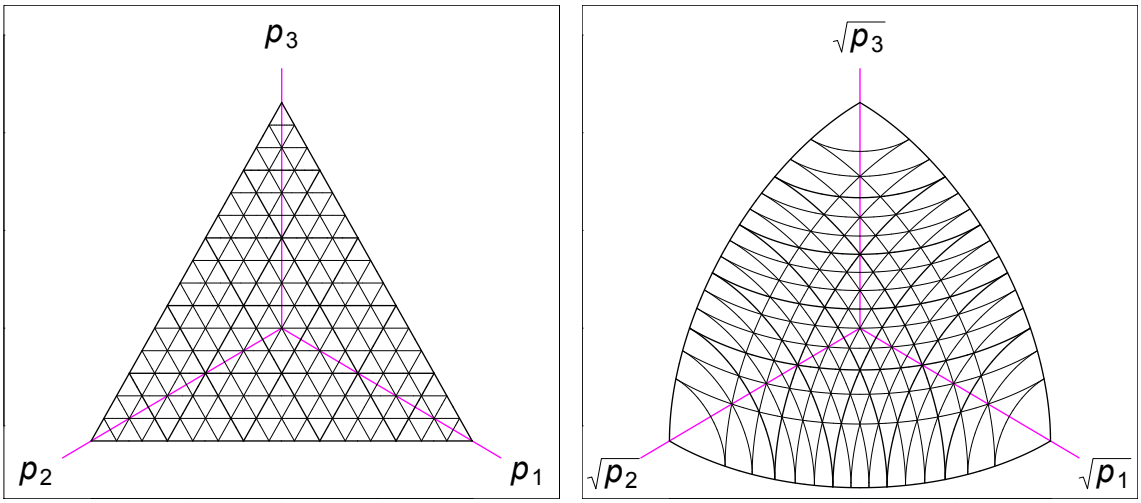

Figure 2.7: Geometric consideration for the chi-squared test statistic.

Here we consider the data as fixed while $T_1, T_2, \ldots, T_m$ are independent random variables with distribution function G_o. Show that

$$\mathbb{E}\big((\widehat{\pi} - \pi)^2\big) \;\leq\; \frac{1}{4m+1} \quad \text{if } \; m \geq 2.$$

Exercise 2.16 (Leading digits). What is the distribution of the leading digit of a random variable? We consider the following sample: From a huge list of towns and villages we picked randomly a list of 305 municipalities. The table below contains the numbers of municipalities whose population size starts with digit $1, 2, \ldots, 9$.

Leading digit	1	2	3	4	5	6	7	8	9
Frequency	107	55	39	22	13	18	13	23	15

(a) Test the null hypothesis that the leading digit of the population size in this 'population' of municipalities is uniformly distributed on the set $\{1, 2, \ldots, 9\}$.

(b) Test the null hypothesis that the leading digit follows Benford's law, i.e.

$$\mathbb{P}(\text{leading digit } = k) \;=\; \log_{10}(1 + 1/k) \quad \text{for } \; k = 1, 2, \ldots, 9.$$

Exercise 2.17 (Benford's law). Behind the Benford distribution in the previous exercise there is a general phenomenon: If X is a random variable with continuous distribution function F on $\mathbb{R}$, and if this distribution is 'quite diffuse', then the random variable $Y := X - \lfloor X \rfloor$ is 'approximately uniformly' distributed on $[0, 1)$. (This vague statement can be made mathematically rigorous.)

Now let $Z > 0$ be a random variable with continuous distribution on $(0, \infty)$. We represent Z as decimal number, that means,

$$Z \;=\; Z_0.Z_1Z_2Z_3\ldots \cdot 10^W \;=\; \big(Z_0 + 10^{-1}Z_1 + 10^{-2}Z_2 + 10^{-3}Z_3 + \ldots\big) \cdot 10^W$$

with digits $Z_0 \in \{1, \ldots, 9\}$, $Z_1, Z_2, Z_3, \ldots \in \{0, 1, \ldots, 9\}$ and an integer exponent W. We assume that the distribution of $X = \log_{10}(Z)$ is 'quite diffuse'. Show that the phenomenon just mentioned implies that

$$\mathbb{P}(Z_0 = k) \;\approx\; \log_{10}(1 + 1/k) \quad \text{for } \; k = 1, 2, \ldots, 9.$$

Remark: Benford's Law is used, for instance, by the internal revenue service to detect manipulations of financial data.

Exercise 2.18. The following table contains the frequencies of deaths in the US population in the twelve months of 1966:

January	$166'761$	July	$159'924$
February	$151'296$	August	$145'184$
March	$164'804$	September	$141'164$
April	$158'973$	October	$154'777$
May	$156'455$	November	$150'678$
June	$149'251$	December	$163'882$

The question is whether the death rate of each month is proportional to its length. One can justify that conditional on the total number N of deaths of US citizens in 1966, the months $X_1, X_2, \ldots, X_N$ in which these people died may indeed be viewed as independent, identically distributed random variables. Now we are interested in the probabilities $p_k = \mathbb{P}(X_i = \text{month } k)$. Formulate and test a suitable null hypothesis with the two methods we have treated in this chapter, the chi-squared goodness-of-fit test at level $\alpha = 0.01$ and simultaneous 99%-confidence intervals for the p_k. How do you interpret the results?

Exercise 2.19. Looking carefully at Figure 2.5, note the particular shape of the smooth parts between consecutive jumps. Each of these corresponds to a function $p \mapsto F_{n,p}(\ell) - F_{n,p}(k-1)$ with certain integers $0 \le k \le \ell \le n$. Show that the function

$$(0,1) \ni p \;\mapsto\; \log\big(F_{n,p}(\ell) - F_{n,p}(k-1)\big)$$

is always concave.

Chapter 3

Numerical Variables: Distribution Functions and Quantiles

3.1 The Empirical Distribution

Again we focus on one variable of a data set with values $X_1, X_2, \ldots, X_n$ in an arbitrary measurable space $(\mathcal{X}, \mathcal{B})$. Now we consider these n values as stochastically independent random variables with unknown distribution P on $\mathcal{X}$, so

$$\mathbb{P}(X_1 \in B_1, X_2 \in B_2, \ldots, X_n \in B_n) \;=\; P(B_1)P(B_2)\cdots P(B_n)$$

for arbitrary measurable sets $B_1, B_2, \ldots, B_n \subset \mathcal{X}$. This is sometimes paraphrased as X_1,X_2, …, X_n being a *sample (of size n) from the distribution P*. Under this assumption one may estimate the distribution P by the *empirical distribution* $\widehat{P}$ of the data which is defined as follows: For a measurable set $B \subset \mathcal{X}$ let

$$\widehat{P}(B) \;:=\; \#\{i \le n : X_i \in B\}/n.$$

This is the relative fraction of data points X_i lying within B. In other words, $\widehat{P}$ is a random discrete probability distribution on $\mathcal{X}$ with weight $\#\{i \le n : X_i = x\}/n$ at an arbitrary point $x \in \mathcal{X}$.

The random variables $1_{[X_i \in B]}$ are stochastically independent with values in $\{0, 1\}$ and expected value $P(B)$. Hence

$$n\widehat{P}(B) = \sum_{i=1}^{n} 1_{[X_i \in B]} \;\sim\; \mathrm{Bin}(n, P(B)).$$

In particular,

$$\mathbb{E}(\widehat{P}(B)) \;=\; P(B) \quad \text{and} \quad \mathrm{Std}(\widehat{P}(B)) \;=\; \sqrt{\frac{P(B)(1-P(B))}{n}} \;\le\; \frac{1}{2\sqrt{n}}.$$

3.2 Distribution Functions and Quantiles

From now on we consider the special case of a numerical variable, i.e. $\mathcal{X} = \mathbb{R}$. Let us first recall the definition and properties of distribution functions.

The distribution function. The distribution P is uniquely characterised by its *distribution function F*. Here

$$F(x) \;:=\; P((-\infty, x]) = \mathbb{P}(X_i \le x) \quad \text{for } x \in \mathbb{R}.$$

This function F has the following properties:

- F is non-decreasing,
- $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to\infty} F(x) = 1$,
- F is right-continuous. Precisely, for arbitrary $x \in \mathbb{R}$,

$$F(x) \;=\; \lim_{s\to x, s>x} F(s) \quad \text{and} \quad F(x-) := \lim_{s\to x, s<x} F(s) \;=\; P((-\infty, x)).$$

The jump size $F(x) - F(x-)$ of F at an arbitrary position x is equal to $P(\{x\})$.

Quantiles. Closely related to the distribution function are so-called *quantiles*. Let $0 < \gamma < 1$. A real number q_γ is called a *γ-quantile of P* if

$$P((-\infty, q_\gamma]) \;\geq\; \gamma \quad \text{and} \quad P([q_\gamma, \infty)) \;\geq\; 1-\gamma.$$

This is equivalent to the requirement that

$$P((-\infty, q_\gamma)) \;\leq\; \gamma \quad \text{and} \quad P((q_\gamma, \infty)) \;\leq\; 1-\gamma.$$

Roughly saying, q_γ divides the real line into a left and a right halfline with approximate probabilities γ and $1-\gamma$, respectively. By means of the distribution function F one may write

$$F(q_\gamma -) \;\leq\; \gamma \;\leq\; F(q_\gamma)$$

or

$$F(x) \;\begin{cases} \leq\; \gamma & \text{for } x < q_\gamma, \\ \geq\; \gamma & \text{for } x \geq q_\gamma. \end{cases}$$

Thus the distribution function F passes the value γ at the point q_γ.

The set of *all* γ-quantiles of P is always a closed interval with boundaries

$$q_{\gamma,1} \;:=\; \min\bigl\{x \in \mathbb{R} : F(x) \geq \gamma\bigr\} \quad \text{and} \quad q_{\gamma,2} \;:=\; \inf\bigl\{x \in \mathbb{R} : F(x) > \gamma\bigr\}.$$

If we talk about *the* γ-quantile of P, then we refer to the midpoint $q_\gamma := (q_{\gamma,1} + q_{\gamma,2})/2$. If F is continuous on $\mathbb{R}$ and strictly increasing on $\{x \in \mathbb{R} : 0 < F(x) < 1\}$, then there exists a unique γ-quantile $q_\gamma = F^{-1}(\gamma)$ with the inverse function $F^{-1} : (0,1) \to \{x \in \mathbb{R} : 0 < F(x) < 1\}$ of F.

Quartiles and median. Special quantiles are the so-called *quartiles*: The

- first quartile: $q_{0.25}$,
- second quartile: $q_{0.50}$,
- third quartile: $q_{0.75}$.

A 50%-quantile is also called a *median* of P. The median is an important feature of the distribution P which may be characterised as follows:

Lemma 3.1 (Characterising the median). *Let X be a random variable with distribution P, where $\mathbb{E}(|X|) < \infty$. For a fixed number $r \in \mathbb{R}$ let*

$$H(r) \;:=\; \mathbb{E}(|X - r|),$$

the mean distance between X and r. This defines a convex function H with limits $H(\pm\infty) = \infty$. Moreover, r minimises H if and only if r is a median of P.

The 'mail box problem'. Lemma 3.1 provides a solution of the following problem: Along a street there are n houses at positions $x_1 < x_2 < \cdots < x_n$. Now we want to place a mail box at position r such that the average distance

$$\frac{1}{n}\sum_{i=1}^{n} |x_i - r|$$

between a house and the mail box becomes minimal. For odd n, the unique optimal position is equal to $x_{(n+1)/2}$. For even n, any position within $[x_{n/2}, x_{n/2+1}]$ is optimal.
This is a consequence of Lemma 3.1 if we consider a random variable X with $\mathbb{P}(X = x_i) = 1/n$ for $1 \le i \le n$. Alternatively one could argue directly: Imagine that the mail box is currently at position r. Now imagine that the mail box is moved a small distance δ to the left or to the right and what this movement would do to the average of all distances. This type of consideration is also useful for solving Exercise 3.4.

Proof of Lemma 3.1. It follows from the triangle inequality that

$$|X - r| \;\le\; |X| + |r| \quad \text{and} \quad |X - r| \;\ge\; |r| - |X|,$$

whence

$$|r| - \mathbb{E}(|X|) \;\le\; H(r) \;\le\; |r| + \mathbb{E}(|X|).$$

In particular, $H(r) \to \infty$ as $|r| \to \infty$.
Convexity of H follows essentially from the fact that $h(x, r) := |x - r|$ is convex in $r \in \mathbb{R}$ for any fixed $x \in \mathbb{R}$: For $r, s \in \mathbb{R}$ and $0 < \lambda < 1$,

$$\begin{aligned} H((1-\lambda)r + \lambda s) \;&=\; \mathbb{E}\big(h(X, (1-\lambda)r + \lambda s)\big) \\ &\le\; \mathbb{E}\big((1-\lambda)h(X, r) + \lambda h(X, s)\big) \;=\; (1-\lambda)H(r) + \lambda H(s). \end{aligned}$$

Now we consider right- and left-sided derivatives of H: For $r < s$,

$$\frac{H(s) - H(r)}{s - r} \;=\; \mathbb{E}\, h(X, r, s)$$

with

$$h(x, r, s) := \frac{|x - s| - |x - r|}{s - r} \;=\; \begin{cases} 1 & \text{if } x \le r, \\ \dfrac{s + r - 2x}{s - r} & \text{if } r \le x \le s, \\ -1 & \text{if } x \ge s. \end{cases}$$

Since $|h(x, r, s)| \le 1$, we may interchange limits and expectation and obtain the formulae

$$\begin{aligned} H'(s-) \;&=\; \mathbb{E}\Big(\lim_{r \uparrow s} h(X, r, s)\Big) \;=\; \mathbb{P}(X < s) - \mathbb{P}(X \ge s) \;=\; 2\,\mathbb{P}(X < s) - 1, \\ H'(r+) \;&=\; \mathbb{E}\Big(\lim_{s \downarrow r} h(X, r, s)\Big) \;=\; \mathbb{P}(X \le r) - \mathbb{P}(X > r) \;=\; 2\,\mathbb{P}(X \le r) - 1. \end{aligned}$$

Note that r is a minimiser of H if and only if $H'(r+) \ge 0$ and $H(r-) \le 0$. This is equivalent to $\mathbb{P}(X \le r) \ge 1/2$ and $\mathbb{P}(X < r) \le 1/2$. In other words, r has to be a median of P. □

Empirical distribution function and order statistics. An estimator for F is the *empirical distribution function* $\widehat{F}$,

$$\widehat{F}(x) \;:=\; \widehat{P}((-\infty, x]) = \#\{i \le n : X_i \le x\}/n.$$

This is a non-decreasing step function. Precisely, let

$$X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$$

be the ordered values X_i. We call $X_{(i)}$ the *i-th order statistic* of our data. Then

$$\widehat{F}(x) \;=\; \frac{i}{n} \quad \text{for } X_{(i)} \le x < X_{(i+1)},\ 0 \le i \le n,$$

where $X_{(0)} := -\infty$ and $X_{(n+1)} := \infty$.

Sample quantiles. By means of the order statistics it is easy to determine *sample quantiles*: A number $\widehat{q}_\gamma$ is called *sample γ-quantile*, if it is a γ-quantile of the empirical distribution $\widehat{P}$. If $n\gamma$ is not an integer,

$$\widehat{q}_\gamma \;=\; X_{(\lceil n\gamma \rceil)}$$

is the unique sample γ-quantile. If $n\gamma$ is an integer, then any number

$$\widehat{q}_\gamma \;\in\; \left[X_{(n\gamma)}, X_{(n\gamma+1)}\right]$$

is a sample γ-quantile. If we talk about *the* sample γ-quantile, then we refer to

$$\widehat{q}_\gamma \;:=\; \left(X_{(\lceil n\gamma \rceil)} + X_{(\lfloor n\gamma+1 \rfloor)}\right)/2.$$

In particular, for $\gamma = 0.5$ we obtain *the* sample median.

The previous definition of sample quantiles is just one out of many proposals. For instance, the statistics software R provides nine different variants; the one described here is type 2.

Example 3.2 (Monthly rents). In Example 1.18 students have been asked about their monthly rent (in CHF). Now we consider the population of all students at the University of Bern in the academic year 2003/2004 which did *not* live for free with their parents or other relatives. Our sample contained $n = 129$ such students, and now we treat these as a random sample from the aforementioned population. Hence we estimate the proportion of students having to pay at most x in the whole population by the corresponding proportion $\widehat{F}(x)$ in our sample. The smallest and largest values are $X_{(1)} = 220\,\text{CHF}$ and $X_{(129)} = 2000\,\text{CHF}$, respectively. Figure 3.1 depicts the empirical distribution function. The graph of $\widehat{F}$ has been augmented by vertical segments at the order statistics $X_{(i)}$. Moreover one sees a horizontal line at height 0.5, and this line is crossed by the graph of $\widehat{F}$ at $\widehat{q}_{0.5} = X_{(65)} = 550\,\text{CHF}$ (middle vertical line).

Ranks. In some statistical procedures the original data X_i are replaced with their *ranks* which are defined as follows:

Suppose that all n values X_i are different. Then we define $R_i := k$ if $X_i = X_{(k)}$. One can also write

$$R_i \;=\; \#\{\ell : X_\ell \le X_i\} \;=\; n\widehat{F}(X_i)$$

or

$$R_i \;=\; \#\{\ell : X_\ell < X_i\} + 1.$$

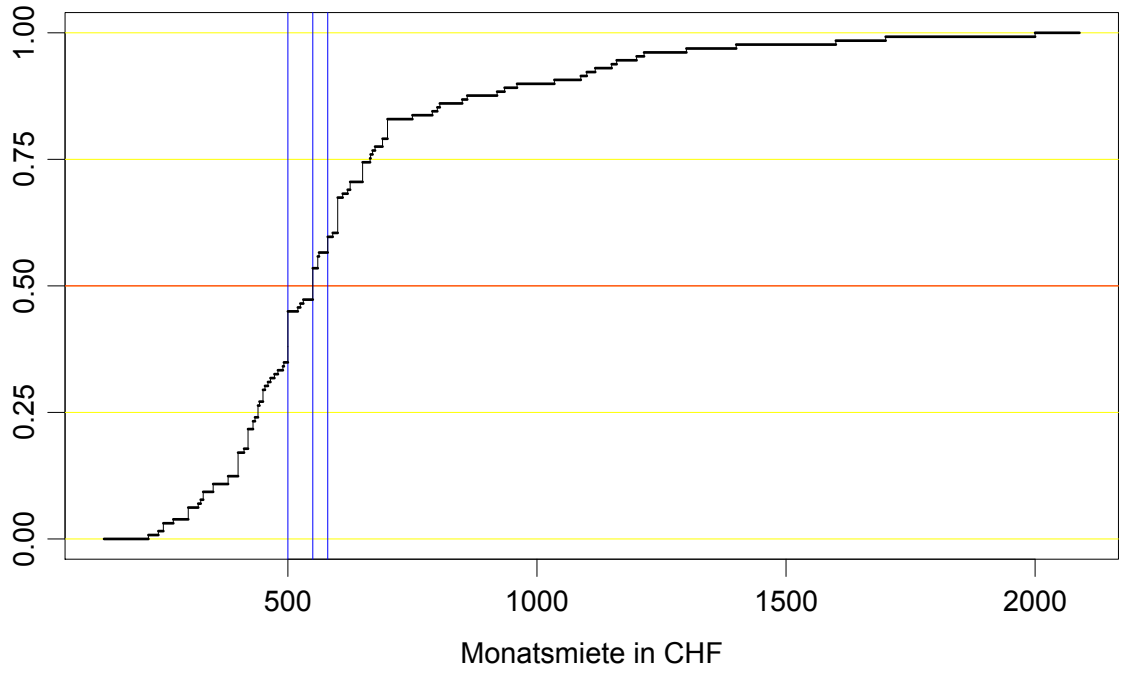

Figure 3.1: Empirical distribution function, sample median and a 95%-confidence interval for $q_{0.5}$.

Here the resulting vector of ranks, $(R_i)_{i=1}^n$, is a permutation of $(i)_{i=1}^n$.
If some values X_i are identical, one uses mean ranks: To the order statistics $X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$ correspond the ranks $1, 2, \ldots, n$. However, if

$$X_{(j-1)} \;<\; X_{(j)} = X_{(j+1)} = \cdots = X_{(k)} \;<\; X_{(k+1)}$$

for certain indices $1 \le j < k \le n$, then we define

$$R_i \;:=\; \frac{j + (j+1) + \cdots + k}{k - j + 1} \;=\; \frac{j+k}{2}$$

for all indices i with $X_i = X_{(j)}$. One can also write

$$\begin{aligned} R_i \;&=\; \big(\#\{\ell : X_\ell < X_i\} + 1 + \#\{\ell : X_\ell \le X_i\}\big)/2 \\ &=\; \big(n\widehat{F}(X_i -) + 1 + n\widehat{F}(X_i)\big)/2 \end{aligned}$$

or

$$R_i \;=\; \#\{\ell : X_\ell < X_i\} + \big(1 + \#\{\ell : X_\ell = X_i\}\big)/2.$$

Note also that the sum of all ranks is always equal to

$$\sum_{i=1}^n R_i \;=\; \sum_{i=1}^n i \;=\; \frac{n(n+1)}{2}.$$

Remark 3.3. If the distribution function F is continuous, the n random variables $X_1, X_2, \ldots, X_n$ are almost surely different, whence $X_{(1)} < X_{(2)} < \cdots < X_{(n)}$. This is equivalent to saying that $\mathbb{P}(X_i = X_j) = \mathbb{P}(X_1 = X_2) = 0$ for arbitrary indices $1 \le i < j \le n$. Here is an elementary proof of the latter equality: For an arbitrary integer $k \ge 2$ we choose real numbers $a_{k,1} < a_{k,2} < \cdots < a_{k,k-1}$ with $F(a_{k,\ell}) = \ell/k$. Setting $a_{k,0} := -\infty$ and $a_{k,k} := \infty$, the

intervals $I_{k,\ell} := (a_{k,\ell-1}, a_{k,\ell}]$, $1 \le \ell \le k$, satisfy the equalities $P(I_{k,\ell}) = k^{-1}$, $1 \le i \le n$. In particular,

$$\begin{aligned}\mathbb{P}(X_1 = X_2) \;&=\; \sum_{\ell=1}^{k} \mathbb{P}(X_1 = X_2 \in I_{k,\ell}) \\ &\le\; \sum_{\ell=1}^{k} \mathbb{P}(X_1 \in I_{k,\ell}, X_2 \in I_{k,\ell}) \;=\; \sum_{\ell=1}^{k} P(I_{k,\ell})^2 \;=\; k^{-1}.\end{aligned}$$

Since k may be chosen arbitrarily large, the probability on the left hand side equals 0.

By means of Fubini's Theorem (see Appendix) one can argue as follows:

$$\mathbb{P}(X_1 = X_2) \;=\; \mathbb{E}\big(\mathbb{P}(X_1 = X_2 \,|\, X_2)\big) \;=\; \mathbb{E}\big(P(\{X_2\})\big) \;=\; 0,$$

since $P(\{x\}) = F(x) - F(x-) = 0$ for arbitrary $x \in \mathbb{R}$. The notation $\mathbb{P}(X_1 = X_2 \,|\, X_2)$ means that X_2 is viewed temporarily as a fixed number, and only X_1 is random with distribution P.

3.3 Confidence Bounds for Quantiles

For quantiles there is a remarkably easy method to compute confidence bounds. For fixed indices $0 \le k < \ell \le n+1$ we view the random interval

$$[X_{(k)}, X_{(\ell)}]$$

as a confidence interval for q_γ. In case of $k = 0$ this corresponds to the upper confidence bound $X_{(\ell)}$, in case of $\ell = n+1$ we compute a lower confidence bound $X_{(k)}$, because $X_{(0)} = -\infty$ and $X_{(n+1)} = \infty$. Otherwise we get a compact confidence interval. The question is how to choose k and ℓ in order to guarantee that

$$\mathbb{P}\big(q_\gamma \in [X_{(k)}, X_{(\ell)}]\big) \;\ge\; 1 - \alpha.$$

Theorem 3.4. *Let q_γ be a γ-quantile of P. For arbitrary indices $0 \le k < \ell \le n+1$,*

(3.1) $\quad \mathbb{P}(X_{(\ell)} \ge q_\gamma) \;\ge\; F_{n,\gamma}(\ell - 1)$ $\qquad$ *with equality if $F(q_\gamma -) = \gamma$,*

(3.2) $\quad \mathbb{P}(X_{(k)} > q_\gamma) \;\le\; F_{n,\gamma}(k - 1)$ $\qquad$ *with equality if $F(q_\gamma) = \gamma$.*

Here $F_{n,\gamma}$ is the distribution function of $\mathrm{Bin}(n, \gamma)$. In particular,

$$\mathbb{P}\big(q_\gamma \in [X_{(k)}, X_{(\ell)}]\big) \;\ge\; F_{n,\gamma}(\ell - 1) - F_{n,\gamma}(k - 1)$$

with equality if F is continuous at q_γ.

Remark 3.5. Suppose that the γ-quantile is not unique, that means, $q_{\gamma,1} < q_{\gamma,2}$. Then $F = \gamma$ on $[q_{\gamma,1}, q_{\gamma,2})$, and the open interval $(q_{\gamma,1}, q_{\gamma,2})$ contains almost surely no observation X_i. Hence for an arbitrary point $q_\gamma \in (q_{\gamma,1}, q_{\gamma,2})$,

$$\begin{aligned}\mathbb{P}\big([q_{\gamma,1}, q_{\gamma,2}] \subset [X_{(k)}, X_{(\ell)}]\big) \;&=\; \mathbb{P}\big(q_\gamma \in [X_{(k)}, X_{(\ell)}]\big) \\ &=\; F_{n,\gamma}(\ell - 1) - F_{n,\gamma}(k - 1).\end{aligned}$$

Proof of Theorem 3.4. Note first that

$$\begin{aligned}\mathbb{P}\big(q_\gamma \in [X_{(k)}, X_{(\ell)}]\big) \;&=\; 1 - \mathbb{P}\big(q_\gamma \notin [X_{(k)}, X_{(\ell)}]\big) \\ &=\; 1 - \mathbb{P}\big(X_{(\ell)} < q_\gamma \text{ or } X_{(k)} > q_\gamma\big) \\ &=\; 1 - \mathbb{P}(X_{(\ell)} < q_\gamma) - \mathbb{P}(X_{(k)} > q_\gamma) \\ &=\; \mathbb{P}(X_{(\ell)} \ge q_\gamma) - \mathbb{P}(X_{(k)} > q_\gamma).\end{aligned}$$

Hence it suffices to prove (3.1) and (3.2).

As to (3.1), in case of $\ell = n+1$ we have $X_{(\ell)} = \infty$, whence both $\mathbb{P}(X_{(\ell)} \geq q_\gamma)$ and $F_{n,\gamma}(\ell-1)$ are equal to 1. Otherwise, $X_{(\ell)} \geq q_\gamma$ if and only if at most $\ell - 1$ observations X_i are strictly less than q_γ. Consequently,

$$\begin{aligned} \mathbb{P}(X_{(\ell)} \geq q_\gamma) &= \mathbb{P}\big(n\widehat{P}((-\infty, q_\gamma)) \leq \ell - 1\big) \\ &= F_{n,F(q_\gamma -)}(\ell - 1) \\ &\geq F_{n,\gamma}(\ell - 1). \end{aligned}$$

Here we used the fact that $n\widehat{P}((-\infty, q_\gamma))$ follows $\mathrm{Bin}(n, F(q_\gamma -))$, the inequality $F(q_\gamma -) \leq \gamma$ and Lemma 2.3. Equality holds if $F(q_\gamma -) = \gamma$.

Analogously one can prove (3.2): In case of $k = 0$ it follows from $X_{(k)} = -\infty$ that both $\mathbb{P}(X_{(k)} > q_\gamma)$ and $F_{n,\gamma}(k-1)$ are equal to 0. Otherwise,

$$\begin{aligned} \mathbb{P}(X_{(k)} > q_\gamma) &= \mathbb{P}\big(n\widehat{P}((-\infty, q_\gamma]) \leq k - 1\big) \\ &= F_{n,F(q_\gamma)}(k - 1) \\ &\leq F_{n,\gamma}(k - 1) \end{aligned}$$

with equality if $F(q_\gamma) = \gamma$. □

Application. To construct a $(1-\alpha)$-confidence interval for q_γ, one should choose indices $0 \leq k < \ell \leq n+1$ such that

$$F_{n,\gamma}(\ell - 1) - F_{n,\gamma}(k - 1) \geq 1 - \alpha. \tag{3.3}$$

This leads to the lower $(1-\alpha)$-confidence bound $X_{(k)}$ for q_γ with

$$k = k_\alpha(n, \gamma) := \max\big\{k \in \{0, 1, \ldots, n\} : F_{n,\gamma}(k-1) \leq \alpha\big\}$$

and the upper $(1-\alpha)$-confidence bound $X_{(\ell)}$ for q_γ with

$$\ell = \ell_\alpha(n, \gamma) := \min\big\{\ell \in \{1, 2, \ldots, n+1\} : F_{n,\gamma}(\ell - 1) \geq 1 - \alpha\big\}.$$

Symmetry considerations yield the equation

$$k_\alpha(n, \gamma) = n + 1 - \ell_\alpha(n, 1 - \gamma).$$

In particular,

$$k_\alpha(n, 0.5) = n + 1 - \ell_\alpha(n, 0.5).$$

A $(1-\alpha)$-confidence interval for q_γ is given by $[X_{(k)}, X_{(\ell)}]$ with the indices $k = k_{\alpha/2}(n, \gamma)$ and $\ell = \ell_{\alpha/2}(n, \gamma)$. But it may be possible to increase k or decrease ℓ without violating (3.3).

Example 3.6 (Monthly rents). In the previous example with $n = 129$ monthly rents of students we want to compute a 95%-confidence interval for the unknown median $q_{0.5}$. Since $F_{n,0.5}(52) < \alpha/2 = 2.5\% < F_{n,0.5}(53)$, we obtain the indices $k_{\alpha/2}(n, 0.5) = 53$ and $\ell_{\alpha/2}(n, 0.5) = n + 1 - k_{\alpha/2}(n, 0.5) = 77$. This leads to the 95%-confidence interval

$$[X_{(53)}, X_{(77)}] = [500\,\text{CHF}, 580\,\text{CHF}]$$

for $q_{0.5}$. The endpoints of this interval are depicted in Figure 3.1, too.

Remark 3.7. So far we know how to construct for fixed $x \in \mathbb{R}$ confidence bounds for $F(x)$, based on the random variable $H := n\widehat{F}(x) \sim \mathrm{Bin}(n, F(x))$. On the other hand we know how to construct confidence bounds for a single quantile q_γ. These two procedures are closely related. For if we consider the one-sided $(1-\alpha)$-confidence bounds

$$\begin{aligned} a_\alpha(x) = a_\alpha(x, \text{Daten}) &:= \inf\bigl\{p \in [0,1] : F_{n,p}(n\widehat{F}_n(x) - 1) < 1 - \alpha\bigr\}, \\ b_\alpha(x) = b_\alpha(x, \text{Daten}) &:= \sup\bigl\{p \in [0,1] : F_{n,p}(n\widehat{F}_n(x)) > \alpha\bigr\} \end{aligned}$$

for $F(x)$, then

$$\begin{aligned} X_{(k_\alpha(n,\gamma))} \le x &\text{ if and only if } b_\alpha(x) > \gamma, \\ X_{(\ell_\alpha(n,\gamma))} > x &\text{ if and only if } a_\alpha(x) < \gamma. \end{aligned}$$

The proof of these equivalences is left to the reader as an exercise.

Remark 3.8 (Distribution of order statistics). The proof of Theorem 3.4 yields an explicit formula for the distribution function of an arbitrary order statistic $X_{(k)}$. Namely,

$$\mathbb{P}(X_{(k)} \le x) \;=\; 1 - F_{n,F(x)}(k-1)$$

for arbitrary $k \in \{1, 2, \ldots, n\}$ and $x \in \mathbb{R}$. For $X_{(k)} \le x$ is equivalent to at least k observations X_i being less than or equal to x. Together with the second part of Lemma 2.3 this yields the formula

$$\mathbb{P}(X_{(k)} \le x) \;=\; \int_0^{F(x)} n\binom{n-1}{k-1} u^{k-1}(1-u)^{n-k}\, du.$$

3.4 Kolmogorov–Smirnov Confidence Bands

In this section we shall derive a $(1-\alpha)$-*confidence band for* F. Precisely, we'll show that for each sample size n and any $\alpha \in (0,1)$ there exists a constant $\kappa_{n,\alpha}$ with the following property:

$$\mathbb{P}_F\bigl(F(x) \in \bigl[\widehat{F}(x) \pm \kappa_{n,\alpha}\bigr] \cap [0,1] \text{ for all } x \in \mathbb{R}\bigr) \;\ge\; 1 - \alpha \tag{3.4}$$

for any distribution function F with equality in case of F being continuous. In other words,

$$\mathbb{P}_F\bigl(\|\widehat{F} - F\|_\infty \le \kappa_{n,\alpha}\bigr) \;\ge\; 1 - \alpha,$$

where $\|h\|_\infty := \sup_{x\in\mathbb{R}} |h(x)|$ denotes the supremum norm of a function $h : \mathbb{R} \to \mathbb{R}$. It will turn out that $\kappa_{n,\alpha}$ is of order $O(n^{-1/2})$ for fixed α. An important tool in this context are so-called quantile transformations which are also essential for computer simulations.

The quantile function. For $0 < u < 1$ let

$$F^{-1}(u) \;:=\; \min\bigl\{x \in \mathbb{R} : F(x) \ge u\bigr\}.$$

This number is well-defined due to the general properties of F. It is just the minimal u-quantile $q_{u,1}$ of the distribution P.

Example 3.9 (Distributions with finite support). For some $m \in \mathbb{N}$ and real numbers $x_1 < x_2 < \ldots < x_m$ let

$$p_i := P\{x_i\} \;>\; 0 \quad \text{for } i = 1, \ldots, m,$$

where $\sum_{i=1}^m p_i = 1$. Then

$$F(x) \;=\; \begin{cases} 0 & \text{for } x < x_1, \\ \sum_{i=1}^j p_i & \text{for } x_j \le x < x_{j+1} \text{ and } 1 \le j < m, \\ 1 & \text{for } x \ge x_m, \end{cases}$$

and

$$F^{-1}(u) \;=\; \begin{cases} x_1 & \text{if } 0 < u \le p_1, \\ x_k & \text{if } \sum_{i=1}^{k-1} p_i < u \le \sum_{i=1}^k p_i \text{ and } 1 < k \le m. \end{cases}$$

Example 3.10 (Exponential distributions). For $b > 0$ let

$$F_b(x) \;:=\; \max\bigl\{1 - e^{-x/b}, 0\bigr\},$$

the distribution function of the *exponential distribution with scale parameter (mean)* b. Here

$$F_b^{-1}(u) \;=\; -b \log(1-u)$$

for arbitrary $u \in (0,1)$.

Lemma 3.11 (Quantile transformation).

(a) *Let U be uniformly distributed on $[0,1]$, that means, $\mathbb{P}(U \in B) = \mathrm{length}(B)$ for arbitrary intervals $B \subset [0,1]$. Then*

$$X \;:=\; F^{-1}(U)$$

defines a random variable with distributions function F.

(b) *Let $U_1, U_2, \ldots, U_n$ be stochastically independent and uniformly distributed on $[0,1]$ with empirical distribution function $\widehat{F}_U$, i.e. $\widehat{F}_U(v) := \#\{i \le n : U_i \le v\}/n$. Then the random function $\mathbb{R} \ni x \mapsto \widehat{F}(x)$ has the same behaviour as the random function*

$$\mathbb{R} \ni x \;\mapsto\; \widehat{F}_U(F(x)).$$

In particular,

$$\mathbb{P}\bigl(\|\widehat{F} - F\|_\infty \le \kappa\bigr) \;\ge\; \mathbb{P}\Bigl(\sup_{v \in [0,1]} \bigl|\widehat{F}_U(v) - v\bigr| \le \kappa\Bigr)$$

for arbitrary $\kappa \ge 0$ with equality if F is continuous. Moreover, the right hand side is continuous in $\kappa \ge 0$.

Concerning part (a) one should mention that $\mathbb{P}(U = 0) = \mathbb{P}(U = 1) = 0$, so $X = F^{-1}(U)$ is well-defined in $\mathbb{R}$ almost surely. Part (a) shows a general recipe for transforming random variables with uniform distribution on $[0,1]$ into random variables with arbitrary given distribution (function). This recipe is used frequently in computer simulations, because computers provide pseudo random numbers with uniform distribution on $[0,1]$.

Proof of Lemma 3.11. The definition of F^{-1} implies the following fact: For arbitrary $x \in \mathbb{R}$ and $u \in (0,1)$,

$$F^{-1}(u) \le x \quad \text{if and only if} \quad F(x) \ge u.$$

This yields part (a), since

$$\mathbb{P}(X \le x) \;=\; \mathbb{P}(F^{-1}(U) \le x) \;=\; \mathbb{P}(U \le F(x)) \;=\; F(x).$$

As to part (b), according to part (a) the random vectors $(X_i)_{i=1}^n$ and $\big(F^{-1}(U_i)\big)_{i=1}^n$ are identically distributed. Thus the random function $\mathbb{R} \ni x \mapsto \widehat{F}(x)$ has the same bevavior as the random function

$$\mathbb{R} \ni x \;\mapsto\; \frac{1}{n}\sum_{i=1}^n 1_{[F^{-1}(U_i)\le x]} \;=\; \frac{1}{n}\sum_{i=1}^n 1_{[U_i\le F(x)]} \;=\; \widehat{F}_U(F(x)).$$

In particular, $\|\widehat{F} - F\|_\infty$ has the same distribution as

$$\sup_{v\in F(\mathbb{R})} \big|\widehat{F}_U(v) - v\big|,$$

und this is obviously not greater than

$$S \;:=\; \sup_{v\in[0,1]} \big|\widehat{F}_U(v) - v\big| \;=\; \sup_{v\in(0,1)} \big|\widehat{F}_U(v) - v\big|.$$

The latter equality follows from $\widehat{F}_U(0) = 0$ almost surely and $\widehat{F}_U(1) = 1$. If F is continuous, then $(0,1) \subset F(\mathbb{R}) \subset [0,1]$, so $\|\widehat{F} - F\|_\infty$ has exactly the same distribution as S.

It remains to show that $\mathbb{P}(S \le \kappa)$ is continuous in $\kappa \ge 0$. In other words, we have to show that $\mathbb{P}(S = \kappa) = 0$ for arbitrary $\kappa \ge 0$. To this end note first that $\widehat{F}_U(v) - v = i/n - v$ on each interval $[U_{(i)}, U_{(i+1)})$, $0 \le i \le n$. Here $U_{(1)} \le U_{(2)} \le \ldots \le U_{(n)}$ are the order statistics of the random variables $U_1, U_2, \ldots, U_n$, and $U_{(0)} := 0$, $U_{(n+1)} := 1$. This yields the equation

$$S \;=\; \max_{i=1,2,\ldots,n} \max\Big(\frac{i}{n} - U_{(i)}, U_{(i)} - \frac{i-1}{n}\Big).$$

In particular,

$$\mathbb{P}(S = \kappa) \;\le\; \sum_{i=1}^n \Big(\mathbb{P}\Big(U_{(i)} = \frac{i}{n} - \kappa\Big) + \mathbb{P}\Big(U_{(i)} = \frac{i-1}{n} + \kappa\Big)\Big) \;=\; 0,$$

because each order statistic $U_{(i)}$ has a continuous distribution function; see Remark 3.8. □

Inspired by part (a) of Lemma 3.11, G. Shorack and J. Wellner (1986) wrote the following little poem:

The Uniform Song

There are continuous distributions,
discrete ones too.
Some are heavy tailed,
and some are skew.
There are logistics and chi squares,
but these we will scorn,
'Cause the loviest of them all
is the Uniform!

Confidence bands. Part (b) of Lemma 3.11 leads to the aforementioned Kolmogorov-Smirnov confidence band[1] for F. Let

$$\kappa_{n,\alpha} \;:=\; \min\Big\{\kappa \ge 0 : \mathbb{P}\Big(\sup_{v\in[0,1]} \big|\widehat{F}_U(v) - v\big| \le \kappa\Big) = 1 - \alpha\Big\}.$$

[1] Andrei N. Kolmogorov (1903-1987) and Vladimir I. Smirnov (1887-1974): famous Russian mathematicians. Kolmogorov was a leading figure in the development of modern probability theory.

Then (3.4) is satisfied. In other words, with confidence $1-\alpha$ one may assume that the unknown graph F is contained within the confidence band

$$\big\{(x,y) : x \in \mathbb{R}, y \in \big[\widehat{F}(x) \pm \kappa_{n,\alpha}\big] \cap [0,1]\big\}.$$

Remark 3.12. The exact distribution of $\sup_{v\in[0,1]} \big|\widehat{F}_U(v) - v\big|$ is derived, for instance, in the monograph of G. Shorack and J. Wellner (1986). There one also finds the limits

$$\lim_{n\to\infty} \mathbb{P}\Big(\sqrt{n} \sup_{v\in[0,1]} \pm\big(\widehat{F}_U(v) - v\big) \geq \eta\Big) \;=\; \exp(-2\eta^2),$$

$$\lim_{n\to\infty} \mathbb{P}\Big(\sqrt{n} \sup_{v\in[0,1]} \big|\widehat{F}_U(v) - v\big| \geq \eta\Big) \;=\; 2\sum_{i=1}^{\infty} (-1)^{i-1} \exp(-2i^2\eta^2)$$

for arbitrary $\eta > 0$. Finally P. Massart (1990) showed that

$$\mathbb{P}\Big(\sup_{v\in[0,1]} \big|\widehat{F}_U(v) - v\big| \geq \kappa\Big) \;\leq\; 2\exp(-2n\kappa^2) \tag{3.5}$$

for arbitrary $n \in \mathbb{N}$ and $\kappa \geq 0$. This yields the inequality

$$\kappa_{n,\alpha} \;\leq\; \tilde{\kappa}_{n,\alpha} := \sqrt{\frac{\log(2/\alpha)}{2n}}.$$

This upper bound is remarkably accurate, so we use it in all our numerical examples as a proxy for $\kappa_{n,\alpha}$.

Further details about $\widehat{F}$ and a proof of a weaker version of (3.5) are provided in Section A.8 of the appendix.

Example 3.13 (Monthly rents). Figure 3.2 shows a 95%-confidence band for F in our data example with $n = 129$ monthly rents of student. Here we used $\tilde{\kappa}_{129,0.05} \approx 0.1196$.

Confidence bands and specific models. The set of all distribution functions G satisfying the inequality $\|\widehat{F} - G\|_\infty \leq \kappa_{n,\alpha}$ comprises a $(1-\alpha)$-confidence region for the unknown true distribution function F. Sometimes we have a particular model $(F_\theta)_{\theta\in\Theta}$ of distribution functions in mind. Then one could determine the set

$$\big\{\eta \in \Theta\colon \|\widehat{F} - F_\eta\|_\infty \leq \kappa_{n,\alpha}\big\}. \tag{3.6}$$

Under the assumption that $F = F_\theta$ for some unknown true parameter $\theta \in \Theta$, the set (3.6) is a $(1-\alpha)$-confidence region for θ. If this set is empty, we may claim with confidence $1-\alpha$ that the model is incorrect, that is, $F \notin \{F_\theta\colon \theta \in \Theta\}$.

The explicit computation of the confidence region (3.6) can be a challenging problem in itself. What is helpful is the fact that for any distribution function G,

$$\|\widehat{F} - G\|_\infty \;=\; \max_{i=1,\ldots,n} \max\Big(\frac{i}{n} - G(X_{(i)}), G(X_{(i)}-) - \frac{i-1}{n}\Big)$$

This can be verified as in the proof of Lemma 3.11 (b), noting that $\widehat{F} - G = i/n - G$ on $[X_{(i)}, X_{(i+1)})$ for $0 \leq i \leq n$. Hence the set (3.6) may be rewritten as

$$\Big\{\eta \in \Theta\colon F_\eta(X_{(i)}) \geq \frac{i}{n} - \kappa_{n,\alpha} \text{ and } F_\eta(X_{(i)}-) \leq \frac{i-1}{n} + \kappa_{n,\alpha} \text{ for } 1 \leq i \leq n\Big\}. \tag{3.7}$$

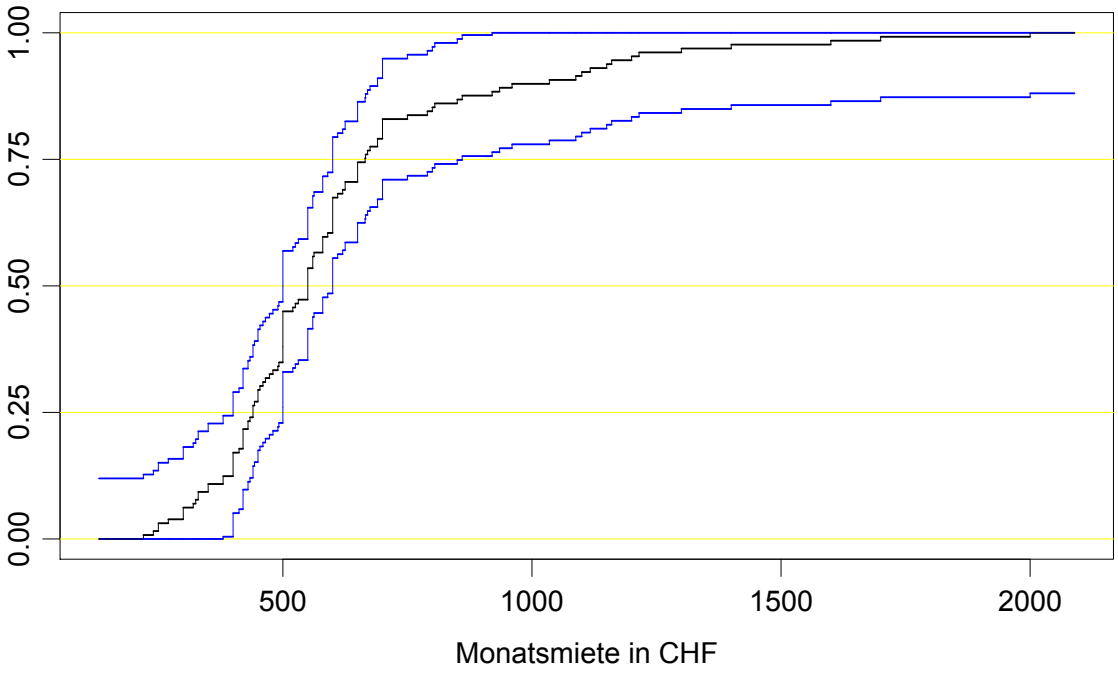

Figure 3.2: A Kolmogorov–Smirnov confidence band for F.

Example 3.14 (Body heights). Numerous empirical studies confirm that in many populations the variable 'body height' follows closely a Gaussian distribution, after dividing the population by gender. This approximation cannot be completely accurate if body heights are measured to whole centimeters only. (Also negative values are obviously impossible, though this is rather irrelevant if the mean is much higher than the standard deviation.) Hence a more precise model starts from a random variable with Gaussian distribution whose values have been rounded to whole centimeters. That results in the following model for the body height X of a randomly chosen person:

$$\mathbb{P}(X \le x) \;=\; \tilde{\Phi}_{\mu,\sigma}(x) := \Phi\Big(\frac{\lfloor x \rfloor + 0.5 - \mu}{\sigma}\Big)$$

for certain unknown parameters $\mu > 0$ und $\sigma > 0$. Whenever $\mu \gg \sigma$, the value of $\tilde{\Phi}_{\mu,\sigma}(0)$ is negligible. The recipe (3.6) and the alternative represention (3.7) lead to the following $(1-\alpha)$-confidence region for the unknown parameter $(\mu, \sigma) \in \mathbb{R} \times (0, \infty)$:

$$\begin{aligned}
C_\alpha \;&:=\; \big\{(m,s)\colon \|\widehat{F} - \tilde{\Phi}_{m,s}\|_\infty \le \kappa_{n,\alpha}\big\} \\
&=\; \Big\{(m,s)\colon \tilde{\Phi}_{m,s}(X_{(i)}) \ge \frac{i}{n} - \kappa_{n,\alpha} \\
&\qquad\qquad \text{and } \tilde{\Phi}_{m,s}(X_{(i)} - 1) \le \frac{i-1}{n} + \kappa_{n,\alpha} \text{ for } 1 \le i \le n\Big\} \\
&=\; \Big\{(m,s)\colon m + s\,\Phi^{-1}\Big(\frac{i}{n} - \kappa_{n,\alpha}\Big) \le X_{(i)} + 0.5 \\
&\qquad\qquad \text{and } m + s\,\Phi^{-1}\Big(\frac{i-1}{n} + \kappa_{n,\alpha}\Big) \ge X_{(i)} - 0.5 \text{ for } 1 \le i \le n\Big\},
\end{aligned}$$

where $\Phi^{-1}(u) := -\infty$ for $u \le 0$ and $\Phi^{-1}(u) := \infty$ for $u \ge 1$. The latter representation of C_α shows that it is the intersection of at most $2n$ halfplanes in $\mathbb{R} \times (0, \infty)$, that is, sets of the form $\{(m,s)\colon am + bs \le c\}$ with certain numbers $a = \pm 1$ and $b, c \in \mathbb{R}$.

As an explicit data example we consider the body heights of men in Example 1.18. This corresponds to a sample with $n = 145$ observations. Figure 3.3 shows contour lines of the function

$(m, s) \mapsto \|\widehat{F} - \tilde{\Phi}_{m,s}\|_\infty$. (Precisely we evaluated this function on a fine grid of 251×251 points within the displayed region and interpolated these values.) The smallest distance of 0.0311 was achieved for $(m, s) = (178.8, 6.39)$, marked by a star. This *minimum distance estimator* of (μ, σ) is an alternative to the more traditional estimator $(\overline{X}, S) = (178.94, 6.24)$ introduced and analyzed in Section 4.1. The fat line corresponds to all parameters (m, s) with $\|\widehat{F} - \tilde{\Phi}_{m,s}\|_\infty = \tilde{\kappa}_{n,0.05} \approx 0.1128$ and surrounds the confidence region $C_{0.05}$. Figure 3.4 shows the empirical distribution function $\widehat{F}$ and the Kolmogorov–Smirnov confidence band (fine lines) together with the estimated function $\tilde{\Phi}_{178.8,6.39}$ (highlighted).

By the way, the confidence region would be substantially smaller if we would have ignored the rounding errors and used the continuous distribution functions $\Phi_{m,s}(x) = \Phi((x-m)/s)$. But this would be an artefact of assuming a wrong model rather than an advantage.

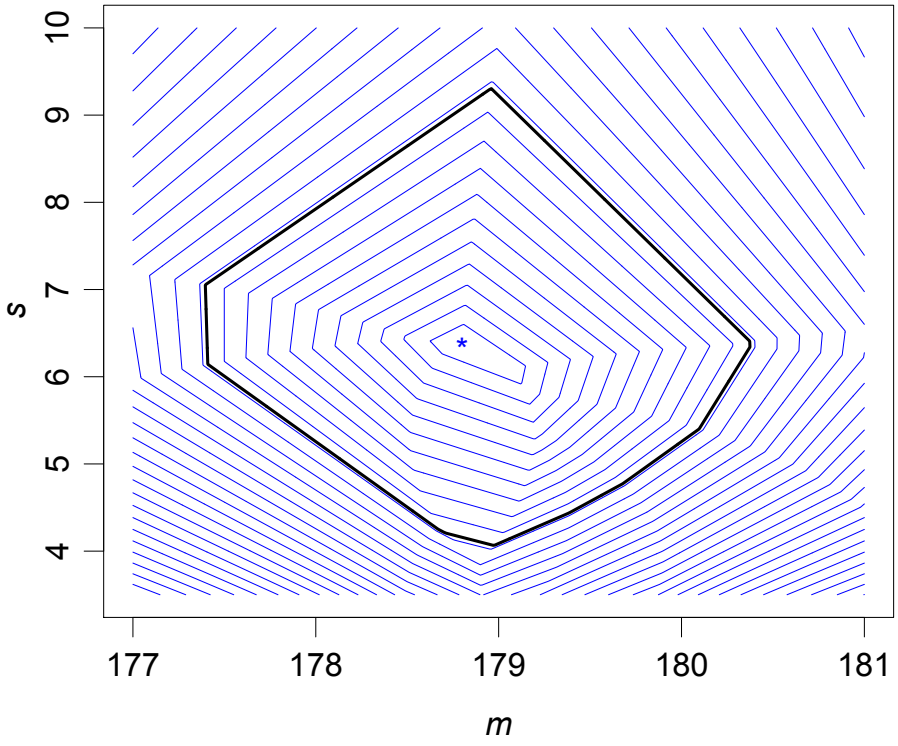

Figure 3.3: A Kolmogorov–Smirnov confidence region for (μ, σ).

3.5 Exercises

Exercise 3.1. Verify that the following functions are distribution functions, and determine the corresponding inverse functions $F^{-1} : (0, 1) \to \mathbb{R}$:

$$F_1(x) := \frac{e^x}{1+e^x}, \qquad F_2(x) := \exp(-\exp(-x)),$$

$$F_3(x) := \frac{1}{2} + \frac{x}{2\sqrt{1+x^2}}, \qquad F_4(x) := \begin{cases} 0 & \text{for } x \le 0, \\ 1 - (1+x^2)^{-b/2} & \text{for } x \ge 0, \quad \text{where } b > 0. \end{cases}$$

Exercise 3.2 (Tails of the standard normal distribution, I). For the distribution function Φ or quantile function Φ^{-1} of the standard normal distribution there are no closed formulae available. But show that

$$1 - \Phi(x) \le \exp(-x^2/2)/2 \quad \text{for } x \ge 0.$$

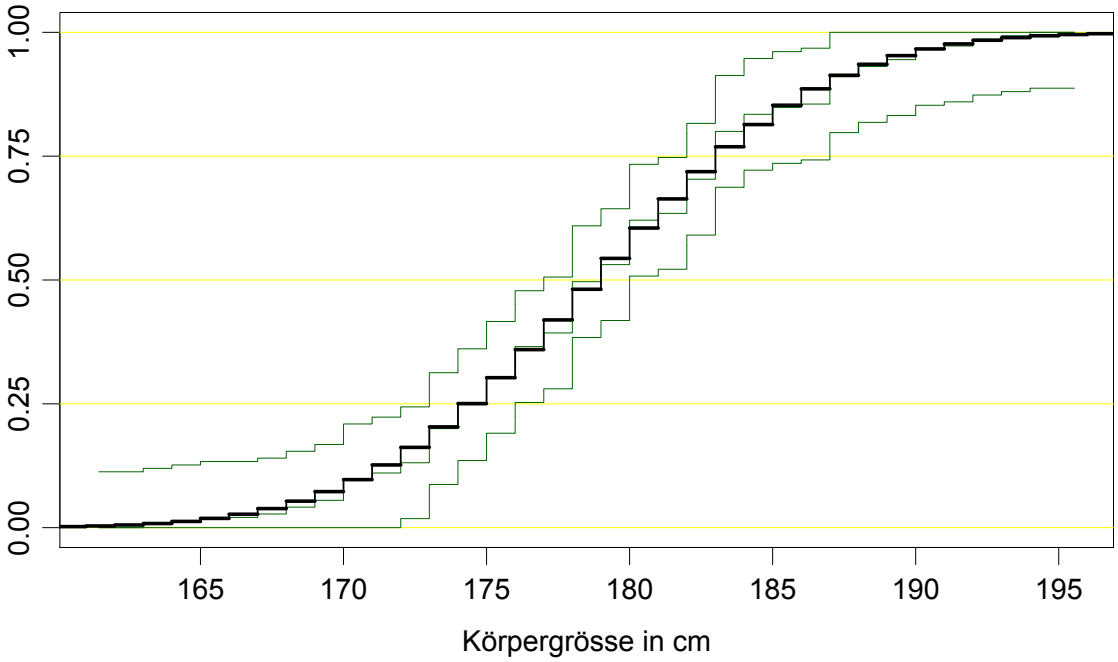

Figure 3.4: A Kolmogorov–Smirnov approximation.

Deduce from that the inequality

$$\Phi^{-1}(1-\alpha) \;\le\; \sqrt{-2\log(2\alpha)} \quad \text{for } 0 < \alpha \le 1/2.$$

Exercise 3.3 (Tails of the standard normal distribution, II). One can approximate $1-\Phi(x)$ quite well by expressions of the form $\phi(x)/h(x)$, where $\phi = \Phi'$ and $h : [0,\infty) \to (0,\infty)$ is differentiable and monotone increasing. Show that $\Delta := \phi/h - (1-\Phi)$ always satisfies the equations $\lim_{x\to\infty} \Delta(x) = 0$ and

$$\Delta'(x) \;=\; \frac{\phi(x)}{h(x)^2}\big(h(x)^2 - xh(x) - h'(x)\big).$$

Specifically, let $h_1(x) := x/2 + \sqrt{1 + x^2/4}$ and $h_2(x) := x/2 + \sqrt{2/\pi + x^2/4}$. Show that

$$\frac{\phi(x)}{h_1(x)} \;\le\; 1-\Phi(x) \;\le\; \frac{\phi(x)}{h_2(x)} \quad \text{for all } x \ge 0$$

Exercise 3.4. In a village with 33 houses, a mail box is to be placed such that the sum of all distances from a house to the mail box becomes minimal. Here we mean distances along the roads on the plan depicted in Figure 3.5. Show that there is precisely one optimal position for the mail box. (It is not necessary to measure distances explicitly. Just consider an arbitrary stretch of road without houses or intersections and determine how the total sum of distances would change if the mail box were positioned somewhere at that stretch and moved by a small amount.)

Exercise 3.5 (An example for quantiles). The bakery of Schilda has to spend an amount of h to produce a loaf of sunday bread, and it is sold at the price $v > h$. According to past experience, the demand X for this type of bread on a sunday morning (i.e. the number of potentially sold items) follows a certain distribution P on $\mathbb{N}_0$. Now the question is how many loafs of sunday bread the bakery should produce to maximise its expected net revenues. (The citizens of Schilda are picky and would never buy old sunday bread!) The result depends on the distribution P of X and the ratio h/v.

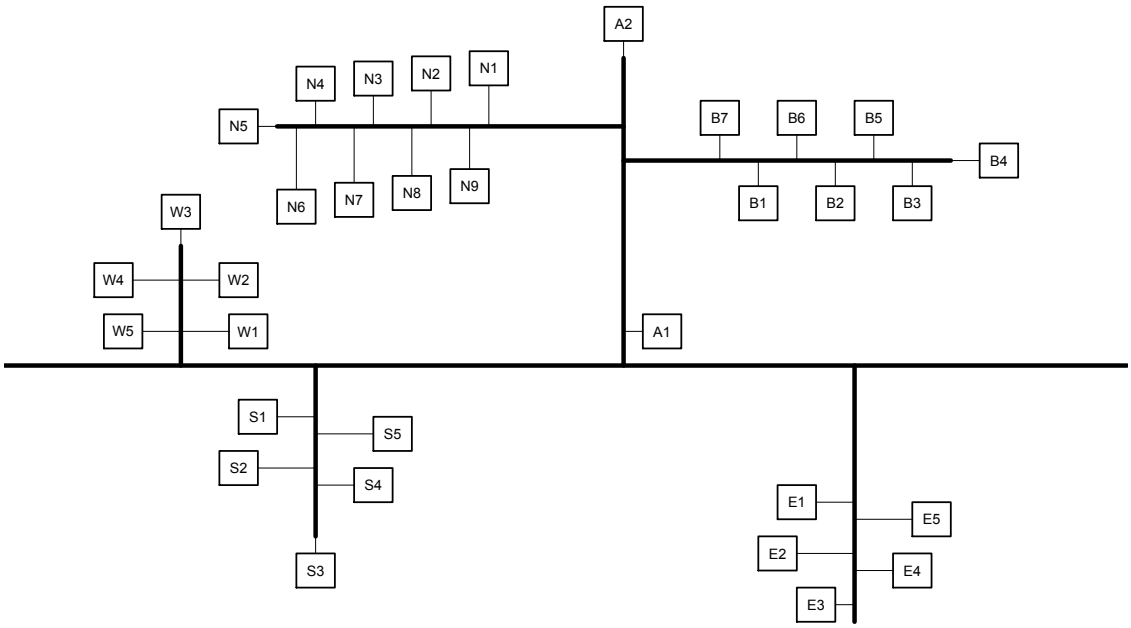

Figure 3.5: A village.

Exercise 3.6. Show that always

$$\sum_{i=1}^{n} R_i \;=\; n(n+1)/2 \quad \text{and} \quad \sum_{i=1}^{n} R_i^2 \;\leq\; n(n+1)(2n+1)/6$$

with equality if and only if the n numbers $X_1, X_2, \ldots, X_n$ are different.

Exercise 3.7. Let $(X_i)_{i=1}^n$ be an arbitrary random vector with n real-valued components satisfying the following two properties:

(i) With probability one, $X_1, X_2, \ldots, X_n$ are different;
(ii) For any permutation σ of $\{1, 2, \ldots, n\}$, the two vectors $(X_{\sigma(i)})_{i=1}^n$ and $(X_i)_{i=1}^n$ are identically distributed.

Show that $R : \{1, 2, \ldots, n\} \to \{1, 2, \ldots, n\}$ with $R(j) = \sum_{i=1}^n 1_{[X_i \leq X_j]}$ is uniformly distributed on the set of all $n!$ permutations of $\{1, 2, \ldots, n\}$.

Exercise 3.8. Show that

$$\mathbb{P}\big(q_{0.5} \in [X_{(1)}, X_{(n)}]\big) \;\geq\; 1 - 2^{1-n}$$

with equality if F is continuous at $q_{0.5}$. How large should n be such that this lower bound is at least 95%?

Exercise 3.9. Let $n \in \mathbb{N} \setminus (4\mathbb{N})$. What do you know about

$$\mathbb{P}\big(q_{0.5} \in [\widehat{q}_{0.25}, \widehat{q}_{0.75}]\big)\,?$$

Exercise 3.10. Suppose one wants to determine a confidence interval for the quantile q_γ based on a sample of size $n = 30$.

(a) Determine for $\gamma \in \{0.25, 0.5, 0.75\}$ *all* 'minimal' index pairs (k, ℓ) such that $[X_{(k)}, X_{(\ell)}]$ is a 90%-confidence interval for q_γ. Here 'minimal' means that (k, ℓ) could not be replaced by $(k+1, \ell)$ or $(k, \ell-1)$. Use Tables 3.1 and 3.2 or write a computer program to solve this task.

(b) Table 3.3 contains the life spans of $n = 30$ house cats in months (ordered values). Determine

a lower 90%-confidence bound for the median life span of house cats. Formulate your result in words.

x	$F_{30,0.25}(x)$	x	$F_{30,0.25}(x)$	x	$F_{30,0.25}(x)$	x	$F_{30,0.25}(x)$
0	0.0002	8	0.6736	16	0.9998	24	1.0000
1	0.0020	9	0.8034	17	0.9999	25	1.0000
2	0.0106	10	0.8943	18	1.0000	26	1.0000
3	0.0374	11	0.9493	19	1.0000	27	1.0000
4	0.0979	12	0.9784	20	1.0000	28	1.0000
5	0.2026	13	0.9918	21	1.0000	29	1.0000
6	0.3481	14	0.9973	22	1.0000	30	1.0000
7	0.5143	15	0.9992	23	1.0000		

Table 3.1: Distribution function $F_{30,0.25}$ of $\mathrm{Bin}(30, 0.25)$.

x	$F_{30,0.5}(x)$	x	$F_{30,0.5}(x)$	x	$F_{30,0.5}(x)$	x	$F_{30,0.5}(x)$
0	0.0000	8	0.0081	16	0.7077	24	0.9998
1	0.0000	9	0.0214	17	0.8192	25	1.0000
2	0.0000	10	0.0494	18	0.8998	26	1.0000
3	0.0000	11	0.1002	19	0.9506	27	1.0000
4	0.0000	12	0.1808	20	0.9786	28	1.0000
5	0.0002	13	0.2923	21	0.9919	29	1.0000
6	0.0007	14	0.4278	22	0.9974	30	1.0000
7	0.0026	15	0.5722	23	0.9993		

Table 3.2: Distribution function $F_{30,0.5}$ of $\mathrm{Bin}(30, 0.5)$.

66.6	89.5	103.2	122.5	140.0	148.4
70.5	96.1	106.2	127.0	140.6	160.1
77.1	96.6	106.9	127.2	143.0	167.7
84.4	97.7	112.0	129.0	144.0	182.0
88.4	102.0	122.2	129.1	145.8	189.0

Table 3.3: Life spans (in months) of $n = 30$ house cats.

Exercise 3.11. Write a program that returns for given $n \in \mathbb{N}$, $\gamma \in (0, 1)$ and $\alpha \in (0, 1)$ the two indices $k = k_\alpha(n, \gamma)$ and $\ell = \ell_\alpha(n, \gamma)$.

Exercise 3.12. In medical science, a measure for a person's weightyness is its 'body mass index'

$$\mathrm{BMI} := \frac{\text{body weight in kg}}{(\text{body height in m})^2}.$$

Persons with $20 \leq \mathrm{BMI} < 25$ are considered as normal, persons with $25 \leq \mathrm{BMI} < 30$ as potentially overweight, and persons with $\mathrm{BMI} \geq 30$ as potentially obese. (Note however that sportive people tend to have higher BMI because of increased muscle and bone mass.)

Obtain some data set containing body heights and weights of various people. Reflect which population these people could represent. Then determine point estimators and 90%-confidence intervals for the three quartiles $q_{0.25}$, $q_{0.5}$ and $q_{0.75}$.

Exercise 3.13 (Approximations for $k_\alpha(n,\gamma)$ and $\ell_\alpha(n,\gamma)$). Let $H \sim \mathrm{Bin}(n,\gamma)$. The Central Limit Theorem implies that $(H - n\gamma)/\sqrt{n\gamma(1-\gamma)}$ follows approximately a standard Gaussian distribution when $n\gamma(1-\gamma) \to \infty$. In particular,

$$F_{n,\gamma}(x) \;=\; \mathbb{P}(H \le x) = \mathbb{P}(H < x+1) \;\approx\; \Phi\Big(\frac{x + 1/2 - n\gamma}{\sqrt{n\gamma(1-\gamma)}}\Big)$$

for $x = 0, 1, \ldots, n$.

(a) Illustrate graphically that the 'continuity correction' $+1/2$ above is indeed reasonable. To this end, compare the exact value $F_{n,\gamma}(x)$ with the approximation $\Phi\big((x + s - np)/\sqrt{np(1-p)}\big)$ for $s = 0, 0.5, 1$.

(b) Use the approximation formula above to define approximations for $k_\alpha(n,\gamma)$ and $\ell_\alpha(n,\gamma)$. Compare these approximations with the exact indices.

Exercise 3.14. Let Y be a real-valued random variable with distribution function G and quantile function G^{-1}.

(a) Express the distribution function F and the quantile function F^{-1} of the following random variables in terms of G and G^{-1}:

(a.1) $X := \lceil Y \rceil$,
(a.2) $X := b^Y$ with $b > 1$,
(a.3) $X := \log_b(Y)$ with $b > 1$, where we assume that $Y > 0$.

(b) Suppose that G has a continuous density $g = G'$. Determine the density $f = F'$ for (a.2-3).

Exercise 3.15 (Kolmogorov–Smirnov band and quantiles). The confidence band for F implies confidence bounds for q_γ, *simultaneously for all* $\gamma \in (0,1)$: Determine indices $0 \le k(\gamma) < \ell(\gamma) \le n+1$, $\gamma \in (0,1)$, such that $\|\widehat{F} - F\|_\infty \le \kappa_{n,\alpha}$ implies that

$$q_\gamma \;\in\; \big[X_{(k(\gamma))}, X_{(\ell(\gamma))}\big] \quad \text{for } 0 < \gamma < 1.$$

Exercise 3.16 (Monte-Carlo simulations of the Kolmogorov–Smirnov statistic). Write a program which yields for given parameters $n \in \mathbb{N}$, $\alpha \in (0,1)$ and $m \in \mathbb{N}$ a Monte-Carlo estimate for the $(1-\alpha)$-quantile of

$$S \;:=\; \sup_{v \in [0,1]} \big|\widehat{F}_U(v) - v\big|$$

in m simulations. Use the special representation of S in the proof of Lemma 3.11.

Exercise 3.17 (Smallest value of the Kolmogorov–Smirnov statistic). Show that the random variable $S = \sup_{v\in[0,1]}\big|\widehat{F}_U(v) - v\big|$ in Lemma 3.11 satisfies

$$\mathbb{P}(S \le \kappa) \;\begin{cases} = \; 0 & \text{if } \kappa \le (2n)^{-1}, \\ > \; 0 & \text{if } \kappa > (2n)^{-1}. \end{cases}$$

Exercise 3.18 (Sample size determination for Kolmogorov–Smirnov bands). Determine a sample size n by means of Massart's inequality (3.5) such that the empirical distribution function $\widehat{F}$ satisfies

$$\big\|\widehat{F}_n - F\big\|_\infty \;\ge\; 0.01$$

with probability at most 0.01.

On the other hand, determine a small constant $\kappa > 0$ (up to five digits) with the property that $\big\|\widehat{F}_n - F\big\|_\infty \ge \kappa$ with probability at most κ, provided that $n \ge 40'000$.

Chapter 4

Numerical Variables: Means and Other Features

4.1 Means and Standard Deviations

This section is primarily about estimation of the *mean* $\mu = \mathbb{E}(X_i)$ of the distribution P, that is the number

$$\mu = \mu(P) \;=\; \int x\, P(dx).$$

For a general function h the integral $\int h(x)\, P(dx)$ stands for the number $\sum_x h(x) \cdot P(\{x\})$ in case of P being a discrete distribution or $\int_{-\infty}^{\infty} h(x) f(x)\, dx$ in case of P having density function f.

We always assume that $\mathbb{E}(X_i^2) < \infty$ which implies $\mathbb{E}(|X_i|)$ being finite, too. Closely related to the mean is the *variance* $\sigma^2 = \mathbb{E}((X_i - \mu)^2)$ of the distribution P. It may be written as

$$\sigma^2 = \sigma(P)^2 \;=\; \int (x-\mu)^2\, P(dx) \;=\; \int x^2\, P(dx) - \mu^2.$$

The *standard deviation* of the distribution P is the square root $\sigma = \sigma(P)$ of the variance.

The mean μ is of interest, for instance, in the following situations:

- The data X_i are the values of a numerical variable within a random sample from a certain population. Then μ is the arithmetic mean of this variable over the whole population. Here, σ^2 is the arithmetic mean of the squared difference between this variable and μ over the whole population.
- The data X_i are repeated measurements with a certain measurement device to determine an unknown parameter μ. The measurement device works correctly if there are no systematic errors, that means, if each single measurement has mean μ. Then σ quantifies the inaccuracy of a single measurement.

A prediction problem. Before we discuss the estimation of μ and σ, let us motivate these quantities by means of a prediction problem. Suppose for the moment that the distribution P is known. We want to predict the value of a future observation X with distribution P by a fixed real number as precisely as possible. One may interpret 'precisely as possible' in many ways. The two most popular are:

- Minimizing the mean absolute prediction error

$$\mathbb{E}(|X - r|).$$

This criterion has been looked at in Lemma 3.1 already, and the median $q_{0.5} = q_{0.5}(P)$ turned out to be the optimal prediction.

- Minimizing the mean squared prediction error

$$\mathbb{E}((X-r)^2).$$

The equation $\mathbb{E}((X-r)^2) = \mathrm{Var}(X) + (r - \mathbb{E}(X))^2 = \sigma^2 + (r-\mu)^2$ shows that the optimal prediction is given by $r = \mu$. The resulting mean squared prediction error is just the variance σ^2.

Point estimation of μ and σ

A canonical estimator for the mean μ of P is the mean of the empirical distribution $\widehat{P}$, and this leads to the *sample mean*

$$\mu(\widehat{P}) \;=\; \frac{1}{n}\sum_{i=1}^{n} X_i \;=:\; \overline{X}.$$

This estimator is unbiased, and its precision increases with n:

$$\mathbb{E}(\overline{X}) \;=\; \mu \quad \text{and} \quad \mathbb{E}\big((\overline{X}-\mu)^2\big) \;=\; \frac{\sigma^2}{n}.$$

In particular, $\mathbb{E}\big(|\overline{X}-\mu|\big) \le \sqrt{\mathbb{E}\big((\overline{X}-\mu)^2\big)} = \sigma/\sqrt{n}$.

Usually not only the mean but also the standard deviation σ and the variance σ^2 are unknown, so we need an estimator for the latter quantities. Again one could replace the unknown distribution P with the empirical distribution $\widehat{P}$. That means, we could estimate the variance σ^2 by

$$\sigma^2(\widehat{P}) \;=\; \frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2 \;=\; \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2.$$

This value, however, is systematically too small. It follows from Exercise 4.1 that the *sample variance*

$$S^2 \;:=\; \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2 \;=\; \frac{1}{n-1}\Big(\sum_{i=1}^{n} X_i^2 - n\overline{X}^2\Big)$$

is an unbiased estimator for the variance σ^2. For large sample sizes, the correction factor $n/(n-1)$ has almost no impact, but for smaller samples it is relevant. The square root S is the so-called *sample standard deviation* and serves as an estimator for the true standard deviation σ.

Both S^2 and S are consistent estimators for σ^2 and σ, respectively. Precisely,

$$\lim_{n\to\infty} \mathbb{E}\big(|S^2-\sigma^2|\big) \;=\; 0 \;=\; \lim_{n\to\infty} \mathbb{E}(|S-\sigma|).$$

This is a consequence of the following version of the weak law of large numbers: For independent, identically distributed random variables $Y_1, Y_2, Y_3, \ldots$ with expected value $\nu \in \mathbb{R}$,

$$\lim_{n\to\infty} \mathbb{E}\big(|\overline{Y}-\nu|\big) \;=\; 0,$$

where $\overline{Y} := n^{-1}\sum_{i=1}^{n} Y_i$. We apply this fact to the random variables $Y_i := (X_i-\mu)^2$ with

expectation $\nu = \sigma^2$: Indeed,

$$\begin{aligned} S^2 - \sigma^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 - \sigma^2 \\ &= \frac{1}{n-1}\sum_{i=1}^{n}\big((X_i - \mu) - (\overline{X} - \mu)\big)^2 - \sigma^2 \\ &= \frac{1}{n-1}\sum_{i=1}^{n} Y_i - \frac{n}{n-1}(\overline{X} - \mu)^2 - \sigma^2 \\ &= \frac{n}{n-1}\big(\overline{Y} - \nu\big) - \frac{n}{n-1}(\overline{X} - \mu)^2 + \frac{\sigma^2}{n-1}. \end{aligned}$$

Hence

$$\begin{aligned} \mathbb{E}\big(|S^2 - \sigma^2|\big) &\le \frac{n}{n-1}\,\mathbb{E}\big(|\overline{Y} - \nu|\big) + \frac{n}{n-1}\,\mathbb{E}\big((\overline{X} - \mu)^2\big) + \frac{\sigma^2}{n-1} \\ &= \frac{n}{n-1}\,\mathbb{E}\big(|\overline{Y} - \nu|\big) + \frac{2\sigma^2}{n-1}, \end{aligned}$$

and this converges to 0 as $n \to \infty$. Moreover,

$$\mathbb{E}(|S - \sigma|) = \mathbb{E}\Big(\frac{|S^2 - \sigma^2|}{S + \sigma}\Big) \le \frac{\mathbb{E}\big(|S^2 - \sigma^2|\big)}{\sigma}.$$

Z-Confidence Bounds for μ

To construct confidence bounds for μ we consider the standardised quantity

$$Z := \frac{\overline{X} - \mathbb{E}(\overline{X})}{\mathrm{Std}(\overline{X})} = \frac{\sqrt{n}(\overline{X} - \mu)}{\sigma}.$$

This random variable Z has mean zero and standard deviation one. Furthermore, it follows from the Central Limit Theorem that for large n it follows approximately a standard Gaussian distribution:

$$\lim_{n\to\infty} \mathbb{P}(r \le Z \le s) = \Phi(s) - \Phi(r)$$

for arbitrary $-\infty \le r < s \le \infty$. If P is a Gaussian distribution itself, i.e. $P = \mathcal{N}(\mu, \sigma^2)$, then Z is a standard Gaussian random variable for any sample size n; see Theorem 4.3. The inequalities $r \le Z \le s$ are equivalent to

$$\overline{X} - \frac{\sigma}{\sqrt{n}}\,s \le \mu \le \overline{X} - \frac{\sigma}{\sqrt{n}}\,r.$$

Hence, if the standard deviation σ is known, we obtain the following confidence regions for μ: The upper confidence bound

$$\overline{X} + \frac{\sigma}{\sqrt{n}}\Phi^{-1}(1-\alpha),$$

the lower confidence bound

$$\overline{X} - \frac{\sigma}{\sqrt{n}}\Phi^{-1}(1-\alpha)$$

or the confidence interval

$$\Big[\overline{X} \pm \frac{\sigma}{\sqrt{n}}\Phi^{-1}(1-\alpha/2)\Big].$$

The confidence level is approximately equal to $1-\alpha$ for large n. In case of observations X_i with normal distribution the confidence level is precisely $1-\alpha$. If we knew only an upper bound $\overline{\sigma}$ for σ, we could replace σ in the bounds above with $\overline{\sigma}$. The resulting confidence level would be (precisely or approximately) at least $1-\alpha$.

Example 4.1 (Measurement errors). Consider a scale which shows a measurement X when putting a probe of weight μ on it. Suppose that from extensive test series it is known or at least plausible that X follows a Gaussian distribution with (unknown) mean μ and known standard deviation $\sigma > 0$. Repeating this measurement n times independently yields independent random variables $X_1, X_2, \ldots, X_n$ with distribution $P = \mathcal{N}(\mu, \sigma^2)$. The standard deviation $\sigma > 0$ is a known feature of the scale which quantifies the uncertainty of a single measurement. The sample mean $\overline{X}$ has the property that

$$\begin{aligned}&\mathbb{P}[\overline{X} \text{ deviates from } \mu \text{ by more than } c]\\ &= \mathbb{P}\big(|\overline{X}-\mu| > c\big) = \mathbb{P}\Big(|Z| > \frac{\sqrt{n}\,c}{\sigma}\Big) = 2\Big(1-\Phi\Big(\frac{\sqrt{n}\,c}{\sigma}\Big)\Big).\end{aligned}$$

In other words, with confidence $1-\alpha$, μ lies within the interval $[\overline{X} \pm c]$, where $c := \sigma\Phi^{-1}(1-\alpha/2)/\sqrt{n}$. If one specifies a value α and a precision c, one can determine a minimal required sample size n as follows:

$$n \geq \frac{\sigma^2\Phi^{-1}(1-\alpha/2)^2}{c^2}.$$

Student Confidence Bounds for μ

The assumption that σ is known is rarely met. An obvious way out is to replace σ in the definition of Z with the sample standard deviation S and to consider the standardised quantity

$$T := \frac{\sqrt{n}(\overline{X}-\mu)}{S}.$$

In other words, one replaces the unknown standard deviation $\sigma/\sqrt{n}$ of $\overline{X}$ with the so-called *standard error* $S/\sqrt{n}$. Indeed, for large sample sizes n, T follows approximately a standard Gaussian distribution, too, because $\mathbb{E}\,|S/\sigma - 1| \to 0$ as $n \to \infty$. The question is, however, what the precise impact of estimating σ is for fixed n.

W. S. Gosset[1] investigated this question for Gaussian observations X_i. Upon request of his employer, he published his results under the pseudonym 'student' and introduced a new class of distributions:

Definition 4.2 (Student's t distributions). Let $Z_0, Z_1, Z_2, \ldots, Z_k$ be stochastically independent and standard Gaussian random variables. *Student's t distribution (student distribution, t distribution) with k degrees of freedom* is defined as the distribution of

$$Z_0 \Big/ \sqrt{\frac{1}{k}\sum_{i=1}^{k} Z_i^2}.$$

The usual symbol for this distribution is t_k. Its β-quantile is denoted with $t_{k;\beta}$.

Remarks on t_k. Student's t distribution has a density given by

$$f_k(x) = C_k(1+x^2/k)^{-(k+1)/2}$$

[1] William S. Gosset (1876-1937): british statistician, employee of the Guinness brewery in Dublin.

with a certain normalising constant $C_k > 0$. The main important fact for us is that f_k is also a bell-shaped function and symmetric around zero. This symmetry implies that $t_{k;1/2} = 0$ and

$$t_{k;1-\beta} \;=\; -t_{k;\beta}.$$

For explicit values one needs suitable software or tables.

The density function f_k is derived in the appendix. There it is also shown that $f_k(0)$ is strictly increasing in k, that $\lim_{k\to\infty} f_k(x) = \phi(x)$ for arbitrary $x \in \mathbb{R}$, and that for $1/2 < \beta < 1$,

$$t_{1;\beta} \;>\; t_{2;\beta} \;>\; t_{3;\beta} \;>\; \cdots \quad \text{with} \quad \lim_{k\to\infty} t_{k;\beta} \;=\; \Phi^{-1}(\beta).$$

Figure 4.1 shows the density functions f_k for $k = 1, 2, 3, 4$ as well as the standard Gaussian density ϕ, where $f_1(0) < f_2(0) < f_3(0) < f_4(0) < \phi(0)$.

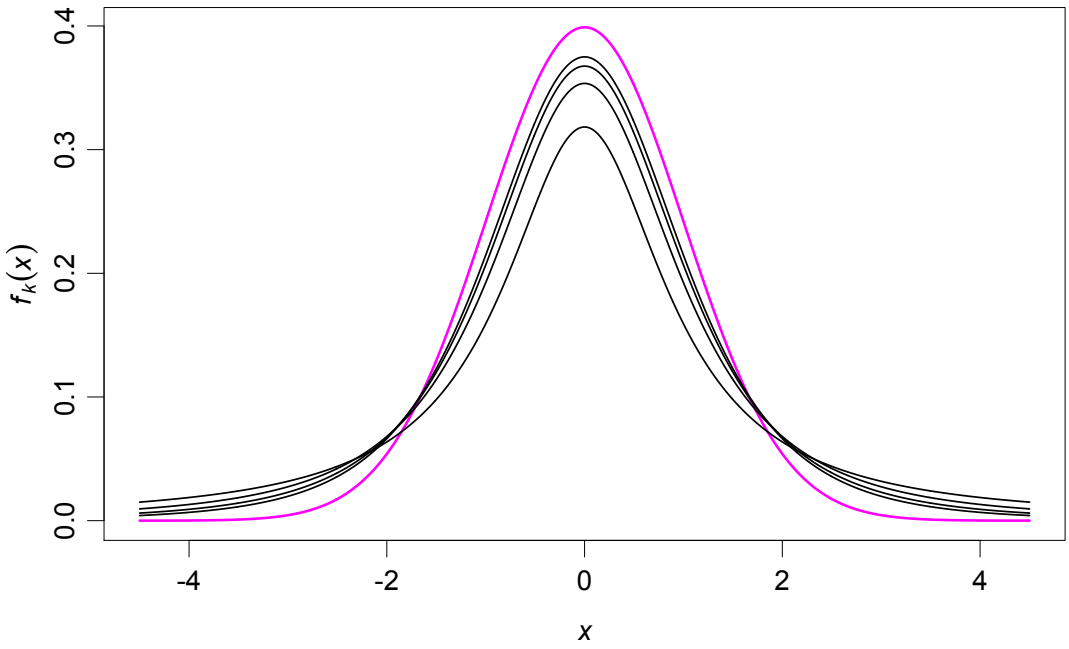

Figure 4.1: Density functions of t_1, t_2, t_3, t_4 and $\mathcal{N}(0, 1)$.

These t and also chi-squared distributions come into play via the following result:

Theorem 4.3 (W. Gosset, R.A. Fisher). *Let $X_1, X_2, \ldots, X_n$ be stochastically independent random variables with distribution $\mathcal{N}(\mu, \sigma^2)$. Then the random pair*

$$\left(\frac{\sqrt{n}(\overline{X} - \mu)}{\sigma}, \frac{S}{\sigma} \right)$$

has the same distribution as

$$\left(Z_1, \sqrt{\frac{1}{n-1} \sum_{i=2}^{n} Z_i^2} \right)$$

with independent, standard Gaussian random variables $Z_1, Z_2, \ldots, Z_n$. In particular, $T = \sqrt{n}(\overline{X} - \mu)/S$ has distribution t_{n-1}, and $(n-1)S^2/\sigma^2$ has distribution χ^2_{n-1}.

Confidence bounds for μ. In case of Gaussian observations X_i, the auxiliary quantity $T = \sqrt{n}(\overline{X} - \mu)/S$ follows a student distribution with $n-1$ degrees of freedom, whence

$$\left.\begin{array}{c} \mathbb{P}\big(T \le t_{n-1;1-\alpha}\big) \\ \mathbb{P}\big(T \ge -t_{n-1;1-\alpha}\big) \\ \mathbb{P}\big(|T| \le t_{n-1;1-\alpha/2}\big) \end{array}\right\} \;=\; 1-\alpha.$$

Solving the inequalities $\pm T \le c$ for μ yields three different $(1-\alpha)$-confidence regions for μ, namely, the lower confidence bound

$$\overline{X} - \frac{S}{\sqrt{n}}\, t_{n-1;1-\alpha},$$

the upper confidence bound

$$\overline{X} + \frac{S}{\sqrt{n}}\, t_{n-1;1-\alpha}$$

and the confidence interval

$$\left[\overline{X} \pm \frac{S}{\sqrt{n}}\, t_{n-1;1-\alpha/2}\right].$$

If the distribution P of the observations X_i is *not* Gaussian, these confidence regions still have *approximate* confidence level $1-\alpha$ as $n \to \infty$.

Example 4.4 (Monthly rents). Once again we consider the monthly rents of students, this time focussing on the mean rent μ of all students in Bern in 2003 (those who did rent a room or an appartment). Suppose that we want to emphasise how expensive a student's life is in Bern. Then it is appropriate to compute a lower confidence bound for μ. Our sample contains $n = 129$ observations with $\overline{X} = 609.128$ and $S = 289.153$. From a table or suitable software we find $t_{128;0.95} = 1.6568$ and obtain the lower confidence bound

$$\overline{X} - \frac{S}{\sqrt{n}}\, t_{n-1;1-\alpha} \;=\; 609.128 - \frac{289.153}{\sqrt{129}}\, 1.6568 \;=\; 565.947.$$

Thus we may claim with confidence about 95% that the mean monthtly rent μ is larger than $565\,\mathrm{CHF}$. The confidence is 'about 95%', because the empirical distribution function of the data indicates clearly a non-Gaussian distribution P.

Proof of Theorem 4.3. With $Z_i := (X_i - \mu)/\sigma$ we may write $X_i = \mu + \sigma Z_i$. The components of the random vector $\boldsymbol{Z} = (Z_i)_{i=1}^n$ are independent and standard Gaussian. With the sample mean $\overline{Z}$ of the Z_i we get $\overline{X} = \mu + \sigma\overline{Z}$. The sample standard deviations $S = S_X$ and S_Z of the X_i and Z_i, respectively, satisfy the equation $S_X = \sigma S_Z$. Consequently,

$$\Big(\frac{\sqrt{n}(\overline{X}-\mu)}{\sigma},\, \frac{S_X}{\sigma}\Big) \;=\; \big(\sqrt{n}\,\overline{Z},\, S_Z\big).$$

Now we employ the *spherical symmetry* of standard Gaussian random vectors: Let $\boldsymbol{B} \in \mathbb{R}^{n\times n}$ be an orthogonal matrix, that means, $\boldsymbol{B}^\top \boldsymbol{B} = \boldsymbol{B}\boldsymbol{B}^\top = \boldsymbol{I}_n$. Then the random vector $\boldsymbol{Z}$ has the same distribution as $\boldsymbol{Y} = (Y_i)_{i=1}^n := \boldsymbol{B}^\top \boldsymbol{Z}$. This follows from the fact that the random vector $\boldsymbol{Z}$ follows the density function

$$f(\boldsymbol{z}) \;:=\; (2\pi)^{-n/2} \exp\big(-\|\boldsymbol{z}\|^2/2\big)$$

on $\mathbb{R}^n$, and the latter is invariant under rotations and reflections of $\boldsymbol{z}$. Now we choose an orthogonal matrix of the form

$$\boldsymbol{B} \;=\; \begin{bmatrix} n^{-1/2} & b_{12} & \cdots & b_{1n} \\ n^{-1/2} & b_{22} & \cdots & b_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ n^{-1/2} & b_{n2} & \ldots & b_{nn} \end{bmatrix} \;=\; [\boldsymbol{b}_1\, \boldsymbol{b}_2\, \ldots\, \boldsymbol{b}_n].$$

In other words, we choose an orthonormal basis $\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_n$ of $\mathbb{R}^n$ such that $\boldsymbol{b}_1$ is equal to $n^{-1/2}(1, 1, \ldots, 1)^\top$. Then

$$Y_1 \;=\; \boldsymbol{b}_1^\top \boldsymbol{Z} \;=\; n^{-1/2} \sum_{i=1}^{n} Z_i \;=\; \sqrt{n}\, \overline{Z}$$

and

$$\sum_{i=1}^{n} (Z_i - \overline{Z})^2 \;=\; \sum_{i=1}^{n} Z_i^2 - n\overline{Z}^2 \;=\; \|\boldsymbol{Z}\|^2 - Y_1^2 \;=\; \|\boldsymbol{Y}\|^2 - Y_1^2 \;=\; \sum_{i=2}^{n} Y_i^2.$$

Consequently,

$$\big(\sqrt{n}\, \overline{Z},\, S_Z\big) \;=\; \left(Y_1,\, \sqrt{\frac{1}{n-1} \sum_{i=2}^{n} Y_i^2}\right).$$

□

An Example of 'Biased Sampling'

In this subsection we discuss a situation in which we draw a sample from a population which is related to but different from the population of interest. More precisely, we consider a population which is assumed to be constant over a longer period concerning life expectancy and family planning of its members. Now we consider the following subpopulations and features:

• Subpopulation 1 of all mothers (i.e. women with at least one child) with completed family planning and the variable $Y =$ 'number of children' with relative proportions

$$q_k \;:=\; \mathbb{P}(Y = k), \quad k = 1, 2, 3, \ldots$$

and mean

$$\nu \;:=\; \mathbb{E}(Y) \;=\; \sum_{k=1}^{\infty} k \cdot q_k,$$

i.e. the mean number of children per mother.

• Subpopulation 2 of all persons whose mothers have completed family planning and the variable $X =$ 'number of siblings' (with the same mother) with relative proportions

$$p_j \;:=\; \mathbb{P}(X = j), \quad j = 0, 1, 2, \ldots$$

and mean

$$\mu \;:=\; \mathbb{E}(X) \;=\; \sum_{j=0}^{\infty} j \cdot p_j,$$

i.e. the mean number of siblings per person.

Now the question is: What is the relationship between the distributions $(q_k)_{k\geq 1}$ and $(p_j)_{j\geq 0}$, in particular, between the corresponding means ν and μ? At first glance one would probably guess that $\nu = \mu + 1$, but we'll see soon that $\nu < \mu + 1$. A mother with k children is represented k times

in subpopulation 2. That means, if we consider a total of N mothers, they have $\sum_{k=1}^{\infty} Nq_k k = N\nu$ children, and $Nq_{j+1}(j+1)$ of these have precisely $j \geq 0$ siblings (with the same mother). Hence

$$p_j \;=\; \frac{q_{j+1}(j+1)}{\nu} \quad \text{and} \quad \frac{q_{j+1}}{\nu} \;=\; \frac{p_j}{j+1} \quad \text{for } j = 0, 1, 2, \ldots.$$

Summing the latter equation over all $j \geq 0$ yields the equation

$$\frac{1}{\nu} \;=\; \sum_{j=0}^{\infty} \frac{p_j}{j+1} \;=\; \mathbb{E}\Big(\frac{1}{X+1}\Big).$$

In particular it follows from Jensen's inequality and strict convexity of the function $0 \leq x \mapsto 1/(x+1)$ that

$$\nu \;=\; \Big(\mathbb{E}\Big(\frac{1}{X+1}\Big)\Big)^{-1} \;<\; \Big(\frac{1}{\mathbb{E}(X)+1}\Big)^{-1} \;=\; \mu + 1$$

unless X is almost surely constant. The latter condition would mean that all mothers have the same number ν of children.

Analyzing samples from subpopulation 2. Suppose one draws a random sample from subpopulation 2 and obtains the X-values $X_1, X_2, \ldots, X_n$. With these values one may compute the estimate $\overline{X}$ and a student confidence interval for μ. With the transformed values $W_i := 1/(X_i+1)$, an estimate for ν is given by

$$\widehat{\nu} \;:=\; \frac{1}{\overline{W}}.$$

By the way, the sample mean $\overline{W}$ may be expressed as follows:

$$\overline{W} \;=\; \sum_{j \geq 0} \frac{\widehat{p}_j}{j+1}$$

with $\widehat{p}_j := H_j/n$ and $H_j := \#\{i \leq n : X_i = j\}$. The probabilities q_k can be estimated by

$$\widehat{q}_k \;:=\; \frac{\widehat{\nu}\widehat{p}_{k-1}}{k}.$$

An approximate $(1-\alpha)$-confidence interval for ν can be constructed by first computing an approximate $(1-\alpha)$-confidence interval for $1/\nu = E(W)$ and then taking the reciprocals of its boundaries:

$$\Big[\Big(\overline{W} + \frac{S_W}{\sqrt{n}}\, t_{n-1;1-\alpha/2}\Big)^{-1},\; \Big(\overline{W} - \frac{S_W}{\sqrt{n}}\, t_{n-1;1-\alpha/2}\Big)_+^{-1}\Big],$$

where $a_+ := \max(a, 0)$. For the sample standard deviation S_W there is also an alternative representation:

$$S_W \;=\; \sqrt{\frac{n}{n-1}\Big(\sum_{j=0}^{\infty} \frac{\widehat{p}_j}{(j+1)^2} \;-\; \overline{W}^2\Big)}$$

Example 4.5. The questionnaires for students (Example 1.18) included the number of siblings (with the same mother). We obtained $n = 260$ values X_i, and it turned out that $\overline{X} = 1.5538$, $S_X = 0.9711$. To compute a 95%-confidence interval for μ we need the 97.5%-quantile of t_{259}. By means of a table or software we obtain $t_{259;0.975} = 1.9692$, so the approximate 95%-confidence interval for μ is equal to

$$\Big[\overline{X} \pm \frac{S_X}{\sqrt{n}}\, t_{n-1;1-\alpha/2}\Big] \;=\; \Big[1.5538 \pm \frac{0.9711}{\sqrt{260}}\, 1.9692\Big] \;=\; [1.4352, 1.6724].$$

However, if we are interested in the distribution of Y in subpopulation 2, we proceed as follows: The absolute frequencies $H_j = \#\{i : X_i = j\}$ and relative frequencies $\widehat{p}_j$ (rounded to four digits) are:

j	0	1	2	3	4	5	6	≥ 7
H_j	22	122	79	28	6	2	1	0
$\widehat{p}_j$	0.0846	0.4692	0.3038	0.1077	0.0231	0.0077	0.0038	0

This leads to $\overline{W} \approx 0.4539$ and $S_W \approx 0.1943$. In place of the naive estimate $\overline{X} + 1 = 2.5538$ for ν we thus obtain

$$\widehat{\nu} \;=\; \frac{1}{0.4539} \;\approx\; 2.2032.$$

For the probabilities q_k we obtain the following estimates $\widehat{q}_k = \widehat{\nu}\widehat{p}_{k-1}/k$ (rounded to four digits):

k	1	2	3	4	5	6	7	≥ 8
$\widehat{q}_k$	0.1864	0.5169	0.2231	0.0593	0.0102	0.0028	0.0012	0

Finally, with $t_{259;0.975} = 1.9692$ we compute the approximate 95%-copnfidence interval

$$\left[\overline{W} \pm \frac{S_W}{\sqrt{n}}\, t_{n-1;1-\alpha/2}\right] \;\approx\; \left[0.4539 \pm \frac{0.1943}{\sqrt{260}}\, 1.9692\right] \;\approx\; [0.4302, 0.4776]$$

forr $1/\nu = \mathbb{E}(W)$. This leads to the approximate 95%-confidence interval

$$\left[\frac{1}{0.4776}, \frac{1}{0.4302}\right] \;\approx\; [2.0937, 2.3248]$$

for its reciprocal $\nu = 1/\,\mathbb{E}(W)$. Note that the latter interval does not contain the naive estimate $\overline{X} + 1 = 2.5538$.

Remark 4.6. The longer one thinks about the previous problem and data example, the more questions come into mind. For instance, we did not specify an explicit population, and our sample is not really a random sample from some population. In particular, one should keep in mind that we only asked university students, i.e. young people studying at a university, so we are talking about a population of mothers whose kids (at least some of them) belong to the latter group. A second problem is that preferences of women concerning family planning may change over time. If one is interested in the more recent trends and wants to avoid the problem of different social classes, one could interview children at preschool age. But here a new problem arises: The corresponding mothers may have further kids in the future, that means, at interview time the final values of X and Y may still be unknown. A possible way out is to ask children about

$$\tilde{X} \;:=\; \text{number of older siblings}$$

(with the same mother). This option will be analyzed in Exercise 4.5.

Bounds for σ

In some applications one is interested in confidence regions for σ as well. For instance, the manufacturer of a measurement device may want an upper bound for the standard deviation σ of a single measurement. If someone wants to prove that a certain measurement method is rather imprecise, a lower confidence bound would be useful.

For the sake of simplicity we restrict ourselves to Gaussian data X_i. According to Theorem 4.3, $(n-1)S^2/\sigma^2$ follows a chi-squared distribution with $n-1$ degrees of freedom. If we denote its

β-quantile with $\chi^2_{n-1;\beta}$, then

$$\left.\begin{array}{c} \mathbb{P}\big((n-1)S^2/\sigma^2 \le \chi^2_{n-1;1-\alpha}\big) \\ \mathbb{P}\big((n-1)S^2/\sigma^2 \ge \chi^2_{n-1;\alpha}\big) \\ \mathbb{P}\big(\chi^2_{n-1;\alpha/2} \le (n-1)S^2/\sigma^2 \le \chi^2_{n-1;1-\alpha/2}\big) \end{array}\right\} \;=\; 1-\alpha.$$

Again one may solve the inequalities within $\mathbb{P}(\cdot)$ for σ and obtains the following $(1-\alpha)$-confidence regions for σ: The lower $(1-\alpha)$-confidence bound

$$S\sqrt{(n-1)/\chi^2_{n-1;1-\alpha}},$$

the upper $(1-\alpha)$-confidence bound

$$S\sqrt{(n-1)/\chi^2_{n-1;\alpha}},$$

and the $(1-\alpha)$-confidence interval

$$\Big[\, S\sqrt{(n-1)/\chi^2_{n-1;1-\alpha/2}}\,,\; S\sqrt{(n-1)/\chi^2_{n-1;\alpha/2}}\,\Big].$$

4.2 Further Features and Robustness

Quantiles, means and standard deviations are particular features which we embed into a more general context now. For the sake of simplicity we focus mostly on empirical features

$$K(X_1, X_2, \ldots, X_n)$$

which quantify certain aspects of the data X_i. Often one may interpret $K(X_1, X_2, \ldots, X_n)$ as a feature $K(\widehat{P})$ of the empirical distribution $\widehat{P}$ of $X_1, X_2, \ldots, X_n$. If we view $X_1, X_2, \ldots, X_n$ as independent random variables with distribution P, then $K(\widehat{P})$ is an estimator for the feature $K(P)$.

In the sequel we describe various features which are used in (descriptive) analyses. Here we distinguish three types:

- Location parameters (centers)
- Scale parameters (measures of spread)
- Shape parameters

Location Parameters

A location parameter $K(X_1, \ldots, X_n)$ is a number which is (i) 'as close as possible' to all X-values or (ii) provides a typical value and the order of magnitude of the X-values.

If one applies an affine transformation to the X-values, the location parameter should change correspondingly. This leads to the following mathematical characterisation of a location parameter: For arbitrary observations $X_1, \ldots, X_n$ and arbitrary constants $a \in \mathbb{R}$, $b > 0$,

$$K(a + bX_1, \ldots, a + bX_n) \;=\; a + bK(X_1, \ldots, X_n).$$

Sample mean. The most popular location parameter is the sample mean $\overline{X}$, i.e. the arithmetic mean of the numbers $X_1, \ldots, X_n$.

Sample quantiles. For any fixed $\gamma \in (0,1)$, the sample quantile $\widehat{q}_\gamma$ is a location parameter.

Trimmed means. Sometimes the largest and smallest values in a sample are suspicious in the sense that they may correspond to erroneous measurements or false answers in questionnaires. In this case one fixes a number $\tau \in (0, 0.5)$, for instance $\tau = 10\%$, and computes the arithmetic mean $\overline{X}_\tau$ of all order statistics $X_{(i)}$ such that $n\tau < i < n + 1 - n\tau$:

$$\overline{X}_\tau \;=\; \frac{1}{n-2k} \sum_{i=k+1}^{n-k} X_{(i)} \quad \text{with } k := \lfloor n\tau \rfloor.$$

For instance, in case of $n = 100$ observations and $\tau = 0.1$ we obtain the trimmed mean $\overline{X}_\tau = \sum_{i=11}^{90} X_{(i)}/80$.

Scale parameters

A scale parameter $K(X_1, \ldots, X_n)$ quantifies (i) the 'typical' distance between the X-values and their 'center' or (ii) the 'typical' distance of the X-values among themselves. Here we only consider sample sizes $n \geq 2$.

Such a feature should remain unchanged if all X-values are shifted to the left or to the right by the same amount, and it should increase by a factor of $b > 0$ if all X-values are multiplied with b. Hence for arbitrary observations $X_1, \ldots, X_n$ and arbitrary constants $a \in \mathbb{R}$, $b > 0$,

$$K(a + bX_1, \ldots, a + bX_n) \;=\; bK(X_1, \ldots, X_n).$$

In addition we require that

$$K(X_1, \ldots, X_n) \;>\; 0 \quad \text{whenever } \#\{X_1, \ldots, X_n\} = n.$$

Range. The simplest scale parameter is the so-called range,

$$X_{(n)} - X_{(1)},$$

which is the distance between smallest and largest observation.

Inter quartile range. A scale parameter which is quite popular in exploratory data analyses is the inter quartile range. It is defined as the distance between the first and third quartile,

$$\mathrm{IQR} \;:=\; \widehat{q}_{0.75} - \widehat{q}_{0.25}.$$

In other words, IQR is the length of the intervals $\big[\widehat{q}_{0.25}, \widehat{q}_{0.75}\big]$ and $\big(\widehat{q}_{0.25}, \widehat{q}_{0.75}\big)$ containing at least and at most, respectively, 50% of all observations.

Sample standard deviation. The sample standard deviation S is a scale parameter, too.

Gini's scale parameter. This feature has been proposed by C. Gini[2]. It is the arithmetic mean of the distances $|X_i - X_j|$ over all pairs of two observations:

$$G \;:=\; \binom{n}{2}^{-1} \sum_{1 \leq i < j \leq n} |X_i - X_j|.$$

[2]Corrado Gini (1884-1965): Italian econometrician. Better known than his scale parameter is the Gini index which measures income inequality.

The sum is over all index pairs (i, j) with $1 \le i < j \le n$, and there are $\binom{n}{2} = n(n-1)/2$ of these.

This definition of Gini's scale parameter is rather intuitive but rather inefficient computationally. An alternative formula for G (Exercise 4.6) is given by

$$G \;=\; \frac{2}{n(n-1)} \sum_{i=1}^{n} (2i - n - 1) X_{(i)}. \tag{4.1}$$

If one computes the order statistics with a suitable sorting method, the computation of G requires only $O(n \log n)$ steps.

Median absolute deviation. Similarly as in case of the standard deviation we want to quantify typical distances from a 'center', this time from the median. Thus we first compute the sample median $M := \mathrm{Median}(X_1, \ldots, X_n)$ of the values X_i and then the sample median of the moduli $|X_i - M|$:

$$\mathrm{MAD} \;:=\; \mathrm{Median}\big(|X_1 - M|, |X_2 - M|, \ldots, |X_n - M|\big).$$

In particular, $|X_i - M| < \mathrm{MAD}$ for at most 50% and $|X_i - M| \le \mathrm{MAD}$ for at least 50% of all observations.

If the distances of the median to the other two quartiles are identical, then one can easily show that $\mathrm{MAD} = \mathrm{IQR}/2$.

Shape parameters

A shape parameter describes aspects of the empirical distribution $\widehat{P}$ such as symmetry with respect to some 'center' which are invariant under arbitrary affine and increasing transformations. Formally, for any sample size $n \ge 2$, for arbitrary observations $X_1, \ldots, X_n$ and arbitrary constants $a \in \mathbb{R}$, $b > 0$,

$$K(a + bX_1, \ldots, a + bX_n) \;=\; K(X_1, \ldots, X_n).$$

In the sequel we introduce briefly two such shape parameters.

Skewness. The mean $\overline{X}$ is the center of gravity in the sense that $\sum_{i=1}^{n} (X_i - \overline{X}) = 0$: Imagine n persons of equal weight taking a seat on a seesaw at positions $X_1, X_2, \ldots, X_n$. If the seesaw's turning point coincides with $\overline{X}$, then the seesaw is balanced.

Now one could quantify non-symmetry of the values X_i around the center $\overline{X}$ by means of the average $n^{-1} \sum_{i=1}^{n} (X_i - \overline{X})^3$. Here deviations from the center are weighted over-proportionally. This average is already invariant with respect to location changes (adding a constant to all observations). To make it also invariant with respect to scale changes (multiplying all observations with a positive number) we divide it by S^3 and obtain the

$$\mathrm{Skewness} \;:=\; \frac{1}{nS^3} \sum_{i=1}^{n} (X_i - \overline{X})^3 \;=\; \frac{1}{n} \sum_{i=1}^{n} \Big(\frac{X_i - \overline{X}}{S} \Big)^3.$$

This may be viewed as an estimator of the theoretical quantity

$$\mathrm{Skewness}(P) \;:=\; \int \Big(\frac{x - \mu(P)}{\sigma(P)} \Big)^3 \, P(dx).$$

We call P 'right-skewed' or 'left-skewed' if $\mathrm{Skewness}(P)$ is strictly positive or strictly negative, respectively.

A good example for right-skewed distributions are Gamma distributions:

Definition 4.7 (Gamma distributions). The *Gamma distribution with shape parameter* $a > 0$ *and scale parameter* $b > 0$ is defined as the probability distribution on $\mathbb{R}$ with density function

$$g_{a,b}(x) := \begin{cases} 0 & \text{for } x \le 0, \\ \dfrac{1}{b\Gamma(a)} \Big(\dfrac{x}{b}\Big)^{a-1} \exp\Big(-\dfrac{x}{b}\Big) & \text{for } x > 0, \end{cases}$$

where $\Gamma(a) := \int_0^\infty t^{a-1} e^{-t}\, dt$. We denote this distribution with $\mathrm{Gamma}(a, b)$.

If $Y \sim \mathrm{Gamma}(a, 1)$ with $a > 0$, then $bY \sim \mathrm{Gamma}(a, b)$ for any $b > 0$, see also Section A.2. That's the reason why we call b a scale parameter. The parameter $a > 0$ determines the shape of the density function $g_{a,b}$. In case of $a < 1$ it has a pole at 0. In case of $a = 1$ it describes an exponential distribution. In case of $a > 1$ it is continuous on $\mathbb{R}$ with unique local and global maximum at $a - 1$, and in case of $a > 2$ it is continuously differentiable on $\mathbb{R}$. The skewness of $\mathrm{Gamma}(a, b)$ equals $2/\sqrt{a}$, see Exercise 4.11.

Kurtosis. The kurtosis is defined as the number

$$\mathrm{Kurtosis} := \frac{1}{nS^4} \sum_{i=1}^n (X_i - \overline{X})^4 - 3 = \frac{1}{n} \sum_{i=1}^n \Big(\frac{X_i - \overline{X}}{S}\Big)^4 - 3.$$

Both skewness and kurtosis are sometimes employed as test statistics to detect non-Gaussian distributions. Note that $\mathrm{Kurtosis}(X_1, X_2, \ldots, X_n)$ may be viewed as an estimator of the theoretical quantity

$$\mathrm{Kurtosis}(P) := \int \Big(\frac{x - \mu(P)}{\sigma(P)}\Big)^4 P(dx) - 3,$$

and in case of a Gaussian distribution P this value equals zero. In general, $\mathrm{Kurtosis}(P) > 0$ or $\mathrm{Kurtosis}(P) < 0$ indicates that P puts more or less mass, respectively, into the tail regions than the corresponding normal distribution $\mathcal{N}(\mu(P), \sigma(P)^2)$.

Exercise 4.12 establishes a connection between skewness, kurtosis and so-called moment-generating functions. Based on that one can easily verify that skewness and kurtosis of $\mathrm{Gamma}(a, b)$ are given by $2/\sqrt{a}$ and $6/a$, respectively; see Exercise 4.14.

Robustness

The sample mean is easier to calculate than the sample median, because there is no need to sort the observations. On the other hand, it is sensitive to 'outliers' in the data. Such 'outliers' may be values which have been entered incorrectly (e.g. omission or wrong positioning of decimal points, non-sensible answers in questionnaires) or values which are really extremely large or small. One single extreme value X_i may cause the mean $\overline{X}$ to be quite far away from the majority of X-values. By way of contrast, the median is rather *robust* with respect to outliers, see Exercise 4.16.

One can quantify robustness of a feature by means of the *breakdown point* which has been introduced by Hampel (1968, 1971) and Donoho and Huber (1982, 1983). Let a_n be the largest integer in $\{0, 1, \ldots, n\}$ with the property that for arbitrary values $X_1, \ldots, X_n$,

$$\sup\big\{\big|K(Y_1, \ldots, Y_n)\big| : Y_i \neq X_i \text{ for at most } a_n \text{ indices } i\big\} < \infty.$$

The *breakdown point* of the feature $K(\cdot)$ is defined as the number

$$\liminf_{n\to\infty} \frac{a_n}{n}.$$

(Here we tacitly assume that $K(\cdot)$ is well-defined for arbitrary sample sizes $n \geq n_o$.) If this number is strictly positive, the feature $K(\cdot)$ is called robust.

The sample mean has breakdown point 0, because $a_n = 0$ for all n. For the sample median we find in Exercise 4.16 that $a_n = \lfloor (n-1)/2 \rfloor$. Hence its breakdown point equals $1/2$. This statement may be generalised for arbitrary sample quantiles.

Lemma 4.8. *For $\gamma \in (0,1)$, the sample quantile $\widehat{q}_\gamma$ has breakdown point*

$$\min(\gamma, 1-\gamma).$$

Proof. Let $\widehat{P}$ be the empirical distribution of the observations $X_1, X_2, \ldots, X_n$, and for a fixed $k \in \{1, 2, \ldots, n\}$ let $\widehat{Q}$ be the empirical distribution of $Y_1, Y_2, \ldots, Y_n$, where $\#\{i : Y_i \neq X_i\} \leq k$. For any interval $B \subset \mathbb{R}$ we have the inequalities

$$\widehat{P}(B) - k/n \;\leq\; \widehat{Q}(B) \;\leq\; \widehat{P}(B) + k/n.$$

In case of $k < n \min(\gamma, 1-\gamma)$ this implies that

$$\widehat{Q}((-\infty, x]) \;\leq\; k/n \;<\; \gamma \quad \text{whenever } x < X_{(1)}$$

and

$$\widehat{Q}([x, \infty)) \;\leq\; k/n \;<\; \gamma \quad \text{and} \qquad \text{whenever } x > X_{(n)}.$$

Thus $\widehat{q}_\gamma(Y_1, \ldots, Y_n)$ will be contained in $[X_{(1)}, X_{(n)}]$, whence a_n is at least $\lceil n \min(\gamma, 1-\gamma) \rceil - 1$. In case of $\gamma \leq 1/2$ and $k > n\gamma$, we choose an arbitrarily small number x and replace $X_1, \ldots, X_k$ with values in $(-\infty, x]$. This leads to $\widehat{q}_\gamma(Y_1, \ldots, Y_n) \leq x$, because $\widehat{Q}((-\infty, x]) > \gamma$. In case of $\gamma > 1/2$ and $k > n(1-\gamma)$, we choose an arbitrarily large number x and replace $X_1, \ldots, X_k$ with values in $[x, \infty)$. This leads to $\widehat{q}_\gamma(Y_1, \ldots, Y_n) \geq x$, because $\widehat{Q}([x, \infty)) > 1 - \gamma$. This shows that a_n is no larger than $\lfloor n \min(\gamma, 1-\gamma) \rfloor$.

These simple considerations show that $a_n/n \to \min(\gamma, 1-\gamma)$ as $n \to \infty$. Exercise 4.17 provides refined bounds for the difference between $\widehat{q}_\gamma(Y_1, \ldots, Y_n)$ and $\widehat{q}_\gamma(X_1, \ldots, X_n)$. □

In connection with scale parameters one considers $\log(K)$ in place of K and restricts one's attention to samples $X_1, \ldots, X_n$ with pairwise different values. Thus, a scale parameter breaks down if modifying some components of the sample leads to arbitrary large values or to values arbitrarily close to zero.

Lemma 4.9. *The IQR has breakdown point $1/4$, and the MAD has breakdown point $1/2$.*

Proof. We only derive the breakdown point of the IQR; proving the statement about the MAD is posed as Exercise 4.18. If $2 \leq n \leq 4$, then $\mathrm{IQR}(X_1, \ldots, X_n) = X_{(n)} - X_{(1)}$, and $a_n = 0$. Let $n \geq 5$, and let $Y_1, \ldots, Y_n$ be the new observations after modifying up to $k \geq 1$ of the observations X_i. Then $\mathrm{IQR}(Y_1, \ldots, Y_n)$ is the length of an interval $[A, B]$ containing at least $\lceil n/2 \rceil$ observations Y_i, so it contains at least $\lceil n/2 \rceil - k$ observations X_i. If $\ell := \lceil n/2 \rceil - k \geq 2$, then

$$\mathrm{IQR}(Y_1, \ldots, Y_n) \;\geq\; \min_{i=1,\ldots,n+1-\ell} (X_{(i+\ell-1)} - X_{(i)}) \;>\; 0,$$

so $\log \mathrm{IQR}(Y_1, \ldots, Y_n)$ stays bounded away from $-\infty$. On the other hand both $(-\infty, A]$ and $[B, \infty)$ contain at least $\lceil n/4 \rceil$ observations Y_i, so they both contain at least $\lceil n/4 \rceil - k$ observations $X_{(i)}$. In case of $\ell := \lceil n/4 \rceil - k \geq 1$ we may conclude that $A \geq X_{(\ell)}$ and $B \leq X_{(n+1-\ell)}$, that means

$$\mathrm{IQR}(Y_1, \ldots, Y_n) \;\leq\; X_{(n+1-\ell)} - X_{(\ell)} \;<\; \infty.$$

This shows that $a_n \geq \min\big(\lceil n/2 \rceil - 2, \lceil n/4 \rceil - 1\big) = \lceil n/4 \rceil - 1$.

Feature		Breakdown point
Mean	$\overline{X}$	0
Quantile	$\widehat{q}_\gamma$	$\min(\gamma, 1-\gamma)$
Trimmed mean	$\overline{X}_\tau$	τ
Range	$X_{(n)} - X_{(1)}$	0
Inter quartile range	IQR	$1/4$
Standard deviation	S	0
Gini's scale parameter	G	0
Median of absolute deviations	MAD	$1/2$

Table 4.1: Breakdown points of some locations and scale parameters.

Finally, if one adds a constant $R > 0$ to the $\lceil n/4 \rceil$ largest order statistics of $X_1, \ldots, X_n$, then $\mathrm{IQR}(X_1, \ldots, X_n)$ increases by $R/2$ or R, depending on $n/4$ being an integer or not. Thus $a_n \leq \lceil n/4 \rceil - 1$. These considerations show that $a_n/n \to 1/4$ as $n \to \infty$. □

Table 4.1 shows which of our location and scale parameters are robust.

4.3 Sign Tests and Related Procedures

In this section we leave temporarily the framework of one numerical variable and consider so-called 'paired samples'. As a by-product we'll obtain procedures to estimate the center of a symmetric distribution.

Sign Tests for Paired Samples

The expression 'paired samples' is somewhat misleading. We consider *one* sample with *two* numerical variables. Thus we observe pairs (Y_1, Z_1), (Y_2, Z_2), …, (Y_n, Z_n). Now the questions is whether the differences

$$X_i \;:=\; Y_i - Z_i$$

tend to be greater than or less than zero. To answer this question one could view the differences X_i as stochastically independent and identically distributed random variables. Then one could apply one of our previous methods to compute confidence bounds for the mean $\mathbb{E}(X_1)$ or particular quantiles of the distribution of X_1.

Example 4.10 (Gossets barley data). In his famous paper about student's t distribution, published in 1908, Gosset illustrated his method with the data in Table 4.2: Each of eleven fields of equal size had been divided into two halves. On one half, people deployed regular barley seeds, on the other half they deployed barley seeds which had undergone a special drying treatment. The measurements are the yield on each half-field (in lbs/acre).

Gosset analysed these data under the assumption that the differences X_i are independent and follow a Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$ with unknown parameters μ and σ^2. One may interpret the mean μ as mean increase of yield when replacing dried barley seeds with regular barley seeds. For μ he obtained the 95%-confidence interval

$$\Big[\overline{X} \,\pm\, \frac{S_X}{\sqrt{n}}\, t_{10;0.975}\Big] \;\approx\; \Big[-33.727 \,\pm\, \frac{66.171}{\sqrt{11}}\, 2.228\Big] \;=\; \big[-78.182, 10.727\big].$$

	Yield			Yield			Yield	
Field	regular	dried	Field	regular	dried	Field	regular	dried
1	1903	2009	5	2108	2180	9	1612	1542
2	1935	1915	6	1961	1925	10	1316	1443
3	1910	2011	7	2060	2122	11	1511	1535
4	2496	2463	8	1444	1482			

Table 4.2: Gosset's barley data.

Thus it is possible that $\mu = 0$. With confidence 95% one may claim that the change of mean yield is no more than 79 lbs/acre.

Sometimes the assumption of independent, identically and normally distributed differences is questionable. For instance, it could be that the differences are independent but not identically distributed. The following lemma describes three equivalent possibilities to describe the null hypothesis that there is no systematic difference between the Y- and Z-values. Here and in the sequel we use the following notation:

$$\boldsymbol{wx} \;:=\; (w_i x_i)_{i=1}^n \quad \text{and} \quad |\boldsymbol{x}| \;:=\; \big(|x_i|\big)_{i=1}^n$$

for vectors $\boldsymbol{w}, \boldsymbol{x} \in \mathbb{R}^n$.

Lemma 4.11 (Sign-symmetry). *Let $\boldsymbol{\xi}$ be a random vector with uniform distribution on $\{-1,1\}^n$ and independent from $\boldsymbol{X}$. In other words, $\boldsymbol{X}, \xi_1, \xi_2, \ldots, \xi_n$ are independent random variables such that $\mathbb{P}(\xi_i = 1) = \mathbb{P}(\xi_i = -1) = 1/2$. Then the following three statements are equivalent:*

(i) *For arbitrary fixed $\boldsymbol{s} \in \{-1,1\}^n$, the random vectors $\boldsymbol{sX}$ and $\boldsymbol{X}$ are identically distributed.*

(ii) *The random vectors $\boldsymbol{\xi X}$ and $\boldsymbol{X}$ are identically distributed.*

(iii) *The random vectors $\boldsymbol{\xi}|\boldsymbol{X}|$ and $\boldsymbol{X}$ are identically distributed.*

Proof of Lemma 4.11. For arbitrary Borel sets $B \subset \mathbb{R}^n$,

$$\mathbb{P}(\boldsymbol{\xi X} \in B) \;=\; \sum_{\boldsymbol{s}\in\{-1,1\}^n} \mathbb{P}(\boldsymbol{\xi} = \boldsymbol{s}, \boldsymbol{sX} \in B) \;=\; 2^{-n} \sum_{\boldsymbol{s}\in\{-1,1\}^n} \mathbb{P}(\boldsymbol{sX} \in B).$$

If Condition (i) is met, all summands $\mathbb{P}(\boldsymbol{sX} \in B)$ on the right hand side are equal to $\mathbb{P}(\boldsymbol{X} \in B)$. Thus $\mathbb{P}(\boldsymbol{\xi X} \in B) = \mathbb{P}(\boldsymbol{X} \in B)$, and Condition (ii) is satisfied, too.

Now we show that the distributions of $\boldsymbol{\xi X}$ and $\boldsymbol{\xi V X}$ coincide whenever $\boldsymbol{V}$ is an arbitrary sign vector of the form $\boldsymbol{V} = \boldsymbol{f}(\boldsymbol{X}) \in \{-1,1\}^n$. For arbitrary Borel sets $B \subset \mathbb{R}^n$,

$$\begin{aligned}
\mathbb{P}(\boldsymbol{\xi V X} \in B) \;&=\; 2^{-n} \sum_{\boldsymbol{s}\in\{-1,1\}^n} \mathbb{P}(\boldsymbol{sVX} \in B) \\
&=\; 2^{-n}\, \mathbb{E}\Big(\sum_{\boldsymbol{s}\in\{-1,1\}^n} 1_{[\boldsymbol{sVX}\in B]} \Big) \\
&=\; 2^{-n}\, \mathbb{E}\Big(\sum_{\boldsymbol{s}\in\{-1,1\}^n} 1_{[\boldsymbol{sX}\in B]} \Big) \\
&=\; 2^{-n} \sum_{\boldsymbol{s}\in\{-1,1\}^n} \mathbb{P}(\boldsymbol{sX} \in B) \\
&=\; \mathbb{P}(\boldsymbol{\xi X} \in B).
\end{aligned}$$

In the third step we used the fact that the mapping $\boldsymbol{s} \mapsto \boldsymbol{sV}$ from $\{-1,1\}^n$ to $\{-1,1\}^n$ is bijective. This shows that, indeed, the random vectors $\boldsymbol{\xi V X}$ and $\boldsymbol{\xi X}$ are identically distributed.

If we set $V_i := 1_{[X_i \geq 0]} - 1_{[X_i < 0]}$, then $\boldsymbol{VX} = |\boldsymbol{X}|$. Hence $\boldsymbol{\xi}|\boldsymbol{X}|$ and $\boldsymbol{\xi X}$ are identically distributed, which implies the equivalence of Conditions (ii) and (iii).

If, on the other hand, $\boldsymbol{V}$ is an arbitrary fixed sign vector, then $\boldsymbol{\xi VX} = \boldsymbol{V\,\xi X}$ and $\boldsymbol{\xi X}$ are identically distributed. Thus Condition (ii) implies Condition (i). □

Null hypothesis H_o (Sign-symmetry). The random vector $\boldsymbol{X} = (X_i)_{i=1}^n$ has a *sign-symmetric distribution*. That means, it satisfies the conditions described in Lemma 4.11.

Example 4.12 (Lectures as a sedative). In a lecture about biometry for students of computer science, $n = 18$ students determined their pulse rate prior at the beginning (Y_i) and after (Z_i) class. Both values are the number of heart beats in one minute. The working hypothesis was that the Y-values are systematically larger than the Z-values, which we would interpret as the lecture having a sedative effect. The null hypothesis H_o that the difference vector $\boldsymbol{X} = \boldsymbol{Y} - \boldsymbol{Z}$ has sign-symmetric distribution is illustrated in Figure 4.2. There one sees graphical displays of the vectors $\boldsymbol{X}$ and $\boldsymbol{\xi}^{(1)}|\boldsymbol{X}|$, $\boldsymbol{\xi}^{(2)}|\boldsymbol{X}|$, …, $\boldsymbol{\xi}^{(8)}|\boldsymbol{X}|$ in random order. Here we simulated eight sign vectors $\boldsymbol{\xi}^{(1)}, \boldsymbol{\xi}^{(2)}, \ldots, \boldsymbol{\xi}^{(8)}$ with uniform distribution on $\{-1,1\}^n$ which are independent mutually and from $\boldsymbol{X}$. The components of $\boldsymbol{X}$ have been ordered such that $|X_1| \leq |X_2| \leq \cdots \leq |X_n|$. The readers should try to find the original data vector $\boldsymbol{X}$ before reading on.

The original data vector is the top right one. If you found it, you may claim with confidence $8/9 \approx 88.9\%$ that the null hypothesis is wrong. For otherwise, the probability of finding the original data vector would have been equal to $1/9$; see also the considerations about Monte-Carlo tests in the last chapter. Formal statistical tests will be applied later.

If one doesn't view the 18 persons as a random sample from some population, one should keep in mind that random fluctuations of the pulse rate (without external influence) are rather different from person to person. Hence it is important that we don't assume identically distributed random variables X_i.

P-values for H_o. In some applications one guesses a priori that the differences X_i tend to be positive or tend to be negative, respectively (one-sided working hypotheses). In other situations one expects a deviation from H_o without an a priori guess about the direction (two-sided working hypothesis). To test H_o we compute for a given test statistic $T : \mathbb{R}^n \to \mathbb{R}$ and depending on our working hypothesis one of the p-values $\pi_\ell(\boldsymbol{X})$, $\pi_r(\boldsymbol{X})$ or $\pi_z(\boldsymbol{X})$. Here

$$\begin{aligned} \pi_\ell(\boldsymbol{x}) &:= 2^{-n}\,\#\big\{\boldsymbol{s} \in \{-1,1\}^n : T(\boldsymbol{s}|\boldsymbol{x}|) \leq T(\boldsymbol{x})\big\}, \\ \pi_r(\boldsymbol{x}) &:= 2^{-n}\,\#\big\{\boldsymbol{s} \in \{-1,1\}^n : T(\boldsymbol{s}|\boldsymbol{x}|) \geq T(\boldsymbol{x})\big\} \end{aligned}$$

and $\pi_z(\boldsymbol{x}) := 2 \cdot \min\{\pi_\ell(\boldsymbol{x}), \pi_r(\boldsymbol{x})\}$ for a fixed vector $\boldsymbol{x} \in \mathbb{R}^n$. These p-values quantify how exceptional the vector $\boldsymbol{x}$ is among all vectors $\tilde{\boldsymbol{x}}$ with $|\tilde{\boldsymbol{x}}| = |\boldsymbol{x}|$. With a random sign vector $\boldsymbol{\xi} \in \{-1,1\}^n$ as in Lemma 4.11 one can also write

$$\begin{aligned} \pi_\ell(\boldsymbol{x}) &= \mathbb{P}\big(T(\boldsymbol{\xi}|\boldsymbol{x}|) \leq T(\boldsymbol{x})\big), \\ \pi_r(\boldsymbol{x}) &= \mathbb{P}\big(T(\boldsymbol{\xi}|\boldsymbol{x}|) \geq T(\boldsymbol{x})\big). \end{aligned}$$

The null hypothesis H_o is rejected at level α if the p-value of our choice (a priori!) is less than or equal to α. Here is a justification for this sign test:

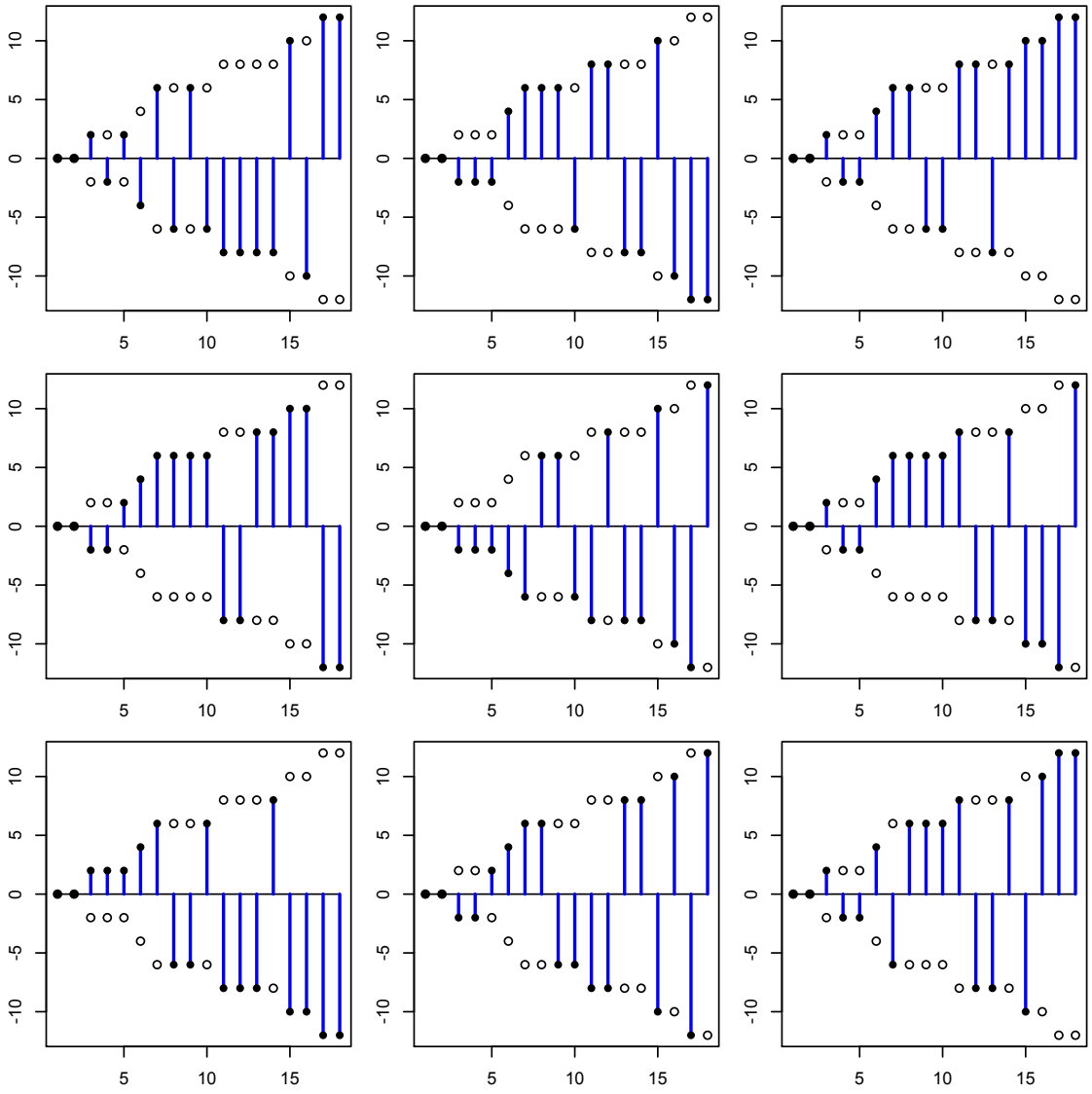

Figure 4.2: The null hypothesis of sign-symmetry.

Lemma 4.13. *Let $\pi(\boldsymbol{X})$ be one of the three p-values above. Under the null hypothesis of sign-symmetry of $\boldsymbol{X}$,*

$$\mathbb{P}\big(\pi(\boldsymbol{X}) \le \alpha\big) \;\le\; \alpha$$

for any $\alpha \in (0,1)$.

Proof of Lemma 4.13. Under H_o,

$$\begin{aligned}\mathbb{P}\big(\pi(\boldsymbol{X}) \le \alpha\big) \;=\; \mathbb{P}\big(\pi(\boldsymbol{\xi}|\boldsymbol{X}|) \le \alpha\big) \;&=\; 2^{-n} \sum_{\boldsymbol{s}\in\{-1,1\}^n} \mathbb{P}\big(\pi(\boldsymbol{s}|\boldsymbol{X}|) \le \alpha\big) \\ &=\; \mathbb{E}\Big(2^{-n} \sum_{\boldsymbol{s}\in\{-1,1\}^n} 1_{[\pi(\boldsymbol{s}|\boldsymbol{X}|)\le\alpha]}\Big).\end{aligned}$$

Thus it suffices to show that

$$2^{-n} \sum_{\boldsymbol{s}\in\{-1,1\}^n} 1_{[\pi(\boldsymbol{s}|\boldsymbol{x}|)\le\alpha]} = \mathbb{P}\big(\pi(\boldsymbol{\xi}|\boldsymbol{x}|) \le \alpha\big) \;\le\; \alpha$$

for arbitrary fixed vectors $\boldsymbol{x} \in \mathbb{R}^n$. To this end we consider the random variable $Y := T(\boldsymbol{\xi}|\boldsymbol{x}|)$. Note that

$$\mathbb{P}(Y \le y) \;=\; 2^{-n}\#\big\{\boldsymbol{s} \in \{-1,1\}^n : T(\boldsymbol{s}|\boldsymbol{x}|) \le y\big\} \;=:\; G_{|\boldsymbol{x}|}(y)$$

for arbitrary $y \in \mathbb{R}$, and

$$\begin{aligned}\pi_\ell(\boldsymbol{x}) \;&=\; G_{|\boldsymbol{x}|}(T(\boldsymbol{x})), \\ \pi_r(\boldsymbol{x}) \;&=\; 1 - G_{|\boldsymbol{x}|}(T(\boldsymbol{x}) -).\end{aligned}$$

Since $\big|\boldsymbol{\xi}|\boldsymbol{x}|\big| = |\boldsymbol{x}|$, one may also write

$$\begin{aligned}\pi_\ell(\boldsymbol{\xi}|\boldsymbol{x}|) \;&=\; G_{|\boldsymbol{x}|}(Y), \\ \pi_r(\boldsymbol{\xi}|\boldsymbol{x}|) \;&=\; 1 - G_{|\boldsymbol{x}|}(Y -), \\ \pi_z(\boldsymbol{\xi}|\boldsymbol{x}|) \;&=\; 2 \cdot \min\{G_{|\boldsymbol{x}|}(Y), 1 - G_{|\boldsymbol{x}|}(Y -)\}.\end{aligned}$$

According to Lemma 1.4, $\mathbb{P}\big(\pi(\boldsymbol{\xi}|\boldsymbol{x}|) \le \alpha\big)$ is always less than or equal to α. □

To compute the p-values above explicitly, one avoids the definition with 2^n summands. Instead one uses special properties of the test statistic $T(\cdot)$, or one employs approximate p-values or computes Monte-Carlo p-values.

Special Sign Tests

In the sequel we consider three specific examples for T and the resulting tests. In all cases the test statistic has the form

$$T(\boldsymbol{x}) \;:=\; \sum_{i=1}^{n} \operatorname{sign}(x_i) B_i \tag{4.2}$$

with certain numbers $B_i = B_i(|\boldsymbol{x}|)$, $1 \le i \le n$. We always assume that $B_i = 0$ whenever $|x_i| = 0$. Then $\operatorname{sign}(x_i)B_i = 2 \cdot 1_{[x_i>0]}B_i - B_i$, so we may write

$$T(\boldsymbol{x}) \;=\; 2\,T_o(\boldsymbol{x}) - B_+$$

with $B_+ := \sum_{i=1}^n B_i$ and

$$T_o(\boldsymbol{x}) \;:=\; \sum_{i=1}^{n} 1_{[x_i>0]} B_i. \tag{4.3}$$

For the explicit computation of p-values, the test statistic $T_o(\boldsymbol{x})$ is often preferable, but its value is typically more difficult to interpret than $T(\boldsymbol{x})$.

Pair	Cross	Self	Pair	Cross	Self	Pair	Cross	Self
1	23.5	17.4	6	21.5	18.6	11	23.3	16.3
2	12.0	20.4	7	22.1	18.6	12	21.0	18.0
3	21.0	20.0	8	20.4	15.3	13	22.1	12.8
4	22.0	20.0	9	18.3	16.5	14	23.0	15.5
5	19.1	18.4	10	21.6	18.0	15	12.0	18.0

Table 4.3: Darwin's plant experiment.

Pearson's sign test. In the simplest case we just consider the signs of the x_i and define

$$T(\boldsymbol{x}) \;:=\; \sum_{i=1}^{n} \mathrm{sign}(x_i).$$

This corresponds to (4.2) with $B_i = 1_{[|x_i|>0]}$. The corresponding sum B_+ equals

$$N = N(|\boldsymbol{x}|) \;:=\; \#\{i \le n : x_i \neq 0\},$$

and $T(\boldsymbol{x}) = 2\,T_o(\boldsymbol{x}) - N$ with

$$T_o(\boldsymbol{x}) \;=\; \#\{i \le n : x_i > 0\}.$$

Here $T_o(\boldsymbol{\xi}|\boldsymbol{x}|)$ has the same distribution as $\sum_{i=1}^{N} 1_{[\xi_i=1]}$, and the latter follows $\mathrm{Bin}(N, 0.5)$. This leads to the p-values

$$\begin{aligned} \pi_\ell(\boldsymbol{x}) \;&=\; F_{N,0.5}(T_o(\boldsymbol{x})), \\ \pi_r(\boldsymbol{x}) \;&=\; 1 - F_{N,0.5}(T_o(\boldsymbol{x}) - 1), \end{aligned}$$

where $F_{N,0.5}$ denotes the distribution function of $\mathrm{Bin}(N, 0.5)$.

Example 4.14 (Darwin's plant experiment). To verify that cross fertilisation leads to stronger plants than self-fertilisation, Charles Darwin (1809-1882) carried through the following experiment: In each of 15 plant pots he grew two plants of the same species, one of which was generated via cross-fertilisation and the other one via self-fertilisation. After a certain time period the heights (in 0.125 inches) of the plants were measured; see Table 4.3. With these data, Darwin approached Karl Pearson.

For the i-th pair let Y_i and Z_i be the heights of the plant generated via cross- and self-fertilisation, respectively. All $n = 15$ differences X_i are non-zero, so $N = 15$. Darwin's one-sided working hypothesis leads to the right-sided p-value which is compared with $\alpha = 0.05$: Our of the $N = 15$ differences $T_o(\boldsymbol{X}) = 13$ turned out to be strictly positive, so

$$\pi_r(\boldsymbol{X}) \;=\; 1 - F_{15,0.5}(12) \;=\; 0.0037.$$

Thus we reject H_o at level 5% (and confirm Darwin's working hypothesis with confidence 95%).

Example 4.15 (Lectures as a sedative, cont.). Out of the $n = 18$ differences, $N = 16$ are different from zero and $T_o(\boldsymbol{X}) = 11$ strictly positive. In view of our one-sided working hypothesis we compute the right-sided p-value

$$\pi_r(\boldsymbol{X}) \;=\; 1 - F_{16,0.5}(10) \;=\; 0.1051.$$

Hence we cannot reject H_o at the standard test level 5%.

The sign-t-test. The simple sign test doesn't take the moduli $|X_i|$ into account, although bigger differences may be more relevant than small ones. Alternatively one could consider the test statistic $T(\boldsymbol{x}) := \sum_{i=1}^n x_i = \sum_{i=1}^n \mathrm{sign}(x_i)|x_i|$. The resulting sign test is considerably more difficult to perform than the simple sign test. On the other hand, one can show that the sign test with this test statistic is at least as sensitive to violations of the null hypothesis as Gosset's method whenever the latter is justified.

Wilcoxon's signed-rank test. A possible compromise between the simple sign test and the sign-t-test is to replace the moduli $|x_1|, |x_2|, \ldots, |x_n|$ with their ranks. Precisely, we only consider the non-zero components of $\boldsymbol{x}$ and define

$$R_i \;:=\; \#\big\{\ell : 0 < |x_\ell| < |x_i|\big\} + \Big(1_{[|x_i|>0]} + \#\big\{\ell : 0 < |x_\ell| = |x_i|\big\}\Big)\Big/2.$$

Then the *signed-rank statistic* of Wilcoxon[3] (1945) is defined as

$$T(\boldsymbol{x}) \;:=\; \sum_{i=1}^n \mathrm{sign}(x_i) R_i.$$

If the non-zero values $|x_i|$ are pairwise different, then the tuple $(R_1, R_2, \ldots, R_n)$ is a permutation of $(1, 2, \ldots, n)$ if $N = n$ and of $(0, \ldots, 0, 1, 2, \ldots, N)$ if $N < n$. In this case we compare $T(\boldsymbol{x})$ with the distribution of the random variable

$$\sum_{i=1}^N \xi_i \cdot i.$$

The explicit computation of p-values is still computer-intensive. But the distribution of $T(\boldsymbol{\xi}|\boldsymbol{x}|)$ may be computed exactly in $O(N^3)$ steps with memory $O(N^2)$. To this end we use the representation (4.3) and the fact that

$$R_+ \;=\; N(N+1)/2,$$

see Exercise 3.6. Hence $T(\boldsymbol{x}) = 2\,T_o(\boldsymbol{x}) - N(N+1)/2$ with

$$T_o(\boldsymbol{x}) \;=\; \sum_{i=1}^n 1_{[x_i>0]} R_i,$$

and

$$T_o(\boldsymbol{\xi}|\boldsymbol{x}|) \;=\; \sum_{i=1}^n 1_{[\xi_i=1]} R_i.$$

The possible values of $T_o(\boldsymbol{x})$ and $T_o(\boldsymbol{\xi}|\boldsymbol{x}|)$ are contained in $\{k/2 : k = 0, 1, \ldots, N(N+1)\}$, and one may write

$$\begin{aligned}
\pi_\ell(\boldsymbol{x}) \;&=\; \mathbb{P}\big(T_o(\boldsymbol{\xi}|\boldsymbol{x}|) \le T_o(\boldsymbol{x})\big) \;=\; G_N(T_o(\boldsymbol{x})),\\
\pi_r(\boldsymbol{x}) \;&=\; \mathbb{P}\big(T_o(\boldsymbol{\xi}|\boldsymbol{x}|) \ge T_o(\boldsymbol{x})\big) \;=\; 1 - G_N(T_o(\boldsymbol{x}) - 1/2),
\end{aligned}$$

where

$$G_j(y) \;:=\; \mathbb{P}\Big(\sum_{i=1}^j 1_{[\xi_i=1]} M_i \le y\Big)$$

[3]Frank Wilcoxon (1892-1965): US-american chemist and statistician; introduced in his paper [30] two new and nowadays widely used statistical tests.

$\boldsymbol{G} = \big(G[k]\big)_{k=0}^{N(N+1)} \leftarrow \big(1\big)_{k=0}^{N(N+1)}$
$m \leftarrow 0$
for $j \leftarrow 1$ to N do
  $m \leftarrow m + 2M_j$
  $\big(G[k]\big)_{k=2M_j}^{m} \leftarrow \Big(\big(G[k]\big)_{k=2M_j}^{m} + \big(G[k-2M_j]\big)_{k=2M_j}^{m}\Big)/2$
  $\big(G[k]\big)_{k=0}^{2M_j-1} \leftarrow \big(G[k]\big)_{k=0}^{2M_j-1}/2$
end for.

Table 4.4: Auxiliary program for Wilcoxon's signed rank test.

for $1 \le j \le N$ with the sorted and strictly positive components $M_1 \le M_2 \le \cdots \le M_N$ of $(R_i)_{i=1}^n$. But

$$\begin{aligned} G_j(y) &= \mathbb{P}\Big(\xi_j = -1 \text{ and } \sum_{i=1}^{j-1} 1_{[\xi_i=1]} M_i \le y\Big) + \mathbb{P}\Big(\xi_j = 1 \text{ and } \sum_{i=1}^{j-1} 1_{[\xi_i=1]} M_i + M_j \le y\Big) \\ &= \big(G_{j-1}(y) + G_{j-1}(y - M_j)\big)/2 \end{aligned}$$

with $G_0(y) := 1_{[0 \le y]}$. With this recursion formula one can compute the tuple $\boldsymbol{G} = \big(G[k]\big)_{k=0}^{N(N+1)}$ with $G[k] := G_N(k/2)$ as a function of N and $(M_i)_{i=1}^N$; see Table 4.4.

Example 4.16 (lectures as a sedative, cont.). In Table 4.5 the data pairs (Y_i, Z_i) are arranged such that the moduli $|X_i|$ increase. In the column with ranks the numbers in brackets correspond to the ranks one would assign without caring about equal values of $|X_i|$. Here $T_o(\boldsymbol{X}) = 108.5$ and $T(\boldsymbol{X}) = 81$. The corresponding exact right-sided p-value equals $\pi_r(\boldsymbol{X}) = 0.0171$. Hence we may claim with confidence 95% that H_o is wrong (and that the lecture had a sedating effect).

Y_i	Z_i	X_i	R_i		$\operatorname{sign}(X_i)$
66	66	0	0	(0)	0
78	78	0	0	(0)	0
54	56	−2	2	(1)	−1
76	78	−2	2	(2)	−1
80	78	2	2	(3)	+1
94	90	4	4	(4)	+1
68	74	−6	6.5	(5)	−1
64	70	−6	6.5	(6)	−1
76	70	6	6.5	(7)	+1
80	74	6	6.5	(8)	+1
64	72	−8	10.5	(9)	−1
66	58	8	10.5	(10)	+1
70	62	8	10.5	(11)	+1
80	72	8	10.5	(12)	+1
82	72	10	13.5	(13)	+1
102	92	10	13.5	(14)	+1
74	62	12	15.5	(15)	+1
90	78	12	15.5	(16)	+1

Table 4.5: Example for the computation of Wilcoxon's signed rank statistic.

Approximate and conservative p-values. In all cases we work with a test statistic of the form (4.2). Here

$$\mathbb{E}\big(T(\boldsymbol{\xi}|\boldsymbol{x}|)\big) \;=\; 0 \quad \text{and} \quad \mathrm{Std}\big(T(\boldsymbol{\xi}|\boldsymbol{x}|)\big) \;=\; \|\boldsymbol{B}\|$$

with the Euclidean norm $\|\boldsymbol{B}\|$ of $\boldsymbol{B} = (B_i)_{i=1}^n$; see the first part of Exercise 4.19. It follows from the second part or the Central Limit Theorem that

$$\big|\pi(\boldsymbol{x}) - \tilde{\pi}(\boldsymbol{x})\big| \;\to\; 0 \quad \text{as} \max_{i=1,\dots,n} |B_i|/\|\boldsymbol{B}\| \to 0$$

with the approximate p-values

$$\begin{aligned}
\tilde{\pi}_\ell(\boldsymbol{x}) \;&:=\; \Phi\big(T(\boldsymbol{x})/\|\boldsymbol{B}\|\big),\\
\tilde{\pi}_r(\boldsymbol{x}) \;&:=\; \Phi\big(-T(\boldsymbol{x})/\|\boldsymbol{B}\|\big) \;=\; 1 - \tilde{\pi}_\ell(\boldsymbol{x}) \quad \text{and}\\
\tilde{\pi}_z(\boldsymbol{x}) \;&:=\; 2\cdot\min\big\{\tilde{\pi}_\ell(\boldsymbol{x}), \tilde{\pi}_r(\boldsymbol{x})\big\} \;=\; 2\,\Phi\big(-|T(\boldsymbol{x})|/\|\boldsymbol{B}\|\big).
\end{aligned}$$

One can also bound the exact p-values from above by the following expressions:

$$\begin{aligned}
\pi_\ell(\boldsymbol{x}) \;&\le\; \exp\Big(-\frac{\min\{T(\boldsymbol{x}),0\}^2}{2\|\boldsymbol{B}\|^2}\Big),\\
\pi_r(\boldsymbol{x}) \;&\le\; \exp\Big(-\frac{\max\{T(\boldsymbol{x}),0\}^2}{2\|\boldsymbol{B}\|^2}\Big),\\
\pi_z(\boldsymbol{x}) \;&\le\; 2\exp\Big(-\frac{T(\boldsymbol{x})^2}{2\|\boldsymbol{B}\|^2}\Big).
\end{aligned}$$

These bounds follow from the second part of Exercise 4.19. They are a special case of a famous inequality due to Hoeffding[4] (1963).

The Center of a Symmetric Distribution

By means of Wilcoxon's signed rank test one may also compute a confidence interval for the unknown median μ of a symmetric distribution P. Precisely, we assume that the random variables $X_1, X_2, \dots, X_n$ are stochastically independent with unknown continuous distribution function F satisfying

$$F(\mu - r) + F(\mu + r) \;=\; 1 \quad \text{for arbitrary } r \in \mathbb{R}; \tag{4.4}$$

in particular, $F(\mu) = 1/2$. Now one can compute $(1-\alpha)$-confidence bounds for μ by applying Wilcoxon's signed rank test to the shifted data vectors

$$\boldsymbol{X} - m \;:=\; (X_i - m)_{i=1}^n$$

for hypothetical values m of μ. Precisely, it follows from assumption (4.4) that $\boldsymbol{X} - \mu$ has a sign-symmetric distribution, and $T(\boldsymbol{X} - \mu)$ has the same distribution as

$$\sum_{i=1}^n \xi_i \cdot i.$$

Hence for an arbitrary threshold c,

$$\left.\begin{aligned}
\mathbb{P}\big(T(\boldsymbol{X}-\mu) \le\;\; c\big)\\
\mathbb{P}\big(T(\boldsymbol{X}-\mu) \ge -c\big)
\end{aligned}\right\} \;=\; \mathbb{P}\Big(\sum_{i=1}^n \xi_i \cdot i \le c\Big).$$

To solve inequalities of type $\pm T(\boldsymbol{X} - \mu) \le c$ for μ, the following representation of Wilcoxon's signed rank statistic is useful:

[4] Wassily Hoeffding (1914-1991): Finnish statistician and probabilist who emigrated 1946 to the USA; one of the founders of nonparametric statistics.

Lemma 4.17 (Tukey[5]). *For a Vektor $\boldsymbol{x} \in \mathbb{R}^n$ with non-zero components,*

$$T(\boldsymbol{x}) \;=\; \tilde{T}(\boldsymbol{x}) := \sum_{1\le i\le j\le n} \operatorname{sign}(x_i + x_j).$$

Under assumption (4.4) the random vector $\boldsymbol{X} - \mu$ satisfies the assumption of Lemma 4.17 with probability one. Hence we may utilise that

$$m \;\mapsto\; \tilde{T}(\boldsymbol{X} - m) = \sum_{1\le i\le j\le n} \operatorname{sign}\Big(\frac{X_i + X_j}{2} - m\Big)$$

is monotone decreasing in m, and that for arbitrary thresholds c,

$$\left.\begin{array}{l} \mathbb{P}\big(\tilde{T}(\boldsymbol{X} - \mu) \le\; c\big) \\ \mathbb{P}\big(\tilde{T}(\boldsymbol{X} - \mu) \ge -c\big) \end{array}\right\} \;=\; \mathbb{P}(2T_n^* - \tilde{n} \le c),$$

where

$$\tilde{n} \;:=\; \frac{n(n+1)}{2} \quad \text{and} \quad T_n^* \;:=\; \sum_{i=1}^{n} 1_{[\xi_i=1]} i.$$

We denote the distribution function of T_n^* with G_n and fix an error probability $\alpha \in (0,1)$. For the unknown center μ we obtain the lower $(1-\alpha)$-confidence bound

$$a_\alpha(\boldsymbol{X}) \;=\; \inf\big\{m \in \mathbb{R} : \tilde{T}(\boldsymbol{X} - m) \le 2G_n^{-1}(1-\alpha) - \tilde{n}\big\},$$

the upper $(1-\alpha)$-confidence bound

$$b_\alpha(\boldsymbol{X}) \;=\; \sup\big\{m \in \mathbb{R} : \tilde{T}(\boldsymbol{X} - m) \ge \tilde{n} - 2G_n^{-1}(1-\alpha)\big\}$$

or the $(1-\alpha)$-confidence interval $\big[a_{\alpha/2}(\boldsymbol{X}), b_{\alpha/2}(\boldsymbol{X})\big]$. To deduce explicit formulae we consider the $\tilde{n}$ pairwise means $(X_i + X_j)/2$, $1 \le i \le j \le n$, in increasing order:

$$W_1 \le W_2 \le \cdots \le W_{\tilde{n}}.$$

For $0 \le k \le \tilde{n}$ and $W_k < m < W_{k+1}$, we obtain $\tilde{T}(\boldsymbol{X} - m) = \tilde{n} - 2k$, where $W_0 := -\infty$ and $W_{\tilde{n}+1} := \infty$. This is less than or equal to $2G_n^{-1}(1-\alpha) - \tilde{n}$ or greater than or less than $\tilde{n} - 2G_n^{-1}(1-\alpha)$ if and only if $k \ge \tilde{n} - G_n^{-1}(1-\alpha)$ or $k \le G_n^{-1}(1-\alpha)$, respectively. Thus

$$a_\alpha(\boldsymbol{X}) \;=\; W_{\tilde{n} - G_n^{-1}(1-\alpha)} \quad \text{and} \quad b_\alpha(\boldsymbol{X}) \;=\; W_{G_n^{-1}(1-\alpha)+1}.$$

Example 4.18 (Gosset's barley data, cont.). For $n = 11$ observations, numerical calculation yield $G_{11}^{-1}(0.95) = 52$ and $G_{11}^{-1}(0.975) = 55$. Since $\tilde{n} = 11 \cdot 12/2 = 66$, we obtain the 95%-confidence interval

$$[W_{11}, W_{56}] \;=\; [-84, 20],$$

which is comparable to the result of student's method. Figure 4.3 shows the function $m \mapsto \tilde{T}(\boldsymbol{X} - m)$ and the resulting one- and twosided 95%-confidence bounds.

[5] John W. Tukey (1915-2000): US-american statistician; co-developer of the fast Fourier transform; important contributions to mathematical and explorative and robust statistics.

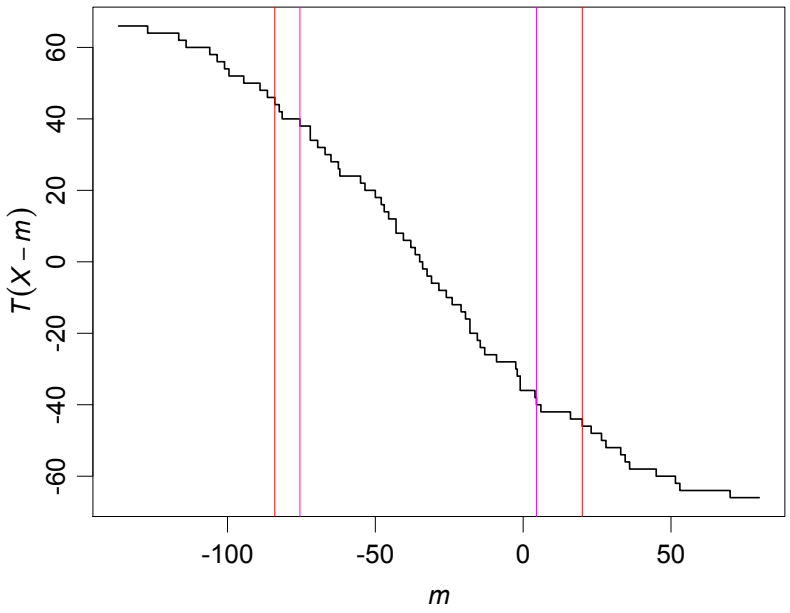

Figure 4.3: The function $m \mapsto \tilde{T}(\boldsymbol{X} - m)$ for Gosset's data.

Proof of Lemma 4.17. Since by assumption $|x_j| > 0$ for all j,

$$R_i(\boldsymbol{x}) \;=\; \#\big\{j : |x_j| < |x_i|\big\} + \frac{1 + \#\big\{j : |x_j| = |x_i|\big\}}{2}.$$

Hence

$$T(\boldsymbol{x}) \;=\; \sum_{i=1}^{n} \operatorname{sign}(x_i) \sum_{j=1}^{n} \Big(1_{[|x_j|<|x_i|]} + \frac{1_{[i=j]} + 1_{[|x_j|=|x_i|]}}{2} \Big) \;=\; \sum_{i=1}^{n} \sum_{j=1}^{n} H_{ij}$$

with

$$H_{ij} \;:=\; \operatorname{sign}(x_i) \Big(1_{[|x_j|<|x_i|]} + \frac{1_{[|x_j|=|x_i|]}}{2} + \frac{1_{[i=j]}}{2} \Big).$$

On the one hand,

$$H_{ii} \;=\; \operatorname{sign}(x_i) \;=\; \operatorname{sign}(x_i + x_i).$$

On the other hand, in case of $i \neq j$ and $|x_j| < |x_i|$,

$$H_{ij} \;=\; \operatorname{sign}(x_i) \;=\; \operatorname{sign}(x_i + x_j) \quad \text{and} \quad H_{ji} \;=\; 0,$$

whereas in case of $i \neq j$ and $|x_j| = |x_i|$,

$$H_{ij} \;=\; \operatorname{sign}(x_i)/2, \quad H_{ji} \;=\; \operatorname{sign}(x_j)/2 \quad \text{and} \quad H_{ij} + H_{ji} \;=\; \operatorname{sign}(x_i + x_j).$$

Consequently,

$$T(\boldsymbol{x}) \;=\; \sum_{1 \le i < j \le n} \big(\underbrace{H_{ij} + H_{ji}}_{=\operatorname{sign}(x_i + x_j)} \big) + \sum_{i=1}^{n} \operatorname{sign}(x_i + x_i) \;=\; \sum_{1 \le i \le j \le n} \operatorname{sign}(x_i + x_j).$$

□

4.4 Asymptotic Considerations and Comparisons

In the previous sections we became familiar with various location parameters which may be considered as an estimator of the 'center' of the underlying distribution P. Now we would like to investigate and compare the precision of these estimators. For a given sample size, this is quite a difficult task, because we know little about their exact distributions. But it turns out that for large sample sizes n and under certain conditions, all these location estimators follow approximately a Gaussian distribution which facilitates a comparison.

We'll start with the sample mean, then consider the sample median, and then we'll introduce a new location parameter which is closely related to Wilcoxon's signed rank test. The proofs are deferred to the end of this section. In all cases we consider random variables Δ_n of the form

$$\Delta_n \;:=\; \sqrt{n}\big(K(X_1, X_2, \ldots, X_n) - K(P)\big)$$

and show that they follow asymptotically, as $n \to \infty$, a Gaussian distribution $\mathcal{N}(0, \tau^2)$ with some $\tau = \tau(P) > 0$. That means, for arbitrary numbers $-\infty \le r < s \le \infty$,

$$\lim_{n\to\infty} \mathbb{P}(r \le \Delta_n \le s) \;=\; \Phi(s/\tau) - \Phi(r/\tau).$$

We indicate this fact briefly as

$$\Delta_n \;\to_d\; \mathcal{N}(0, \tau^2),$$

see also Section A.3 in the appendix. At this point we recommend Exercises 4.26 and 4.27.

The mean. As mentioned already, the sample mean $\overline{X}$ has the following property as an estimator of $\mu = \mu(P)$: By virtue of the Central Limit Theorem,

$$\sqrt{n}(\overline{X} - \mu) \;\to_d\; \mathcal{N}(0, \sigma^2). \tag{4.5}$$

The median and quantiles. Also for the sample quantile $\widehat{q}_\gamma$ and the corresponding quantile $q_\gamma = q_\gamma(P)$ one may derive such a limit theorem under certain regularity assumptions.

Theorem 4.19 (Asymptotic normality of sample quantiles). *For a fixed $\gamma \in (0,1)$ let q_γ be the unique γ-quantile of the distribution P. Precisely, let F be differentiable at q_γ with derivative $F'(q_\gamma) > 0$. Then*

$$\sqrt{n}(\widehat{q}_\gamma - q_\gamma) \;\to_d\; \mathcal{N}(0, \sigma_\gamma^2)$$

with

$$\sigma_\gamma \;:=\; \frac{\sqrt{\gamma(1-\gamma)}}{F'(q_\gamma)}.$$

The Hodges–Lehmann estimator for μ. Suppose that the distribution function F of P is continuous and symmetric around $\mu \in \mathbb{R}$, that means, $F(\mu+r) = 1-F(\mu-r)$ for arbitrary $r \in \mathbb{R}$. In this case, μ is a median as well as the mean (if it exists) of the distribution P. Hence one could estimate μ by $\overline{X}$ or $\widehat{q}_{0.5}$. Here is an alternative location parameter: The absolute value of Wilcoxon's signed-rank statistic $\tilde{T}(\boldsymbol{X} - m)$ becomes minimal if, any only if, m is a sample median of the pairwise means $(X_i + X_j)/2$ (so-called Walsh-means), $1 \le i \le j \le n$. This suggests an estimator which has been proposed and analyzed by Hodges and Lehmann[6] (1963):

$$\widehat{\mu}_W \;:=\; \mathrm{Median}\Big(\frac{X_i + X_j}{2} \;:\; 1 \le i < j \le n\Big).$$

[6] Joseph L. Hodges (1922-2000) and Erich L. Lehmann (1917-2009): Mathematical statisticians at Berkeley University.

We omit index pairs (i, j) with $i = j$. This simplifies the analysis somewhat; the impact is negligible. This Hodges–Lehmann estimator is asymptotically Gaussian, too:

Theorem 4.20 (Asymptotic normality of the Hodges–Lehmann estimator). *Let $F = F_o(\cdot - \mu)$ with a distribution function F_o with bounded and continuous density $f_o = F_o'$ such that $f_o(-x) = f_o(x)$ for all $x \in \mathbb{R}$. Then*

$$\sqrt{n}(\widehat{\mu}_W - \mu) \;\to_d\; \mathcal{N}(0, \sigma_W^2)$$

with

$$\sigma_W \;:=\; \Big(\sqrt{12}\int_{-\infty}^{\infty} f_o(x)^2\,dx\Big)^{-1}.$$

Comparing the three estimators. Suppose that P has a density f of the form $f(x) = f_o(x-\mu)$, where f_o is bounded, continuous and even. In addition let $f_o(0) > 0$ and $\sigma^2 = \int_{-\infty}^{\infty} f_o(x)x^2\,dx < \infty$. Then the three random quantities $\sqrt{n}(\overline{X} - \mu)$, $\sqrt{n}(\widehat{q}_{0.5} - \mu)$ and $\sqrt{n}(\widehat{\mu}_W - \mu)$ are asymptotically Gaussian with mean 0 and standard deviations

$$\sigma, \quad \sigma_{0.5} \;=\; (2f_o(0))^{-1} \quad \text{and} \quad \sigma_W \;=\; \Big(\sqrt{12}\int_{-\infty}^{\infty} f_o(x)^2\,dx\Big)^{-1},$$

respectively. Specifically, suppose that $P = \mathcal{N}(\mu, \sigma^2)$. Then it follows from Exercises 4.28 and 4.29 that

$$\frac{\sigma_{0.5}}{\sigma} \;=\; \sqrt{\pi/2} \;\approx\; 1.2533 \quad \text{and} \quad \frac{\sigma_W}{\sigma} \;=\; \sqrt{\pi/3} \;\approx\; 1.0233.$$

Hence, in case of Gaussian observations X_i, the Hodges–Lehmann estimator is substantially more precise than the sample median, and it is only slightly less precise than the sample mean.

Note that $\sigma_{0.5}/\sigma$ and σ_W/σ can be arbitrarily close to 0, see Exercise 4.30. At the end of this section we'll show that always

$$\frac{\sigma_W}{\sigma} \;\le\; \sqrt{125/108} \;\approx\; 1.0758. \tag{4.6}$$

Equality holds if and only if $f_o(x) = \tau^{-1} f_*(\tau^{-1}x)$ for some $\tau > 0$ with

$$f_*(x) \;=\; 0.75\,\max(1 - x^2, 0).$$

This is the Epanechnikov density which will play an important role in connection with kernel density estimators; see Chapter 5.

Now we prove the previous statements. Unless stated differently, asymptotic statements always refer to $n \to \infty$. The proofs of Theorems 4.19 and 4.20 are based on a more general fact:

Theorem 4.21 (Asymptotic normality of random quantiles). *For $n \ge 1$ let $\widehat{G}_n$ be a random distribution function, and let G be a fixed distribution function. Suppose that the following two properties are satisfied:*

(i) *G is differentiable at a point $x_o \in \mathbb{R}$ with*

$$G(x_o) \;\in\; (0,1) \quad \textit{and} \quad G'(x_o) \;>\; 0.$$

(ii) *There exists a $\tau > 0$ such that for any sequence $(x_n)_{n\ge1}$ with limit x_o,*

$$\sqrt{n}\big(\widehat{G}_n(x_n) - G(x_n)\big) \;\to_d\; \mathcal{N}(0, \tau^2).$$

For $n \ge 1$ let M_n be a random variable such that

$$\widehat{G}_n(M_n -) \;\le\; G(x_o) \;\le\; \widehat{G}_n(M_n).$$

Then

$$\sqrt{n}(M_n - x_o) \;\to_d\; \mathcal{N}\big(0, \tau^2/G'(x_o)^2\big).$$

Proof of Theorem 4.21. According to Exercise 4.27 (a), it suffices to show that for an arbitrary fixed number $r \in \mathbb{R}$,

$$\mathbb{P}\big(\sqrt{n}(M_n - x_o) \le r\big) \;\to\; \Phi(rG'(x_o)/\tau).$$

But

$$\sqrt{n}(M_n - x_o) \;\le\; r \quad \text{if and only if} \quad M_n \;\le\; x_n := x_o + r/\sqrt{n}.$$

Moreover, the assumption on M_n implies that

$$\big[\widehat{G}_n(x_n) > G(x_o)\big] \;\subset\; [M_n \le x_n] \;\subset\; \big[\widehat{G}_n(x_n) \ge G(x_o)\big].$$

Note also that

$$\begin{aligned}
\widehat{G}_n(x_n) \;>\; G(x_o) \quad &\text{if and only if} \quad \sqrt{n}\big(\widehat{G}_n(x_n) - G(x_n)\big) \;>\; y_n, \\
\widehat{G}_n(x_n) \;\ge\; G(x_o) \quad &\text{if and only if} \quad \sqrt{n}\big(\widehat{G}_n(x_n) - G(x_n)\big) \;\ge\; y_n,
\end{aligned}$$

where

$$y_n \;:=\; \sqrt{n}\big[G(x_o) - G\big(x_o + r/\sqrt{n}\big)\big] \;\to\; -rG'(x_o).$$

Consequently, it follows from $\sqrt{n}\big(\widehat{G}_n(x_n) - G(x_n)\big) \to_d \mathcal{N}(0, \tau^2)$ that

$$\begin{aligned}
\limsup_{n\to\infty} \mathbb{P}\big(\sqrt{n}(M_n - x_o) \le r\big) \;&\le\; \limsup_{n\to\infty} \mathbb{P}\Big(\sqrt{n}\big(\widehat{G}_n(x_n) - G(x_n)\big) \ge y_n\Big) \\
&=\; \limsup_{n\to\infty} \Phi(-y_n/\tau) \;=\; \Phi(rG'(x_o)/\tau), \\
\liminf_{n\to\infty} \mathbb{P}\big(\sqrt{n}(M_n - x_o) \le r\big) \;&\ge\; \liminf_{n\to\infty} \mathbb{P}\Big(\sqrt{n}\big(\widehat{G}_n(x_n) - G(x_n)\big) > y_n\Big) \\
&=\; \liminf_{n\to\infty} \Phi(-y_n/\tau) \;=\; \Phi(rG'(x_o)/\tau),
\end{aligned}$$

see also Exercise 4.27 (a). □

Proof of Theorem 4.19. We apply Theorem 4.21 with $\widehat{G}_n := \widehat{F}$, $G := F$ and $x_o := q_\gamma$. Indeed, for any sequence $(x_n)_{n\ge1}$ in $\mathbb{R}$ with limit q_γ,

$$\sqrt{n}\big(\widehat{F}(x_n) - F(x_n)\big) \;\to_d\; \mathcal{N}(0, \gamma(1-\gamma))$$

by the Central Limit Theorem. Precisely, the random variable on the left-hand side may be written as $\sum_{i=1}^n Y_{ni}$ with $Y_{ni} := n^{-1/2}(1_{[X_i \le x_n]} - F(x_n))$, and $\mathbb{E}(Y_{ni}) = 0$, $|Y_{ni}| \le 1/\sqrt{n}$, and $\sum_{i=1}^n \mathrm{Var}(Y_{ni}) = F(x_n)(1 - F(x_n)) \to \gamma(1-\gamma)$. Thus the assumptions of Theorem 4.21 are satisfied with $\tau = \sqrt{\gamma(1-\gamma)}$, and $\tau/G'(x_o) = \sigma_\gamma$. □

An essential tool for the proof of Theorem 4.20 is the following special case of a general result due to Hoeffding (1948) about so-called U-statistics:

Lemma 4.22 (Hoeffding). *Let*

$$U \;:=\; \binom{n}{2}^{-1} \sum_{1\le i<j\le n} h(X_i, X_j)$$

with independent, identically distributed random variables $X_1, X_2, \ldots, X_n \in \mathbb{R}$ and a measurable function $h : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, where $h(x, y) = h(y, x)$ for all $x, y \in \mathbb{R}$ and $\mathbb{E}\big(h(X_1, X_2)^2\big) < \infty$. With $h_1(x) := \mathbb{E}\, h(x, X_2)$ and $h_0 := \mathbb{E}\, h(X_1, X_2) = \mathbb{E}\, h_1(X_1)$,

$$U \;=\; h_0 + \frac{2}{n}\sum_{i=1}^n \big(h_1(X_i) - h_0\big) + R,$$

where

$$\mathbb{E}(R^2) \;\le\; \frac{2\,\mathrm{Var}(h(X_1, X_2))}{n(n-1)}.$$

Proof of Lemma 4.22. The proof is essentially an application of Fubini's Theorem as presented in Section A.5 in the appendix. First of all, with the distribution P of X_1 we can write $h_1(x) = \int h(x,y)\,P(dy)$, and h_0 is equal to $\mathbb{E}\,h(X_1,X_2) = \int\int h(x,y)\,P(dy)\,P(dx) = \int h_1(x)\,P(dx) = \mathbb{E}\,h_1(X_1)$. Now we define

$$h_2(x,y) \;:=\; h(x,y) - h_1(x) - h_1(y) + h_0.$$

This function h_2 is also symmetric in its two arguments, and

$$\int h_2(x,y)\,P(dy) \;=\; h_1(x) - h_1(x) - h_0 + h_0 \;=\; 0.$$

In particular, $\mathbb{E}\,h_2(X_1,X_2) = 0$. More generally, for arbitrary indices i,j,k,ℓ with $i \neq j$, $k \neq \ell$ and measurable functions $g : \mathbb{R}\times\mathbb{R} \to \mathbb{R}$ with $\mathbb{E}\big(g(X_k,X_\ell)^2\big) < \infty$ we get the equation

$$\mathbb{E}\big(h_2(X_i,X_j)g(X_k,X_\ell)\big) \;=\; 0 \quad \text{if } \{i,j\} \neq \{k,\ell\}. \tag{4.7}$$

For in case of $\{i,j\}\cap\{k,\ell\} = \emptyset$, independence of (X_i,X_j) and (X_k,X_ℓ) implies the equalities $\mathbb{E}\big(h_2(X_i,X_j)g(X_k,X_\ell)\big) = \mathbb{E}\,h_2(X_i,X_j)\,\mathbb{E}\,g(X_k,X_\ell) = 0$. In case of $k = i$ and $\ell \notin \{i,j\}$,

$$\begin{aligned}\mathbb{E}\big(h_2(X_i,X_j)g(X_k,X_\ell)\big) \;&=\; \int\int\int h_2(x,y)g(x,z)\,P(dy)\,P(dx)\,P(dz)\\ &=\; \int\int g(x,z)\Big(\int h_2(x,y)\,P(dy)\Big)\,P(dx)\,P(dz)\\ &=\; 0.\end{aligned}$$

Hence if we write $h(x,y) = h_0 + (h_1(x)-h_0) + (h_1(y)-h_0) + h_2(x,y)$, then we obtain the representation

$$\begin{aligned}U \;&=\; \binom{n}{2}^{-1} \sum_{1\le i<j\le n} \big(h_0 + (h_1(X_i)-h_0) + (h_1(X_j)-h_0) + h_2(X_i,X_j)\big)\\ &=\; h_0 + \frac{2}{n}\sum_{i=1}^n (h_1(X_i)-h_0) + R\end{aligned}$$

with

$$R \;:=\; \binom{n}{2}^{-1} \sum_{1\le i<j\le n} h_2(X_i,X_j).$$

According to (4.7), the random variables $h_2(X_i,X_j)$, $1 \le i < j \le n$, are centered and uncorrelated, so

$$\mathbb{E}(R^2) \;=\; \binom{n}{2}^{-2} \sum_{1\le i<j\le n} \mathrm{Var}(h_2(X_i,X_j)) \;=\; \frac{2\,\mathrm{Var}(h_2(X_1,X_2))}{n(n-1)}.$$

Moreover it follows from (4.7) that

$$\mathrm{Var}(h(X_1,X_2)) \;=\; 2\,\mathrm{Var}(h_1(X)) + \mathrm{Var}(h_2(X_1,X_2)),$$

so $\mathrm{Var}(h_2(X_1,X_2)) \le \mathrm{Var}(h(X_1,X_2))$. □

Proof of Theorem 4.20. The estimator $\widehat{\mu}_W$ is a median of the empirical distribution function $\widehat{G}_n$ of the $m_n = \binom{n}{2}$ Walsh-means $(X_i+X_j)/2$, $1 \le i < j \le n$. The latter function is given by

$$\widehat{G}_n(x) \;=\; m_n^{-1} \sum_{1\le i<j\le n} 1_{[(X_i+X_j)/2\le x]} \;=\; m_n^{-1} \sum_{1\le i<j\le n} 1_{[X_i+X_j\le 2x]},$$

and this is an estimator of

$$G(x) := \mathbb{E}\,\widehat{G}(x) \;=\; \mathbb{P}(X_1 + X_2 \le 2x).$$

Without loss of generality let $\mu = 0$. Otherwise just replace X_i with $X_i - \mu$. In that case, the X_i have distribution function F_o, density function f_o, and they are symmetrically distributed around 0. Now we want to verify the conditions of Theorem 4.21 with $x_o = 0$.

First of all, by means of Fubini's theorem,

$$G(x) \;=\; \mathbb{E}\,\mathbb{P}(X_1 + X_2 \le 2x \,|\, X_1) \;=\; \mathbb{E}\,F_o(2x - X_1) \;=\; \mathbb{E}\,F_o(2x + X_1),$$

because the distribution of X_1 is symmetric around 0. In particular,

$$G(0) \;=\; \mathbb{E}\,F_o(X_1) \;=\; \int_0^1 u\,du \;=\; 1/2,$$

because $F_o(X_1)$ is uniformly distributed on $[0,1]$. To verify the latter claim, recall from Chapter 3 that X_1 is distributed like $F_o^{-1}(U)$ with $U \sim \mathrm{Unif}[0,1]$. Hence, continuity of F_o implies that $F_o(X_1)$ is distributed like $F_o(F_o^{-1}(U)) = U$. Furthermore, by dominated convergence,

$$G'(0) \;=\; \lim_{x\to 0} \mathbb{E}\Big(\frac{F_o(2x + X_1) - F_o(X_1)}{x}\Big) \;=\; 2\,\mathbb{E}\,f_o(X_1) \;=\; 2\int f_o(t)^2\,dt,$$

because

$$\lim_{x\to 0} \frac{F_o(2x + X_1) - F_o(X_1)}{x} \;=\; 2f_o(X_1) \quad \text{and} \quad \Big|\frac{F_o(2x + X_1) - F_o(X_1)}{x}\Big| \;\le\; 2\|f_o\|_\infty.$$

Moreover, if $(x_n)_n$ is any sequence converging to 0, then

$$\widehat{G}(x_n) \;=\; m_n^{-1} \sum_{1\le i<j\le n} h_n(X_i, X_j) \quad \text{with} \quad h_n(x,y) \;:=\; 1_{[x+y\le 2x_n]}.$$

Now we apply Hoeffding's Lemma. With

$$h_{n,1}(x) \;:=\; \mathbb{E}\,h_n(x, X_2) \;=\; F_o(2x_n - x)$$

we may write $\mathbb{E}\,h_n(X_1, X_2) = \mathbb{E}\,h_{n,1}(X_1) = G(x_n)$, and

$$\widehat{G}(x_n) \;=\; G(x_n) + \frac{2}{n}\sum_{i=1}^n \big(F_o(2x_n - X_i) - G(x_n)\big) + R_n$$

where $\mathbb{E}\,|R_n| = O(n^{-1})$. But

$$\big|F_o(2x_n - X_i) - G(x_n)\big| \;\le\; 1,$$

and another application of dominated convergence yields that

$$\begin{aligned}
\mathrm{Var}\big(F_o(2x_n - X_i)\big) \;&=\; \mathbb{E}\big(F_o(2x_n - X_i)^2\big) - G(x_n)^2 \\
&\to\; \mathbb{E}\big(F_o(-X_1)^2\big) - 1/4 \\
&=\; \mathbb{E}\big(F_o(X_1)^2\big) - 1/4 \;=\; \int_0^1 u^2\,du - 1/4 \;=\; 1/12.
\end{aligned}$$

Thus, it follows from the Central Limit Theorem that

$$\sqrt{n}\big(\widehat{G}(x_n) - G(x_n)\big) \;\to_d\; \mathcal{N}(0, \tau^2) \quad \text{with } \tau := 2/\sqrt{12}.$$

Consequently, the conditions of Theorem 4.21 are satisfied, and $\tau/G'(x_o) = \sigma_W$. □

Proof of (4.6)**.** Maximizing σ_W^2/σ^2 is equivalent to minimizing $12\sigma^2\big(\int_{-\infty}^{\infty} f_o(x)^2\,dx\big)^2$. If one replaces $f_o(x)$ with $\tau^{-1}f_o(\tau^{-1}x)$ for some $\tau > 0$, then this product remains unchanged. Thus we may fix an arbitrary value for $\sigma^2 = \int_{-\infty}^{\infty} f_o(x)x^2\,dx$ and minimise $\int_{-\infty}^{\infty} f_o(x)^2\,dx$ under this constraint. In addition we have to remember the constraints that f_o has to be non-negative, even, bounded and continuous with $\int_{-\infty}^{\infty} f_o(x)\,dx = 1$.

To this end we use Lagrange's method: We consider a linear combination of the three integrals, namely,

$$\int_{-\infty}^{\infty} f(x)^2\,dx + a\int_{-\infty}^{\infty} f(x)x^2\,dx - b\int_{-\infty}^{\infty} f(x)\,dx \;=\; \int_{-\infty}^{\infty}\big(f(x)^2 - f(x)(b-ax^2)\big)\,dx$$

for constants $a, b > 0$, and minimise this expression over *all* measurable functions $f : \mathbb{R} \to [0,\infty)$. Elementary calculations show that

$$f(x)^2 - f(x)(b-ax^2) \;\geq\; -\max(b-ax^2,0)^2/4$$

with equality if and only if $f(x) = f_*(x) := \max(b-ax^2,0)/2$. Especially for $a = b = 1.5$ we obtain $f_*(x) = 0.75\max(1-x^2,0)$, a bounded, continuous and even probability density. Hence we know that

$$\int_{-\infty}^{\infty} f_o(x)^2\,dx \;\geq\; \int_{-\infty}^{\infty} f_*(x)^2\,dx$$

for any probability density f_o with the additional property that the integral $\int_{-\infty}^{\infty} f_o(x)x^2\,dx$ is equal to $\int_{-\infty}^{\infty} f_*(x)x^2\,dx$. Equality holds if and only if $f_o = f_*$ almost everywhere.

These considerations show that σ_W^2/σ^2 becomes maximal for $f_o = f_*$, and elementary calculations yield the value $125/108$. □

4.5 Exercises

Exercise 4.1. Show that

$$\mathbb{E}\Big(\sum_{i=1}^{n}(X_i - \overline{X})^2\Big) \;=\; (n-1)\sigma^2.$$

Exercise 4.2 (Upper bounds for the variance). Let X be a random variable with values in a given interval $[a,b] \subset \mathbb{R}$. Show that

$$\mathrm{Var}(X) \;\leq\; (\mathbb{E}(X)-a)(b-\mathbb{E}(X)) \;\leq\; (b-a)^2/4.$$

Exercise 4.3 (An inequality for student quantiles). We consider stochastically independent random variables Z and Y, where Z is standard Gaussian while $Y > 0$ and $\mathbb{E}(Y) = 1$.

(a) Show that

$$\mathbb{P}\big(Z/\sqrt{Y} > t\big) \;=\; \mathbb{E}\big(\Phi(-t\sqrt{Y})\big)$$

for arbitrary $t \in \mathbb{R}$. For that purpose use Fubini's theorem (Section A.5 in the appendix) or consider the special case that Y has a discrete distribution.

(b) Suppose that $\mathbb{P}(Y \neq 1) > 0$. Show that

$$\mathbb{P}\big(Z/\sqrt{Y} > t\big) \;>\; \Phi(-t)$$

for each $t > 0$. Investigate for that purpose the function $[0,\infty) \ni y \mapsto \Phi(-t\sqrt{y})$ and apply Jensen's inequality (Section A.6 in the appendix).

(c) Deduce now that

$$t_{k;\beta} \;>\; \Phi^{-1}(\beta)$$

for arbitrary $k \in \mathbb{N}$ and $\beta \in (1/2,1)$.

Exercise 4.4. Use the data in Exercise 3.10 to compute an approximate 90%-confidence interval for the mean life time of house cats.

Exercise 4.5. In our example for biased sampling we considered a population of mothers with the variable $Y =$ number of children and the relative proportions $q_k = \mathbb{P}(Y = k)$ for $k = 1, 2, 3, \ldots$. Now we consider the population of corresponding children and the variable

$$\tilde{X} \;:=\; \text{number of older siblings (with same mother)}$$

with the relative proportions $\tilde{p}_j = \mathbb{P}(\tilde{X} = j)$ for $j = 0, 1, 2, \ldots$. Establish a relationship between the distributions $(\tilde{p}_j)_{j\geq 0}$ and $(q_k)_{k\geq 1}$. Show also that $\nu = 1/\tilde{p}_0$.

When interviewing $n = 173$ young people we obtained the following absolute frequencies $\tilde{H}_j = \#\{i : \tilde{X}_i = j\}$:

j	0	1	2	3	4	5	≥ 6
$\tilde{H}_j$	83	56	23	6	3	2	0

Compute point estimates for the probabilities q_k and for ν. Further compute a 95%-confidence interval for ν.

Exercise 4.6. Verify formula (4.1).

Exercise 4.7 (Norming of scale parameters). The various scale parameters $K = K(X_1, \ldots, X_n)$ may be viewed as point estimators of certain parameters $K(P)$. Suppose the unknown distribution P is equal to $\mathcal{N}(\mu, \sigma^2)$, so $F(r) = \Phi((r - \mu)/\sigma)$. Which parameters $K(P)$ correspond to (i) the inter quartile range IQR, (ii) the median MAD of absolute deviations and (iii) Gini's scale parameter G? How should one modify these scale parameters such that they estimate σ correctly?

Exercise 4.8 (The range as an estimator). Show that the range $X_{(n)} - X_{(1)}$ is a consistent estimator for

$$\text{range}(P) \;:=\; q_1(P) - q_0(P),$$

where $q_0(P) := \inf\{r \in \mathbb{R} : F(r) > 0\}$ and $q_1(P) := \sup\{r \in \mathbb{R} : F(r) < 1\}$. More precisely, one should show that

$$\mathbb{P}\big([X_{(1)}, X_{(n)}] \subset [q_0(P), q_1(P)]\big) \;=\; 1 \;=\; \lim_{n\to\infty} \mathbb{P}\big([r_0, r_1] \subset [X_{(1)}, X_{(n)}]\big)$$

for arbitrary fixed numbers $q_0(P) < r_0 < r_1 < q_1(P)$.

Exercise 4.9 (L-statistics). A feature of the form

$$L(X_1, \ldots, X_n) \;:=\; \sum_{i=1}^{n} w_i X_{(i)}$$

with fixed constants $w_1, w_2, \ldots, w_n \in \mathbb{R}$ is called an L-statistic.

(a) Show that sample mean, τ-trimmed mean, sample-γ-quantile, range, inter quartile range and Gini's scale parameter are special L-statistics.

(b) Under which condition on the w_i is $L(X_1, \ldots, X_n)$ a location or scale parameter, respectively?

Exercise 4.10 (L-statistics as estimators). We consider an L-statistic of the form

$$L = L(X_1, \ldots, X_n) \;:=\; \frac{1}{n} \sum_{i=1}^{n} w\Big(\frac{i - 0.5}{n}\Big) X_{(i)}$$

with a certain function $w : (0, 1) \to \mathbb{R}$. Which parameter $L(P)$ is estimated by $L(X_1, \ldots, X_n)$ if one assumes P to have a density function f? Hint: $(i - 0.5)/n = \big(\widehat{F}(X_{(i)}) + \widehat{F}(X_{(i)} -)\big)/2$.

Exercise 4.11 (Gamma distributions, I). Let $Y \sim \mathrm{Gamma}(a,1)$ with $a > 0$. Show that

$$\mathbb{E}(Y^k) \;=\; \Gamma(a+k)/\Gamma(a)$$

for arbitrary $k > 0$. Now let $P = \mathrm{Gamma}(a,b)$ with $a, b > 0$. Show that

$$\mu(P) \;=\; ab, \quad \sigma(P) \;=\; \sqrt{a}\, b$$

and

$$\mathrm{Skewness}(P) \;=\; 2/\sqrt{a}.$$

Exercise 4.12 (Moment-generating functions and shape parameters). Let X be a random variable with distribution P on $\mathbb{R}$. The moment-generating function of X (of P) is defined as the function $\mathbb{R} \ni t \mapsto m_X(t) := \mathbb{E}\exp(tX) \in (0,\infty]$. Suppose that for some number $t_o > 0$ both $m_X(t_o)$ and $m_X(-t_o)$ are finite.

(a) Show that the latter assumption is equivalent to $\mathbb{E}\exp(t_o|X|) < \infty$. Then show that $\mathbb{E}(|X|^k) < \infty$ for all $k \in \mathbb{N}$ and that

$$m_X(t) \;=\; \sum_{k=0}^{\infty} \mathbb{E}(X^j) t^j/j! \;<\; \infty \quad \text{for all } t \in [-t_o, t_o].$$

In particular, m_X is arbitrarily often differentiable on the interval $[-t_o, t_o]$, and the k-th derivative satisfies the equation

$$\mathbb{E}(X^k) \;=\; m_X^{(k)}(0).$$

This explains the name 'moment-generating function'.

(b) Show that

$$\log m_X(t) \;=\; \mu(P)t + \sigma(P)^2 t^2/2 + O(t^3) \quad \text{as } t \to 0.$$

(c) Show that the standardised random variable $Z := (X - \mu(P))/\sigma(P)$ satisfies

$$\log m_Z(t) \;=\; t^2/2 + \mathrm{Skewness}(P) t^3/6 + \mathrm{Kurtosis}(P) t^4/24 + O(t^5) \quad \text{as } t \to 0.$$

Hint to parts (b) and (c): Use the Taylor-expansion $\log(1+y) = y - y^2/2 + O(y^3)$ as $y \to 0$. Apply this to $y = m_X(t) - 1$ and $y = m_Z(t) - 1$.

Exercise 4.13 (Moments of the standard Gaussian distribution). Show that

$$\mathbb{E}\exp(tZ) \;=\; \exp(t^2/2)$$

for a standard Gaussian random variable Z and $t \in \mathbb{R}$. Now determine by means of Exercise 4.12 the moments $\mathbb{E}(Z^k)$, $k \in \mathbb{N}$.

Exercise 4.14 (Gamma distributions, II). Let X be a random variable with distribution $P = \mathrm{Gamma}(a,b)$, $a, b > 0$. Show that its moment-generating function (Exercise 4.12) is given by

$$m_X(t) \;=\; \begin{cases} (1-bt)^{-a} & \text{if } t < 1/b, \\ \infty & \text{else.} \end{cases}$$

Show that $Z := (X - \mu(P))/\sigma(P)$ satisfies

$$\log m_Z(t) \;=\; \sum_{k=2}^{\infty} a^{1-k/2} t^k/k.$$

Determine now skewness and kurtosis of P.

Exercise 4.15. Imagine a real sample $X_1, \ldots, X_n$ such that $\widehat{q}_{0.25} < \widehat{q}_{0.5} < \widehat{q}_{0.75}$. Let

$$K(X_1, \ldots, X_n) \;:=\; \log\Big(\frac{\widehat{q}_{0.75} - \widehat{q}_{0.5}}{\widehat{q}_{0.5} - \widehat{q}_{0.25}}\Big).$$

Is this a location, scale or shape parameter? What does it measure?

Exercise 4.16 (Robustness of the median). Imagine a data set with $n = 11$ real values $X_1, \ldots, X_n$. How large or small may $\widehat{q}_{0.5}$ get if one replaces an arbitrary value X_i by an arbitrary different number? (Formulate your result by means of the order statistics $X_{(i)}$.) Generalise your finding to arbitrary sample sizes n and arbitrary numbers k of observations which may be modified.

Exercise 4.17 (Robustness of quantiles). Refine the considerations in the proof of Lemma 4.8 as follows: Determine for $k \in \{1, 2, \ldots, n\}$ a maximal index $\ell = \ell(k, n) \in \{0, 1, \ldots, n\}$ and a minimal index $m = m(k, n) \in \{1, \ldots, n, n+1\}$ such that

$$\widehat{q}_\gamma(Y_1, Y_2, \ldots, Y_n) \;\in\; [X_{(\ell)}, X_{(m)}]$$

whenever $\#\{i : Y_i \neq X_i\} \leq k$.

Exercise 4.18. Show that the median of absolute deviations has breakdown point $1/2$.

Exercise 4.19 (Sign tests and Hoeffding's inequality). Let $\boldsymbol{b} \in \mathbb{R}^n$ be a fixed unit vector, i.e. $\|\boldsymbol{b}\|^2 := \sum_{i=1}^n b_i^2 = 1$, and let $\boldsymbol{\xi}$ be uniformly distributed on $\{-1, 1\}^n$. Now we investigate the random variable $T := \sum_{i=1}^n \xi_i b_i$.

(a) Show that $\mathbb{E}\, h(T) = 0$ for any odd function $h : \mathbb{R} \to \mathbb{R}$. In particular, $\mathbb{E}(T^k) = 0$ for $k = 1, 3, 5, \ldots$.

Show that $\mathbb{E}(T^2) = 1$ and $\mathbb{E}(T^4) = 3 + \sum_{i=1}^n b_i^4 \leq 3 + \|\boldsymbol{b}\|_\infty^2$, where $\|\boldsymbol{b}\|_\infty$ stands for $\max_{i=1,\ldots,n} |b_i|$.

(b) Show that for arbitrary $s \in \mathbb{R}$,

$$\log \mathbb{E} \exp(sT) \;=\; \sum_{i=1}^n \log\cosh(sb_i) \;\begin{cases} \leq\; s^2/2, \\ \geq\; \big(1 - \tanh(s\|\boldsymbol{b}\|_\infty)^2\big) s^2/2. \end{cases} \tag{4.8}$$

Now show that for arbitrary $c \geq 0$ and $s \geq 0$,

$$\mathbb{P}(T \geq c) \;\leq\; \mathbb{E} \exp(sT - sc) \;\leq\; \exp(s^2/2 - sc).$$

Deduce from this inequality that

$$\mathbb{P}(T \geq c) \;\leq\; \exp(-c^2/2) \quad \text{and} \quad \mathbb{P}(T \leq -c) \;\leq\; \exp(-c^2/2).$$

Hint for (4.8): $h(x) := \log\cosh(x)$ satisfies the equations $h(0) = h'(0) = 0$ and $h''(x) = 1 - \tanh(x)^2$.

Exercise 4.20. Let $r_1, r_2, \ldots, r_n$ be the ranks of real numbers $x_1, x_2, \ldots, x_n$. According to Exercise 3.6, $\sum_{i=1}^n r_i = n(n+1)/2$ and $\sum_{i=1}^n r_i^2 \leq n(n+1)(2n+1)/6$ with equality if the numbers $x_1, x_2, \ldots, x_n$ are pairwise different.

Deduce from these facts and Exercise 4.19 that the two-sided p-value for Wilcoxon's signed-rank test satisfies the inequality

$$\pi_z(\boldsymbol{X}) \;\leq\; 2 \exp\Big(-\frac{3\, T(\boldsymbol{X})^2}{N(N+1)(2N+1)}\Big).$$

1960	1970	1960	1970	1960	1970
10.1	20.4	10.6	22.1	8.2	10.2
4.9	9.8	11.5	13.7	17.3	24.7
12.4	15.4	11.1	12.7	8.6	13.3
10.0	18.4	4.4	3.9	13.0	14.0
9.3	11.1	11.7	16.9	9.1	16.2
7.9	8.2	4.5	12.6	8.1	17.8
17.7	13.1	11.0	15.6	10.8	14.7
12.5	12.6	8.9	7.9	4.4	11.2
6.4	14.9	3.8	10.5	14.2	15.3
6.6	11.4	6.2	5.5	3.3	6.6

Table 4.6: Murder rates in 30 US-american cities.

Athlete	Best time for 200 m	Best time for 100 m
L. Christie	20.09	9.97
J. Regis	20.32	10.31
M. Rosswess	20.51	10.40
A. Carrott	20.76	10.56
T. Bennett	20.90	10.92
A. Mafe	20.94	10.64
D. Reid	21.00	10.54
P. Snoddy	21.14	10.85
L. Stapleton	21.17	10.71
C. Jackson	21.19	10.56

Table 4.7: Best times of ten sprinters.

Exercise 4.21. Table 4.6 shows the murder rates in the years 1960 and 1970 for a random sample of $n = 30$ cities in the southern USA. (These rates are the number of murders per 100'000 inhabitants). How could one try to verify that the murder rates of these two years are systematically different? Analyze the data with a test level of $\alpha = 0.01$. If you prefer work without any statistics software, Table 3.2 might be useful.

Exercise 4.22. Table 4.7 contains the record times (in seconds) of ten sprinters from Great Britain over 200 m and 100 m for the year 1988. These are all runners who finished the distance of 200 m in less than 21.20 seconds and provided a personal record for 100 m as well. One could guess that the average spead over 200 m is higher than over 100 m because the starting phase has less influence. On the other hand, it is possible that over a distance of 200 m one has to run more economically than over 100 m. Analyze the data with a suitable method and test level $\alpha = 0.05$ and draw a conclusion, if possible. Do you see a potential problem with the choice of the athletes?

Exercise 4.23. Consider again the data from Exercise 4.21. Let P be the distribution of the variable 'murder rate 1960 minus murder rate 1970' for all cities in the southern USA.

(a) Compute a 95%-confidence interval for the median μ of P by means of the method in Section 3.3.

(b) Now compute a 95%-confidence interval for the median μ of P under the idealised assumption, that P has a continuous distribution function and is symmetric around μ. Take into account that the data in Table 4.6 have been rounded to one decimal digit.

Exercise 4.24. Determine the breakdown point of the Hodges–Lehmann estimator.

Exercise 4.25. In addition to $\widehat{\mu}_W$, Bickel and Lehmann (1976) proposed the scale parameter

$$\widehat{\sigma}_W \;:=\; \mathrm{Median}\big(|X_i - X_j| : 1 \le i < j \le n\big).$$

Write a program to compute both $\widehat{\mu}_W$ and $\widehat{\sigma}_W$. How should a factor $c > 0$ be chosen such that $c\widehat{\sigma}_W$ estimates the standard deviation of a Gaussian distribution correctly?

Exercise 4.26 (Uniform convergence). Let $(h_n)_n$ be a sequence of monotone increasing functions $h_n : \mathbb{R} \to [0,1]$ which converges pointwise to a monotone increasing function $h : \mathbb{R} \to [0,1]$. Further let h be continuous with $\lim_{x\to-\infty} h(x) = 0$ and $\lim_{x\to\infty} h(x) = 1$. Show that even

$$\lim_{n\to\infty} \sup_{x\in\mathbb{R}} \big|h_n(x) - h(x)\big| \;=\; 0.$$

Exercise 4.27 (Convergence in distribution). Let $Z_1, Z_2, Z_3, \ldots$ be real-valued random variables, and let Q be a probability distribution on $\mathbb{R}$ with continuous distribution function H.

(a) Show that the following statements are equivalent:

(a.1) For arbitrary $r \in \mathbb{R}$,

$$\lim_{n\to\infty} \mathbb{P}(Z_n \le r) \;=\; H(r).$$

(a.2) For arbitrary $r \in \mathbb{R}$,

$$\lim_{n\to\infty} \mathbb{P}(Z_n < r) \;=\; H(r).$$

(a.3)

$$\lim_{n\to\infty} \sup_{\text{intervals } B\subset\mathbb{R}} \big|\mathbb{P}(Z_n \in B) - Q(B)\big| \;=\; 0.$$

(b) In addition to $Z_1, Z_2, Z_3, \ldots$ let $A_1, A_2, A_3, \ldots$ and $S_1, S_2, S_3, \ldots$ be further random variables such that $A_n \to_p 0$ and $S_n \to_p 1$. That means, for arbitrary $\delta > 0$,

$$\lim_{n\to\infty} \mathbb{P}(|A_n| \ge \delta) \;=\; 0 \;=\; \lim_{n\to\infty} \mathbb{P}(|S_n - 1| \ge \delta).$$

Show that statements (a.1-3) remain valid if Z_n is replaced with $\tilde{Z}_n := A_n + S_n Z_n$.

Exercise 4.28 (Comparing three estimators). Suppose we replace each random variable $X_i \sim f_o$ with τX_i, where τ is a fixed positive constant. What is the effect of this on μ and f_o? Show that all asymptotic variances σ^2, $\sigma^2_{0.5}$ and σ^2_W increase by a factor of τ^2.

Exercise 4.29 (Comparing three estimators). Compute the three asymptotic variances σ^2, $\sigma^2_{0.5}$ and σ^2_W for the density function f_o of the

(a) standard normal distribution, $f_o(x) = \phi(x) = (2\pi)^{-1/2}\exp(-x^2/2)$,
(b) Laplace distribution, $f_o(x) = \exp(-|x|)/2$,
(c) logistic distribution, $f_o(x) = e^x/(e^x+1)^2$,
(d) Epanechnikov distribution, $f_o(x) = 0.75\,\max(1-x^2, 0)$.

Hint for (c): For a random variable X with logistic distribution, its moment-generating function is given by $\mathbb{E}(e^{tX}) = \pi t/\sin(\pi t)$ for $|t| < 1$. (This can be proved by means of the Residual Theorem from Complex Analysis.) Now use Exercise 4.12 (a).

Exercise 4.30 (Comparing three estimators). Compute the three asymptotic variances σ^2, $\sigma^2_{0.5}$ and σ^2_W for the density function $f_o(x) := (1-\epsilon)\phi(x) + \epsilon^2\phi(\epsilon x)$, where ϕ is the standard Gaussian density and $\epsilon \in (0,1)$. This is the density of the mixture distribution $(1-\epsilon)\mathcal{N}(0,1) + \epsilon\mathcal{N}(0,\epsilon^{-2})$. Deduce that $\sigma_{0.5}/\sigma \to 0$ and $\sigma_W/\sigma \to 0$ as $\epsilon \to 0+$.

Exercise 4.31 (Other representations of variances). To illustrate Hoeffding's Lemma 4.22 we consider the variance $\sigma^2 = \sigma^2(P)$ and the sample variance S^2. Show that

$$\sigma^2(P) \;=\; \mathbb{E}\Big(\frac{(X_1 - X_2)^2}{2}\Big) \quad \text{and} \quad S^2 \;=\; \binom{n}{2}^{-1} \sum_{1 \le i < j \le n} \frac{(X_i - X_j)^2}{2}.$$

Show also that

$$S^2 \;=\; \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 + R,$$

where $\mathbb{E}(R^2) \le \big(\mathbb{E}((X_1 - \mu)^4) + \sigma^4\big)/(n(n-1))$.

Exercise 4.32 (Sequential computation of sample mean and variance). Let $X_1, X_2, X_3, \ldots$ be real numbers, and for $n = 1, 2, 3, \ldots$ let

$$\bar{X}_n \;:=\; \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{and} \quad S_n^2 \;:=\; \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

with $S_1^2 := 0$.

(a) Write $\bar{X}_{n+1}$ as a simple function of $\bar{X}_n$ and $\Delta_n := X_{n+1} - \bar{X}_n$.

(b) Write S_{n+1}^2 as a simple function of S_n^2 and Δ_n^2.

Chapter 5

Numerical Variables: Density Estimation and Model Diagnostics

In Chapter 4 we considered various characteristics of the distribution P. Now we return to visualizing the empirical distribution $\widehat{P}$ and estimation of the entire distribution P, this time under the stronger assumption that P is given by a density function f. In addition we describe methods wo check graphically or test whether a given model for P is plausible.

5.1 Histograms and Density Functions

The empirical distribution of a numerical variable is often visualised by means of *histograms*. In the sequel we shall explain this method, discuss its advantages and disadvantages and finally analyze it as an estimator of an underlying density function. Then we'll describe and analyze an alternative method, kernel density estimators.

Histograms

From the graph of the empirical distribution function $\widehat{F}$ one can, in principle, recover all order statistics $X_{(i)}$. Thus apart from the order of observations one loses no information. This is definitely an advantage over the much more popular histograms. The latter are closely related to bar plots and have been introduced by K. Pearson at the end of the 19th century.

One chooses pairwise disjoint, bounded and nondegenerate intervals $B_1, B_2, \ldots, B_K$ covering all sample values X_i; for instance,

$$(a_0, a_1],\ (a_1, a_2],\ (a_2, a_3],\ \ldots, (a_{K-1}, a_K]$$

with $a_0 < a_1 < a_2 < \cdots < a_K$ and $X_{(1)}, X_{(n)} \in (a_0, a_K]$. Then one determines for $k = 1, 2, \ldots, K$ the absolute frequencies $H(B_k) := \#\{i : X_i \in B_k\}$ and the relative frequencies $\widehat{P}_n(B_k) = H(B_k)/n$.

Now one draws for each interval B_k a rectangle with horizontal baseline B_k, and vertically it goes from zero to a certain height. For the latter there are two conventions:

Convention 1: The height equals $H(B_k)$.

Convention 2: The height equals $\widehat{P}_n(B_k)/\lambda(B_k)$.

Here $\lambda(B_k)$ denotes the length of the interval B_k. Convention 2 yields a rectangle whose area is identical with the relative frequency $\widehat{P}_n(B_k)$.

If all intervals B_k have the same length, both conventions yield the same picture up to a scaling factor in vertical direction. Otherwise one should use convention 2. The latter prevents distortions, because people tend to perceive the areas rather than the heights of the rectangles. Moreover, with convention 2 it is easier to compare histograms of different data sets, even if the sample sizes differ substantially.

Example 5.1. Suppose the sample contains $n = 20$ X-values lying in one of the following five intervals: $(150, 160]$, $(160, 170]$, $(170, 175]$, $(175, 180]$, $(180, 190]$. Let the corresponding absolute frequencies be 2, 5, 3, 6 and 4. Figure 5.1 shows the resulting histograms with both conventions.

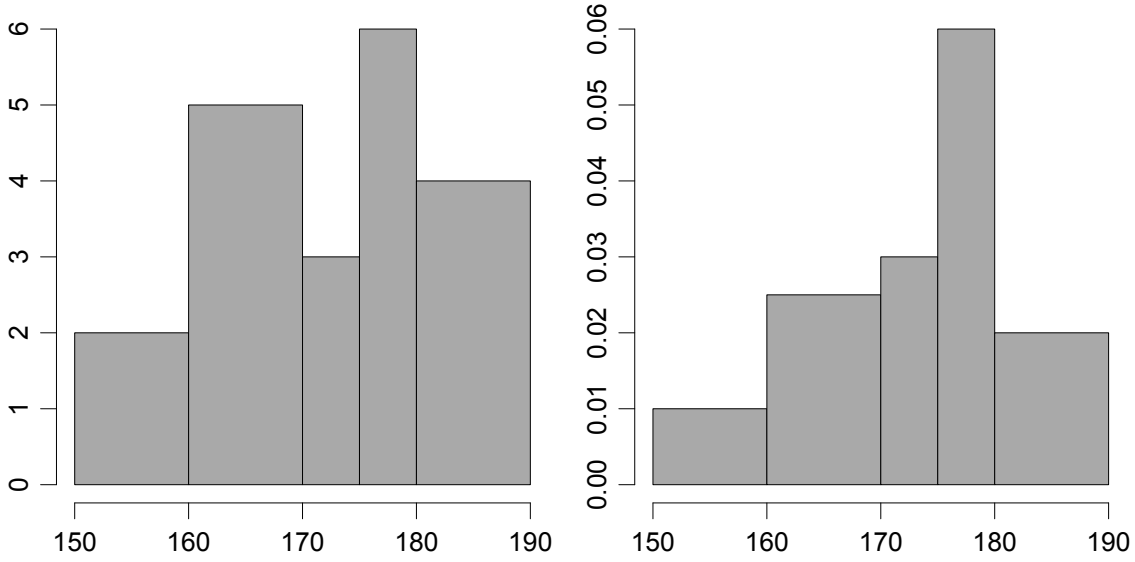

Figure 5.1: A histogram with convention 1 (left) and 2 (right).

Histograms provide an impression how many sample values are in which region. However the picture depends strongly on the choice of intervals. Even if we restrict ourselves to intervals of the same length, varying the boundary values may alter the histogram drastically. An additional problem are the boundary points of the intervals. There is no way of seeing whether a boundary point belongs to the left or right adjacent interval.

Example 5.2. For the data in Example 1.18, Figure 5.2 shows four histograms of the $n = 113$ body heights (in cm) of female students. In the upper row we used intervals of length 3, in the lower row intervals of length 5.

Density Functions

Recall that we consider $X_1, X_2, \ldots, X_n$ as independent random variables with unknown distribution P, distribution function F and *density function* f. That means $f : \mathbb{R} \to [0, \infty)$ is an integrable function with $\int_{-\infty}^{\infty} f(x)\, dx = 1$, and for arbitrary intervals $B \subset \mathbb{R}$,

$$P(B) \;=\; \int_B f(x)\, dx = \int_{\inf(B)}^{\sup(B)} f(x)\, dx.$$

Equivalently,

$$F(r) \;=\; \int_{-\infty}^{r} f(x)\, dx$$

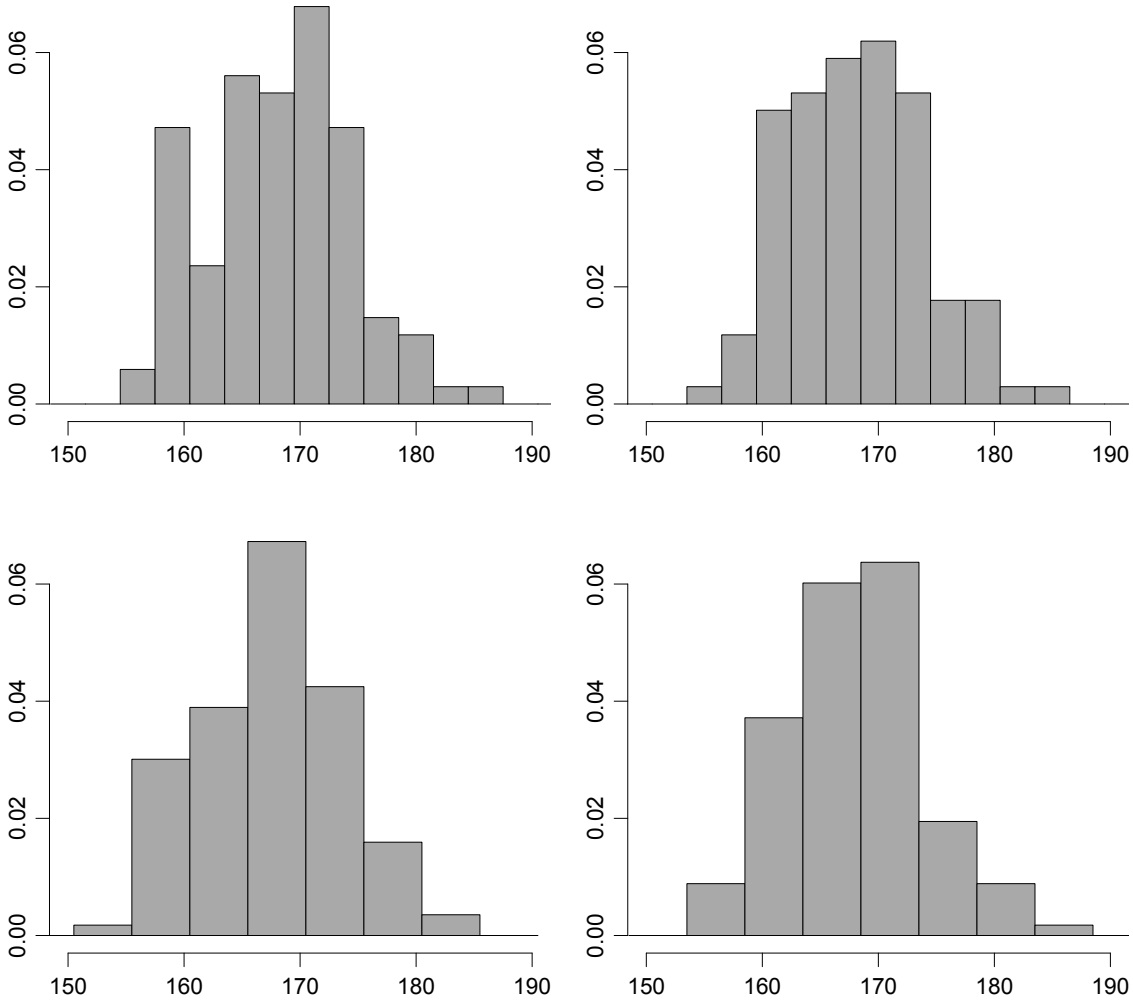

Figure 5.2: Four histograms of one data vector.

for arbitrary $r \in \mathbb{R}$.

Distributions with densities are idealised models for real distributions. For a distribution P with density f,

$$P(\{x\}) \;=\; 0 \quad \text{for arbitrary } x \in \mathbb{R}.$$

For any continuity point x of f and nondegenerate intervals $B \subset \mathbb{R}$,

$$\frac{P(B)}{\lambda(B)} \;\to\; f(x) \quad \text{as } \inf(B), \sup(B) \to x,$$

and

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

Thus a possible estimator for $f(x)$ would be

$$\widehat{f}(x) \;:=\; \frac{\widehat{P}_n(B_n(x))}{\lambda(B_n(x))}$$

with a certain interval $B_n(x)$ containing x.

As we'll see later, estimating the density function f is substantially more difficult than estimating the distribution function F. The quality of an arbitrary density estimator $\widehat{f} = \widehat{f}(\cdot \mid \text{data})$ at a point x will be quantified by the root mean squared error,

$$\mathrm{RMSE}(x) \;:=\; \sqrt{\mathbb{E}\big(\big(\widehat{f}(x) - f(x)\big)^2\big)};$$

see also Chapter 1. As usual, this quantity is split into bias and standard deviation. Precisely,

$$\mathrm{RMSE}(x) \;=\; \sqrt{\mathrm{Bias}(x)^2 + \mathrm{SD}(x)^2}$$

with

$$\begin{aligned} \mathrm{Bias}(x) \;&:=\; \mathbb{E}\big(\widehat{f}(x)\big) - f(x) \qquad (\text{Bias of } \widehat{f}(x)), \\ \mathrm{SD}(x) \;&:=\; \sqrt{\mathrm{Var}\big(\widehat{f}(x)\big)} \qquad (\text{standard deviation of } \widehat{f}(x)). \end{aligned}$$

The empirical distribution $\widehat{P}$ is an unbiased estimator of P in the sense that $\mathbb{E}(\widehat{P}(B)) = P(B)$ for any Borel set $B \subset \mathbb{R}$. For the density function f there is no unbiased estimator, and one should try to balance both error sourced Bias^2 and SD^2. Typically, a decrease of $\mathrm{Bias}(x)^2$ leads to an increase of $\mathrm{SD}(x)^2$ and vice versa.

5.2 Histograms as Density Estimators

One may interpret histograms following convention 2 as estimators for the density function f. Precisely, the histogram produced with intervals B_1, B_2, …, B_K corresponds to the histogram function $\widehat{f}$ with

$$\widehat{f}(x) \;=\; \frac{\widehat{P}(B_k)}{\lambda(B_k)} \quad \text{for } x \in B_k, 1 \le k \le K$$

and $\widehat{f}(x) = 0$ for $x \notin \bigcup_{k=1}^K B_k$.
In the special case that $B_k = (a_{k-1}, a_k]$ with real numbers $a_0 < a_1 < \cdots < a_K$ one may write

$$\widehat{f}(x) \;=\; \frac{\widehat{F}(a_k) - \widehat{F}(a_{k-1})}{a_k - a_{k-1}} \quad \text{for } x \in (a_{k-1}, a_k], 1 \le k \le K.$$

Thus one approximates the non-differentiable distribution function $\widehat{F}$ by a continuous, piecewise linear function whose left-sided derivative is equal to the histogram function $\widehat{f}$.

Example 5.3. In Figure 5.2 we saw already four different histograms of $n = 113$ observations (body heights in cm). Figure 5.3 depicts the underlying empirical distribution function and the corresponding four approximations by a continuous, piecewise linear function.

From now on we restrict our attention to histograms with intervals of equal length. For an offset $a \in \mathbb{R}$ and an interval length $h > 0$ we consider the intervals

$$B_{a,h,z} \;:=\; (a + zh, a + zh + h] \quad (z \in \mathbb{Z})$$

and define

$$\widehat{f}(x) = \widehat{f}_{a,h}(x) \;:=\; \frac{\widehat{P}(B_{a,h,z})}{h} \quad \text{for } x \in B_{a,h,z}, z \in \mathbb{Z}.$$

Here $\mathrm{Bias}(x)^2$ tends to be larger and $\mathrm{SD}(x)^2$ tends to be smaller if h is increased. The following theorem provides explicit inequalities and approximations for $\mathrm{Bias}(x)$, $\mathrm{SD}(x)$ and $\mathrm{RMSE}(x)$ under certain regularity assumptions on f.

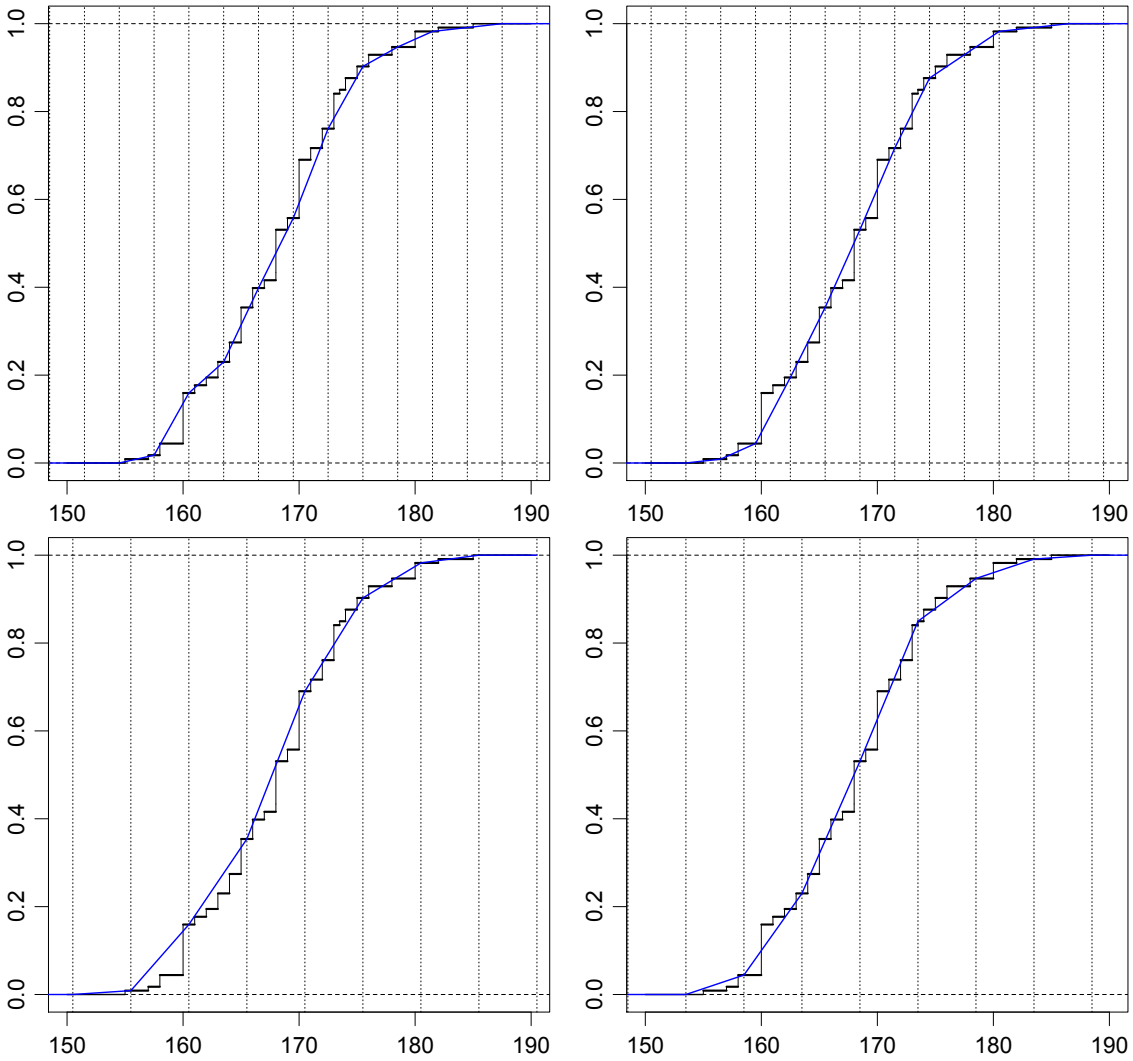

Figure 5.3: Empirical distribution function with four approximations.

Theorem 5.4 (Precision of histograms). *Let $\widehat{f}$ be the histogram function $\widehat{f}_{a,h}$. Suppose that f is differentiable with $f \le M_0$ and $|f'| \le M_1$. Then*

$$\begin{aligned} \mathrm{Bias}(x) &= \left(f'(x) S\Big(\frac{x-a}{h}\Big) + r_1(x,a,h)\right) h, \\ |\mathrm{Bias}(x)| &\le \frac{M_1 h}{2}, \\ \mathrm{SD}(x)^2 &= \frac{f(x) + r_2(x,a,h)}{nh} \le \frac{M_0}{nh}, \end{aligned}$$

where $r_1(x,a,h), r_2(x,a,h) \to 0$ as $h \downarrow 0$, uniformly in $a \in \mathbb{R}$, and

$$S(y) := \lceil y \rceil - y - 0.5.$$

In case of $h = Cn^{-1/3}$ for some constant $C > 0$,

$$\mathrm{RMSE}(x) \le \tilde{C} n^{-1/3}$$

with $\tilde{C} := \sqrt{M_1^2 C^2/4 + M_0/C}$.

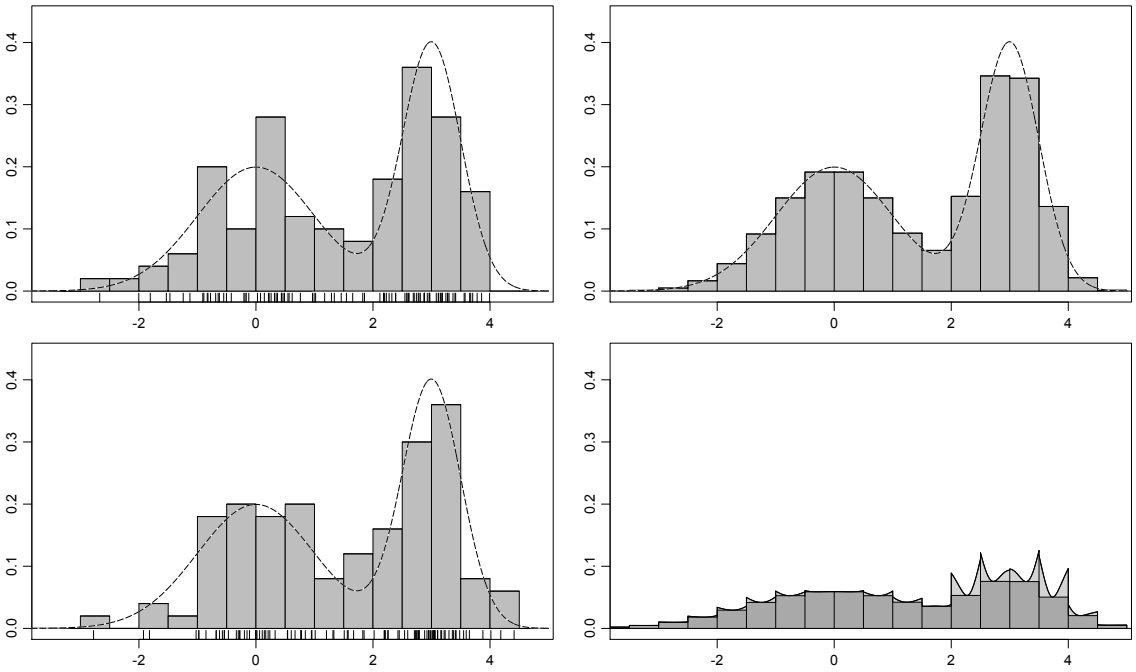

Figure 5.4: Two histograms $\widehat{f}$ together with $\mathbb{E}(\widehat{f})$, SD and RMSE for $h = 0.5$.

This theorem shows that for suitable interval length $h = h(n)$ and well-behaved densities f the estimation error $\widehat{f}(x) - f(x)$ is of order $O_p(n^{-1/3})$. Roughly saying this means: To decrease the estimation error by a factor of 2 one has to increase the sample size by a factor of 8. For a decrease by a factor of 10 one needs even $1000\,n$ instead of n observations.

The 'sawtooth function' S in Theorem 5.4 is periodic, namely, $S(z + u) = 0.5 - u$ for arbitrary $z \in \mathbb{Z}$ and $u \in (0, 1]$.

Figures 5.4 and 5.5 illustrate the previous considerations. In both figures we consider two simulated data sets of size $n = 100$. On the left hand side one sees for offset $a = 0$ and a certain interval length $h > 0$ the corresponding histograms of both samples. The samples themselves are depicted by line plots just below the horizontal axis. The underlying true density function is depicted as a dashed line. On the right hand side one sees the corresponding expected value, $x \mapsto \mathbb{E}\big(\widehat{f}(x)\big)$. Below that one sees $x \mapsto \mathrm{SD}(x)$ (darker area, step function) as well as $x \mapsto \mathrm{RMSE}(x)$ (total gray area). One sees clearly that for larger h the main contribution to $\mathrm{RMSE}(x)$ is the systematic error $\mathrm{Bias}(x)$. For smaller h the standard deviation $\mathrm{SD}(x)$ plays a major role.

Proof of Theorem 5.4. Let $x \in B_{a,h,z}$ for some $z \in \mathbb{Z}$. Then $\widehat{f}(x) = \widehat{P}(B_{a,h,z})/h$, and

$$\mathrm{Bias}(x) \;=\; \frac{P(B_{a,h,z})}{h} - f(x) \;=\; \frac{1}{h}\int_{a+hz}^{a+hz+h} (f(y) - f(x))\,dy.$$

Now we write $x = a+hz+hu$ and $y = a+hz+hv$ for certain $u, v \in (0, 1]$. Then $y = x+h(v-u)$, and

$$\mathrm{Bias}(x) \;=\; \int_0^1 \big(f(x + h(v - u)) - f(x)\big)\,dv.$$

The mean value theorem for differentiable functions and the definition of $f'(x)$ imply that

$$|f(x + t) - f(x)| \;\le\; M_1|t| \quad \text{and} \quad f(x + t) - f(x) = \big(f'(x) + \rho(x, t)\big)\,t$$

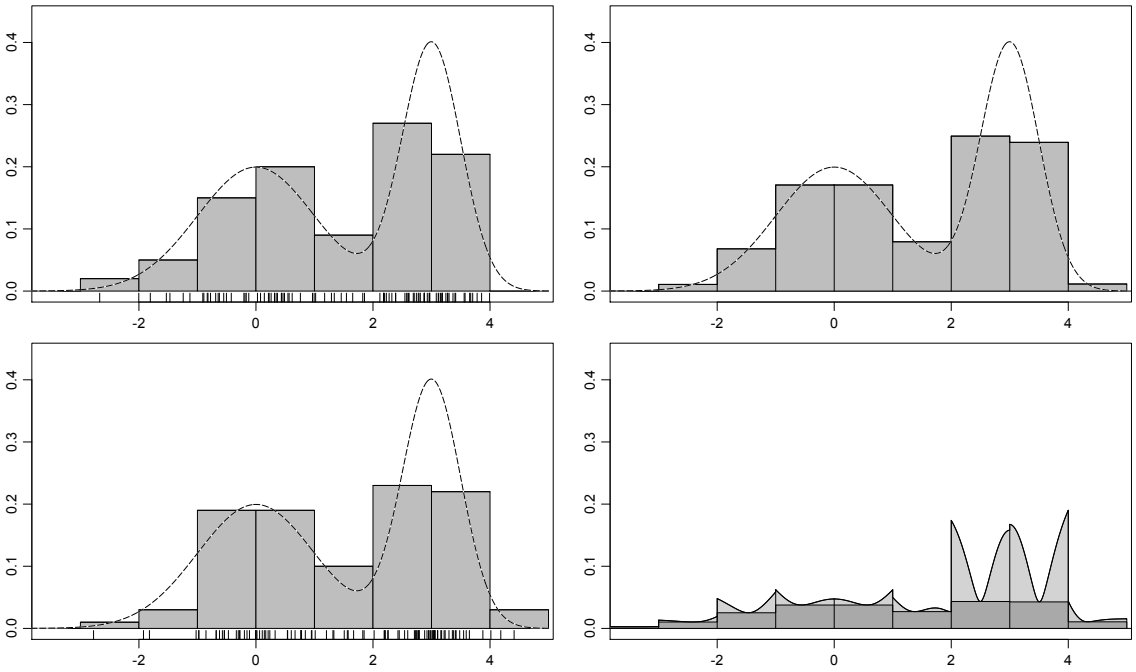

Figure 5.5: Two histograms $\widehat{f}$ together with $\mathbb{E}(\widehat{f})$, SD and RMSE for $h = 1.0$.

for arbitrary $t \in \mathbb{R}$, where $\lim_{t\to 0} \rho(x,t) = 0$. Consequently,

$$|\mathrm{Bias}(x)| \;\le\; \int_0^1 M_1 h|v-u|\,dv \;\le\; M_1 h/2.$$

On the other hand,

$$\begin{aligned}\mathrm{Bias}(x) \;&=\; \int_0^1 \big(f'(x) + \rho(x, h(v-u))\big)(v-u)\,dv\; h \\ &=\; \big(f'(x)(0.5-u) + r_1(x,a,h)\big)\, h,\end{aligned}$$

where

$$|r_1(x,a,h)| \;\le\; \int_0^1 \big|\rho(x,h(v-u))\big||v-u|\,dv \;\le\; \sup_{t\in[-h,h]} |\rho(x,t)|/2.$$

Moreover,

$$0.5-u \;=\; 0.5 - \frac{x-a}{h} + z \;=\; 0.5 - \frac{x-a}{h} + \Big\lceil \frac{x-a}{h} \Big\rceil - 1 \;= S\Big(\frac{x-a}{h}\Big).$$

Concerning the standard deviation $\mathrm{SD}(x)$, the fact that $n\widehat{P}(B_{a,h,z})$ follows $\mathrm{Bin}(n, P(B_{a,h,z}))$ implies that

$$\begin{aligned}\mathrm{SD}(x)^2 \;=\; \frac{\mathrm{Var}(\widehat{P}(B_{a,h,z}))}{h^2} \;&=\; \frac{P(B_{a,h,z})(1-P(B_{a,h,z}))}{nh^2} \\ &=\; \frac{\mathbb{E}\big(\widehat{f}(x)\big)\big(1 - h\,\mathbb{E}\big(\widehat{f}(x)\big)\big)}{nh}.\end{aligned}$$

On the one hand, $\mathbb{E}\big(\widehat{f}(x)\big) = h^{-1}\int_{a+hz}^{a+hz+h} f(y)\,dy \le M_0$, so

$$\mathrm{SD}(x)^2 \;\le\; \frac{\mathbb{E}\big(\widehat{f}(x)\big)}{nh} \;\le\; \frac{M_0}{nh}.$$

On the other hand,

$$\mathbb{E}\big(\widehat{f}(x)\big)\big(1 - h\,\mathbb{E}\,\big(\widehat{f}(x)\big)\big) \;=\; f(x) + r_2(x,a,h)$$

with $|r_2(x,a,h)| \le |\mathrm{Bias}(x)| + \mathbb{E}\big(\widehat{f}(x)\big)^2 h \le (M_1/2 + M_0^2)h$.
The inequality for $\mathrm{RMSE}(x)$ in case of $h = Cn^{-1/3}$ follows from plugging in the upper bounds for $\mathrm{Bias}(x)^2$ and $\mathrm{SD}(x)^2$. □

5.3 Kernel Density Estimation

Starting from histograms we derive now another class of density estimators.

Consideration 1. Theorem 5.4 shows that for the histogram estimator $\widehat{f}_{a,h}$, the function $x \mapsto \mathrm{RMSE}(x)$ tends to be larger at the boundaries and smaller in the middle of the intervals $B_{a,h,z}$. The reason for that phenomenon is the special sawtooth-shape of the bias. Thus to estimate $f(x)$ at a particular point x by means of a histogram, one should choose the intervals such that x is a midpoint of one of them. This consideration leads to the estimator

$$\widehat{f}_h(x) \;:=\; \widehat{f}_{x-h/2,h}(x) \;=\; \frac{\widehat{F}(x+h/2) - \widehat{F}(x-h/2)}{h}.$$

Alternatively one may write

$$\widehat{f}_h(x) \;=\; \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h}\, 1_{[x-h/2<X_i\le x+h/2]} \;=\; \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h}\, R\Big(\frac{x-X_i}{h}\Big)$$

with

$$R(y) \;:=\; 1_{[-0.5\le y<0.5]}.$$

Consideration 2. When computing a histogram function $\widehat{f}_{a,h}$, one has to choose suitable parameters $a \in \mathbb{R}$ and $h > 0$. How should one choose the offset a for a given interval length $h > 0$? Indeed, different choices of a may lead to rather different functions $\widehat{f}_{a,h}$; see Figure 5.2. A possible way out is to average the histogram over various choices for a. Precisely we consider

$$\widehat{f}_h(x) \;:=\; \frac{1}{h}\int_b^{b+h} \widehat{f}_{a,h}(x)\,da \tag{5.1}$$

for an arbitrary real number b. Since $\widehat{f}_{a+z_o h,h} = \widehat{f}_{a,h}$ for arbitrary integers z_o, the choice of b has no effect. But one can show that

$$\widehat{f}_h(x) \;=\; \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h}\Delta\Big(\frac{x-X_i}{h}\Big) \tag{5.2}$$

with

$$\Delta(y) \;:=\; \max\big(1-|y|,0\big),$$

see Exercise 5.3.
Both considerations lead to kernel density estimators which are defined as follows:

Definition 5.5 (Kernel density estimator). Let $K : \mathbb{R} \to \mathbb{R}$ be an integrable function such that $\int_{-\infty}^{\infty} K(y)\,dy = 1$. The *kernel density estimator* with kernel function K and bandwidth $h > 0$ is defined as the data-dependent function $\widehat{f}_h$ with

$$\widehat{f}_h(x) = \widehat{f}_h(x, \text{data}) \;:=\; \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i).$$

Here K_h is a rescaled version of K,

$$K_h(y) \;:=\; \frac{1}{h}\,K\Big(\frac{y}{h}\Big).$$

It follows from K integrating to one that $\int_{-\infty}^{\infty} K_h(x)\,dx \;=\; \int_{-\infty}^{\infty} \widehat{f}_h(x)\,dx \;=\; 1$ for arbitrary bandwidths $h > 0$. In case of a continuous kernel function K, the estimator $\widehat{f}_h$ is continuous, too. In case of $K \geq 0$, the estimator $\widehat{f}_h \geq 0$ is a probability density. At this point it may seem a strange idea to use kernel functions with negative values, but later we shall see the potential benefits of such a choice.

Examples. Consideration 1 yielded the rectangular kernel R with

$$R(y) \;:=\; 1_{[-0.5 \leq y < 0.5]}.$$

Consideration 2 led to the triangular kernel Δ with

$$\Delta(y) \;:=\; \max(1 - |y|, 0).$$

Other examples are the Gaussian kernel $\phi = \Phi'$, i.e.

$$\phi(y) \;=\; (2\pi)^{-1/2} \exp(-y^2/2),$$

and the Epanechnikov kernel K_o with

$$K_o(y) \;:=\; 0.75 \cdot \max(1 - y^2, 0).$$

Connection between $\widehat{P}$ and $\widehat{f}_h$. The empirical distribution $\widehat{P}$ is the arithmetic mean of the n probability distributions $\delta_{X_1}, \delta_{X_2}, \ldots, \delta_{X_n}$, where $\delta_{X_i}(B) := 1_{[X_i \in B]}$. For the computation of $\widehat{f}_h$ with non-negative kernel K each point mass δ_{X_i} is replaced with a probability distribution with density $K_h(\cdot - X_i)$. Figure 5.6 shows the kernel density estimator $\widehat{f}_h$ resulting from $n = 8$ observations, where $h = 0.6$ and $K = \phi$. One sees the functions $n^{-1}K_h(\cdot - X_i)$ as well as their sum $\widehat{f}_h$.

A physical interpretation. Kernel density estimators $\widehat{f}_h$ with Gaussian kernel ϕ have an explicit interpretation in Physics: Imagine the real axis as an arbitrarily long thin wire. At time zero, each point[1] X_i is heated up to a certain temperature while the rest of the wire has a constant lower temperature. Now one lets the heat diffuse along the wire. Measuring temperature and time in appropriate units, $\widehat{f}_h(x)$ is the difference between actual and initial temperature at point $x \notin \{X_1, X_2, \ldots, X_n\}$ at time $h^2 > 0$. Behind this is the fact that $\widehat{f}_{\sqrt{t}}(x)$, viewed as a function of $(t, x) \in (0, \infty) \times \mathbb{R}$ solves the heat equation.

As in case of histogram estimators, $\text{Bias}(x)^2$ tends to be increasing and $\text{SD}(x)^2$ tends to be decreasing in $h > 0$. Again one may derive explicit bounds and approximations for bias and standard deviation of $\widehat{f}_h$ under certain regularity assumptions on f.

[1] an infinitesimal neighborhood thereof

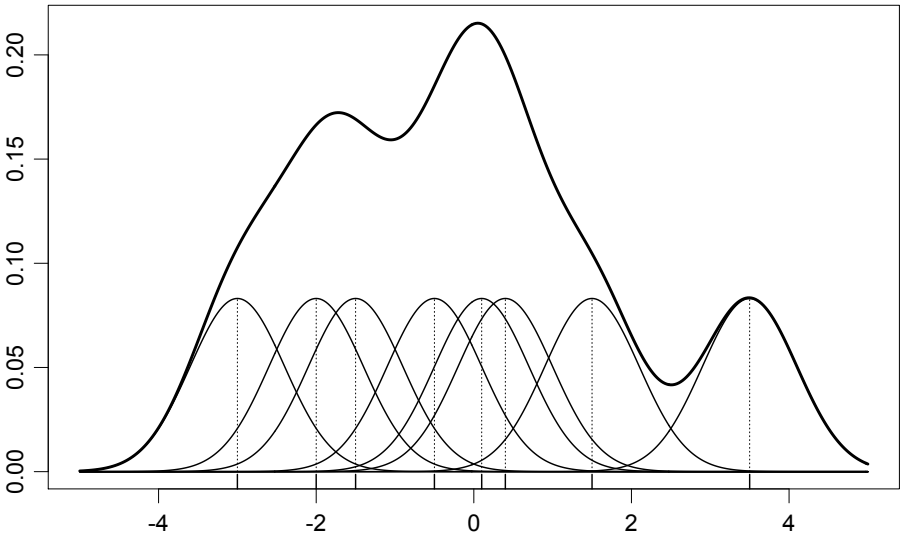

Figure 5.6: From $\widehat{P}$ to $\widehat{f}$.

Theorem 5.6 (Precision of kernel density estimators). *Let $\widehat{f}$ be the kernel density estimator $\widehat{f}_h$ with kernel function $K \geq 0$ and bandwidth $h > 0$. Suppose that f is twice differentiable with $f \leq M_0$ and $|f''| \leq M_2$. Further let $\int_{-\infty}^{\infty} yK(y)\,dy = 0$, and suppose that both $C_{\mathrm{B}} := 2^{-1}\int_{-\infty}^{\infty} y^2 K(y)\,dy$ and $C_{\mathrm{SD}} := \int_{-\infty}^{\infty} K(y)^2\,dy$ are finite. Then*

$$\begin{aligned} \mathrm{Bias}(x) &= \big(C_{\mathrm{B}} f''(x) + r_1(x,h)\big)\, h^2, \\ |\mathrm{Bias}(x)| &\leq C_{\mathrm{B}} M_2\, h^2, \\ \mathrm{SD}(x)^2 &= \frac{C_{\mathrm{SD}} f(x) + r_2(x,h)}{nh} \leq \frac{C_{\mathrm{SD}} M_0}{nh}, \end{aligned}$$

where $\lim_{h\downarrow 0} r_j(x,h) = 0$ for $j = 1, 2$.
In case of $h = Cn^{-1/5}$ for some constant $C > 0$,

$$\mathrm{RMSE}(x) \leq \tilde{C} n^{-2/5}$$

with $\tilde{C} := \sqrt{C_{\mathrm{B}}^2 M_2^2 C^4 + C_{\mathrm{SD}} M_0 / C}$.

In case of a sufficiently smooth function f we obtain a density estimator with $\mathrm{RMSE}(x) = O(n^{-2/5})$ which is substantially better than the rate $O(n^{-1/3})$ for histogram estimators. The aforementioned examples of kernel functions K fulfil the requirements in Theorem 5.6.

To illustrate the preceding considerations and to compare kernel density estimators with histograms, we consider the same two data sets as in Figures 5.4 and 5.5. Figures 5.7 and 5.8 show for a fixed bandwidth $h > 0$ the following functions: On the left hand side one sees the kernel density estimators $\widehat{f}_h$ with triangular kernel Δ. On the upper right hand side one sees the corresponding expected value, $x \mapsto \mathbb{E}(\widehat{f}(x))$. On the lower right side one sees $x \mapsto \mathrm{SD}(x)$ (darker part) and $x \mapsto \mathrm{RMSE}(x)$ (total gray area).

Again one sees clearly that for smaller bandwidths h, $\mathrm{RMSE}(x)$ is caused mainly by the standard deviation $\mathrm{SD}(x)$, while for larger bandwidths h the systematic error $|\mathrm{Bias}(x)|$ starts dominating. Particularly interesting is a comparison of Figures 5.5 and 5.7. This shows clearly the improvement by averaging the histogram estimator $\widehat{f}_{a,h}$ with respect to the offset parameter a; see Consideration 2. For bandwidth $h = 2$ (Figure 5.8) the systematic error is rather large. Nevertheless the two local maxima and the local minimum in between are still detected with reasonably high probability.

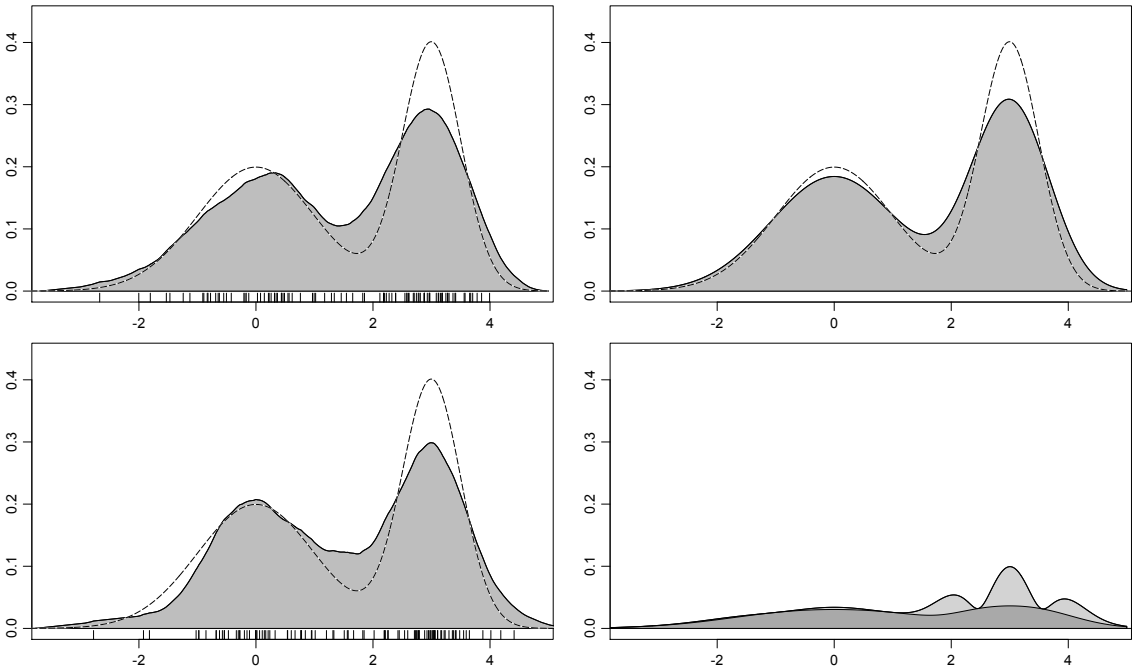

Figure 5.7: Two kernel density estimators $\widehat{f}$ as well as $\mathbb{E}(\widehat{f})$, SD and RMSE for $h = 1.0$.

Proof of Theorem 5.6. Since $X_1, X_2, \ldots, X_n$ are independent and identically distributed, the same is true for fixed $x \in \mathbb{R}$ and the random variables $K_h(x - X_i)$, $1 \le i \le n$. For the arithmetic mean $\widehat{f}_h(x) = n^{-1} \sum_{i=1}^n K_h(x - X_i)$ this implies that

$$\begin{aligned} \mathbb{E}\big(\widehat{f}_h(x)\big) &= \mathbb{E}(K_h(x - X_1)), \\ \mathrm{Var}\big(\widehat{f}_h(x)\big) &= \frac{1}{n} \mathrm{Var}\big(K_h(x - X_1)\big) \\ &= \frac{1}{n}\Big(\mathbb{E}(K_h(x - X_1)^2) - \mathbb{E}\big(\widehat{f}_h(x)\big)^2\Big). \end{aligned}$$

Moreover,

$$\begin{aligned} \mathbb{E}\big(K_h(x - X_1)^j\big) &= \int_{-\infty}^{\infty} \frac{1}{h^j} K\Big(\frac{x - z}{h}\Big)^j f(z)\, dz \\ &= h^{1-j} \int_{-\infty}^{\infty} K(y)^j f(x - hy)\, dy \end{aligned}$$

for $j \in \mathbb{N}$. Here we used the transformation $y = (x - z)/h$, so $z = x - hy$ and $dz = -h\, dy$. For the bias of $\widehat{f} = \widehat{f}_h$ this implies the expression

$$\mathrm{Bias}(x) = \int_{-\infty}^{\infty} K(y)\big(f(x - hy) - f(x)\big)\, dy.$$

It follows from Taylor's formula that

$$f(x + t) - f(x) = f'(x)t + 2^{-1} f''(\xi(x, t))\, t^2$$

for a suitable point $\xi(x, t)$ in the interval $\big[x \pm |t|\big]$, and $f''(\xi(x, t)) \to f''(x)$ as $t \to 0$. (The latter

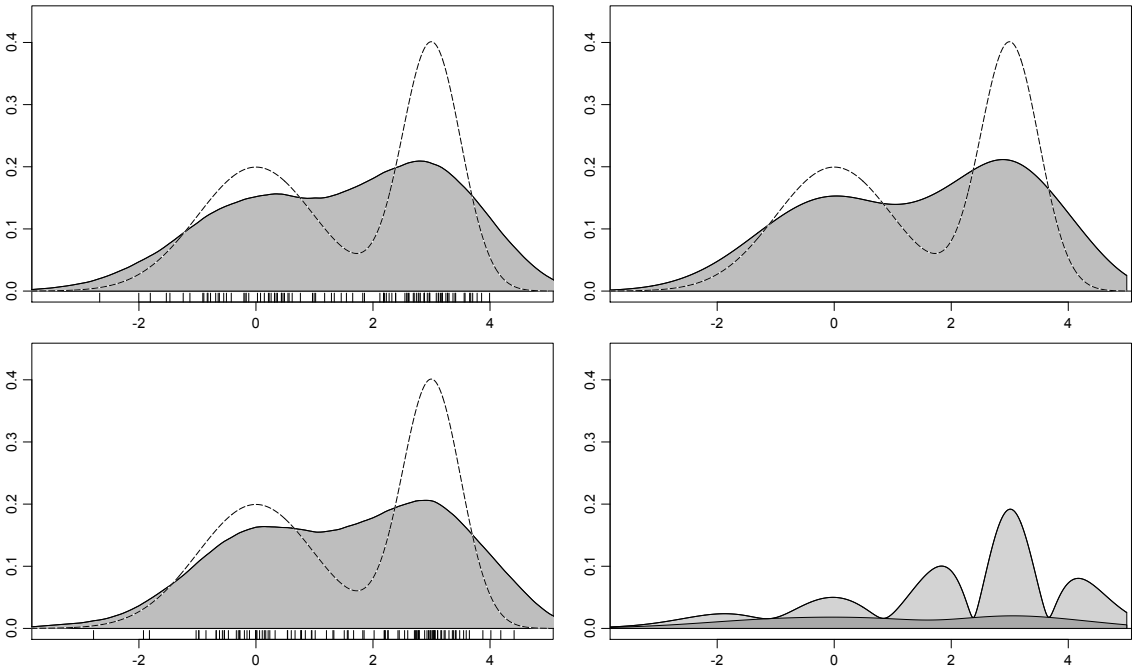

Figure 5.8: Two kernel density estimators $\widehat{f}$ as well as $\mathbb{E}(\widehat{f})$, SD and RMSE for $h = 2.0$.

claim is true even if f'' is not continuous.) Consequently,

$$\begin{aligned}\mathrm{Bias}(x) &= -f'(x)h\int_{-\infty}^{\infty} yK(y)\,dy + \frac{h^2}{2}\int_{-\infty}^{\infty} y^2K(y)f''(\xi(x,-hy))\,dy \\ &= \frac{h^2}{2}\int_{-\infty}^{\infty} y^2K(y)f''(\xi(x,-hy))\,dy,\end{aligned}$$

because $\int_{\infty}^{\infty} yK(y)\,dy = 0$. In particular, $|f''| \le M_2$ implies that $|\mathrm{Bias}(x)| \le C_{\mathrm{B}}M_2h^2$, and by dominated convergence, $r_1(x,h) := h^{-2}\,\mathrm{Bias}(x) - C_{\mathrm{B}}f''(x)$ converges to zero as $h \downarrow 0$.
For the standard deviation $\mathrm{SD}(x)$ we obtain the expression

$$\mathrm{SD}(x)^2 = \frac{1}{nh}\left(\int_{-\infty}^{\infty} K(y)^2 f(x-hy)\,dy - h\,\mathbb{E}\big(\widehat{f}_h(x)\big)^2\right).$$

Obviously the right hand side is not greater than

$$\frac{1}{nh}\int_{-\infty}^{\infty} K(y)^2 f(x-hy)\,dy \le \frac{C_{\mathrm{SD}}M_0}{nh}.$$

On the other hand, $0 \le \mathbb{E}\big(\widehat{f}_h(x)\big) \le M_0$, and it follows from dominated convergence that

$$\lim_{h\downarrow 0}\int_{-\infty}^{\infty} K(y)^2 f(x-hy)\,dy = C_{\mathrm{SD}}f(x).$$

Thus $r_2(x,h) := nh\,\mathrm{SD}(x)^2 - C_{\mathrm{SD}}f(x)$ converges to zero as $h \downarrow 0$. □

Choosing the bandwidth h. In explicit applications it is not clear how to choose $h > 0$. If the kernel density estimators are used as an exploratory tool to visualise the empirical distribution of the data, one should plot $\widehat{f}_h$ for different values of $h > 0$ to get a feeling for $\widehat{P}$.

There is a huge literature on data-driven choice of bandwidths, i.e. $h = h(\text{data}) > 0$. One could even try to choose different bandwidths at different positions, leading to estimators $\widehat{f}(x) = \widehat{f}_{h(x,\text{data})}(x, \text{data})$. Subsequently we describe three proposals for the choice of a (global) bandwidth.

Normal distributions as gold standard. Under the implicit assumption that P is similar to a Gaussian distribution, we choose

$$h \;=\; \frac{\text{IQR(data)}}{2\Phi^{-1}(3/4)}\, h(n).$$

The factor $\text{IQR(data)}/(2\Phi^{-1}(3/4))$ is a robust estimator of the Here $h(n) > 0$ is chosen such that the kernel density estimator $\widehat{f}_{h(n)}$ would be 'optimal' in case of P being the standard normal distribution. Possible notions of optimality would be to minimze $\sup_{x\in\mathbb{R}} \text{RMSE}(x)^2$ or $\int_{\mathbb{R}} \text{RMSE}(x)^2\, dx$ in case of f being the standard Gaussian density ϕ. Since we do not have explicit formulae for $\text{RMSE}(x)^2$, we could replace it with its approximation provided by Theorem 5.6. Indeed, if $h(n) = Cn^{-1/5}$ with $C > 0$, then

$$\lim_{n\to\infty}\; n^{4/5}\text{RMSE}(x)^2 \;=\; C_{\text{B}}^2\phi''(x)^2C^4 + C_{\text{SD}}\phi(x)C^{-1}.$$

Thus we could choose C such that the supremum or the integral of this limit over the real line becomes minimal.

Kolmogorov-Smirnov criterion. In addition to the true distribution function F and the empirical distribution function $\widehat{F}$ we consider the distribution function $\widehat{F}_h$ of the kernel density estimator $\widehat{f}_h$, i.e.

$$\widehat{F}_h(r) \;:=\; \int_{-\infty}^{r} \widehat{f}_h(x)\, dx.$$

Since $\mathbb{E}\,\|\widehat{F} - F\|_\infty = O(n^{-1/2})$, we choose for a constant $c > 0$ (for instance, $c = 0.5$) the bandwidth $h = h(\text{data})$ as large as possible such that

$$\big\|\widehat{F} - \widehat{F}_h\big\|_\infty \;\le\; \frac{c}{\sqrt{n}}.$$

Weak smoothing of $\widehat{F}$. If our main goal is to visualise the empirical distribution $\widehat{P}$, that means, to get an impression in which regions there are few or many observations, we may use rather small bandwidths. Suppose we are working with the Gaussian kernel $K = \phi$. If we view the data as fixed, $\widehat{f}_h$ is the density function of $\widehat{X} + hZ$ with stochastically independent random variables $\widehat{X} \sim \widehat{P}$ and $Z \sim \mathcal{N}(0,1)$. In particular, $\mathbb{E}(\widehat{X} + hZ) = \overline{X}$ and

$$\text{Var}(\widehat{X} + hZ) \;=\; \sigma(\widehat{P})^2 + h^2 \;=\; (1 - n^{-1})S^2 + h^2.$$

Hence choosing $h = n^{-1/2}S$ leads to a distribution with mean $\overline{X}$ and variance S^2. The resulting density estimators tend to have too many local minima and maxima. Nevertheless one gets a good impression about the empirical distribution of the data.

Optimal (nonnegative) kernels. Theorem 5.6 suggests to minimise both quantities $C_{\text{B}}(K) = 2^{-1}\int_{-\infty}^{\infty} y^2K(y)\, dy$ and $C_{\text{SD}}(K) = \int_{-\infty}^{\infty} K(y)^2\, dy$. If one would replace $K(y)$ with $K_{\text{new}}(y) = \tau^{-1}K(\tau^{-1}y)$ for some $\tau > 0$, we would obtain the characteristics $C_{\text{B}}(K_{\text{new}}) = \tau^2 C_{\text{B}}(K)$ and $C_{\text{SD}}(K_{\text{new}}) = \tau^{-1}C_{\text{SD}}(K)$, and $\widehat{f}_h$ with kernel K_{new} would be equal to $\widehat{f}_{\tau h}$ with the original kernel K. Thus one may fix an arbitrary value for $C_{\text{B}}(K)$ and try to minimise $C_{\text{SD}}(K)$ under that constraint. This problem was already solved in the proof of (4.6), and it turned out that the Epanechnikov kernel K_o or any rescaled version of it is optimal.

Kernels of higher order. If we consider kernels K with possibly negative values, the preceding results may be generalised as follows:

Theorem 5.7. *Let $\widehat{f}$ be the kernel density estimator $\widehat{f}_h$ with bandwidth $h > 0$ and kernel function K. Suppose that for an even integer $J \geq 2$ the density f is J times differentiable with $f \leq M_0$ and $|f^{(J)}| \leq M_J$. Further suppose that*

$$\int_{-\infty}^{\infty} y^j K(y)\, dy \;=\; 0 \quad \text{for } j = 1, \ldots, J-1,$$

and that both $\bar{C}_{\mathrm{B}} := (J!)^{-1} \int_{-\infty}^{\infty} y^J |K(y)|\, dy$ and $C_{\mathrm{SD}} := \int_{-\infty}^{\infty} K(y)^2\, dy$ are finite. With $C_{\mathrm{B}} := (J!)^{-1} \int_{-\infty}^{\infty} y^J K(y)\, dy$,

$$\begin{aligned} \mathrm{Bias}(x) \;&=\; \big(C_{\mathrm{B}} f^{(J)}(x) + r_1(x,h)\big)\, h^J, \\ |\mathrm{Bias}(x)| \;&\leq\; \bar{C}_{\mathrm{B}} M_J\, h^J, \\ \mathrm{SD}(x)^2 \;&=\; \frac{C_{\mathrm{SD}} f(x) + r_2(x,h)}{nh} \;\leq\; \frac{C_{\mathrm{SD}} M_0}{nh}, \end{aligned}$$

where $\lim_{h \downarrow 0} r_j(x,h) = 0$ for $j = 1, 2$.
In case of $h = C n^{-1/(2J+1)}$ for some constant $C > 0$,

$$\mathrm{RMSE}(x) \;\leq\; \tilde{C} n^{-J/(2J+1)}$$

with $\tilde{C} := \sqrt{\bar{C}_{\mathrm{B}}^2 M_J^2 C^{2J} + C_{\mathrm{SD}} M_0 / C}$.

The proof of this theorem is analogous to the proof of Theorem 5.6. This time one uses Taylor's formula

$$f(x+t) - f(x) \;=\; \sum_{j=1}^{J-1} \frac{f^{(j)}(x)}{j!}\, t^j + \frac{f^{(J)}(\xi(x,t))}{J!}\, t^J$$

for a suitable point $\xi(x,t)$ within $[x \pm |t|]$, where $f^{(J)}(\xi(x,t)) \to f^{(J)}(x)$ as $t \to 0$. Now

$$\mathrm{Bias}(x) \;=\; \frac{h^J}{J!} \int_{-\infty}^{\infty} y^J K(y) f^{(J)}(\xi(x, -hy))\, dy.$$

A kernel function K with the property that $\int_{-\infty}^{\infty} y^j K(y)\, dy = 0$ for $1 \leq j < J$ is called a *kernel of order J*. Hence Theorem 5.6 refers to kernels of order tow. An example for a kernel of order four is given by the 'sombrero function' K_* with

$$K_*(y) \;:=\; \frac{3 - y^2}{2}\, \phi(y). \tag{5.3}$$

Figure 5.9 shows its graph.

Computation and visualisation of kernel density estimators. The explicit computation of $\widehat{f}_h$ at a single point x is straightforward. Less obvious is the computation and visualisation of the whole function $\widehat{f}_h$. Depending on the kernel function K there are different options.

In case of the Gaussian kernel ϕ or the sombrero function K_* in (5.3) the function $\widehat{f}_h$ is smooth. Thus it suffices to compute $\widehat{f}$ on a fine grid of points and interpolate linearly.

Now we describe a particular method to compute $\widehat{f}_h$ in case of the triangular kernel Δ. Each summand $\Delta((x - X_i)/h)/(nh)$ of $\widehat{f}_h(x)$ is a continuous and piecewise linear function of x with

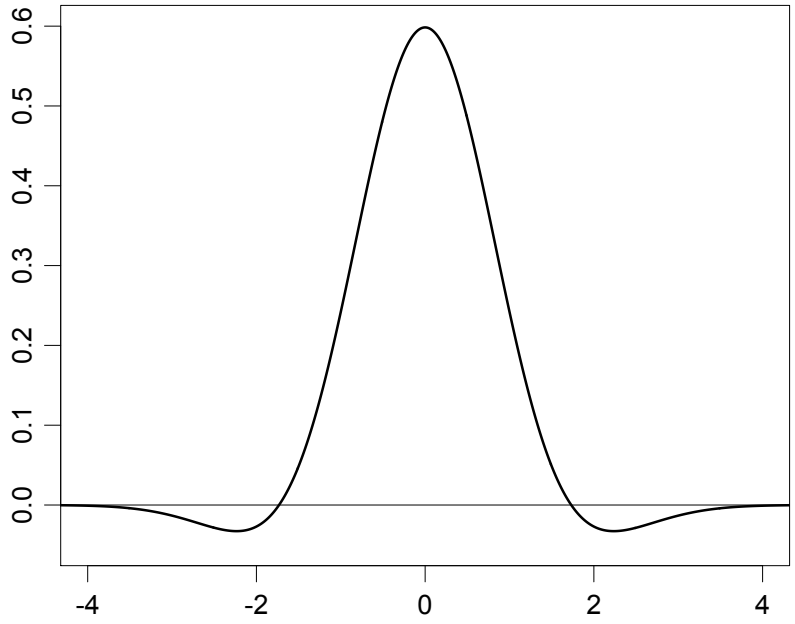

Figure 5.9: The sombrero kernel function.

changes of slope at the three points $X_i - h, X_i, X_i + h$. Hence $\widehat{f}_h$ is continuous and piecewise linear with potential changes of slope at the points in

$$\{X_i - h, X_i, X_i + h : 1 \le i \le n\}.$$

If $y_1 < y_2 < \ldots < y_m$ denote the $m \le 3n$ different elements of the latter set, then $\widehat{f}_h = 0$ on $(-\infty, y_1] \cup [y_m, \infty)$, and it suffices to compute $\widehat{f}_h(y_j)$ for $1 < j < m$. Other values can be obtained via linear interpolation.

For the computation of $(\widehat{f}_h(y_j))_{j=1}^m$ we consider now the left-sided derivative $\widehat{f}_h{}'(y-)$ of $\widehat{f}_h$ at a point y. Note that $\widehat{f}_h(y_1) = 0$ and

$$\widehat{f}_h(y_j) \;=\; \widehat{f}_h(y_{j-1}) + (y_j - y_{j-1})\widehat{f}_h{}'(y_j -) \quad \text{for } j = 2, 3, \ldots, m.$$

Once we have computed $\big(\widehat{f}_h{}'(y_j -)\big)_{j=2}^m$, the vector $(\widehat{f}_h(y_j))_{j=1}^m$ may be computed easily in $O(n)$ steps. An explicit expression for the derivative $\widehat{f}_h{}'(y-)$ is

$$\begin{aligned}
\widehat{f}_h{}'(y-) \;&=\; \frac{1}{nh} \sum_{i=1}^n \lim_{x \uparrow y} \frac{\Delta((y - X_i)/h) - \Delta((x - X_i)/h)}{y - x} \\
&=\; \frac{1}{nh^2} \sum_{i=1}^n \big(1_{[X_i - h < y \le X_i]} - 1_{[X_i < y \le X_i + h]}\big) \\
&=\; \frac{1}{nh^2} \sum_{i=1}^n \big(1_{[X_i - h < y]} - 2 \cdot 1_{[X_i < y]} + 1_{[X_i + h < y]}\big) \\
&=\; \frac{1}{nh^2}\, D(y)
\end{aligned}$$

with

$$D(y) \;:=\; \#\{i : X_{(i)} - h < y\} - 2\#\{i : X_{(i)} < y\} + \#\{i : X_{(i)} + h < y\}.$$

The vectors $\tilde{\boldsymbol{X}} := (X_{(i)})_{i=1}^{n+1}$ of order statistics and $\boldsymbol{y} := (y_j)_{j=2}^m$ may be determined in $O(n \log n)$ steps. Then it is possible to compute $\boldsymbol{D} := (D(y_j))_{j=2}^m$ in $O(n)$ steps. For with $X_{(0)} = -\infty$ and

```
ℓ1 ← 0
ℓ2 ← 0
ℓ3 ← 0
for j ← 2 to m do
     while X_(ℓ1+1) + h < y_j do
          ℓ1 ← ℓ1 + 1
     end
     while X_(ℓ2+1) < y_j do
          ℓ2 ← ℓ2 + 1
     end
     while X_(ℓ3+1) − h < y_j do
          ℓ3 ← ℓ3 + 1
     end
     D(y_j) ← ℓ1 + ℓ3 − 2ℓ2
end
```

Table 5.1: Auxiliary code for kernel densitiy estimator with triangular kernel.

$X_{(n+1)} = \infty$ one may write

$$D(y_j) \;=\; \ell_{j,1} - 2\ell_{j,2} + \ell_{j,3},$$

where

$$\begin{aligned}
\ell_{j,1} &:= \max\bigl\{i \in \{0,1,\dots,n+1\} : X_{(i)} + h < y_j\bigr\},\\
\ell_{j,2} &:= \max\bigl\{i \in \{0,1,\dots,n+1\} : X_{(i)} < y_j\bigr\},\\
\ell_{j,3} &:= \max\bigl\{i \in \{0,1,\dots,n+1\} : X_{(i)} - h < y_j\bigr\}.
\end{aligned}$$

Table 5.1 contains corresponding pseudocode.

Remark 5.8. Kernel density estimators have been introduced by M. Rosenblatt (1956) and E. Parzen (1962), and numerous authors extended this method and theory. Optimality of the Epanechnikov kernel was proved by V.A. Epanechnikov (1969). The problem of bandwidth choice is still a topic of research; see the review paper of M. C. Jones, J. S. Marron and S. J. Sheather (1996). The monograph B. W. Silverman (1986) presents various approaches to density estimation.

5.4 Checking Model Assumptions

In some applications it is important to check whether P belongs to a given class $(P_\theta)_{\theta\in\Theta}$ of distributions. For instance, often people are wondering whether P is a normal distribution, that means $P = P_\theta$ with $\theta = (\mu,\sigma) \in \Theta = \mathbb{R}\times(0,\infty)$ and $P_\theta = \mathcal{N}(\mu,\sigma^2)$. To check such an assumption, one could compute histograms or kernel density estimators to check its plausibility. But these methods tend to be rather imprecise; in particular, it is difficult to investigate the tails of distributions in this fashion.

Just some bad news at the very beginning: It is impossible to verify that a particular model $(P_\theta)_{\theta\in\Theta}$ is adequate. Nevertheless it is possible to check the plausibility of a certain model. Sometimes one can also falsify a certain model with given confidence by means of statistical tests.

In the sequel we assume that the distributions P_θ are given by continuous distribution functions F_θ. We describe two graphical methods and a formal test for the plausibility of the model $(P_\theta)_{\theta\in\Theta}$.

We start with some considerations about order statistics. According to Lemma 3.11, $(X_i)_{i=1}^n$ has the same distribution as $\bigl(F^{-1}(U_i)\bigr)_{i=1}^n$ with independent random variables $U_1, U_2, \dots, U_n$ which

are uniformly distributed on $[0,1]$. For the order statistics $X_{(k)}$ of the X_i and $U_{(k)}$ of the U_i this implies that

$$(X_{(k)})_{k=1}^n \text{ and } \big(F^{-1}(U_{(k)})\big)_{k=1}^n$$

have the same distribution. If F is continuous, then $F(F^{-1}(u)) = u$ for all $u \in (0,1)$, so

$$\big(F(X_{(k)})\big)_{k=1}^n \text{ and } (U_{(k)})_{k=1}^n$$

have the same distribution. In Exercise 5.6 it is shown that

$$\mathbb{E}(U_{(k)}) = u_k := \frac{k}{n+1} \quad \text{and} \tag{5.4}$$

$$\mathrm{Var}(U_{(k)}) = \frac{u_k(1-u_k)}{n+2} \le \frac{1}{4(n+2)}. \tag{5.5}$$

P-P-plots. Under the assumption that $P = P_\theta$ for an unknown parameter $\theta \in \Theta$, let $\widehat{\theta} = \widehat{\theta}(\text{data})$ be a corresponding estimator. In view of (5.4) and (5.5) we expect that $F_{\widehat{\theta}}(X_{(k)}) \approx u_k$. Thus we consider a scatter plot of the points $\big(u_k, F_{\widehat{\theta}}(X_{(k)})\big)$, $1 \le k \le n$, a so-called P-P-plot (probability-probability-plot). If indeed $P_{\widehat{\theta}}$ is a good approximation of P, these points should be close to the straight line $\{(x,x) : x \in \mathbb{R}\}$.

Q-Q-plots, version 1. In view of (5.4) and (5.5), $X_{(k)}$ should be close to $F_{\widehat{\theta}}^{-1}(u_k)$ whenever $P_{\widehat{\theta}}$ is a good approximation of P. Thus we consider a scatter plot of the points $\big(F_{\widehat{\theta}}^{-1}(u_k), X_{(k)}\big)$, $1 \le k \le n$, a so-called Q-Q-plot (quantile-quantile-plot). Again these points should be close to the straight line $\{(x,x) : x \in \mathbb{R}\}$.

Q-Q-plots, version 2. Consider the special case of a location-scale family. That means, $\Theta = \mathbb{R} \times (0,\infty)$, and for $\theta = (\mu,\sigma)$ let $F_\theta(r) = F_o((r-\mu)/\sigma)$ with a given continuous distribution function F_o. Then $F_\theta^{-1}(v) = \mu + \sigma F_o^{-1}(v)$, and $(X_{(k)})_{k=1}^n$ has the same distribution as $\big(\mu + \sigma F_o^{-1}(U_{(k)})\big)_{k=1}^n$. Thus we consider a scatter plot of the points

$$\big(F_o^{-1}(u_k), X_{(k)}\big) \qquad \text{(version 2a)}$$

or of the points

$$\Big(F_o^{-1}(u_k), \frac{X_{(k)} - \widehat{\mu}}{\widehat{\sigma}}\Big) \qquad \text{(version 2b)},$$

$1 \le k \le n$. Under the assumption that $P = P_{\mu,\sigma}$, these n points should be close to the straight line $\big\{(x, \mu + \sigma x) : x \in \mathbb{R}\big\}$ (version 2a) or $\{(x,x) : x \in \mathbb{R}\}$ (version 2b).
In case of $P_{\mu,\sigma} = \mathcal{N}(\mu,\sigma^2)$, obvious estimators for μ and σ would be $\widehat{\mu} = \overline{X}$ and $\widehat{\sigma} = S$, respectively. Alternatively one could use robust estimators such as, say, $\widehat{\mu} = \mathrm{Median}(X_1,\ldots,X_n)$ and $\widehat{\sigma} := \mathrm{MAD}(X_1,\ldots,X_n)/\Phi^{-1}(0.75)$.

An informal test. The requirement that the points of a P-P-plot or Q-Q-plot (version 1 or 2b) should be close to the straight line $\{(x,x) : x \in \mathbb{R}\}$ is rather vague, of course. To get a feeling for the typical shape of such a plot in case of the location-scale family being correct, one could simulate data vectors $(\tilde{X}_i)_{i=1}^n$ with independent components following F_o and compare their P-P- or Q-Q-plots with the corresponding plot based on the original data. If the latter plot is substantially different from the plots for the simulated data, this is evidence against our model assumption.

Formal tests. Statistical tests of the null hypothesis that P belongs to a given location-scale family are easy to implement. Let $T = T(X_1, X_2, \ldots, X_n)$ be a test statistic which is also a shape parameter. That means, $T(a + bX_1, a + bX_2, \ldots, a + bX_n) = T(X_1, X_2, \ldots, X_n)$ for arbitrary $a \in \mathbb{R}$ and $b > 0$. This constraint is natural since we are mainly interested in the shape of the distribution, not the particular parameter (μ, σ). Moreover, the distribution of T under the null hypothesis does not depend on the particular parameter (μ, σ). With the distribution function

$$G_o(r) := \mathbb{P}\big(T(\tilde{X}_1, \tilde{X}_2, \ldots, \tilde{X}_n) \le r\big)$$

for stochastically independent random variables $\tilde{X}_1, \tilde{X}_2, \ldots, \tilde{X}_n$ with distribution F_o, a p-value of the null hypothesis that P belongs to $(P_\theta)_{\theta\in\Theta}$ is given by $1 - G_o(T-)$. Obviously one can also devise Monte-Carlo versions of this test.

Concerning the test statistic T, the following choices would be closely related with P-P-plots and Q-Q-plots, respectively:

$$T_1 := \max_{k=1,2,\ldots,n} \Big| F_o\Big(\frac{X_{(k)} - \widehat{\mu}}{\widehat{\sigma}}\Big) - u_k \Big|,$$
$$T_2 := \frac{1}{n}\sum_{k=1}^{n} \Big| \frac{X_{(k)} - \widehat{\mu}}{\widehat{\sigma}} - F_o^{-1}(u_k) \Big|.$$

But the reader may design different quantities. The only important point is that $\widehat{\mu}(\cdot)$ and $\widehat{\sigma}(\cdot)$ should be location and scale parameters, respectively. Then any function of $\big((X_k - \widehat{\mu})/\widehat{\sigma}\big)_{k=1}^n$ is automatically a shape parameter.

Example 5.9 (Log-Returns). Let K_i be the value of a share or a portfolio of shares by the end of the i-th trading day. A simple model in financial mathematics assumes that the *log-returns* $X_i := \log_{10}(K_{i+1}/K_i)$ are stochastically independent random variables with distribution $\mathcal{N}(\mu, \sigma^2)$, where $\mu \in \mathbb{R}$ and $\sigma > 0$ are unknown parameters. In particular, this model is used when determining the value of option by means of the celebrated Black–Scholes formula.

Figure 5.10 shows on the left hand side the log-values $\log_{10}(K_i)$ of a certain portfolio of German shares at all 3246 trading days in the years 1981-1993. On the right hand side one sees the corresponding $n = 3245$ log-Returns X_i.

Figure 5.11 shows for the Gaussian location-scale model the resulting P-P-plot (left), based on the sample median $\widehat{\mu} = 2.785 \cdot 10^{-4}$ and $\widehat{\sigma} = \mathrm{MAD}/\Phi^{-1}(0.75) = 3.272 \cdot 10^{-3}$. On the right hand side one sees the corresponding Q-Q-plot (version 2b). While the P-P-plot looks okay at first glance, the Q-Q-plot shows strong deviations from the straight line $\{(x, x) : x \in \mathbb{R}\}$. Comparisons with simulated data vectors (not shown here) show clearly that these deviations are significant. A formal test based on the test statistic $T_1 = 3.042 \cdot 10^{-2}$ yields Monte-Carlo p-values smaller than 10^{-4} (if the number of simulations is sufficiently high). Hence the assumption of independent, identically distributed log-returns with Gaussian distribution is wrong with high confidence.

A model which seems to fit the data much better assumes the distribution function F_o of a student distribution t_g. Trying different degrees g of freedom showed that for $g = 3$ (or $g = 4$) the approximation is quite good. Figure 5.12 shows the corresponding P-P- and Q-Q-plot. Here we used the scale parameter $\widehat{\sigma} = \mathrm{IQR}/(2t_{g;0.75})$.

In Section 8.3 we shall look at this data example once more and show that the log-returns are significantly *stochastically dependent*.

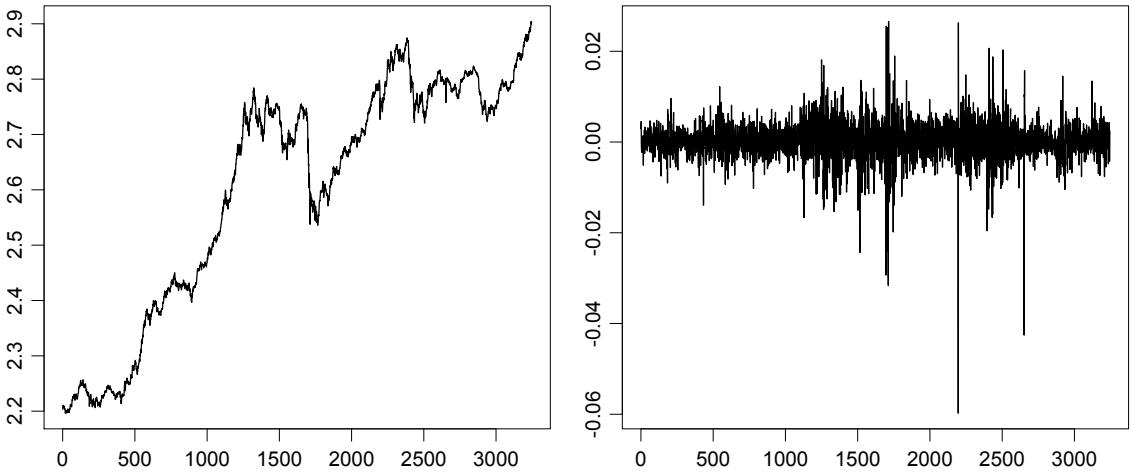

Figure 5.10: Log-transformed share values and log-returns.

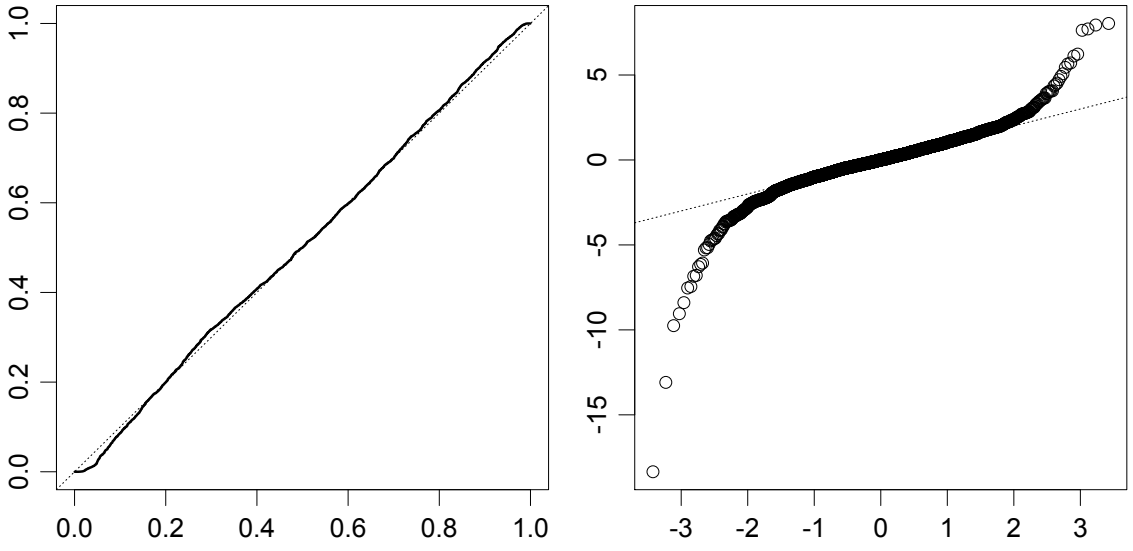

Figure 5.11: P-P- and Q-Q-plots for normal distributions and log-returns.

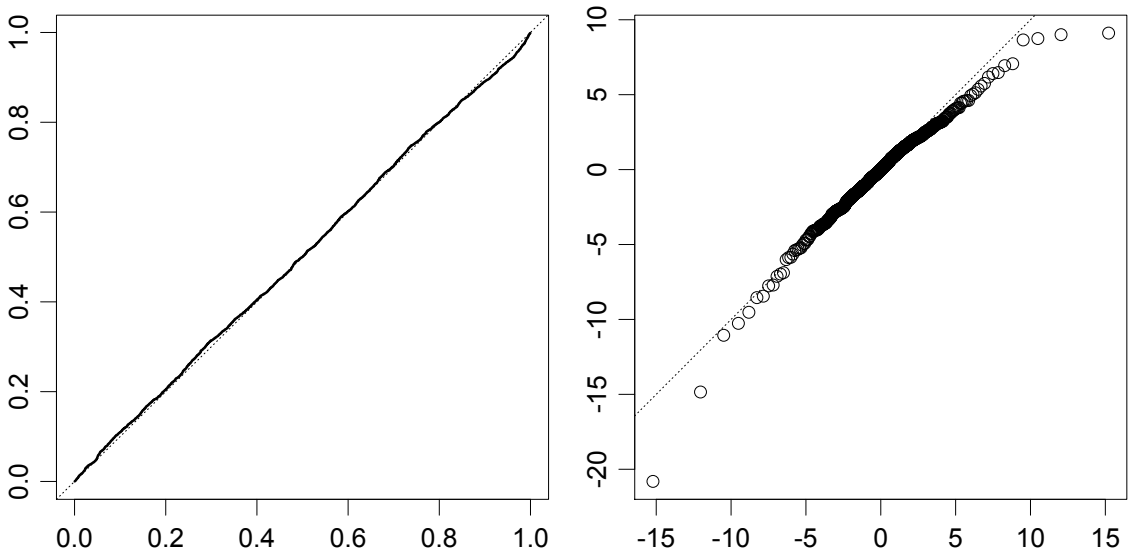

Figure 5.12: P-P- and Q-Q-Plot for t_3 distribution and log-returns.

5.5 Exercises

Exercise 5.1. For fixed $k \in \mathbb{N}_0$ let

$$F(r) \;:=\; \begin{cases} 1 - e^{-r} \sum_{i=0}^{k} \frac{r^i}{i!} & \text{for } r \geq 0, \\ 0 & \text{for } r \leq 0. \end{cases}$$

Show that F is a distribution function and determine the corresponding density function f. (Do you recognise a standard distribution?)

Exercise 5.2. Suppose we lost the data $X_1, X_2, \ldots, X_n$. Even the sample size is unknown, the only available information are two histograms following convention 2; see Figure 5.13. The first histogram corresponds to the intervals $(0, 1]$, $(1, 2]$, $\ldots$, $(4, 5]$, the second one to the intervals $(-0.5, 0.5]$, $(0.5, 1.5]$, $(1.5, 2.5]$, $\ldots$, $(4.5, 5.5]$.

(a) What do you learn about the graph of the empirical distribution function $\widehat{F}$ from a single histogram? Deduce from each histogram a picture with shaded regions indicating where the graph of $\widehat{F}$ has to be, ignoring for the moment the other histogram.

(b) Now combine the information from both histograms. Determine a histogram corresponding to the intervals $(-0.5, 0]$, $(0, 0.5]$, $(0.5, 1]$, $\ldots$, $(4.5, 5]$, $(5, 5.5]$, and provide a refined picture for the graph of $\widehat{F}$.

Exercise 5.3 (Histograms and triangular kernel). Show that the histogram estimator $\widehat{f}_{a,h}$ can be written as

$$\widehat{f}_{a,h}(x) \;=\; \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h}\, g_{a,h}(x, X_i)$$

with a certain function $g_{a,h} : \mathbb{R} \times \mathbb{R} \to \{0, 1\}$ such that $g_{a+z_o h,h}(x, y) = g_{a,h}(x, y)$ for arbitrary $x, y \in \mathbb{R}$ and $z_o \in \mathbb{Z}$. Then verify formula (5.2) for the average histogram $\widehat{f}_h(x)$ in (5.1) by showing that for any $b \in \mathbb{R}$,

$$\frac{1}{h} \int_b^{b+h} g_{a,h}(x, y)\, da \;=\; \max\Big(1 - \frac{|x - y|}{h}, 0\Big).$$

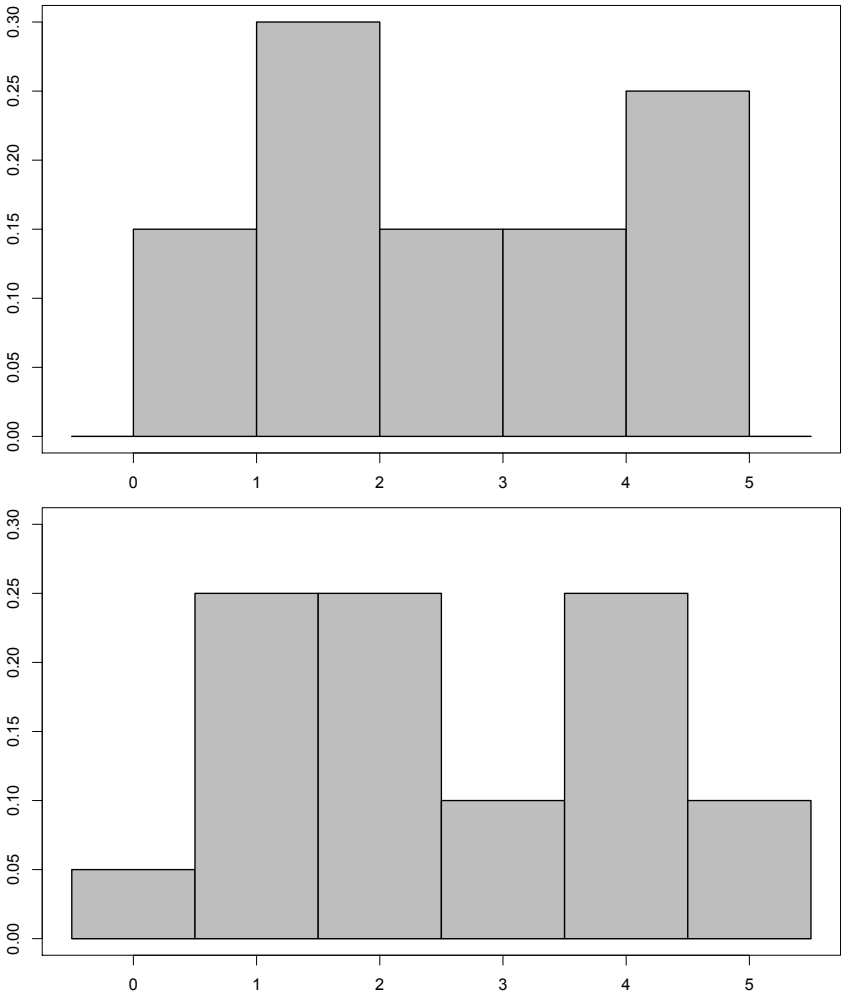

Figure 5.13: Two histograms of a lost data set.

Exercise 5.4 (Estimation of f'). Let $\widehat{f}_h$ be the kernel density estimator with a continuously differentiable kernel function K. Show that under the assumptions of Theorem 5.6,

$$\sup_{x\in\mathbb{R}} \mathbb{E}\big|\widehat{f}_h{}'(x) - f'(x)\big| \;=\; O(n^{-1/5}),$$

provided that $h = Cn^{-1/5}$ for some constant $C > 0$ and

$$\int_{-\infty}^{\infty} K'(y)^2\,dy \;<\; \infty, \quad \lim_{y\to\pm\infty} K(y) \;=\; 0.$$

Exercise 5.5 (Order of the sombrero kernel). Show that the sombrero kernel $K_*(y) = (3 - y^2)\phi(y)/2$ is a kernel of order 4. That means,

$$\int y^j K_*(y)\,dy \;=\; \begin{cases} 1 & \text{for } j = 0, \\ 0 & \text{for } 1 \le j < 4. \end{cases}$$

Determine a kernel of order 6. Hint: Exercise 2.5.

Exercise 5.6 (Moments of uniform order statistics). Let $U_{(1)} < U_{(2)} < \cdots < U_{(n)}$ be the order statitics of independent random variables $U_1, U_2, \ldots, U_n$ with uniform distribution on $[0, 1]$. Show that for $k \in \{1, 2, \ldots, n\}$,

$$\mathbb{E}(U_{(k)}) \;=\; u_k := \frac{k}{n+1} \quad \text{and} \quad \mathrm{Var}(U_{(k)}) \;=\; \frac{u_k(1-u_k)}{n+2} \;\le\; \frac{1}{4(n+2)}.$$

Hint: Remark 3.8 shows that $U_{(k)}$ has density function $f_{k-1,n-k}$ on $[0, 1]$, where generally

$$f_{\ell,m}(u) \;:=\; \frac{(\ell+m+1)!}{\ell!\,m!}\,u^\ell(1-u)^m$$

for $\ell, m \in \mathbb{N}_0$. Since $\int_0^1 f_{k-1,n-k}(u)\,du = 1$, we obtain the equation

$$\int_0^1 u^\ell(1-u)^m\,du \;=\; \frac{\ell!\,m!}{(\ell+m+1)!}.$$

Now compute $\mathbb{E}(U_{(k)})$ and $\mathbb{E}(U_{(k)}^2)$.

Exercise 5.7 (Exponential distributions). How could one check the model assumption that P is an exponential distribution with unknown mean $b > 0$, graphically or formally? Here $F_b(r) = \max(1 - \exp(-r/b), 0)$.

Exercise 5.8 (Q-Q-curves). For growing sample size n the Q-Q-plot (version 2a) resembles more and more the curve $(0, 1) \ni u \mapsto \big(F_o^{-1}(u), F^{-1}(u)\big)$. Plot this curve for $F_o = \Phi$ and the distribution function F of $P = \mathrm{Gamma}(a, 1)$ with different shape parameters $a > 0$ and $P = t_k$ with different degrees $k \ge 1$ of freedom.

Chapter 6

Comparing Samples

Quite often one is analyzing two or more samples corresponding to several studies or experiments. The question is whether these samples are substantially different with respect to a particular variable. In the present chapter we focus on numerical variables. Let $X_{ki} \in \mathbb{R}$ be our i-th observation in the k-th sample. Here $1 \leq k \leq K$ and $1 \leq i \leq n_k$. We consider all $N = n_1 + n_2 + \cdots + n_K$ observations as stochastically independent random variables and assume that X_{ki} follows an unknown distribution P_k with distribution function F_k. Now the question is whether and how these distributions $P_1, P_2, \ldots, P_K$ differ.

Sometimes we consider only one sample with a numerical variable and a categorical variable. Precisely, let (X_1, G_1), (X_2, G_2), ..., (X_N, G_N) be our observations with values $X_i \in \mathbb{R}$ and $G_i \in \{g_1, g_2, \ldots, g_K\}$. Now we may reorganise these observations such that $(X_{ki})_{i=1}^{n_k}$ contains all values X_j such that $G_j = g_k$. Assuming that the observations (X_j, G_j) are stochastically independent and identically distributed random variables, the vector $(n_k)_{k=1}^{K}$ has a multinomial distribution. If we condition on the random variables G_j, the random variables X_{ki} are stochastically independent, and P_k is the conditional distribution of X_j, given $G_j = g_k$. The question whether G_j and X_j are stochastically dependent is equivalent to the question whether the conditional distributions $P_1, P_2, \ldots, P_K$ differ.

In the next section we describe a simple graphical method to compare several (sub)samples. Then we focus on the case of $K = 2$ (sub)samples or distributions. In that context we also introduce the important concept of stochastic order. Finally we consider settings with $K \geq 3$ samples.

6.1 Box Plots and Box-Whisker Plots

In principle one could visualise and compare the K samples $\boldsymbol{X}_k = (X_{ki})_{i=1}^{n_k}$ by means of empirical distribution functions, histograms or kernel density estimators. But this may be cumbersome, in particular, if K is larger than three, say. John W. Tukey introduced a very simple but useful graphical display, the so-called box plots and box-and-whiskers plots.

Box plots. For a single sample $\boldsymbol{X}_k$ we determine five features, namely its minimum (Q_0), its first quartile (Q_1), its median (Q_2), its third quartile (Q_3) and its maximum (Q_4). These five features are now depicted as follows: The vertical axis corresponds to the potential values. Now we draw a rectangle with bottom line at Q_1 and top line at Q_3. This box is divided into two subboxes by an additional line at Q_3. In addition we draw a single line from Q_1 to the minimum Q_0 and from Q_3 to the maximum Q_4.

Despite this reduction to only five features, box plots provide a good first impression of the empirical distribution of the values X_{ki}, $1 \leq i \leq n_k$. In particular, the height of the box is equal to the

inter quartile range $\mathrm{IQR} = Q_3 - Q_1$.

Drawing the box plots for all K samples $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_K$ side by side allows us to spot empirical differences between the samples, for instance, differences of medians or inter quartile ranges. Whether or not such differences are significant has to be analyzed by means of different methods.

Example 6.1. Suppose that the sorted components of $\boldsymbol{X}_1$ are 0, 1, 5, 6, 7, 7, 8, 10, 14 and 18 while the sorted components of $\boldsymbol{X}_2$ are -3, 2, 4.5, 6, 7, 7.5, 8, 8.5, 11 and 15, so $n_1 = n_2 = 10$. The 2×5 characteristics are then

	Q_0	Q_1	Q_2	Q_3	Q_4
$\boldsymbol{X}_1$	0	5	7	10	18
$\boldsymbol{X}_2$	-3	4.5	7.25	8.5	15

The corresponding box plots are shown on the left hand side of Figure 6.1.

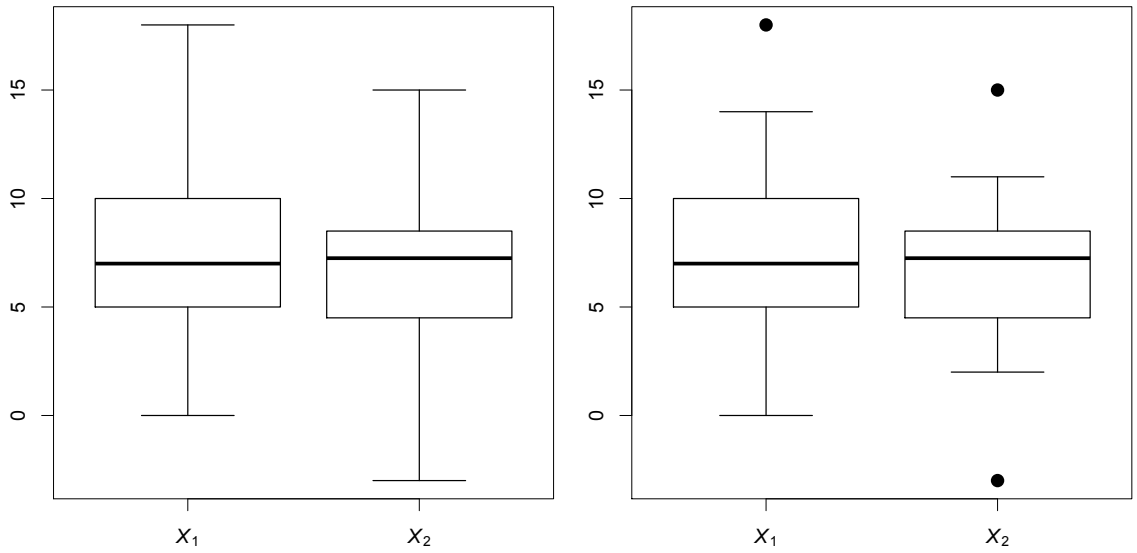

Figure 6.1: Box plot (left) and box-whiskers plot (right) for a simple data example.

Box-and-whiskers plots. A potential weakness of box plots is missing information for the ranges outside $[Q_1, Q_3]$. To represent observations in this region more precisely one defines a sample value as

- ‘suspiciously small’ if it is smaller than $Q_1 - 1.5 \cdot \mathrm{IQR}$,
- ‘suspiciously large’ if it is larger than $Q_3 + 1.5 \cdot \mathrm{IQR}$,
- ‘non-suspicious’ if it lies within $\big[Q_1 - 1.5 \cdot \mathrm{IQR}, Q_3 + 1.5 \cdot \mathrm{IQR}\big]$.

Now we replace the single line between minimum Q_0 and first quartile Q_1 by a single line from the *smallest non-suspicious observation* to Q_1. Analogously the single line between third quartile Q_3 and maximum Q_4 is replaced by a single line from Q_3 to the *largest non-suspicious observation*. Suspiciously small or large observations, if there are any, are depicted as single points.

Example 6.2 (Example 6.1 continued). In the sample $\boldsymbol{X}_1$, values outside of $\big[5 - 1.5 \cdot 5, 10 + 1.5 \cdot 5\big] = [-2.5, 17.5]$ are suspicious. This concerns only the component 18; the smallest non-suspicious value is 0 while the largest non-suspicious value is 14. In the sample $\boldsymbol{X}_2$, values outside of $\big[4.5 - 1.5 \cdot 4, 8.5 + 1.5 \cdot 4\big] = [-1.5, 14.5]$ are suspicious. This concerns the values -3 and 15; the smallest non-suspicious value is 2 while the largest non-suspicious value is 11. The corresponding box-whiskers plots are shown on the right hand side of Figure 6.1.

Remark 6.3. The thresholds $Q_1 - 1.5 \cdot \mathrm{IQR}$ and $Q_3 + 1.5 \cdot \mathrm{IQR}$ themselves are *not* drawn. They serve only to classify observations as non-suspicious or suspicious. For small samples or in case of many tied observations the box(-and-whiskers) plot may degenerate in the sense that certain single lines are missing or the median coincides with the first or third quartile.

The factor 1.5 for the IQR may be motivated as follows: In Exercise 6.1 it is shown that the sample mean lies always within the interval

$$\Big[\frac{Q_0 + Q_1 + Q_2 + Q_3}{4}, \frac{Q_1 + Q_2 + Q_3 + Q_4}{4}\Big].$$

to guarantee that the sample mean is at least within the box, i.e. the interval $[Q_1, Q_3]$, the following two inequalities should be satisfied:

$$\begin{aligned} Q_0 &\geq 3Q_1 - Q_2 - Q_3 = Q_1 - \mathrm{IQR} - (Q_2 - Q_1), \\ Q_4 &\leq 3Q_3 - Q_2 - Q_1 = Q_3 + \mathrm{IQR} + (Q_3 - Q_2). \end{aligned}$$

If the median (Q_2) is precisely the midpoint between first and third quartile, these conditions read

$$\begin{aligned} Q_0 &\geq Q_1 - 1.5 \cdot \mathrm{IQR}, \\ Q_4 &\leq Q_3 + 1.5 \cdot \mathrm{IQR}. \end{aligned}$$

Hence observations outside these thresholds are potentially problematic.

We end this section with multiple box-(whisker) plots for two larger data sets.

Example 6.4 (Income of professional baseball players). We consider a data set with the annual incomes of $N = 263$ US-american baseball players in the professional league. In addition to the variable $X =$ income (in 1000 USD) this data set contains an ordinal variable $G =$ years specifiying the number of seasons (including the current one) a player has played professionally. Since only 25 players have been playing for more than 14 years, we combine these observations into one category. The multiple box-whiskers plot of X with respect to this modified group variable G is shown in Figure 6.2. One sees clearly that the income tends to increase during the first three-four years. After that no clear trend is visible. Note also that in various groups there are remarkably large values. There are some 'freshmen' earning more than several senior guys.

The existence of very large big values, the non-existence of very small values and the fact that in many groups the median is closer to the first than to the third quartile indicate that the empirical distributions of incomes are right-skewed. Considering $\log_{10}(X)$ in place of X ameliorates these non-symmetries, and the development of incomes during the first few years becomes easier to grasp.

Example 6.5 (Hamburg-Marathon 2000). Now we consider the net running times (X, in hours) of the $N = 13049$ finishers of the Hamburg Marathon 2000. (About 16000 people registered.)

At first we show the empirical distribution function of this variable X in Figure 6.4. The fastest runner reached the finish line after 2 hours, 11 minutes and 6 seconds; the slowest runner came in after 5 hours, 32 minutes and 21 seconds. The median of all running times equals $X_{(6525)}$, 3 hours, 52 minutes and 10 seconds. From the organisers' viewpoint this distribution function is very interesting. It helps planning the size of facilities for showering, drinking and food. Another interesting phenomenon is the subtle kink of the empirical distribution function at 3, 3.5 and 4. Presumably this is due to the fact that many runners are wearing a watch and planning to reach the finish line in a little less than 3 hours or 3.5 hours or 4 hours.

Now we are interested in the dependence of the running time on the participants' age, for male and female runners separately. The data set contains a variable 'Altersklasse (age class)'. For the $N_M = 11203$ men this variable has the following potential values:

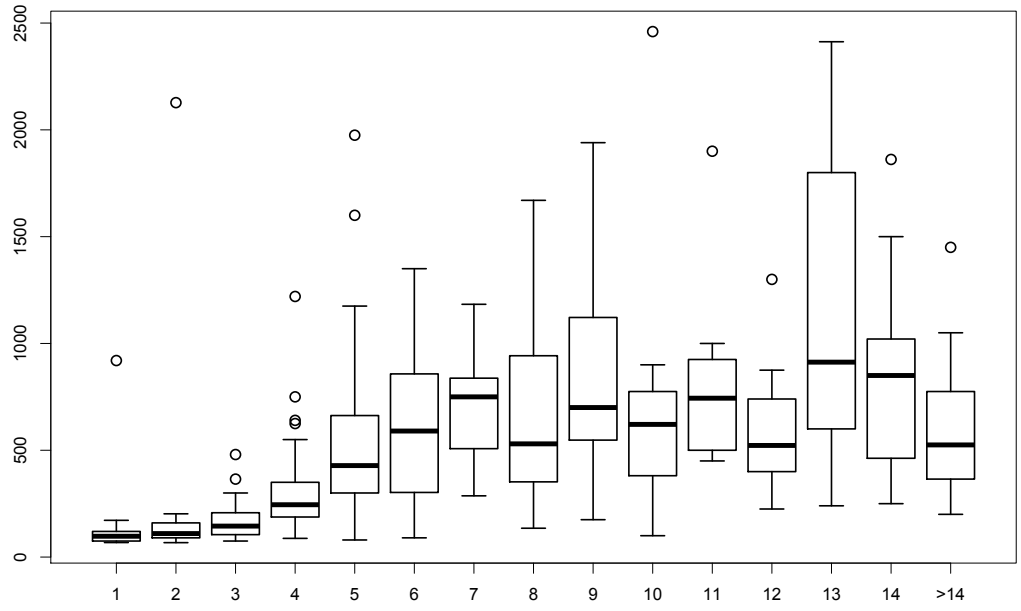

Figure 6.2: Box-whisker plots of annual incomes of baseball players versus experience.

MJ	:	Participant turned 18 or 19 during 2000,
MH	:	Participant turned 20-29 during 2000,
M30	:	Participant turned 30-34 during 2000,
M35	:	Participant turned 35-39 during 2000,
⋮		⋮
M75	:	Participant turned 75-79 during 2000.

The oldest participant was born in 1923. Since only two men started in age class M75, we combine age classes M70 and M75 to a new class M70+. Figure 6.5 shows the corresponding multiple box plot of the men's running times versus age class.

Interestingly the median is not monotone increasing with age. In age class MJ it is higher then in age class MH, and in age class MH it is still higher than in age classes M30, M35, M40 and M40. Over the latter groups the median is almost constant while for the higher age classes it starts increasing. This phenomenon is wellknown from sport science. Even professional runners reach their peak performance over long distances at age between 25 and 35 years.

Now we show in Figure 6.6 the running times of the $N_W = 1846$ women: Here there were the analogous age classes WJ, WH, W30, ..., W65. (The oldest female participant was born in 1931.) Since age class W65 comprised only six runners, we combined age classes W60 and W65 to a new class W60+. Again we see an almost constant median in the age classes W30, W35 and W40.

6.2 Comparing Two Means

Now we consider $K = 2$ samples $\boldsymbol{X}_1, \boldsymbol{X}_2$ and assume that the X_{ki} have unknown mean μ_k and unknown but finite standard deviation σ_k. Obvious estimators of μ_k and σ_k are the sample mean $\overline{X}_k := n_k^{-1} \sum_{i=1}^{n_k} X_{ki}$ and the sample standard deviation $S_k := \big((n_k - 1)^{-1} \sum_{i=1}^{n_k} (X_{ki} - \overline{X}_k)^2\big)^{1/2}$. Here $\mathbb{E}(\overline{X}_k) = \mu_k$, so

$$\mathbb{E}(\overline{X}_1 - \overline{X}_2) \;=\; \mu_1 - \mu_2,$$

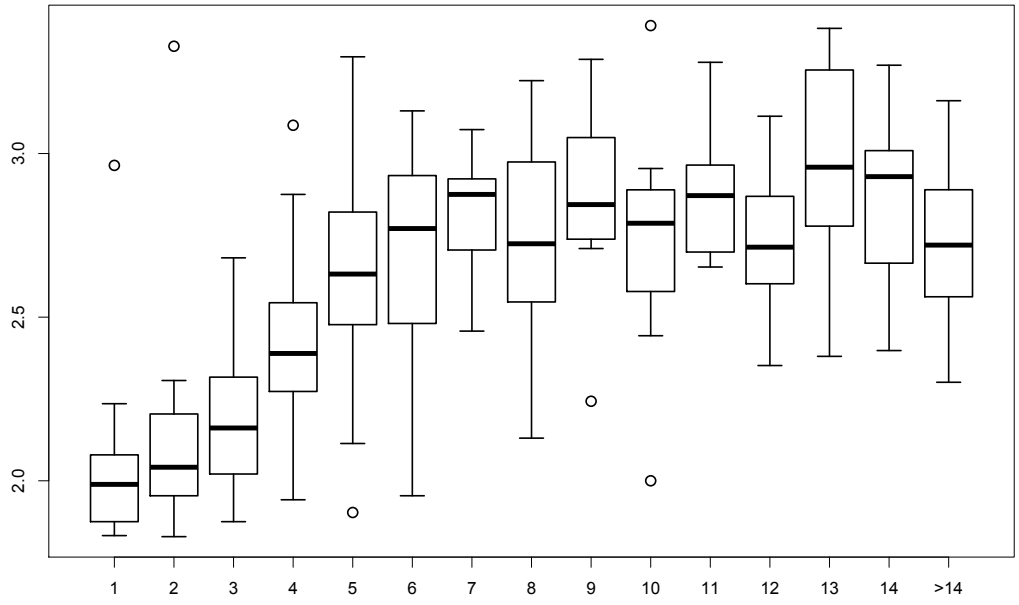

Figure 6.3: Box-whisker plots of $\log_{10}$-incomes of baseball players versus experience.

and

$$\mathrm{Std}(\overline{X}_1 - \overline{X}_2) \;=\; \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

In case of Gaussian observations X_{ki}, Theorem 4.3 of Gosset–Fisher implies that the four random variables $\overline{X}_1$, $\overline{X}_2$, S_1 and S_2 are stochastically independent, where $\overline{X}_k \sim \mathcal{N}\big(\mu_k, \sigma_k^2/n_k\big)$ and $(n_k-1)S_k^2/\sigma_k^2 \sim \chi^2_{n_k-1}$. This will be used subsequently.

Case 1: Identical standard deviations σ_1 and σ_2. If all $N = n_1 + n_2$ observations X_{ki} have one and the same standard deviation σ, then

$$\mathrm{Std}(\overline{X}_1 - \overline{X}_2) \;=\; \sigma\sqrt{n_1^{-1} + n_2^{-1}}.$$

A possible estimator for σ is given by

$$\widehat{\sigma} \;:=\; \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{N-2}}. \tag{6.1}$$

In case of Gaussian observations,

$$(N-2)\,\widehat{\sigma}^2/\sigma^2 \;\sim\; \chi^2_{N-2},$$

and it is stochastically independent from $\overline{X}_1 - \overline{X}_2$. Hence

$$\frac{\overline{X}_1 - \overline{X}_2 - \mu_1 + \mu_2}{\widehat{\sigma}\sqrt{n_1^{-1} + n_2^{-1}}} \;\sim\; t_{N-2}.$$

This implies the following confidence regions for $\mu_1 - \mu_2$: The lower bound

$$\overline{X}_1 - \overline{X}_2 \;-\; \widehat{\sigma}\sqrt{n_1^{-1} + n_2^{-1}}\, t_{N-2;1-\alpha},$$

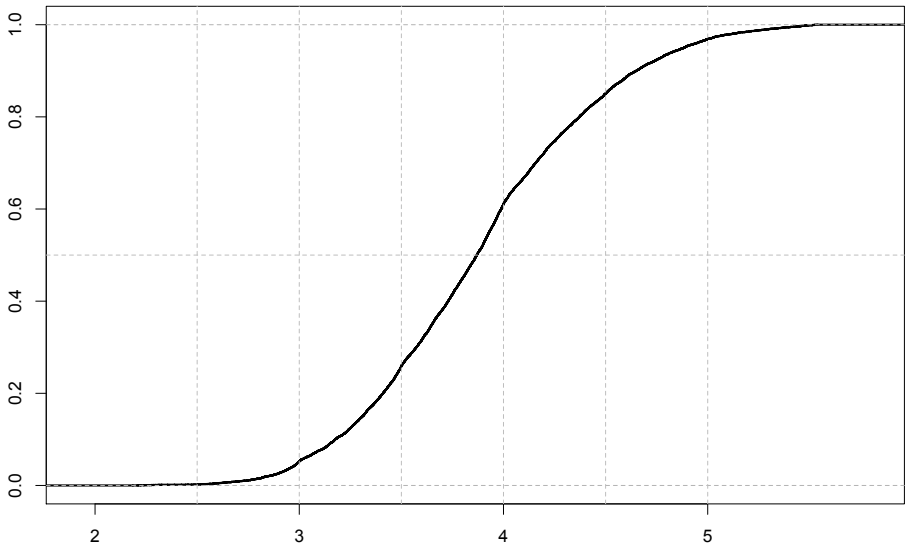

Figure 6.4: Empirical distribution function of the net running times (in hours), Hamburg Marathon 2000

the upper confidence bound

$$\overline{X}_1 - \overline{X}_2 \, + \, \widehat{\sigma}\sqrt{n_1^{-1} + n_2^{-1}}\, t_{N-2;1-\alpha}$$

or the confidence interval

$$\big[\overline{X}_1 - \overline{X}_2 \, \pm \, \widehat{\sigma}\sqrt{n_1^{-1} + n_2^{-1}}\, t_{N-2;1-\alpha/2}\big].$$

The confidence level is precisely $1-\alpha$ if the observations X_{ki} have Gaussian distribution. Otherwise the confidence level is approximately $1-\alpha$ as $\min(n_1, n_2) \to \infty$.

Case 2: Welch's method for arbitrary standard deviations σ_1 and σ_2. For the general case we mentioned already that $\overline{X}_1 - \overline{X}_2$ has expected value $\mu_1 - \mu_2$ and standard deviation $\tau := \sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$. In case of Gaussian observations we know that the standard error

$$\widehat{\tau} \, := \, \sqrt{S_1^2/n_1 + S_2^2/n_2}$$

and the estimator $\overline{X}_1 - \overline{X}_2$ are stochastically independent, and one can show that the standardised quantity

$$\frac{\overline{X}_1 - \overline{X}_2 \, - \, \mu_1 + \mu_2}{\widehat{\tau}}$$

is *approximately* student-distributed with

$$k = k(n_1, n_2, \sigma_1, \sigma_2) \, := \, \frac{\tau^4}{\sigma_1^4/(n_1^2(n_1-1)) + \sigma_2^4/(n_2^2(n_2-1))}$$

degrees of freedom; see below. This number k involves the unknown parameters σ_1, σ_2, so it is estimated by $\widehat{k} = k(n_1, n_2, S_1, S_2)$. In general, neither k nor $\widehat{k}$ is an integer. In this case one uses a generalised definition of student-distributions (see Section A.7 in the appendix), or one rounds these numbers to the next integer below. Finally this leads to the following confidence regions for $\mu_1 - \mu_2$ with approximate confidence level $(1-\alpha)$: The lower bound

$$\overline{X}_1 - \overline{X}_2 \, - \, \widehat{\tau}\, t_{\widehat{k};1-\alpha},$$

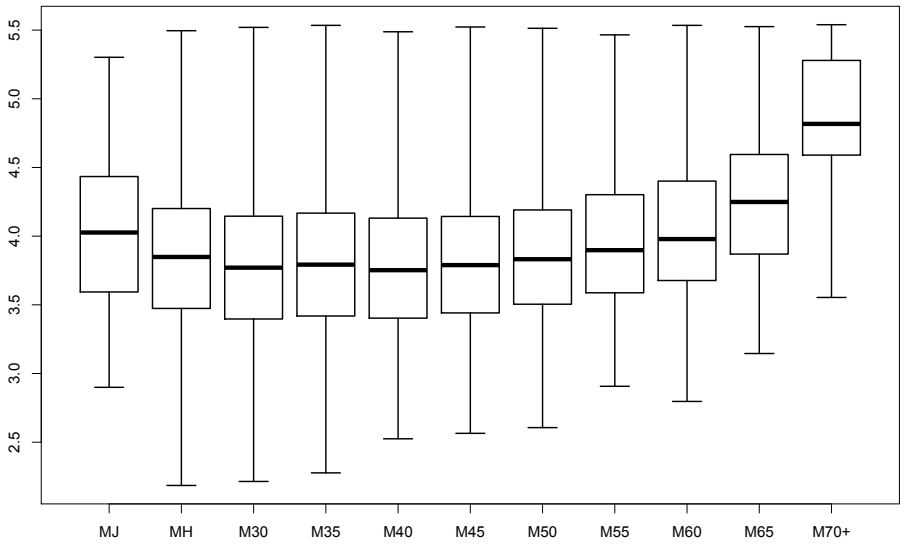

Figure 6.5: Multiple box plot of men's running times (in hours) versus age class.

the upper bound

$$\overline{X}_1 - \overline{X}_2 + \widehat{\tau}\, t_{\widehat{k};1-\alpha}$$

or the confidence interval

$$\big[\overline{X}_1 - \overline{X}_2 \pm \widehat{\tau}\, t_{\widehat{k};1-\alpha/2}\big].$$

Justifications of Welch's method. At first let us consider the student distribution t_k, i.e. the distribution of

$$Z_0 \Big/ \sqrt{\frac{1}{k}\sum_{i=1}^{k} Z_i^2}$$

with stochastically independent, standard Gaussian random variables Z_0, Z_1, Z_2, Z_3, ... The random quantity $k^{-1}\sum_{i=1}^{k} Z_i^2$ has mean 1, variance $2/k$ and is approximately Gaussian as $k \to \infty$; see Exercise 6.2.

With these considerations in mind we consider the ratio $(\overline{X}_1 - \overline{X}_2 - \mu_1 + \mu_2)/\widehat{\tau}$. In case of Gaussian observations it is distributed as

$$Z_0 \Big/ \sqrt{\frac{1}{\tau^2}\Big(\frac{\sigma_1^2}{n_1(n_1-1)}\sum_{i=1}^{n_1-1} Z_i^2 + \frac{\sigma_2^2}{n_2(n_2-1)}\sum_{i=n_1}^{n_1+n_2-2} Z_i^2\Big)}.$$

The term inside the square root is a random variable with mean 1, with variance

$$\frac{1}{\tau^4}\Big(\frac{2\sigma_1^4}{n_1^2(n_1-1)} + \frac{2\sigma_2^4}{n_2^2(n_2-1)}\Big) \;=\; \frac{2}{k(n_1,n_2,\sigma_1,\sigma_2)},$$

and as $\min(n_1, n_2) \to \infty$ it is approximately normally distributed. □

Example 6.6 (North-south gradient of body height.). As a numerical example for Welch's method we consider the mean body height μ_1 of all Swiss men and μ_2 of all men in northern Germany at age around 18–30. When interviewing $n_1 = 145$ male students at the university of Bern we obtained $\overline{X}_1 = 178.938$ and $S_1 = 6.2363$. Interviewing $n_2 = 26$ male students at the University of Lübeck yielded $\overline{X}_2 = 183.962$ and $S_2 = 7.5497$. This entails the estimator

$$\overline{X}_1 - \overline{X}_2 \;=\; -5.024$$

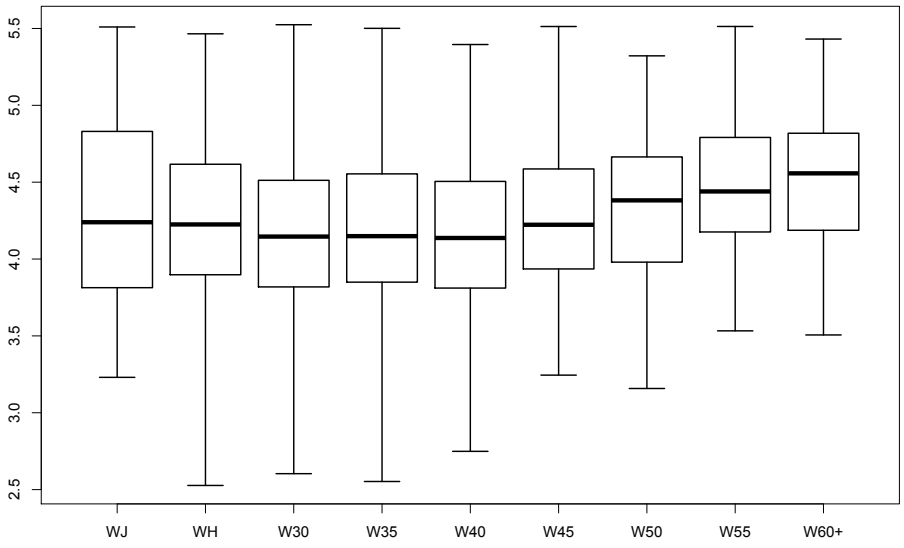

Figure 6.6: Multiple box plot of women's running times (in hours) versus age class.

for $\mu_1 - \mu_2$, and its standard deviation τ is estimated by the standard error

$$\widehat{\tau} \;=\; \sqrt{\frac{6.2363^2}{145} + \frac{7.5497^2}{26}} \;=\; 1.5686.$$

For $\widehat{k}$ we obtained here the (rounded) value 31, and $t_{31;0.975} = 2.0395$. Hence an approximate 95%-confidence interval for the difference $\mu_1 - \mu_2$ is given by

$$[-5.024 \,\pm\, 1.5686 \cdot 2.0395] \;=\; [-8.223, -1.825].$$

Thus we may claim with confidence about 95% that (a) the mean body height μ_1 is smaller than the mean body height μ_2 and that (b) the absolute difference is between 1.8 and 8.3 cm. (The problem that we didn't obtain true random samples from the two populations is ignored here.)

6.3 Stochastic Order

Before treating additional procedures to compare samples we introduce the important concept of stochastic order. In what follows we always consider random variables X_1, X_2 with distributions P_1, P_2 and distribution functions F_1, F_2. The vague statement that X_1 tends to be smaller than X_2 may be specified in various equivalent ways:

Lemma 6.7. *The following four statements are equivalent:*

(i) *For arbitrary $x \in \mathbb{R}$,*

$$F_1(x) \geq F_2(x).$$

(ii) *For arbitrary $u \in (0,1)$,*

$$F_1^{-1}(u) \;\leq\; F_2^{-1}(u).$$

(iii) *There exists a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ with random variables $\tilde{X}_1 \sim P_1$ and $\tilde{X}_2 \sim P_2$ such that $\tilde{X}_1 \leq \tilde{X}_2$ almost surely.*

(iv) *For any monotone increasing and bounded or non-negative function $h : \mathbb{R} \to \mathbb{R}$,*

$$\mathbb{E}\, h(X_1) \;\leq\; \mathbb{E}\, h(X_2).$$

The proof of this lemma is Exercise 6.6. The conditions stated therein lead to the following definition:

Definition 6.8 (Stochastic order). The distribution P_1 is *stochastically smaller than or equal to* the distribution P_2 if the conditions stated in Lemma 6.7 are satisfied. Sometimes we also say that the random variable X_1 or distribution function F_1 is stochastically smaller than or equal to the random variable X_2 or distribution function F_2. We write briefly $P_1 \leq_{\text{st.}} P_2$ or $X_1 \leq_{\text{st.}} X_2$ or $F_1 \leq_{\text{st.}} F_2$.
If in addition $P_1 \neq P_2$, that means, $F_1(x) > F_2(x)$ for ar least one $x \in \mathbb{R}$, then we call P_1 or X_1 or F_1 stochastically smaller than P_2 or X_2 or F_2. The corresponding brief notation is $P_1 <_{\text{st.}} P_2$ or $X_1 <_{\text{st.}} X_2$ or $F_1 <_{\text{st.}} F_2$.

Examples. Subsequently we give some examples for stochastic order. The respective proofs are left to the reader as exercises.

(a) Let Z be a real-valued random variable. For real constants μ_1, μ_2, $\mu_1 + Z <_{\text{st.}} \mu_2 + Z$ if and only if $\mu_1 < \mu_2$.

(b) Let Z be a real-valued random variable with density function f_o such that $f_o(-x) = f_o(x)$ for all $x \geq 0$ and f_o is non-increasing on $[0, \infty)$. For instance let $Z \sim \mathcal{N}(0, 1)$. For real constants μ_1, μ_2, $|\mu_1 + Z| <_{\text{st.}} |\mu_2 + Z|$ if and only if $|\mu_1| < |\mu_2|$.

(c) For $n \in \mathbb{N}$ and $p_1, p_2 \in [0, 1]$, $\text{Bin}(n, p_1) <_{\text{st.}} \text{Bin}(n, p_2)$ if and only if $p_1 < p_2$.

(d) $\text{Bin}(n_1, p) <_{\text{st.}} \text{Bin}(n_2, p)$ for $0 < p \leq 1$ and natural numbers $n_1 < n_2$.

(e) For $\theta > 0$ let F_θ be defined as in Lemma 2.7, where $\#\{k \geq 0 : w_k > 0\} \geq 2$. For $\theta_1, \theta_2 > 0$, $F_{\theta_1} <_{\text{st.}} F_{\theta_2}$ if and only if $\theta_1 < \theta_2$.

(f) Let X_1 and X_2 be random variables with density functions f_1 and f_2 on $\mathbb{R}$, respectively. Suppose that $f_2(x) = h(x) f_1(x)$ for some non-decreasing function $h : \mathbb{R} \to [0, \infty)$. Then $X_1 \leq_{\text{st.}} X_2$.

6.4 Smirnov's Test for Empirical Distribution Functions

Now we consider $N = n_1 + n_2$ stochastically independent random variables X_{ki} for $k = 1, 2$ and $i = 1, 2, \ldots, n_k$, where X_{ki} follows distribution function F_k. The corresponding empirical distribution functions are denoted with $\widehat{F}_k$, i.e.

$$\widehat{F}_k(x) \;:=\; \frac{1}{n_k} \sum_{i=1}^{n_k} 1_{[X_{ki} \leq x]}.$$

In many applications one would like to verify the working hypothesis that $F_1 >_{\text{st.}} F_2$ with a certain confidence. Strictly speaking this is impossible, unless one assumes that either $F_1 <_{\text{st.}} F_2$ or $F_1 \equiv F_2$ or $F_1 >_{\text{st.}} F_2$. For the null hypothesis

$$H_o \,:\, F_1 \leq_{\text{st.}} F_2$$

there exist reasonable tests one of which we'll describe now. The first characterisation of stochastic order in Lemma 6.7 motivates the test statistic

$$T_{\text{Sm}} \;:=\; \sup_{x \in \mathbb{R}} \big(\widehat{F}_2(x) - \widehat{F}_1(x) \big).$$

The following result shows how to construct critical values or p-values for this test statistic.

Lemma 6.9. *Let M be uniformly distributed on the set of all subsets of $\{1, 2, \ldots, N\}$ with n_2 elements. For $\ell \in \{1, 2, \ldots, N\}$ let*

$$H_\ell \;:=\; \frac{N}{n_1 n_2}\,\#(M \cap \{1, \ldots, \ell\}) - \frac{\ell}{n_1}.$$

Under the null hypothesis H_o above, for arbitrary $c \geq 0$,

$$\mathbb{P}(T_{\mathrm{Sm}} \geq c) \;\leq\; \mathbb{P}\Big(\max_{\ell=1,2,\ldots,N} H_\ell \geq c\Big).$$

Equality holds true if F_1 and F_2 are identical and continuous.
In the special case that $n_1 = n_2 = n$,

$$\mathbb{P}\Big(\max_{\ell=1,2,\ldots,N} H_\ell \geq c\Big) \;=\; \binom{N}{n + \lceil nc \rceil}\Big/\binom{N}{n}.$$

With $G^{\mathrm{Sm}}_{n_1,n_2}(c) := \mathbb{P}\big(\max_\ell H_\ell \geq c\big)$, a p-value for the null hypothesis H_o above is given by $G^{\mathrm{Sm}}_{n_1,n_2}(T_{\mathrm{Sm}})$. In case of $n_1 = n_2 = n$ this equals

$$\binom{N}{n + nT_{\mathrm{Sm}}}\Big/\binom{N}{n}.$$

Proof of Lemma 6.9. Let $U_1, U_2, \ldots, U_N$ be independent random variables with uniform distribution on $[0, 1]$. It follows from Lemma 3.11 that the observations X_{ki} have the same distribution as

$$\tilde{X}_{ki} \;:=\; \begin{cases} F_1^{-1}(U_i) & \text{if } k = 1, \\ F_2^{-1}(U_{n_1+i}) & \text{if } k = 2. \end{cases}$$

Under the null hypothesis, $F_1 \geq F_2$ pointwise, and T_{Sm} has the same distribution as

$$\begin{aligned}
&\sup_{x\in\mathbb{R}}\Big(\frac{1}{n_2}\sum_{j=n_1+1}^{N} 1_{[F_2^{-1}(U_j)\leq x]} - \frac{1}{n_1}\sum_{i=1}^{n_1} 1_{[F_1^{-1}(U_i)\leq x]}\Big) \\
&\quad= \sup_{x\in\mathbb{R}}\Big(\frac{1}{n_2}\sum_{j=n_1+1}^{N} 1_{[U_j\leq F_2(x)]} - \frac{1}{n_1}\sum_{i=1}^{n_1} 1_{[U_i\leq F_1(x)]}\Big) \\
&\quad\leq \sup_{x\in\mathbb{R}}\Big(\frac{1}{n_2}\sum_{j=n_1+1}^{N} 1_{[U_j\leq F_1(x)]} - \frac{1}{n_1}\sum_{i=1}^{n_1} 1_{[U_i\leq F_1(x)]}\Big) \\
&\quad= \sup_{v\in F_1(\mathbb{R})}\Big(\frac{1}{n_2}\sum_{j=n_1+1}^{N} 1_{[U_j\leq v]} - \frac{1}{n_1}\sum_{i=1}^{n_1} 1_{[U_i\leq v]}\Big) \\
&\quad\leq \sup_{v\in[0,1]}\Big(\frac{1}{n_2}\sum_{j=n_1+1}^{N} 1_{[U_j\leq v]} - \frac{1}{n_1}\sum_{i=1}^{n_1} 1_{[U_i\leq v]}\Big) \;=:\; T^o_{\mathrm{Sm}}.
\end{aligned}$$

Obviously equality holds if F_1 and F_2 are identical and continuous.

Now we investigate the random variable T^o_{Sm} in more detail. With probability one the values $U_1, U_2, \ldots, U_N$ are pairwise different. If $U_{(1)} < U_{(2)} < \ldots < U_{(N)}$ denote the corresponding

order statistics and $R_1, R_2, \ldots, R_N$ the corresponding ranks, then

$$\begin{aligned} T^o_{\mathrm{Sm}} &= \max_{\ell=1,2,\ldots,N}\Big(\frac{1}{n_2}\sum_{j=n_1+1}^{N} 1_{[U_j\le U_{(\ell)}]} - \frac{1}{n_1}\sum_{i=1}^{n_1} 1_{[U_i\le U_{(\ell)}]}\Big) \\ &= \max_{\ell=1,2,\ldots,N}\Big(\frac{1}{n_2}\sum_{j=n_1+1}^{N} 1_{[R_j\le \ell]} - \frac{1}{n_1}\sum_{i=1}^{n_1} 1_{[R_i\le \ell]}\Big) \\ &= \max_{\ell=1,2,\ldots,N}\Big(\frac{1}{n_2}\sum_{j=n_1+1}^{N} 1_{[R_j\le \ell]} - \frac{1}{n_1}\Big(\ell - \sum_{i=n_1+1}^{N} 1_{[R_i\le \ell]}\Big)\Big) \\ &= \max_{\ell=1,2,\ldots,N}\Big(\frac{N}{n_1 n_2}\,\#(M\cap\{1,\ldots,\ell\}) - \frac{\ell}{n_1}\Big) \end{aligned}$$

with the random set $M := \{R_{n_1+1}, \ldots, R_N\}$. In the third step we used the fact that $(R_i)_{i=1}^N$ contains the numbers $1, 2, \ldots, N$ whence $\sum_{i=1}^N 1_{[R_i\le\ell]} = \ell$. For symmetry reasons the order of the numbers $R_1, R_2, \ldots, R_N$ is completely at random (Exercise 3.7), and this implies that M is uniformly distributed on the set of subsets of $\{1, 2, \ldots, N\}$ with n_2 elements.

It remains to derive the distribution of T^o_{Sm} in the special case of $n_1 = n_2 = n$. To this end we define $H_0 := 0$ and note that the tuple $H = (H_\ell)_{\ell=0}^N$ lies in the set

$$\mathcal{H}_N := \big\{(h_\ell)_{\ell=0}^N : h_0 = 0 \text{ and } |h_\ell - h_{\ell-1}| = 1/n \text{ for } 1\le\ell\le N\big\}.$$

Precisely, H is uniformly distributed on the set $\big\{h \in \mathcal{H}_N : h_N = 0\big\}$. The latter set consists of $\binom{N}{n}$ different tuples, because one has to specify n "time points" $\ell \in \{1, 2, \ldots, N\}$ at which $h_\ell - h_{\ell-1} = 1/n$. The question is how many tuples $h \in \mathcal{H}_N$ with $h_N = 0$ reach or exceed a given value $c \in \{1/n, 2/n, \ldots, 1\}$. Here we apply the mirror principle: For an arbitrary tuple $h \in \mathcal{H}_N$ we define a tuple $\tilde{h} = (\tilde{h}_\ell)_{\ell=0}^N$ via

$$\tilde{h}_\ell := \begin{cases} h_\ell & \text{if } \max_{j\le\ell} h_j < c, \\ c - (h_\ell - c) & \text{if } \max_{j\le\ell} h_j \ge c. \end{cases}$$

This defines a bijective mapping $h \mapsto \tilde{h}$ from $\mathcal{H}_N$ to $\mathcal{H}_N$, and

$$h_N = 0 \text{ and } \max_{\ell\le N} h_\ell \ge c \quad \text{if and only if} \quad \tilde{h}_N = 2c.$$

Figure 6.7 illustrates this mapping: For $n_1 = n_2 = 50$ and $c = 7/50$ one sees a tuple $h \in \mathcal{H}_{100}$ and its image $\tilde{h}$. In general there are precisely $\binom{N}{n+nc}$ tuples $h \in \mathcal{H}_N$ with $h_N = 2c$. For h_N equals $2c$ if and only if $h_\ell - h_{\ell-1} = 1/n$ for $n + nc$ "time points" ℓ while $h_\ell - h_{\ell-1} = -1/n$ for $n - nc$ "time points" ℓ. These considerations show that indeed

$$\mathbb{P}\Big(\max_{\ell=1,2,\ldots,N} H_\ell \ge c\Big) = \binom{N}{n+nc}\Big/\binom{N}{n}$$

for $c = 0, 1/n, 2/n, \ldots, 1$. □

Example 6.10. Smirnov's test is even applicable if instead of the precise $N = n_1 + n_2$ values X_{ki} one knows only their ranks. Here is an explicit example: For a particular type of sports (e.g. long distance swimming) a new training method has been developed. To verify its usefulness, $N = 2n$ test persons are divided randomly into two groups of equal size n. People in group 1 train as usual while people in group 2 use the new method. After a certain time the N people

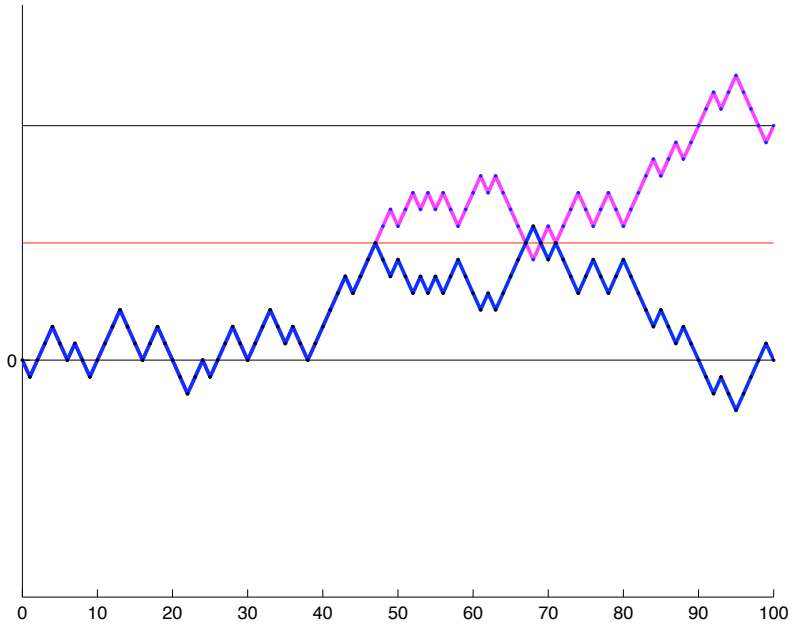

Figure 6.7: The mirror principle.

compete in a swimming race and their order of reaching the finish line is noted. With the set $M \subset \{1, 2, \ldots, N\}$ of ranks from people in group 2 we compute the test statistic

$$T_{\mathrm{Sm}} := \max_{\ell=1,2,\ldots,N} H_\ell \quad \text{with} \quad H_\ell = \frac{2}{n} \#(M \cap \{1, \ldots, \ell\}) - \frac{1}{n}.$$

Under the null hypothesis that the two methods of training are equivalent, $\mathbb{P}(T_{\mathrm{Sm}} \geq c) = G^{\mathrm{Sm}}_{n,n}(c)$. Under the working hypothesis that people from group 2 tend to be faster than people from group 1, M tends to contain smaller numbers. Then T_{Sm} tends to take larger values, so $G^{\mathrm{Sm}}_{n,n}(T)$ is a suitable p-value for the null hypothesis.

6.5 Rank Sum Tests

Starting from $K = 2$ samples $\boldsymbol{X}_1, \boldsymbol{X}_2$ and the corresponding distribution functions we want to quantify to what extent $\boldsymbol{X}_1$ contains larger values than $\boldsymbol{X}_2$.

To motivate this goal let us think again about two groups of sports people who may gain certain scores X_{ki} in a certain discipline (e.g. decathlon). The question is how to assess which team performs better. There is no canonical answer. One could compute Smirnov's test statistic, but this is difficult to explain to non-statisticians. Another obvious measure would be the difference $\overline{X}_1 - \overline{X}_2$ of sample means. But this quantity is sensitive to outliers. For instance, it could happen that all members of group 1 reach higher scores than all but one member of group 2. But if the best result in group 2 is very large, the difference $\overline{X}_1 - \overline{X}_2$ could be negative, see the left part of Figure 6.8. To avoid this problem one could take the difference of sample medians. But this quantity could be too non-sensitive. For instance, it could happen that the sample medians are identical although many scores in group 1 are strictly larger than a majority of scores in group 2 and many scores in group 2 are strictly smaller than a majority of scores in group 1, see the right part of Figure 6.8.

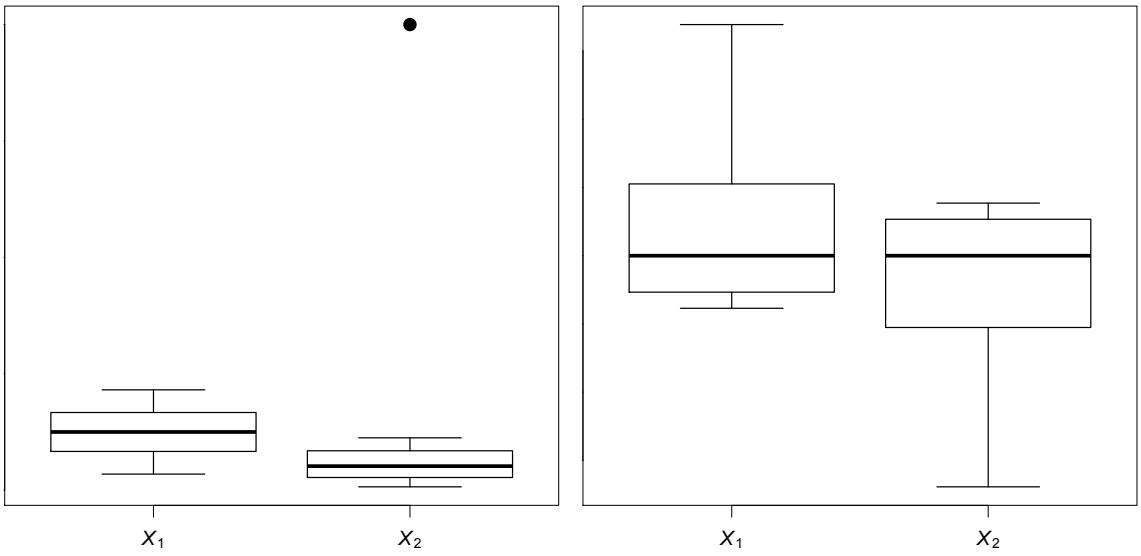

Figure 6.8: Problems with "fair comparisons" of two samples.

The Mann–Whitney U-statistic. To quantify the tendency of $\boldsymbol{X}_1$ having larger components than $\boldsymbol{X}_2$ one could compare each component of $\boldsymbol{X}_1$ with each component of $\boldsymbol{X}_2$ and count how many times the former value is larger than the latter. This leads to the test statistic introduced by H.B. Mann and D. Whitney (1947):

$$T_U \;:=\; \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} h(X_{1i}, X_{2j})$$

with

$$h(x,y) \;:=\; \begin{cases} 1 & \text{if } x > y, \\ 1/2 & \text{if } x = y, \\ 0 & \text{if } x < y. \end{cases}$$

In other words, the statistic

$$\widehat{u} \;:=\; \frac{T_U}{n_1 n_2}$$

specifies the probability that a randomly chosen component of $\boldsymbol{X}_1$ is larger than a randomly chosen component of $\boldsymbol{X}_2$. One may view $\widehat{u}$ as an estimator of the theoretical quantity

$$\mathbb{E}(\widehat{u}) \;=\; u(F_1, F_2) := \mathbb{E}\, h(X_1, X_2)$$

with stochastically independent random variables $X_1 \sim F_1$ and $X_2 \sim F_2$. Exercise 6.12 establishes a connection between Smirnov's test statistic T_{Sm} and the rescaled Mann–Whitney statistic $\widehat{u}$.

Wilcoxon's rank sum statistic. The test statistics T_U and $\widehat{u}$ have a clear interpretation and are convenient to analyze theoretically. For explicit calculations the following rank sum statistic T_W, proposed by F. Wilcoxon (1945), has its merits: On combines $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ to a pooled sample

$$\boldsymbol{X} \;=\; (X_{11}, X_{12}, \ldots, X_{1n_1},\, X_{21}, X_{22}, \ldots, X_{2n_2}).$$

with $N = n_1 + n_2$ components. For this one determines the ranks $R_1, R_2, \ldots, R_N$ and defines

$$T_W \;:=\; \sum_{i=1}^{n_1} R_i,$$

that is the sum of ranks of X_{1i}, $1 \le i \le n_1$, within the pooled sample. Computing this test statistic is rather easy, because the ranks may be determined in $O(N \log N)$ steps via a suitable sorting routine.

The connection between T_W and T_U. We have the simple equation

$$T_W \;=\; \frac{n_1(n_1+1)}{2} + T_U.$$

For

$$R_i \;=\; \#\{j \le N : X_j < X_i\} + \frac{1}{2} + \frac{\#\{j \le N : X_j = X_i\}}{2} \;=\; \frac{1}{2} + \sum_{j=1}^{N} h(X_i, X_j),$$

so

$$\begin{aligned} T_W \;&=\; \sum_{i=1}^{n_1}\Big(\frac{1}{2} + \sum_{j=1}^{N} h(X_i, X_j)\Big) \\ &=\; \frac{n_1}{2} + \sum_{i=1}^{n_1}\sum_{j=1}^{n_1} h(X_i, X_j) + \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} h(X_{1i}, X_{2j}) \\ &=\; \frac{n_1}{2} + \sum_{i=1}^{n_1}\sum_{j=1}^{n_1} h(X_i, X_j) + T_U. \end{aligned}$$

But

$$\begin{aligned} \sum_{i=1}^{n_1}\sum_{j=1}^{n_1} h(X_i, X_j) \;&=\; \sum_{i=1}^{n_1} \underbrace{h(X_i, X_i)}_{=1/2} + \sum_{1 \le i < j \le n_1} \underbrace{\big(h(X_i, X_j) + h(X_j, X_i)\big)}_{=1} \\ &=\; \frac{n_1}{2} + \frac{n_1(n_1-1)}{2} \;=\; \frac{n_1^2}{2}, \end{aligned}$$

whence $T_W = n_1/2 + n_1^2/2 + T_U = n_1(n_1+1)/2 + T_U$ indeed.

Wilcoxon's rank sum test. With the test statistic T_W or T_U one could perform the permutation tests introduced in a later chapter, and this would lead to exact p-values for the null hypothesis

$$H_o \,:\, F_1 \equiv F_2.$$

Now we make the simplifying assumption that the distribution functions F_1, F_2 are continuous. Under the null hypothesis H_o, $(R_1, R_2, \ldots, R_N)$ is uniformly distributed on the set of all permutations of $(1, 2, \ldots, N)$. Hence for arbitrary numbers x,

$$\mathbb{P}(T_U \le x) \;=\; G_{n_1,n_2}(x),$$

where

$$\begin{aligned} G_{n_1,n_2}(x) \;&:=\; \mathbb{P}\Big(\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} 1_{[\Pi(i) > \Pi(n_1+j)]} \le x\Big) \\ &=\; \mathbb{P}\Big(\sum_{i=1}^{n_1} \Pi(i) \le \frac{n_1(n_1+1)}{2} + x\Big) \end{aligned}$$

with a random permutation Π of $\{1, 2, \ldots, N\}$. This distribution function has an interesting symmetry property: If we define $\tilde{\Pi}$ via $\tilde{\Pi}(i) := N + 1 - \Pi(i)$, then $\tilde{\Pi}$ is also a random permutation of $\{1, 2, \ldots, N\}$. Replacing Π with $\tilde{\Pi}$ in the definition of $G_{n_1,n_2}(\cdot)$ and noting that $1_{[\tilde{\Pi}(i)>\tilde{\Pi}(n_1+j)]} = 1 - 1_{[\Pi(i)>\Pi(n_1+j)]}$ shows that

$$1 - G_{n_1,n_2}(x-) \;=\; G_{n_1,n_2}(n_1 n_2 - x)$$

for arbitrary x.

Depending on the user's working hypothesis, one of the following p-values are useful: For the working hypothesis that $F_1 <_{\text{st.}} F_2$ the left-sided p-value

$$\pi_\ell(\boldsymbol{X}_1, \boldsymbol{X}_2) \;:=\; G_{n_1,n_2}(T_U),$$

and for the working hypothesis that $F_1 >_{\text{st.}} F_2$ the right-sided p-value

$$\pi_r(\boldsymbol{X}_1, \boldsymbol{X}_2) \;:=\; 1 - G_{n_1,n_2}(T_U -) \;=\; G_{n_1,n_2}(n_1 n_2 - T_U).$$

If one only wants to verify that $F_1 \not\equiv F_2$, without any a priori guess of the direction, the two-sided p-value would be appropriate, i.e. the minimum of the two one-sided p-values times two. Indeed one may view the one-sided p-values as p-values for more general null hypotheses. One can show (Exercise 6.8) that for arbitrary numbers $\alpha \in (0, 1)$,

$$\mathbb{P}\big(\pi_\ell(\boldsymbol{X}_1, \boldsymbol{X}_2) \le \alpha\big) \;\le\; \alpha \quad \text{if } F_1 \ge_{\text{st.}} F_2, \tag{6.2}$$

$$\mathbb{P}\big(\pi_r(\boldsymbol{X}_1, \boldsymbol{X}_2) \le \alpha\big) \;\le\; \alpha \quad \text{if } F_1 \le_{\text{st.}} F_2. \tag{6.3}$$

Thus $\pi_\ell(\boldsymbol{X}_1, \boldsymbol{X}_2)$ is a p-value for the null hypothesis that $F_1 \ge_{\text{st.}} F_2$, and $\pi_r(\boldsymbol{X}_1, \boldsymbol{X}_2)$ is a p-value for the null hypothesis that $F_1 \le_{\text{st.}} F_2$.

If the distribution functions F_1 and F_2 are not necessarily continuous, one should perform a permutation test as described later or redefine the p-values as $\pi_\ell(\boldsymbol{X}_1, \boldsymbol{X}_2) = G_{n_1,n_2}(\lceil T_U \rceil)$ and $\pi_r(\boldsymbol{X}_1, \boldsymbol{X}_2) = G_{n_1,n_2}(n_1 n_2 - \lfloor T_U \rfloor)$. The latter p-values tend to be conservative (i.e. larger) than p-values obtained with permutations tests, so one is on the safe side.

Computation of the distribution function G_{n_1,n_2}. If the sample sizes n_1 and n_2 are large, one may use the fact that G_{n_1,n_2} describes approximately a Gaussian distribution with mean $n_1 n_2/2$ and variance $n_1 n_2 (N+1)/12$, see also Exercise 6.10 and Section A.9 in the appendix. This yields the approximation

$$G_{n_1,n_2}(x) \;\approx\; \Phi\Big(\frac{x - n_1 n_2/2 + 0.5}{\sqrt{n_1 n_2 (N+1)/12}}\Big) \quad \text{for } x \in \mathbb{Z}.$$

The extra summand '$+0.5$' in the enumerator is a correction taking into account that G_{n_1,n_2} is a distribution on the integers; this enhances the quality of the approximation considerably.

For the exact computation of G_{n_1,n_2} one may use the following recursion formula: For arbitrary integers x,

$$G_{n_1,n_2}(x) \;=\; \frac{n_1}{N} G_{n_1-1,n_2}(x - n_2) + \frac{n_2}{N} G_{n_1,n_2-1}(x).$$

To verify this, we construct the random tuple $\big(\Pi(1), \Pi(2), \ldots, \Pi(n)\big)$ in two steps. In the first step we choose a random position J at which the number N should be. Then we fill the remaining $N-1$ positions from left to right with the components of a random permutation $\tilde{\Pi}$ of $\{1, \ldots, N-1\}$, independently from J. In case of $J \le n_1$, which happens with probability n_1/N,

$$\sum_{i=1}^{n_1} \sum_{j=1}^{n_2} 1_{[\Pi(i)>\Pi(n_1+j)]} \;=\; n_2 + \sum_{i=1}^{n_1-1} \sum_{j=1}^{n_2} 1_{[\tilde{\Pi}(i)>\tilde{\Pi}(n_1-1+j)]},$$

and the double sum on the right hand side has distribution function G_{n_1-1,n_2}. In case of $J > n_1$, which happens with probability n_2/N,

$$\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} 1_{[\Pi(i)>\Pi(n_1+j)]} \;=\; \sum_{i=1}^{n_1}\sum_{j=1}^{n_2-1} 1_{[\tilde\Pi(i)>\tilde\Pi(n_1+j)]},$$

und the double sum on the right hand side follows G_{n_1,n_2-1}.

Example 6.11 (Hamburg Marathon, continuation of Example 6.5). We want to test whether the performance of younger male participants (age classes MJ and MH) is significantly different from the performance of mature male participants (age classes M40 and M45), where we aim for a confidence of 99% which corresponds to the test level $\alpha = 1\%$. The data set contains the results of $n_1 = 1551$ runners in age classes MJ and MH as well as $n_2 = 3399$ runners in age classes M40 and M45.

The data analysis involves $n_1 n_2 = 5271849$ pairwise comparisons, yielding $T_U = 2786811$ and $\widehat{u} = T_U/(n_1 n_2) \approx 0.5286$. For sample sizes n_k at least 50, the approximations by means of Gaussian distributions are excellent. Here this leads to the one-sided p-values

$$\begin{aligned}
\pi_\ell \;&=\; \Phi\Big(\frac{T_U - n_1 n_2/2 + 0.5}{\sqrt{n_1 n_2 (N+1)/12}}\Big) \;\approx\; \Phi(3.2353) \;\approx\; 0.9994,\\
\pi_r \;&=\; \Phi\Big(\frac{n_1 n_2/2 - T_U + 0.5}{\sqrt{n_1 n_1 (N+1)/12}}\Big) \;\approx\; \Phi(-3.2353) \;\approx\; 0.0006,
\end{aligned}$$

whence $\pi_z = 2 \cdot \pi_r \approx 0.0012$. Hence we may conclude with confidence 99% that the two age groups perform differently. The analysis indicates that the mature runners tend to be faster than the younger runners.

Remark 6.12. Wilcoxon's rank sum test (i.e. the Mann–Whitney test) is also applicable if the observations X_{ki} are values of an ordinal variable. But in such situations one tends to have numerous identical sample values, and the test should be executed as a permutation test as described later.

Confidence bounds for a shift parameter. Wilcoxon's rank sum test may be inverted, too, to obtain confidence regions for an unknown shift parameter. Suppose that $F_2 \equiv F$ and $F_1(x) = F(x-\mu)$ for $x \in \mathbb{R}$ with an unknown continuous distribution function F and an unknown real shift parameter μ.

A special case of this model are Gaussian observations X_{ki} with unknown standard deviation $\sigma > 0$ and unknown means $\nu + \mu$ for $k = 1$ and ν for $k = 2$.

We define

$$T_U(m) \;:=\; \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} h(X_{1i} - m, X_{2j}) \;=\; \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} h(X_{1i} - X_{2j}, m)$$

for arbitrary $m \in \mathbb{R}$, so $T_U(0) = T_U$. Then $T_U(\mu)$ follows the distribution function G_{n_1,n_2}. Moreover, $m \mapsto T_U(m)$ is non-increasing in $m \in \mathbb{R}$. Hence the inequality

$$\mathbb{P}\big(T_U(\mu) \le G^{-1}_{n_1,n_2}(1-\alpha)\big) \;\ge\; 1-\alpha$$

yields the lower $(1-\alpha)$-confidence bound

$$a_\alpha = a_\alpha(\text{data}) \;:=\; \inf\big\{m \in \mathbb{R} : T_U(m) \le G^{-1}_{n_1,n_2}(1-\alpha)\big\}.$$

Analogously, since $n_1n_2 - T_U(\mu) \sim G_{n_1,n_2}$, too, we obtain the upper $(1-\alpha)$-confidence bound

$$b_\alpha = b_\alpha(\text{data}) \;:=\; \sup\bigl\{m \in \mathbb{R} : T_U(m) \geq n_1n_2 - G^{-1}_{n_1,n_2}(1-\alpha)\bigr\}.$$

If

$$M_1 \leq M_2 \leq \cdots \leq M_{n_1n_2}$$

denote the sorted differences $X_{1i} - X_{2j}$, where $1 \leq i \leq n_1, 1 \leq j \leq n_2$, augmented by $M_0 := -\infty$ and $M_{n_1n_2+1} := \infty$, then $T_U(m) = n_1n_2 - s$ for $M_s < m < M_{s+1}$ and $0 \leq s \leq n_1n_2$. This implies that

$$a_\alpha \;=\; M_{k(\alpha)} \quad \text{and} \quad b_\alpha \;=\; M_{\ell(\alpha)},$$

where

$$\begin{aligned} k(\alpha) = k(\alpha, n_1, n_2) \;&:=\; n_1n_2 - G^{-1}_{n_1,n_2}(1-\alpha), \\ \ell(\alpha) = \ell(\alpha, n_1, n_2) \;&:=\; G^{-1}_{n_1,n_2}(1-\alpha) + 1 \;=\; n_1n_2 + 1 - k(\alpha). \end{aligned}$$

A corresponding estimator for μ is given by the median $\widehat{\mu}$ of these values $M_1, M_2, \ldots, M_{n_1n_2}$. It satisfies the equality $T_U(\widehat{\mu}) = n_1n_2/2$.

6.6 Multiple Tests and Comparisons of Several Samples

Multiple Tests. In some statistical analyzes one is testing several null hypotheses H_1, H_2, …, H_m simultaneously. For $j = 1, 2, \ldots, m$ let $\pi_j = \pi_j(\text{data})$ be a p-value for H_j. Suppose we test all null hypotheses at level α. If we generate a list of all null hypotheses whose p-values are smaller than or equal to α, then the probability that at least one of them is true exceeds α in general.

Suppose one wants to ensure that the probability of committing some error of the first kind is at most α. For this purpose one should *adjust* the single p-values suitably. Our goal is to construct adjusted p-values $\bar{\pi}_j = \bar{\pi}_j(\text{data})$ such that with the unknown set

$$\mathcal{J}_o \;:=\; \bigl\{j \in \{1, 2, \ldots, m\} : H_j \text{ is true}\bigr\}$$

we can guarantee that

$$\mathbb{P}\bigl(\bar{\pi}_j \leq \alpha \text{ for at least one } j \in \mathcal{J}_o\bigr) \;\leq\; \alpha. \tag{6.4}$$

Example 6.13. Suppose we want to compare $K \geq 2$ samples $\boldsymbol{X}_1$, $\boldsymbol{X}_2$, …, $\boldsymbol{X}_K$. Precisely, for $k, \ell \in \{1, 2, \ldots, K\}$ with $k \neq \ell$ we want to test the null hypothesis

$$H_{k,\ell} \;:\; F_k \leq_{\text{st.}} F_\ell.$$

For this purpose we compute the right-sided p-value $\pi_{k,\ell} = \pi_{k,\ell}(\boldsymbol{X}_k, \boldsymbol{X}_\ell)$ for $H_{k,\ell}$ with Smirnov's test or Wilcoxon's rank sum test. Then we replace these $m = K(K-1)$ single p-values $\pi_{k,\ell}$ with adjusted p-values $\bar{\pi}_{k,\ell}$ satisfying (6.4). Then we may claim with confidence $1-\alpha$ that *all* null hypotheses $H_{k,\ell}$ with $\bar{\pi}_{k,\ell} \leq \alpha$ are false. In view of our choice of test statistics, $\bar{\pi}_{k,\ell} \leq \alpha$ indicates that $F_k >_{\text{st.}} F_\ell$, but this is a claim one can never prove with nontrivial confidence.

The Bonferroni adjustment. To achieve (6.4) one could replace the single p-values with

$$\bar{\pi}_j \;:=\; m\pi_j \quad \text{or} \quad \bar{\pi}_j \;:=\; \min(m\pi_j, 1).$$

In both cases

$$\begin{aligned}&\mathbb{P}\big(\bar{\pi}_j \le \alpha \text{ for at least one } j \in \mathcal{J}_o\big)\\ &\quad\le \sum_{j\in\mathcal{J}_o} \mathbb{P}\left(\bar{\pi}_j \le \alpha\right) \;=\; \sum_{j\in\mathcal{J}_o} \mathbb{P}\left(\pi_j \le \alpha/m\right) \;\le\; \#\mathcal{J}_o \cdot \alpha/m \;\le\; \alpha.\end{aligned}$$

Truncating the adjusted p-values at 1 is only for cosmetic reasons.

Holm's adjustment. The Bonferroni adjustment tends to be very conservative in the sense that too many errors of the second kind are committed and our list of false null hypotheses turns out too short. A refined adjustment method has been developed by S. Holm (1979): At first the null hypotheses are rearranged such that the corresponding p-values are non-decreasing. Let $H_{(1)}, H_{(2)}, \ldots, H_{(m)}$ be the rearranged null hypotheses with corresponding p-values $\pi_{(1)} \le \pi_{(2)} \le \cdots \le \pi_{(m)}$. Then one replaces $\pi_{(j)}$ with

$$\bar{\pi}_{(j)} \;:=\; \max_{i\le j} \min\big((m+1-i)\pi_{(i)}, 1\big).$$

Obviously $\bar{\pi}_{(j)} \le \max_{i\le j} \min(m\pi_{(i)}, 1) = \min(m\pi_{(j)}, 1)$ with equality for $j = 1$. Thus the Bonferroni adjustment is more conservative than Holm's adjustment. If $m\pi_{(1)}$ exceeds α, both methods reject none of the null hypotheses.

Proof of (6.4) **for Holm's method.** Let $m_o = \#\mathcal{J}_o$, the number of true hypotheses be strictly positive. After rearrangement let $H_{(J(1))}, \ldots, H_{(J(m_o))}$ be the true null hypotheses with random indices $J(1) < \cdots < J(m_o)$. Then the probability that $\bar{\pi}_j \le \alpha$ for at least one $j \in \mathcal{J}_o$ equals

$$\begin{aligned}&\mathbb{P}\big(\bar{\pi}_{(J(a))} \le \alpha \text{ for at least one } a \in \{1, \ldots, m_o\}\big)\\ &\quad= \mathbb{P}\big(\bar{\pi}_{(J(1))} \le \alpha\big)\\ &\quad\le \mathbb{P}\big((m+1-J(1))\,\pi_{(J(1))} \le \alpha\big)\\ &\quad\le \mathbb{P}\big(m_o\pi_{(J(1))} \le \alpha\big)\\ &\quad= \mathbb{P}\big(\pi_j \le \alpha/m_o \text{ for at least one } j \in \mathcal{J}_o\big)\\ &\quad\le \alpha.\end{aligned}$$

Here in the first step we used the inequalities $\bar{\pi}_{(1)} \le \bar{\pi}_{(2)} \le \cdots \le \bar{\pi}_{(m)}$, in the second step the inequality $\bar{\pi}_{(j)} \ge (m+1-j)\pi_{(j)}$, and in the third step the fact that $J(1) \le m + 1 - m_o$. □

Example 6.14 (Hamburg Marathon, continuation of Example 6.5). Now we want to compare more than two age groups. To limit the total number of comparisons we combine the original twelve age classes to broader classes:

M18-29	:	$n_1 = 1551$	(MJ, MH)
M30-39	:	$n_2 = 4289$	(M30, M35)
M40-49	:	$n_3 = 3399$	(M40, M45)
M50-59	:	$n_4 = 1502$	(M50, M55)
M60+	:	$n_5 = 460$	(M60, M65, M70, M75)

Now we analyze these data as in example Example 6.13. In the following table one sees for each pair (k, ℓ) of two different indices $k, \ell \in \{1, 2, 3, 4, 5\}$ the normed Mann–Whitney statistic $\widehat{u}_{k,\ell} = T_U(\boldsymbol{X}_k, \boldsymbol{X}_\ell)/(n_k n_\ell)$ as well as the right-sided approximative p-value

$$\pi_{k,\ell} \;=\; \Phi\Big(\frac{n_k n_\ell/2 - T_U(\boldsymbol{X}_k, \boldsymbol{X}_\ell) + 0.5}{\sqrt{n_k n_\ell (n_k + n_\ell + 1)/12}}\Big),$$

all values rounded to five digits. Entries with p-value no larger than 1% are highlighted:

M18-29	**0.52819** **0.00049**	**0.52862** **0.00061**	0.47794 0.98258	0.37205 1
0.47181 0.99951	*M30-39*	0.49975 0.51532	0.44732 1	0.34349 1
0.47138 0.99939	0.50025 0.48468	*M40-49*	0.44630 1	0.34104 1
0.52206 0.01742	**0.55268** **0.00000**	**0.55370** **0.00000**	*M50-59*	0.38882 1
0.62795 **0.00000**	**0.65651** **0.00000**	**0.65896** **0.00000**	**0.61118** **0.00000**	*M60+*

To adjust these $K(K-1) = 20$ p-values we have to sort them. This leads to the following numbers (again rounded to five digits):

j	≤ 6	7	8	9	10	≥ 11
$\pi_{(j)}$	0.00000	0.00049	0.00061	0.01742	0.48468	> 0.5
$\bar{\pi}_{(j)}$ via Bonf.	0.00000	0.00982	0.01215	0.34842	1	1
$\bar{\pi}_{(j)}$ via Holm	0.00000	0.00688	0.00790	0.20905	1	1

Replacing the original p-values with the Holm-adjusted ones leads to the following table:

M18-29	**0.52819** **0.00688**	**0.52862** **0.00790**	0.47794 1	0.37205 1
0.47181 1	*M30-39*	0.49975 1	0.44732 1	0.34349 1
0.47138 1	0.50025 1	*M40-49*	0.44630 1	0.34104 1
0.52206 0.20905	**0.55268** **0.00000**	**0.55370** **0.00000**	*M50-59*	0.38882 1
0.62795 **0.00000**	**0.65651** **0.00000**	**0.65896** **0.00000**	**0.61118** **0.00000**	*M60+*

Now one may claim with confidence 99% that $H_{k,\ell}$ is false for the following combinations (k, ℓ) at least: $k = 1$ and $\ell = 2, 3$; $k = 4$ and $\ell = 2, 3$; $k = 5$ and $\ell = 1, 2, 3, 4$. This indicates that the running times in the highest age group tend to be higher than in the four others, that the running times in age group M50-59 are higher than in groups M30-39 and M40-49, and that the running times in age group M18-29 are higher than in groups M30-39 and M40-49.

At this point one should say something about the modelling of the data and interpretation of tests: Each marathon race has its own special features. The participants are influenced by various factors such as weather and ground conditions, inclinations, spirits of audience and many more, some of which are random. Insofar the assumption of fixed distribution functions $F_1, F_2, \ldots, F_K$ is certainly unrealistic. Alternatively one could postulate the existence of *random* distribution functions $F_1, F_2, \ldots, F_K$ and assume that the observed running times are independent with $X_{ki} \sim F_k$ *conditional* on the functions F_k. The null hypothesis $H_{k,\ell}$ could be modified to read $F_k \leq_{\mathrm{st.}} F_\ell$ *almost surely*.

Remark 6.15. In various text books, different procedures are proposed for the comparison of $K \geq 3$ samples. When comparing means, these are so-called F-tests and Analyses of Variance. An analogue of Wilcoxon's rank sum test is the Kruskal–Wallis test. A disadvantage of these procedures is that they can only reject the null hypothesis that all means or distribution functions are identical. But they don't provide further information as to which samples differ significantly in what way from others. We encountered a similar problem in connection with the chi-squared goodness-of-fit test.

6.7 Exercises

Exercise 6.1 (Bounds for the mean). Suppose that you know only the box-plot of the observation vector $\boldsymbol{X} = (X_i)_{i=1}^n$, i.e. its minimum Q_0, its three sample quartiles Q_1, Q_2, Q_3 and its maximum Q_4. Even the sample size n is unknown. Show that

$$\frac{Q_0 + Q_1 + Q_2 + Q_3}{4} \;\leq\; \bar{X} \;\leq\; \frac{Q_1 + Q_2 + Q_3 + Q_4}{4}.$$

Hint 1: The quantities $Q_0, \ldots, Q_4$ remain the same if each component of $\boldsymbol{X}$ is replaced with k copies, leading to a sample of size kn. Thus you may assume without loss of generality that the sample size n is an arbitrarily large multiple of 4.

Hint 2: One can solve this exercise via the formula

$$\overline{X} \;=\; X_{(1)} + \int_{X_{(1)}}^{X_{(n)}} (1 - \widehat{F}(x))\, dx.$$

But then one should verify the latter.

Exercise 6.2 (Chi-squared distributions). Let S_k^2 have distribution χ_k^2. Determine the mean and standard deviation of S_k^2 (Hint: Exercise 2.5).

Show by virtue of the Central Limit Theorem that the standardised random variable

$$(S_k^2 - \mathbb{E}(S_k^2))/\operatorname{Std}(S_k^2)$$

is asymptotically standard Gaussian as $k \to \infty$.

Exercise 6.3 (Welch's method). Show that the number

$$k(n_1, n_2, \sigma_1, \sigma_2) \;:=\; \frac{(\sigma_1^2/n_1 + \sigma_2^2/n_2)^2}{\sigma_1^4/(n_1^2(n_1-1)) + \sigma_2^4/(n_2^2(n_2-1))}$$

is always contained in the interval $\big[\min(n_1 - 1, n_2 - 1), N - 2\big]$.

Verify that $\widehat{\tau}^{-1}(\overline{X}_1 - \overline{X}_2 - \mu_1 + \mu_2)$ has asymptotic distribution t_{n_1-1} and that $k(n_1, n_2, \sigma_1, \sigma_2) \to n_1 - 1$ as $\sigma_2/\sigma_1 \to 0$.

Exercise 6.4 (Example for Welch's method). Consider once more Exercise 3.12. Compute with that data an approximate 95%-confidence interval for the difference $\mu_1 - \mu_2$, where μ_1 and μ_2 denote the mean BMI of women and men, respectively, in the whole population.

Exercise 6.5 (Combining several estimators). One specific goal of this exercise is to show that the special variance estimator $\widehat{\sigma}^2$ in (6.1) has a certain optimality property.

(a) Let $Y_1, Y_2, \ldots, Y_K$ be stochastically independent random variables with unknown mean $\nu = \mathbb{E}(Y_k)$. Further let $\operatorname{Var}(Y_k) = c_k \tau^2$ with unknown $\tau > 0$ but given factors $c_1, c_2, \ldots, c_K > 0$. Now we consider estimators for ν of type

$$\widehat{\nu} \;:=\; \sum_{k=1}^{K} w_k Y_k$$

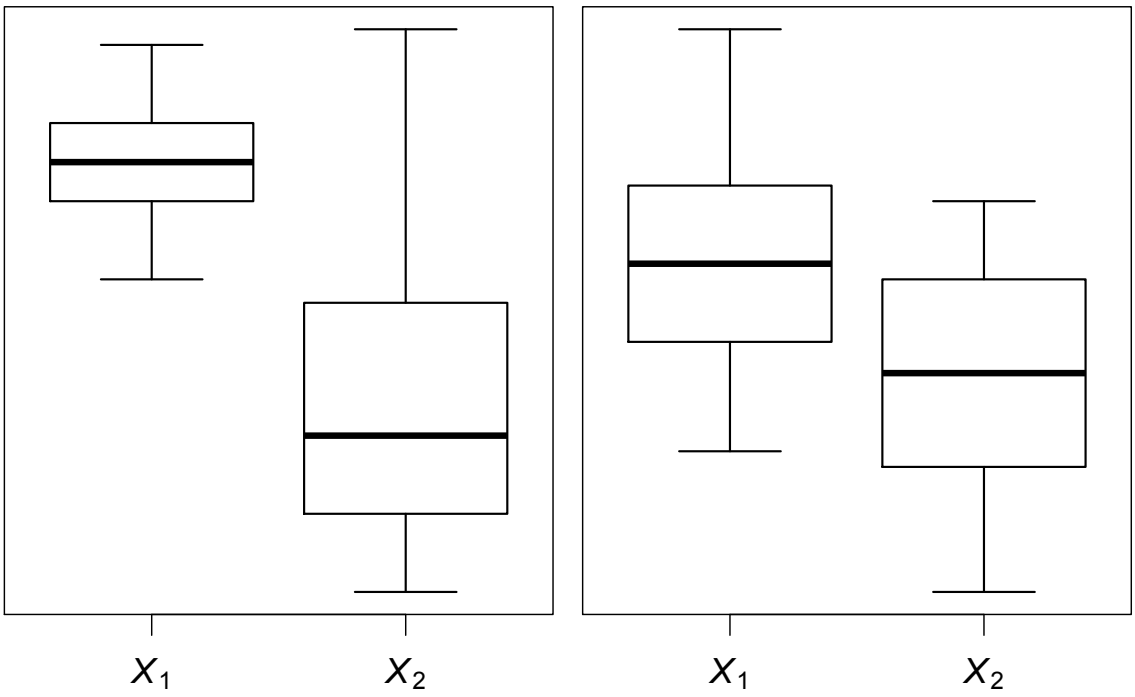

Figure 6.9: Box plots for Exercise 6.7.

with certain weights $w_1, w_2, \ldots, w_K$. Determine weights w_k such that $\mathbb{E}(\widehat{\nu}) = \nu$ and $\mathrm{Var}(\widehat{\nu})$ is minimal. (If you cannot handle the general case, try the special case $K = 2$.)

(b) Now we consider stochastically independent random variables X_{ki}, $1 \le k \le K$, $1 \le i \le n_k$. Suppose that $X_{ki} \sim \mathcal{N}(\mu_k, \sigma^2)$ with unknown parameters $\mu_1, \mu_2, \ldots, \mu_K \in \mathbb{R}$ and $\sigma^2 > 0$. Consider the sample variances $S_k^2 := (n_k - 1)^{-1} \sum_{i=1}^{n_k} (X_{ki} - \overline{X}_k)^2$. Combine these sample variances by means of part (a) to a good unbiased estimator for σ^2. What can you say about its distribution?

Exercise 6.6. Prove Lemma 6.7. The implication "(i) $\Longrightarrow$ (ii)" is a consequence of the definition of quantile functions. For the implication "(ii) $\Longrightarrow$ (iii)" one can use the quantile transformation, see Lemma 3.11.

Exercise 6.7 (Mann–Whitney U-statistic and box plots). Figure 6.9 shows two examples of box plots of two samples $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ with unknown sample sizes. Determine lower and upper bounds for the normalised Mann–Whitney U-statistic $\widehat{u}$.

Exercise 6.8. Verify inequalities (6.2) and (6.3) with a suitable coupling, similarly as in the proof of Lemma 6.9.

Exercise 6.9 (Linear permutation statistics, I). For an integer $N \ge 2$ let Π be uniformly distributed on the set of all permutations of $\{1, 2, \ldots, N\}$. For fixed vectors $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^N$ we consider the random sum $T := \sum_{i=1}^N a_i b_{\Pi(i)}$.

(i) Show that

$$\mathbb{E}(T) \;=\; N \overline{a}\,\overline{b}$$

with $\overline{v} := N^{-1} \sum_{i=1}^N v_i$ for $\boldsymbol{v} = \boldsymbol{a}, \boldsymbol{b}$.

(ii) Show that

$$\mathrm{Var}(T) \;=\; \frac{(\|\boldsymbol{a}\|^2 - N\overline{a}^2)(\|\boldsymbol{b}\|^2 - N\overline{b}^2)}{N-1}.$$

Hint: Consider first the special case $\overline{a} = \overline{b} = 0$.

(iii) Show that $T - \mathbb{E}(T)$ and $\mathbb{E}(T) - T$ have the same distribution whenever $b_i + b_{N+1-i}$ is constant in $i \in \{1, 2, \ldots, N\}$. Hint: If you are helpless, look at Exercise 8.1.

Exercise 6.10. Apply the results of Exercise 6.9 to Wilcoxon's rank sum test: Show that

$$\mathbb{E}(T_W) \;=\; \frac{n_1(N+1)}{2}, \quad \mathbb{E}(T_U) \;=\; \frac{n_1 n_2}{2}$$

and

$$\mathrm{Var}(T_W) = \mathrm{Var}(T_U) \;=\; \frac{n_1 n_2 (N+1)}{12},$$

if the underlying distribution functions F_1 and F_2 are identical and continuous.

Exercise 6.11. In a medical study a physiological parameter (urinary thromboglobulin excretion) has been measured for 12 diabetes patients and 12 healthy people. The question is whether there is a systematic difference between diabetes patients and healthy people with respect to this parameter.

(a) Compute the test statistics T_W and T_U for the explicit data:

diabetes	11.5	12.1	16.1	17.8	24.0	28.8
	33.9	40.7	51.3	56.2	61.7	69.2
no diabetes	4.1	6.3	7.8	8.5	8.9	10.4
	11.5	12.0	13.8	17.6	24.3	37.2

(b) Compute an approximate two-sided p-value for this data example.

Exercise 6.12 (Connection between Smirnov's and the Mann–Whitney statistic). Let $\boldsymbol{X}_1 = (X_{1i})_{i=1}^{n_1}$ and $\boldsymbol{X}_2 = (X_{2j})_{j=1}^{n_2}$ be data vectors such that the pooled sample

$$\boldsymbol{X} = (X_{11}, \ldots, X_{1n_1}, X_{21}, \ldots, X_{2n_2})$$

consists of N different numbers. With the empirical distribution functions $\widehat{F}_1$ and $\widehat{F}_2$ of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$, Smirnov's test statistic equals

$$\max_{\ell=1,2,\ldots,N} \big(\widehat{F}_2(X_\ell) - \widehat{F}_1(X_\ell)\big).$$

Show that

$$\frac{1}{N} \sum_{\ell=1}^{N} \big(\widehat{F}_2(X_\ell) - \widehat{F}_1(X_\ell)\big) \;=\; \frac{T_U}{n_1 n_2} - \frac{1}{2}.$$

Exercise 6.13 (Two-sided binomial tests as multiple tests). Let $H \sim \mathrm{Bin}(n, p)$ with given parameter $n \in \mathbb{N}$ and unknown parameter $p \in [0, 1]$. For a given $p_o \in (0, 1)$ one could test the null hypothesis that $p = p_o$ with the two-sided p-value

$$\pi_z(H) \;:=\; 2 \cdot \min\big\{F_{n,p_o}(H), 1 - F_{n,p_o}(H-1)\big\},$$

where F_{n,p_o} stands for the distribution function of $\mathrm{Bin}(n, p_o)$. Refine and re-interpret this procedure as a multiple test: Show that $\pi_\ell(H) := F_{n,p_o}(H)$ is a p-value for the null hypothesis $H_1 : p \geq p_o$ and that $\pi_r(H) := 1 - F_{n,p_o}(H-1)$ is a p-value for the null hypothesis $H_2 : p \geq p_o$. What conclusion may be drawn in case of $\pi_z(H) \leq \alpha$ with a confidence of $1 - \alpha$?

Chapter 7

Odds Ratios and Two-by-Two Tables

This chapter is about special but very important topics. These include the comparison of two probability parameters or potential dependencies between two dichotomous variables. In both cases so-called odds ratios play an important role, and the data analysis involves two-by-two tables.

7.1 Comparing Two Binomial Parameters

For $k = 1, 2$ let $p_k \in (0, 1)$ be the probability of a certain event A_k, for instance the success of a certain medical treatment. To quantify the difference between p_1 and p_2 one could consider the difference $p_1 - p_2$ or the ratio p_1/p_2. As we shall see later, one should rather consider the *odds* $p_k/(1 - p_k)$ of event A_k and infer something about the *odds ratio*

$$\rho := \frac{p_1}{1-p_1} \Big/ \frac{p_2}{1-p_2} = \frac{p_1(1-p_2)}{(1-p_1)p_2}.$$

Since $(0, 1) \ni p \mapsto p/(1 - p) \in (0, \infty)$ is continuous and strictly increasing,

$$\rho \left\{ \begin{matrix} > \\ = \\ < \end{matrix} \right\} 1 \quad \text{if and only if} \quad p_1 \left\{ \begin{matrix} > \\ = \\ < \end{matrix} \right\} p_2.$$

Suppose that for estimating the two probabilities p_1, p_2 two independent random variables $H_1 \sim \mathrm{Bin}(n_1, p_1)$ and $H_2 \sim \mathrm{Bin}(n_2, p_2)$ are available.

An explicit example is a randomised clinical trial in which $N = n_1 + n_2$ test persons are assigned randomly to two groups: All n_k persons in group k receive treatment k, and with H_k we denote the number of successes in this group. If we view the N test persons as a random sample from a large population, the model above is plausible with p_k denoting the probability of success with treatment k for a randomly chosen person from the population.

As in the first chapter we summarise the data as a two-by-two table:

H_1	$n_1 - H_1$	n_1
H_2	$n_2 - H_2$	n_2
H_+	$N - H_+$	N

The row sums n_1, n_2 are fixed, but the column sums $H_+ = H_1 + H_2$ and $N - H_+$ are random. As we'll see later, the conditional distribution of H_1, given H_+, depends only on N, n_1, H_+ and ρ.

7.2 Correlation of two Binary Variables

Consider a random experiment yielding two binary random variables $X \in \{x_1, x_2\}$ and $Y \in \{y_1, y_2\}$.
As an explicit example consider a population of humans. Now we choose randomly a person from the population and determine two binary features X and Y, for instance the presence or absence of a certain genetic disposition ($X = x_1, x_2$) and the presence or absence of a certain disease ($Y = y_1, y_2$).
The joint distribution of X and Y is given by the four probabilities $p_{11}, p_{12}, p_{21}, p_{22}$ with

$$p_{k\ell} := \mathbb{P}(X = x_k, Y = y_\ell).$$

These may be arranged as a two-by-to table:

p_{11}	p_{12}	p_{1+}
p_{21}	p_{22}	p_{2+}
p_{+1}	p_{+2}	1

with the row sums $p_{k+} = p_{k1} + p_{k2} = \mathbb{P}(X = x_k)$ and the columns sums $p_{+\ell} = p_{1\ell} + p_{2\ell} = \mathbb{P}(Y = y_\ell)$. The corresponding odds ratio is defined as

$$\rho = \frac{p_{11}p_{22}}{p_{12}p_{21}}.$$

Possible interpretations are

$$\rho = \frac{\text{Odds}(X = x_1 \,|\, Y = y_1)}{\text{Odds}(X = x_1 \,|\, Y = y_2)} = \frac{\text{Odds}(Y = y_1 \,|\, X = x_1)}{\text{Odds}(Y = y_1 \,|\, X = x_2)},$$

because $\mathbb{P}(X = x_k \,|\, Y = y_\ell) = p_{k\ell}/p_{+\ell}$ and $\mathbb{P}(Y = y_\ell \,|\, X = x_k) = p_{k\ell}/p_{k+}$. In case of $\rho \neq 1$ we talk about a true association or dependence between X and Y. This is justified by the following lemma (Exercise 7.2).

Lemma 7.1. ***(a)*** *For arbitrary indices $k, \ell \in \{1, 2\}$, the following three statements are equivalent:*

(a.1) $\rho = 1$.
(a.2) $p_{k\ell} = p_{k+}p_{+\ell}$.
(a.3) *X and Y are stochastically independent.*

(b) *For arbitrary indices $k, \ell \in \{1, 2\}$ with $k \neq \ell$, the following three inequalities are equivalent:*

(b.1) $\rho >_{(<)} 1$.
(b.2) $p_{kk} >_{(<)} p_{k+}p_{+k}$.
(b.2) $p_{k\ell} <_{(>)} p_{k+}p_{+\ell}$.

Suppose we observe N independent copies $(X_1, Y_1), (X_2, Y_2), \ldots, (X_N, Y_N)$ of (X, Y). Now we determine the absolute frequencies $H_{k\ell} := \#\{i \leq N : X_i = x_k, Y_i = y_\ell\}$ and arrange them as a two-by-two table:

H_{11}	H_{12}	H_{1+}
H_{21}	H_{22}	H_{2+}
H_{+1}	H_{+2}	N

Here we complemented the table by the row sums $H_{k+} = H_{k1} + H_{k2} = \#\{i \leq N : X_i = x_k\}$ and the column sums $H_{+\ell} = H_{1\ell} + H_{2\ell} = \#\{i \leq N : Y_i = y_\ell\}$. The quadruple $(H_{11}, H_{12}, H_{21}, H_{22})$ follows a multinomial distribution with parameters N and $(p_{11}, p_{12}, p_{21}, p_{22})$.

Consider once more the explicit example of a population with two binary features X and Y, where X refers to a certain genetic disposition and Y to a certain disease. Suppose that in a cross-sectional study these two features are determined for a sample from the population of size N. Under the assumption that the sample has been drawn completely at random we obtain a random quadrupel $(H_{11}, H_{12}, H_{21}, H_{22})$ with the stated distribution.

If in this example the relative frequencies p_{1+} or p_{+1} are very small, other types of study are more appropriate: In a cohort study one recruits n_1 people with $X = x_1$ and n_2 people with $X = x_2$. Then $H_{k1} \sim \mathrm{Bin}(n_k, p_k)$ with $p_k := p_{k1}/p_{k+}$, $H_{k2} = n_2 - H_{k1}$, and the entries H_{11}, H_{21} are stochastically independent. Thus we obtain data as in Section 7.1, and

$$\frac{p_1(1-p_2)}{(1-p_1)p_2} \;=\; \frac{p_{11}p_{22}}{p_{12}p_{21}}. \tag{7.1}$$

The same is true in case of a case-control study: Here one recruits n_1 persons with the disease under consideration ('cases', $Y = y_1$) and n_2 persons without this disease ('controls', $Y = y_2$). Then $H_{1\ell} \sim \mathrm{Bin}(n_\ell, p_\ell)$ with $p_\ell := p_{1\ell}/p_{+\ell}$, $H_{2\ell} = n_\ell - H_{1\ell}$, and the entries H_{11}, H_{12} are stochastically independent. Again, (7.1) holds true.

7.3 Confidence Bounds for Odds Ratios

We consider a two-by-two table

H_{11}	H_{12}	H_{1+}
H_{21}	H_{22}	H_{2+}
H_{+1}	H_{+2}	N

with fixed total sum $N = H_{1+} + H_{2+} = H_{+1} + H_{+2}$. Moreover we assume that we are in one of the following two situations:

Situation 1 (Section 7.1): The row sums H_{1+}, H_{2+} are fixed numbers, and the entries H_{11}, H_{21} are stochastically independent with $H_{k1} \sim \mathrm{Bin}(H_{k+}, p_k)$, where $0 < p_k < 1$. Moreover, $H_{k2} = H_{k+} - H_{k1}$. Here we consider the odds ratio $\rho = p_1(1-p_2)/((1-p_1)p_2)$.

Situation 2 (Section 7.2): The quadruple $(H_{11}, H_{12}, H_{21}, H_{22})$ has a multinomial distribution with parameters N and $(p_{11}, p_{12}, p_{21}, p_{22})$. Here we consider $\rho = p_{11}p_{22}/(p_{12}p_{21})$.

The empirical odds ratio is defined as

$$\widehat{\rho} \;:=\; \frac{H_{11}H_{22}}{H_{12}H_{21}}.$$

To avoid division by zero, some authors propose to add 0.5 to each entry $H_{k\ell}$. This is a point estimator of the underlying true odds ratio ρ.

Instead of a point estimator $\widehat{\rho}$ we now derive confidence bounds for ρ. Similarly as in Chapter 1 we consider the conditional distribution of H_{11}, given the row and columns sums. Since $H_{+2} = N - H_{+1}$ and $H_{2+} = N - H_{1+}$ it suffices to condition on the pair (H_{1+}, H_{+1}). As we'll show soon, this conditional distribution is of the following type:

Definition 7.2 (Exponentially weighted hypergeometric distributions). For integers $N \geq 1$ and $\ell, n \in \{0, 1, \ldots, N\}$ we define

$$f_{\rho,N,\ell,n}(x) \;:=\; C^{-1}_{\rho,N,\ell,n}\, \frac{\rho^x}{x!(\ell-x)!(n-x)!(N-\ell-n+x)!}$$

if $x \in \{\max(0, \ell+n-N), \ldots, \min(\ell, n)\}$, and $f_{\rho,N,\ell,n}(x) := 0$ otherwise, where

$$C_{\rho,N,\ell,n} := \sum_{j=\max(0,\ell+n-N)}^{\min(\ell,n)} \frac{\rho^j}{j!(\ell-j)!(n-j)!(N-\ell-n+j)!}.$$

The corresponding distribution function is denoted by $F_{\rho,N,\ell,n}$.

Remark 7.3. The probability weights above may be rewritten as

$$f_{\rho,N,\ell,n}(x) = \tilde{C}^{-1}_{\rho,N,\ell,n}\binom{\ell}{x}\binom{N-\ell}{n-x}\rho^x = \tilde{C}^{-1}_{\rho,N,n,\ell}\binom{n}{x}\binom{N-n}{\ell-x}\rho^x$$

with suitable norming constants $\tilde{C}_{\rho,N,\ell,n}, \tilde{C}_{\rho,N,n,\ell}$. In case of $\rho = 1$ we obtain the hypergeometric distribution $\mathrm{Hyp}(N,\ell,n) = \mathrm{Hyp}(N,n,\ell)$. This explains the name 'exponentially weighted hypergeometric distributions'.

Lemma 7.4. *In the situations 1 and 2 just described, for arbitrary integers $\ell, n \in \{0,1,\ldots,N\}$ with $\mathbb{P}(H_{+1} = \ell, H_{1+} = n) > 0$ and $x \geq 0$,*

$$\mathbb{P}(H_{11} = x \,|\, H_{+1} = \ell, H_{1+} = n) = f_{\rho,N,\ell,n}(x).$$

Proof of Lemma 7.4. Generally $\mathbb{P}(H_{11} = x \,|\, H_{+1} = \ell, H_{1+} = n)$ is equal to

$$\frac{\mathbb{P}(H_{11} = x, H_{21} = \ell - x, H_{12} = n - x, H_{22} = N - \ell - n + x)}{\mathbb{P}(H_{+1} = \ell, H_{1+} = n)},$$

and the denominator $\mathbb{P}(H_{+1} = \ell, H_{1+} = n)$ equals

$$\sum_{j=\max(0,\ell+n-N)}^{\min(\ell,n)} \mathbb{P}(H_{11} = j, H_{21} = \ell - j, H_{12} = n - j, H_{22} = N - \ell - n + j).$$

In situation 1 it suffices to consider $n = H_{1+}$, and $H_{2+} = N - n$. Moreover, H_{11} and H_{21} are stochastically independent with $H_{11} \sim \mathrm{Bin}(n, p_1)$, $H_{21} \sim \mathrm{Bin}(N-n, p_2)$. Thus

$$\begin{aligned}
&\mathbb{P}(H_{11} = j, H_{21} = \ell - j, H_{12} = n - j, H_{22} = N - \ell - n + j) \\
&\quad = \mathbb{P}(H_{11} = j, H_{21} = \ell - j) \\
&\quad = \mathbb{P}(H_{11} = j)\,\mathbb{P}(H_{21} = \ell - j) \\
&\quad = \binom{n}{j} p_1^j (1-p_1)^{n-j} \binom{N-n}{\ell-j} p_2^{\ell-j} (1-p_2)^{N-n-\ell+j} \\
&\quad = C\,\frac{\rho^j}{j!(\ell-j)!(n-j)!(N-\ell-n+j)!}
\end{aligned}$$

with $C := n!(N-n)!(1-p_1)^n(1-p_2)^{N-n-\ell}$. Consequently, $\mathbb{P}(H_{1+} = n, H_{+1} = \ell)$ equals $C \cdot C_{\rho,N,\ell,n}$, and $\mathbb{P}(H_{11} = x \,|\, H_{+1} = \ell, H_{1+} = n) = f_{\rho,N,\ell,n}(x)$.

In situation 2 the definition of the multinomial distribution yields the formula

$$\begin{aligned}
&\mathbb{P}(H_{11} = j, H_{21} = \ell - j, H_{12} = n - j, H_{22} = N - \ell - n + j) \\
&\quad = \frac{N!}{j!(\ell-j)!(n-j)!(N-\ell-n+j)!}\, p_{11}^j p_{21}^{\ell-j} p_{12}^{n-j} p_{22}^{N-\ell-n+j} \\
&\quad = C\,\frac{\rho^j}{j!(\ell-j)!(n-j)!(N-\ell-n+j)!}
\end{aligned}$$

with $C := N!\, p_{21}^{\ell} p_{12}^{n} p_{22}^{N-\ell-n}$. Thus $\mathbb{P}(H_{+1} = \ell, H_{1+} = n)$ equals $C \cdot C_{\rho,N,\ell,n}$, and again $\mathbb{P}(H_{11} = x \,|\, H_{+1} = \ell, H_{1+} = n) = f_{\rho,N,\ell,n}(x)$. □

Lemma 7.4 shows that at least in situations 1 and 2, the conditional distribution function of H_{11}, given that $H_{+1} = \ell$ and $H_{1+} = n$, is equal to $F_{\rho,N,\ell,n}$. Together with Lemma 1.4 this leads to exact confidence bounds for ρ. For

$$\begin{aligned}
&\mathbb{P}\big(F_{\rho,N,H_{+1},H_{1+}}(H_{11}) \le \alpha\big) \\
&\quad = \sum_{\ell,n=0}^{N} \mathbb{P}(H_{+1} = \ell, H_{1+} = n)\, \mathbb{P}\big(F_{\rho,N,\ell,n}(H_{11}) \le \alpha \,\big|\, H_{+1} = \ell, H_{1+} = n\big) \\
&\quad \le \sum_{\ell,n=0}^{N} \mathbb{P}(H_{+1} = \ell, H_{1+} = n)\, \alpha \\
&\quad = \alpha,
\end{aligned}$$

and analogously,

$$\mathbb{P}\big(F_{\rho,N,H_{+1},H_{1+}}(H_{11} - 1) \ge 1 - \alpha\big) \;\le\; \alpha.$$

Further it follows from Lemma 2.7 that

$$\begin{aligned}
\big\{\rho \in (0,\infty) : F_{\rho,N,H_{+1},H_{1+}}(H_{11}) > \alpha\big\} &= (0, b_\alpha), \\
\big\{\rho \in (0,\infty) : F_{\rho,N,H_{+1},H_{1+}}(H_{11} - 1) < 1 - \alpha\big\} &= (a_\alpha, \infty).
\end{aligned}$$

Here $b_\alpha = b_\alpha(N, H_{+1}, H_{1+}, H_{11})$ is the unique solution $\rho \in (0,\infty)$ of the equation

$$F_{\rho,N,H_{+1},H_{1+}}(H_{11}) \;=\; \alpha$$

provided that $H_{11} < \min(H_{+1}, H_{1+})$. Otherwise we set $b_\alpha := \infty$. This is an upper $(1-\alpha)$-confidence bound for ρ. Moreover, $a_\alpha = a_\alpha(N, H_{+1}, H_{1+}, H_{11})$ is the unique solution $\rho \in (0,\infty)$ of the equation

$$F_{\rho,N,H_{+1},H_{1+}}(H_{11} - 1) \;= 1 - \alpha$$

provided that $H_{11} > \max(0, H_{+1} + H_{1+} - N)$. Otherwise we set $a_\alpha := 0$. This is a lower $(1-\alpha)$-confidence bound for ρ.

Example 7.5. In a randomised study, thirty patients with a certain skin rash received pills with a new drug or a placebo. The working hypothesis was that the new drug has a positive impact on the patients' skin. The treatment results turned out as follows:

	Improvement	No improvem.	
New drug	12	3	15
Placebo	5	10	15
	17	13	30

Now we consider the probabilities p_1 and p_2 of an improvement with the new drug and with placebo, respectively, in the population of all people with the given skin rash. To verify possibly the working hypothesis, we compute a lower 95%-confidence bound for the odds ratio ρ. To this end we consider the function $\rho \mapsto F_{\rho,N,H_{+1},H_{1+}}(H_{11} - 1) = F_{\rho,30,17,15}(11)$. Figure 7.1 shows this function and the resulting lower 95%-confidence bound $a_{0.05}(30, 17, 15, 12) \approx 1.531$. Thus we may claim with confidence 95% that the odds of an improvement with the new drug are at least 1.53 times the odds with placebo only. This confirms the working hypothesis.

By the way, Fisher's exact test yields the (right-sided) p-value $1 - F_{1,30,17,15}(11) \approx 0.0127$.

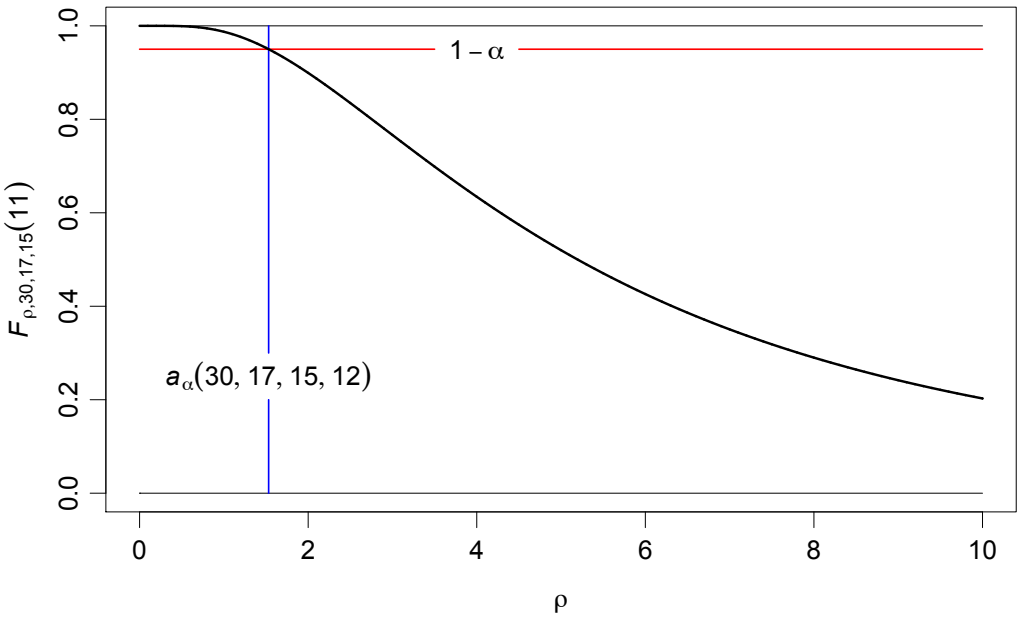

Figure 7.1: Example for the computation of a lower confidence bound for ρ.

Remark 7.6 (Connection with Fisher's exact test). These confidence bounds for ρ are closely related to Fisher's exact test. Namely, the lower bound $a_\alpha(N, H_{+1}, H_{1+}, H_{11})$ for ρ is larger than 1 if and only if the right-sided p-value $1 - F_{N,H_{+1},H_{1+}}(H_{11} - 1)$ is smaller than α. Analogously the upper bound $b_\alpha(N, H_{+1}, H_{1+}, H_{11})$ is smaller than 1 if and only if the left-sided p-value $F_{N,H_{+1},H_{1+}}(H_{11})$ is smaller than α.

Remark 7.7 (Warning). It is not always clear whether for a given two-by-two table there exists a well-defined underlying odds ratio ρ. There are situations in which Fisher's exact test is applicable while a proper definition of ρ and thus a clean interpretation of $\widehat{\rho}$ or the confidence bounds for ρ is unclear. For Fisher's exact test one only needs to justify that for given row and column sums the entry H_{11} follows a hypergeometric distribution with parameters N, H_{+1} and H_{1+}.

Two examples in which Fisher's exact test is applicable without an obvious definition of an odds ratio ρ are provided in Exercises 1.3 and 1.14.

7.4 Simpson's Paradox

In connection with several two-by-two tables sometimes an interesting phenomenon occurs: If one combines and analyses several data sets without taking their origin into account, one could end up with a two-by-two table contradicting the analyses of the single data sets. This phenomenon has been described, among others, by E. E. Simpson (1951). We illustrate it with a well-known data example.

Example 7.8 (Admissions to graduate programs at UC Berkeley). In the year 1973 the University of California at Berkeley hit the headlines because for graduate programs the admission numbers among men were substantially higher than among women. More precisely, 44% of the 8442 male applicants were admitted, but only 35% of the 4321 female applicants. The underlying data had been analyzed by Bickel et al. (1975) among others. In particular they considered the admission numbers of the six largest departments. The absolute numbers are listed in the first four columns of Table 7.1. Column 5 provides the empirical odds ratios for admission of men versus women (rounded to five digits). In addition one sees 95%-confidence intervals for an underlying true

Dept.	Men		Women		$\widehat{\rho}$	$a_{0.025}$	$b_{0.025}$
	Admitted	Not admitted	Admitted	Not admitted			
A	512	313	89	19	0.3496	0.1970	0.5920
B	353	207	17	8	0.8028	0.2945	2.0040
C	120	205	202	391	1.1329	0.8452	1.5163
D	138	279	131	244	0.9214	0.6790	1.2505
E	53	138	94	299	1.2212	0.8065	1.8385
F	22	351	24	317	0.8281	0.4333	1.5756
Total	1198	1493	557	1278	1.8409	1.6214	2.0912

Table 7.1: Admission numbers at UC Berkeley 1973.

odds ratio. Admittedly the definition of the latter is rather dubious, but these confidence intervals indicate potential associations between gender and admission.

Surprisingly, only in two departments the empirical odds ratio is slightly larger than 1, and clearly smaller than the overall empirical odds ratio $\widehat{\rho} = 1.8409$. In four departments the empirical odds ratio turned out to be smaller than 1, and in one case this difference was even significant at test level $\alpha = 5\%$. This stunning discrepancy results from the fact that women tended to apply for subjects with low overall admission rates while men tended to apply for programs with higher odds of admission.

The computation of an empirical odds ratio or even a confidence interval for the total numbers wouldn't make sense, because the decisions about acceptance or rejection are made within departments and follow different criteria, respectively. Also the populations of potential applicants are presumably different from department to department.

If for a single department the empirical odds ratio is significantly different from one, this indicates a true association between gender and admission. This could be due to systematic differences in the qualification of male and female applicants. It does not prove a gender bias on behalf of the department.

7.5 Exercises

Exercise 7.1. We consider the odds ratio ρ for two probabilities $p_1, p_2 \in (0, 1)$. Show that both p_1/p_2 and $(1 - p_2)/(1 - p_1)$ are always between 1 and ρ. Furthermore, show that

$$|\log(\rho)| \geq 4|p_1 - p_2|.$$

Sketch the set of all pairs (p_1, p_2) with $\rho = 0.5$ and $\rho = 2$, respectively.

Exercise 7.2. Prove Lemma 7.1.

Exercise 7.3. Subsequently we describe three cross-sectional studies. Define in each case a suitable odds ratio and decide whether a lower or upper bound or a confidence interval would be appropriate. Then analyze the data with test level $\alpha = 5\%$ and formulate a conclusion. The data are given in Table 7.2.

(a) To verify a possible association between acute bronchitis during early childhood and respiratory diseases during adolescence, $n = 1319$ fourteen year old children and their parents had been interviewed. On the one hand, the parents were asked whether their child suffered from

(a)	Cough	No cough
Acute bronchitis	26	44
No acute bronch.	247	1002

(b)	Male	Female
Right-handed	934	1070
Left-handed	113	92

(c)	Herniated discs	No hern. discs
Professional driver	4	4
No profess. driver	13	77

Table 7.2: Data for Exercise 7.3.

Victim	Suspect	Death penalty	Prison sentence
white	white	53	414
	black	11	37
black	white	0	16
	black	4	139

Table 7.3: Data example for Simpson's paradox.

acute bronchitis during the first five years. The second question was whether the child coughed frequently during the day or night.

(b) When interviewing $n = 2209$ Australians at age of 25-34 years, they were asked about their gender (male/female) and handedness (right-handed/left-handed), among other things.

(c) A cross-sectional study among older male employees was conducted to investigate whether working as a driver (buses or trucks) increases the risk of herniated vertebral discs.

Exercise 7.4. Table 7.3 contains numbers related to law suits for the years 1976-1987 in the state of Florida, US (see Agresti, 2002, and Radelet and Pierce, 1991). The underlying raw data are all law suits involving murder with the three dichotomous variables $X =$ color of skin of suspect (black or white), $Y =$ penalty (death penalty or prison sentence) and $Z =$ color of skin of victim. Discuss empirical associations between two of these three variables, with and without splitting the data by means of the remaining third variable.

Exercise 7.5. Construct a fictitious data example for Simpson's paradox: Suppose one compares a new medical treatment M1 with a standard treatment M2 in two randomised clinical trials, one in a hospital K1 and one in a hospital K2. Suppose method M1 is indeed better than method M2. Invent two two-by-two tables confirming this fact. But try to choose the numbers such that the sum of the two tables yields an empirical odds ratio contradicting this fact. This may happen if, for instance, hospital K1 tends to take over the more difficult cases and at the same time applies method M1 more frequently than method M2.

Chapter 8

Tests for Association

The two previous chapters covered the association between a categorical or binary and another variable. The latter was numerical in Chapter 6 and binary in Chapter 7. In the present chapter we treat testing for association between two variables in a rather general framework. We start with a rather abstract testing principle in the next section and then turn to permutation tests.

8.1 A General Principle of Nonparametric Tests

Both the sign tests in Section 4.3 as well as the permutation tests treated later are special cases of a rather general testing paradigm. The starting point is a data set $D(\omega) \in \mathcal{D}$ which was derived from raw data $\omega \in \Omega$. We consider a finite group $\mathcal{G}$ of bijective mappings $g : \mathcal{D} \to \mathcal{D}$. That means, with two mappings $g, h \in \mathcal{G}$, their composition $h \circ g$, i.e. the mapping $d \mapsto h(g(d))$, as well as the inverse mapping g^{-1} belong to $\mathcal{G}$, too.[1] Now we discuss a special property of the distribution of D.

Lemma 8.1 ($\mathcal{G}$-Invariance). *Let G be uniformly distributed on $\mathcal{G}$, stochastically independent from D. Then the following two conditions are equivalent:*

(i) *For arbitrary fixed $g \in \mathcal{G}$, the random variables $g(D)$ and D are identically distributed.*

(ii) *The random variables $G(D)$ and D are identically distributed.*

Null hypothesis H_o ($\mathcal{G}$-Invariance) The random variable D is *$\mathcal{G}$-invariant (in distribution)*. That means, it satisfies the conditions described in Lemma 8.1.

Example 8.2 (Sign tests). Let the data set D be a difference vector $\boldsymbol{X} = \boldsymbol{Y} - \boldsymbol{Z} \in \mathbb{R}^n$ as in Section 4.3. For an arbitrary sign vector $\boldsymbol{s} \in \{-1, 1\}^n$ we consider the bijection

$$\boldsymbol{x} \;\mapsto\; g_{\boldsymbol{s}}(\boldsymbol{x}) := (s_i x_i)_{i=1}^n$$

from $\mathbb{R}^n$ to $\mathbb{R}^n$. For an additional sign $\boldsymbol{t} \in \{-1, 1\}^n$,

$$g_{\boldsymbol{t}} \circ g_{\boldsymbol{s}} \;=\; g_{\boldsymbol{ts}}$$

with the coordinate-wise product $\boldsymbol{ts} = (t_i s_i)_{i=1}^n$. Thus $\mathcal{G} := \left\{g_{\boldsymbol{s}} : \boldsymbol{s} \in \{-1, 1\}^n\right\}$ is an Abelian group of 2^n bijective mappings, and sign symmetry is equivalent to $\mathcal{G}$-invariance.

[1] Again we hide measurability issues: To $\mathcal{D}$ belongs a σ-field $\mathcal{B}$, D is a $(\mathcal{D}, \mathcal{B})$-valued random variable, and all mappings $g \in \mathcal{G}$ are $\mathcal{B}$-$\mathcal{B}$-measurable.

Proof of Lemma 8.1. We argue similarly as in the proof of Lemma 4.11. For any measurable set $B \subset \mathcal{D}$,

$$\mathbb{P}(G(D) \in B) \;=\; \sum_{g \in \mathcal{G}} \mathbb{P}(G = g, g(D) \in B) \;=\; \frac{1}{\#\mathcal{G}} \sum_{g \in \mathcal{G}} \mathbb{P}(g(D) \in B).$$

Hence Condition (i) implies Condition (ii).

For any fixed $h \in \mathcal{G}$,

$$\mathbb{P}\big(h(G(D)) \in B\big) \;=\; \mathbb{P}\big((h \circ G)(D) \in B\big) \;=\; \mathbb{P}(G(D) \in B),$$

because $h \circ G$ is uniformly distributed on $\mathcal{G}$, too; see Exercise 8.1. Consequently, Condition (ii) implies that $\mathbb{P}(h(D) \in B) = \mathbb{P}(D \in B)$, and this entails Condition (i). □

Exact tests of H_o. To test H_o we choose a test statistic $T : \mathcal{D} \to \mathbb{R}$ and compute one of the three p-values $\pi_\ell(D)$, $\pi_r(D)$ or $\pi_z(D)$, depending on our working hypothesis. Here

$$\begin{aligned} \pi_\ell(d) \;&:=\; \#\big\{g \in \mathcal{G} : T(g(d)) \le T(d)\big\}/\#\mathcal{G}, \\ \pi_r(d) \;&:=\; \#\big\{g \in \mathcal{G} : T(g(d)) \ge T(d)\big\}/\#\mathcal{G} \end{aligned}$$

and $\pi_z(d) := 2 \cdot \min\big\{\pi_\ell(d), \pi_r(d)\big\}$ for an arbitrary data set $d \in \mathcal{D}$. With a random variable G having uniform distribution on $\mathcal{G}$ we may also write

$$\begin{aligned} \pi_\ell(d) \;&=\; \mathbb{P}\big(T(G(d)) \le T(d)\big), \\ \pi_r(d) \;&=\; \mathbb{P}\big(T(G(d)) \ge T(d)\big). \end{aligned}$$

Lemma 8.3. *Let $\pi(D)$ be one of the three p-values just described. Under the null hypothesis H_o,*

$$\mathbb{P}\big(\pi(D) \le \alpha\big) \;\le\; \alpha$$

for arbitrary $\alpha \in [0, 1]$.

Proof of Lemma 8.3. The proof is almost identical with the proof of Lemma 4.13. Under H_o,

$$\mathbb{P}\big(\pi(D) \le \alpha\big) \;=\; \#\mathcal{G}^{-1} \sum_{g \in \mathcal{G}} \mathbb{P}\big(\pi(g(D)) \le \alpha\big) \;=\; \mathbb{E}\Big(\#\mathcal{G}^{-1} \sum_{g \in \mathcal{G}} 1_{[\pi(g(D)) \le \alpha]}\Big).$$

Hence it suffices to show that for any fixed data set $d \in \mathcal{D}$,

$$\#\mathcal{G}^{-1} \sum_{g \in \mathcal{G}} 1_{[\pi(g(d)) \le \alpha]} = \mathbb{P}\big(\pi(G(d)) \le \alpha\big) \;\le\; \alpha.$$

To this end we consider the random variable $X := T(G(d))$. Then

$$\mathbb{P}(X \le x) \;=\; \#\big\{g \in \mathcal{G} : T(g(d)) \le y\big\}/\#\mathcal{G} \;=:\; F_d(x)$$

for arbitrary $x \in \mathbb{R}$. This distribution function $F_d(\cdot)$ remains the same if we replace d with any transformation $h(d)$, $h \in \mathcal{G}$. For the mapping $g \mapsto g \circ h$ is bijective from $\mathcal{G}$ to $\mathcal{G}$, see Exercise 8.1. In particular, $F_{G(d)}(\cdot) \equiv F_d(\cdot)$ and

$$\begin{aligned} \pi_\ell(G(d)) \;&=\; F_d(X), \\ \pi_r(G(d)) \;&=\; 1 - F_d(X-), \\ \pi_z(G(d)) \;&=\; 2 \cdot \min\{F_d(X), 1 - F_d(X-)\}. \end{aligned}$$

Consequently it follows from Lemma 1.4 that $\mathbb{P}\big(\pi(G(d)) \le \alpha\big) \le \alpha$. □

Monte Carlo tests of H_o. Sometimes the computation of the exact p-values $\pi_\ell(D), \pi_r(D)$ is too involved. A possible way out are Monte Carlo p-values: We simulate random variables $G^{(1)}$, $G^{(2)}, \ldots, G^{(m)}$ which are independent, uniformly distributed on $\mathcal{G}$ and independent from D. Then we compute

$$\begin{aligned}\widehat{\pi}_\ell(D) &:= \frac{\#\big\{s \in \{1,2,\ldots,m\} : T(G^{(s)}(D)) \le T(D)\big\} + 1}{m+1},\\ \widehat{\pi}_r(D) &:= \frac{\#\big\{s \in \{1,2,\ldots,m\} : T(G^{(s)}(D)) \ge T(D)\big\} + 1}{m+1}\end{aligned}$$

or $\widehat{\pi}_z(D) := 2 \cdot \min\big\{\widehat{\pi}_\ell(D), \widehat{\pi}_r(D)\big\}$. As shown in the next lemma, these Monte Carlo p-values are a viable surrogate for the exact ones.

Lemma 8.4. *Let $\widehat{\pi}(D)$ be one of the p-values just defined (with true random variables $G^{(s)}$). Under the null hypothesis H_o,*

$$\mathbb{P}\big(\widehat{\pi}(D) \le \alpha\big) \;\le\; \frac{\lfloor (m+1)\alpha \rfloor}{m+1} \;\le\; \alpha$$

for arbitrary $\alpha \in [0,1]$.

Proof of Lemma 8.4. Let $G^{(0)}$ be an additional random variable with uniform distribution on $\mathcal{G}$ and independent from D, $G^{(1)}$, …, $G^{(m)}$. Under H_o, the random data sets $G^{(0)}(D)$ and D are identically distributed. Hence the tuple

$$\big(T(D), T(G^{(1)}(D)), \ldots, T(G^{(m)}(D))\big)$$

has the same distribution as

$$\big(T(G^{(0)}(D)), T(G^{(1)} \circ G^{(0)}(D)), \ldots, T(G^{(m)} \circ G^{(0)}(D))\big).$$

But the tuples $(G^{(0)}, G^{(1)} \circ G^{(0)}, \ldots, G^{(m)} \circ G^{(0)})$ and $(G^{(0)}, G^{(1)}, \ldots, G^{(m)})$ are identically distributed: For arbitrary elements $g_0, g_1, \ldots, g_m$ of $\mathcal{G}$,

$$\begin{aligned}&\mathbb{P}\big(G^{(0)} = g_0, G^{(1)} \circ G^{(0)} = g_1, \ldots, G^{(m)} \circ G^{(0)} = g_m\big)\\ &\quad= \mathbb{P}\big(G^{(0)} = g_0, G^{(1)} = g_1 \circ g_0^{-1}, \ldots, G^{(m)} = g_m \circ g_0^{-1}\big) \;=\; (\#\mathcal{G})^{-(m+1)}.\end{aligned}$$

Hence the tuple $\big(T(D), T(G^{(1)}(D)), \ldots, T(G^{(m)}(D))\big)$ has the same distribution as

$$(T_0, T_1, \ldots, T_m) \;:=\; \big(T(G^{(0)}(D)), T(G^{(1)}(D)), \ldots, T(G^{(m)}(D))\big).$$

The latter satisfies the assumption of Lemma 2.11 whence

$$\mathbb{P}\big(\widehat{\pi}_r(D) \le \alpha\big) \;\le\; \frac{\lfloor (m+1)\alpha \rfloor}{m+1}.$$

With $-T$ in place of T we obtain the analogous inequality for $\widehat{\pi}_\ell(D)$. Then we may argue that two-sided p-value $\widehat{\pi}(D)$ satisfies

$$\begin{aligned}\mathbb{P}\big(\widehat{\pi}_z(D) \le \alpha\big) \;&\le\; \mathbb{P}\big(\widehat{\pi}_\ell(D) \le \alpha/2\big) + \mathbb{P}\big(\widehat{\pi}_r(D) \le \alpha/2\big)\\ &\le\; \frac{2\lfloor (m+1)\alpha/2 \rfloor}{m+1} \;\le\; \frac{\lfloor (m+1)\alpha \rfloor}{m+1}.\end{aligned}$$

□

8.2 Permutation Tests

Now we consider two variables X and Y with values in $\mathcal{X}$ and $\mathcal{Y}$, respectively. We would like to verify, if correct, that there is a true association between these two variables. The starting point is a data set with N data pairs (X_1, Y_1), (X_2, Y_2), ..., (X_N, Y_N), i.e. with two data vectors $\boldsymbol{X} = (X_i)_{i=1}^N \in \mathcal{X}^N$ and $\boldsymbol{Y} = (Y_i)_{i=1}^N \in \mathcal{Y}^N$.
In what follows $\mathcal{S}_N$ stands for the set of all permutations of $\{1, 2, \ldots, N\}$. For an arbitrary tuple $\boldsymbol{y} = (y_i)_{i=1}^N$ and a permutation $\sigma \in \mathcal{S}_N$ we write

$$\sigma\boldsymbol{y} \;:=\; (y_{\sigma(i)})_{i=1}^N.$$

The null hypothesis that there is *no* true association between X- and Y-values may be formalised as follows:

Null hypothesis H_o (Exchangeability) The vector $\boldsymbol{Y} = (Y_i)_{i=1}^N$ is *exchangeable with respect to* $\boldsymbol{X} = (X_i)_{i=1}^N$ *(in distribution)*. That means, for an arbitrary fixed permutation $\sigma \in \mathcal{S}_N$, the data sets $(\boldsymbol{X}, \sigma\boldsymbol{Y})$ and $(\boldsymbol{X}, \boldsymbol{Y})$ have the same distribution.

Example 8.5 (Stochastic independence). Let the data pairs $(X_1, Y_1), (X_2, Y_2), \ldots, (X_N, Y_N)$ be independent and identically distributed random variables. The working hypothesis is that X_1 and Y_1 are stochastically dependent. If they are stochastically independent, the data vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$ satisfy the null hypothesis H_o.

Suppose that $X_1, X_2, \ldots, X_N$ are fixed given values, for instance, N different and consecutive time points or N dosages of a certain substance in increasing order. Then one could simplify the null hypothesis as follows:

Null hypothesis H_o' (Exchangeability) The vector $\boldsymbol{Y} = (Y_i)_{i=1}^N$ is *exchangeable (in distribution)*. That means, for any fixed permutation $\sigma \in \mathcal{S}_N$, the data vectors $\sigma\boldsymbol{Y}$ and $\boldsymbol{Y}$ are identically distributed.

Example 8.6 (Independent, identically distributed random variables). Let $X_1 < X_2 < \cdots < X_N$ be fixed time points, and at time X_i one observes a random variable $Y_i \in \mathcal{Y}$. Now we want to verify, if the Y-values are really time-dependent. This could mean, for instance, that there is a certain trend, or that two consecutive observations are stochastically dependent. If the random variables $Y_1, Y_2, \ldots, Y_N$ are independent and identically distributed, the data vector $\boldsymbol{Y}$ satisfies the null hypothesis H_o'.

Remark 8.7. Both null hypotheses H_o and H_o' are special cases of $\mathcal{G}$-invariance as introduced in Section 8.1. In case of H_o we consider a data set $D = (\boldsymbol{X}, \boldsymbol{Y})$ in $\mathcal{D} = \mathcal{X}^N \times \mathcal{Y}^N$, and $\sigma \in \mathcal{S}_N$ induces a bijection

$$(\boldsymbol{x}, \boldsymbol{y}) \;\mapsto\; g_\sigma(\boldsymbol{x}, \boldsymbol{y}) := (\boldsymbol{x}, \sigma\boldsymbol{y})$$

from $\mathcal{X}^N \times \mathcal{Y}^N$ to $\mathcal{X}^N \times \mathcal{Y}^N$. In case of H_o' we consider only the data vector $D = \boldsymbol{Y}$ in $\mathcal{D} = \mathcal{Y}^N$, and $\sigma \in \mathcal{S}$ induces the bijection

$$\boldsymbol{y} \;\mapsto\; g_\sigma(\boldsymbol{y}) := \sigma\boldsymbol{y}$$

from $\mathcal{Y}^N$ to $\mathcal{Y}^N$. In both cases one can easily verify that for two permutations $\sigma, \tau \in \mathcal{S}_N$,

$$g_\tau \circ g_\sigma \;=\; g_{\sigma\circ\tau}.$$

Thus $\mathcal{G} := \{g_\sigma : \sigma \in \mathcal{S}_N\}$ is indeed a group of bijections.

An equivalent description of the null hypothesis H_o reads as follows: The original data set $(\boldsymbol{X}, \boldsymbol{Y})$ has the same distribution as $(\boldsymbol{X}, \Pi\boldsymbol{Y})$, where Π is uniformly distributed on $\mathcal{S}_N$ and stochastically independent from $(\boldsymbol{X}, \boldsymbol{Y})$.

Analogously, the null hypothesis H_o' is equivalent to the following statement: With Π as above, the original data vector $\boldsymbol{Y}$ has the same distribution as $\Pi\boldsymbol{Y}$.

Permutation tests. The null hypothesis H_o may be tested as described in Section 8.1. One chooses a test statistic $T : \mathcal{X}^N \times \mathcal{Y}^N \to \mathbb{R}$ and computes one of the p-values $\pi_\ell = \pi_\ell(\boldsymbol{X}, \boldsymbol{Y})$, $\pi_r = \pi_r(\boldsymbol{X}, \boldsymbol{Y})$ or $\pi_z = \pi_z(\boldsymbol{X}, \boldsymbol{Y}) = 2\min\{\pi_\ell, \pi_r\}$. Here

$$\begin{aligned}\pi_\ell(\boldsymbol{x}, \boldsymbol{y}) &:= \#\big\{\sigma \in \mathcal{S}_N : T(\boldsymbol{x}, \sigma\boldsymbol{y}) \le T(\boldsymbol{x}, \boldsymbol{y})\big\}\big/N! \\ &= \mathbb{P}\big(T(\boldsymbol{x}, \Pi\boldsymbol{y}) \le T(\boldsymbol{x}, \boldsymbol{y})\big), \\ \pi_r(\boldsymbol{x}, \boldsymbol{y}) &:= \#\big\{\sigma \in \mathcal{S}_N : T(\boldsymbol{x}, \sigma\boldsymbol{y}) \ge T(\boldsymbol{x}, \boldsymbol{y})\big\}\big/N! \\ &= \mathbb{P}\big(T(\boldsymbol{x}, \Pi\boldsymbol{y}) \ge T(\boldsymbol{x}, \boldsymbol{y})\big)\end{aligned}$$

for arbitrary tuples $\boldsymbol{x} \in \mathcal{X}^n$ and $\boldsymbol{y} \in \mathcal{Y}^n$ while Π is a random variable with uniform distribution on $\mathcal{S}_N$.

When testing H_o', the latter formulae become simpler in that we work with test statistics $T : \mathcal{Y}^N \to \mathbb{R}$ and omit the arguments $\boldsymbol{X}$ and $\boldsymbol{x}$, respectively.

Since the cardinality $N!$ of $\mathcal{S}_N$ is huge for moderate or large sample sizes N, we often resort to Monte Carlo p-values.

In principle only two questions remain to be answered: (i) Which test statistic $T(\boldsymbol{X}, \boldsymbol{Y})$ quantifies the potential deviations from H_o we are interested in? This depends on our working hypothesis, of course. (ii) How can we compute the p-values for our given test statistic $T(\boldsymbol{X}, \boldsymbol{Y})$ without going through all $N!$ permutations in $\mathcal{S}_N$? This issue is relevant whenever one wants to avoid Monte Carlo p-values.

8.3 Binary Variables: Trends and Runs

We consider first the null hypothesis H_o' of exchangeability for a random vector $\boldsymbol{Y}$ with components $Y_i \in \{0, 1\}$. Explicitly one may think of equidistant time points $X_0 < X_1 < \cdots < X_N$, and Y_i indicates whether in the time interval $(X_{i-1}, X_i]$ a certain catastrophe (e.g. an earthquake) occurred ($Y_i = 1$) or not ($Y_i = 0$). Potential questions are whether

(i) the frequency of such catastrophes tends to increase or decrease over time,

(ii) these events occur in clusters or, on the contrary, are distributed rather evenly.

Tests for monotone trends. In order to quantify for a vector $\boldsymbol{y} \in \{0, 1\}^N$ to what extent indices i with $y_i = 1$ tend to be rather small or rather large, one can use the test statistic

$$T(\boldsymbol{y}) := \sum_{i=1}^{N} y_i \cdot i.$$

The computation of the resulting p-values may be achieved with Wilcoxon's rank sum test. For it is shown in Exercise 8.2 that the random set $\big\{i \in \{1, 2, \ldots, N\} : y_{\Pi(i)} = 1\big\}$ has the same distribution as $\big\{\Pi(1), \ldots, \Pi(y_+)\big\}$, where $y_+ := \sum_{i=1}^N y_i$. Hence for arbitrary $x \in \mathbb{R}$,

$$\mathbb{P}(T(\Pi\boldsymbol{y}) \le x) \;=\; \mathbb{P}\Big(\sum_{i=1}^{y_+} \Pi(i) \le x\Big) \;=\; G_{y_+, N-y_+}\Big(x - \frac{y_+(y_+ + 1)}{2}\Big)$$

with the distribution functions $G_{n_1,n_2}(\cdot)$ for Wilcoxon'x rank sum test in Section 6.5. This leads to

$$\begin{aligned}\pi_\ell(\boldsymbol{y}) &= G_{y_+,N-y_+}\Big(T(\boldsymbol{y}) - \frac{y_+(y_+ + 1)}{2}\Big),\\ \pi_r(\boldsymbol{y}) &= 1 - G_{y_+,N-y_+}\Big(T(\boldsymbol{y}) - \frac{y_+(y_+ + 1)}{2} - 1\Big).\end{aligned}$$

Tests for clustering or even distribution. Now we want to judge for a vector $\boldsymbol{y} \in \{0,1\}^N$ whether the indices i with $y_i = 1$ (or $y_i = 0$) tend to form clusters or tend to be well separated, leading to a rather even distribution. This may be achieved with the runs test statistic

$$T(\boldsymbol{y}) := \sum_{i=1}^{N-1} 1_{[y_i \neq y_{i+1}]}.$$

We encountered it already in Example 1.17 in Section 1.4. A 'run' in $\boldsymbol{y}$ is a maximal block of adjacent indices i with identical values y_i. Thus $T(\boldsymbol{y}) + 1$ is the number of 'runs' in $\boldsymbol{y}$.

When applying this test statistic to our random vector $\boldsymbol{Y}$, we expect rather small values in case of clustering of 'events' (indices i with $y_i = 1$) and rather larger values in case of evenly distributed 'events'.

The distribution of $T(\Pi\boldsymbol{y})$ is none of the well-known discrete distributions, but it can be easily computed numerically. Thus the exact computation of the p-values π_ℓ and π_r poses no problem. The expected value and standard deviation of $T(\Pi\boldsymbol{y})$ are derived in Exercise 8.3.

Lemma 8.8. *Let $\boldsymbol{y} \in \{0,1\}^N$ with $0 < y_+ < N$. Then for integers $k \geq 1$,*

$$\begin{aligned}&\mathbb{P}(T(\Pi\boldsymbol{y}) = 2k-1)\\ &\quad= 2\binom{y_+ - 1}{k-1}\binom{N - y_+ - 1}{k-1}\Big/\binom{N}{y_+},\\ &\mathbb{P}(T(\Pi\boldsymbol{y}) = 2k)\\ &\quad= \left[\binom{y_+ - 1}{k}\binom{N - y_+ - 1}{k-1} + \binom{y_+ - 1}{k-1}\binom{N - y_+ - 1}{k}\right]\Big/\binom{N}{y_+}.\end{aligned}$$

Proof of Lemma 8.8. Instead of considering all permutations of a vector with y_+ entries 1 and $N - y_+$ entries 0 we just consider the $\binom{N}{y_+}$ possible vectors which may result from that. Each such vector $\tilde{\boldsymbol{y}}$ consists of $T(\tilde{\boldsymbol{y}}) + 1$ blocks of adjacent zeros or adjacent ones.

To cut a series of y_+ ones into k blocks one has to 'activate' $k-1$ of the $y_+ - 1$ interspaces. For instance, a series of $y_+ = 7$ ones may be cut into $k = 3$ blocks as follows:

$$(1\,1\,1\,1\,1\,1\,1) \rightsquigarrow (1\,1\,|\,1\,|\,1\,1\,1\,1).$$

For this cutting there exist $\binom{y_+ - 1}{k-1}$ possibilities. Analogously there are $\binom{N - y_+ - 1}{k-1}$ possibilities to cut a series of $N - y_+$ zeros into k blocks.

The equation $T(\tilde{\boldsymbol{y}}) = 2k - 1$ is equivalent to $\tilde{\boldsymbol{y}}$ consisting of $2k$ blocks, namely k blocks of ones and k blocks of zeros. Having specified these blocks already, one only has to combine them via alternate merging. For this step there are two possibilities, depending on whether one starts with a block or ones or a block of zeros. Here is an example for $y_+ = 7$, $N - y_+ = 5$ and $k = 3$:

$$\binom{1\,1\,1\,1\,1\,1\,1}{0\,0\,0\,0\,0} \rightsquigarrow \binom{1\,1\,|\,1\,|\,1\,1\,1\,1}{0\,|\,0\,0\,|\,0\,0} \rightsquigarrow \begin{cases}(1\,1\,|\,0\,|\,1\,|\,0\,0\,|\,1\,1\,1\,1\,|\,0\,0)\\ \quad\text{or}\\ (0\,|\,1\,1\,|\,0\,0\,|\,1\,|\,0\,0\,|\,1\,1\,1\,1).\end{cases}$$

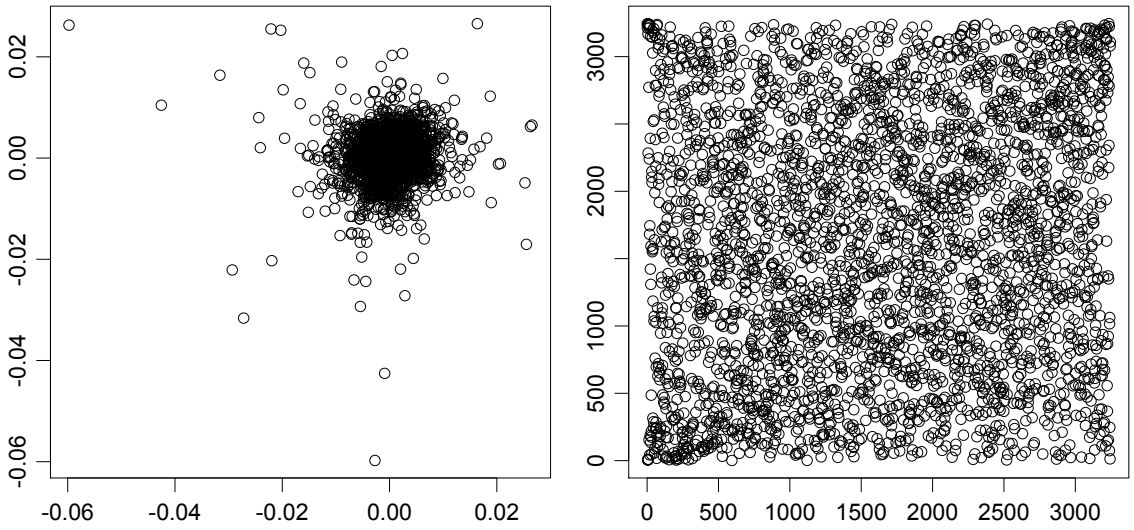

Figure 8.1: log-return today versus log-return tomorrow, raw values (left) and ranks (right)

Thus there are

$$2\binom{y_+ - 1}{k-1}\binom{N - y_+ - 1}{k-1}$$

possible vectors $\tilde{\boldsymbol{y}} \in \{0,1\}^N$ with $\tilde{y}_+ = y_+$ and $T(\tilde{\boldsymbol{y}}) = 2k - 1$.

The equation $T(\tilde{\boldsymbol{y}}) = 2k$ means that either $\tilde{\boldsymbol{y}}$ contains $k+1$ blocks of ones which are separated by k blocks of zeros, or $\tilde{\boldsymbol{y}}$ contains $k+1$ blocks of zeros which are separated by k blocks of ones. Thus there are

$$\binom{y_+ - 1}{k}\binom{N - y_+ - 1}{k-1} + \binom{y_+ - 1}{k-1}\binom{N - y_+ - 1}{k}$$

possible vectors $\tilde{\boldsymbol{y}} \in \{0,1\}^N$ with $\tilde{y}_+ = y_+$ and $T(\tilde{\boldsymbol{y}}) = 2k$. □

Example 8.9 (Log-returns, continued). We re-analyze the data of Example 5.9 with the share index values K_i on 3246 consecutive trading days. In Figure 5.10 we saw already the time series $(\log_{10}(K_i))_{i=1}^{3256}$ of the log-transformed values and the times series of the $N = 3245$ log-returns

$$L_i \;:=\; \log_{10}(K_{i+1}/K_i).$$

Figure 8.1 shows on the left hand side a scatter plot of the $N-1$ pairs (L_i, L_{i+1}) for $1 \le i < N$, and on the right hand side the same picture after replacing the values $L_1, L_2, \ldots, L_N$ with their ranks. Both plots suggest that it is difficult to predict a log-return from the previous one. Nevertheless the density of points close to the first main diagonal and in the upper left corner seems to be slightly increased.

A true time-dependency may be verified as follows: We reduce the log-returns to the following binary quantities:

$$Y_i^{(1)} \;:=\; 1_{[L_i > 0]} \quad \text{and} \quad Y_i^{(2)} \;:=\; 1_{[|L_i| > M]}$$

with $M := \mathrm{Median}(|L_1|, \ldots, |L_N|)$. The vector $\boldsymbol{Y}^{(1)}$ specifies trading days with increase or decrease of the share index value. The vector $\boldsymbol{Y}^{(2)}$ emphasises more the volatility of the share index values, i.e. the modulus of the log-returns. If the original vector $(L_i)_{i=1}^N$ would be exchangeable in distribution, both binary vectors $\boldsymbol{Y}^{(1)}, \boldsymbol{Y}^{(2)}$ would inherit this property.

However, $T(\boldsymbol{Y}^{(1)}) = 1494$ with $Y_+^{(1)} = 1734$ ones and $N - Y_+^{(1)} = 1511$ zeros, and by means of Lemma 8.8 we obtain the following p-values: $\pi_\ell(\boldsymbol{Y}^{(1)}) = 1.0852 \cdot 10^{-5}$ und $\pi_r(\boldsymbol{Y}^{(1)}) = 0.999$,

so

$$\pi_z(\boldsymbol{Y}^{(1)}) \;=\; 2.1704 \cdot 10^{-5}.$$

This shows that the zeros and ones tend to occur in clusters somewhat, although this effect is hardly visible on short time intervals. This confirms our empirical finding of an increased density of points close to the first main diagonal in Figure 8.1 (right hand side).

Also the vector $\boldsymbol{Y}^{(2)}$ violates exchangeability with high confidence: Here $T(\boldsymbol{Y}^{(2)}) = 1494$ again, where $Y_+^{(2)} = 1622$ and $n - Y_+^{(2)} = 1623$. This yields the p-values $\pi_\ell(\boldsymbol{Y}^{(1)}) = 3.4474 \cdot 10^{-6}$, $\pi_r(\boldsymbol{Y}^{(1)}) > 0.999$, so

$$\pi_z(\boldsymbol{Y}^{(1)}) \;=\; 6.8948 \cdot 10^{-6}.$$

This confirms our impression from Figure 5.10. There seem to be periods of high and periods of low volatility of log-returns, and this leads to an increased density of points in three of the four corners in Figure 8.1 (right hand side).

8.4 Categorical Variables: Contingency Tables

Now we consider two categorical variables

$$X \;\in\; \{x_1, x_2, \ldots, x_K\} \quad \text{and} \quad Y \;\in\; \{y_1, y_2, \ldots, y_L\}.$$

Thus the data pairs (X_i, Y_i) have only KL potential values, and we summarise the data as a *contingency table*: For $k \in \{1, \ldots, K\}$ and $\ell \in \{1, \ldots, L\}$ we define

$$H_{k,\ell} = H_{k,\ell}(\boldsymbol{X}, \boldsymbol{Y}) \;:=\; \#\big\{i \in \{1, \ldots, N\} : X_i = x_k \text{ and } Y_i = y_\ell\big\}.$$

Then the contingency table has the general form

	y_1	y_2	$\cdots$	y_L
x_1	$H_{1,1}$	$H_{1,2}$	$\cdots$	$H_{1,L}$
x_2	$H_{2,1}$	$H_{2,2}$	$\cdots$	$H_{2,L}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$
x_K	$H_{K,1}$	$H_{K,2}$	$\cdots$	$H_{K,L}$

Often it is complemented with the row sums

$$H_{k,+} \;:=\; \sum_{\ell=1}^{L} H_{k,\ell} \;=\; \#\big\{i \in \{1, \ldots, N\} : X_i = x_k\big\}$$

and column sums

$$H_{+,\ell} \;:=\; \sum_{k=1}^{K} H_{k,\ell} \;=\; \#\big\{i \in \{1, \ldots, N\} : Y_i = y_\ell\big\}.$$

This leads to

	y_1	y_2	$\cdots$	y_L	
x_1	$H_{1,1}$	$H_{1,2}$	$\cdots$	$H_{1,L}$	$H_{1,+}$
x_2	$H_{2,1}$	$H_{2,2}$	$\cdots$	$H_{2,L}$	$H_{2,+}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
x_K	$H_{K,1}$	$H_{K,2}$	$\cdots$	$H_{K,L}$	$H_{K,+}$
	$H_{+,1}$	$H_{+,2}$	$\cdots$	$H_{+,L}$	N

In case of two binary variables, i.e. $K = L = 2$, we get a two-by-two table again.

Fisher's exact tests. A possible test statistic for the null hypothesis H_o would be $T(\boldsymbol{X}, \boldsymbol{Y}) = H_{k,\ell}$ for a fixed index pair $(k, \ell) \in \{1, \ldots, K\} \times \{1, \ldots, L\}$. It follows from Exercise 8.5 that for given data $(\boldsymbol{X}, \boldsymbol{Y})$, the random variable $T(\boldsymbol{X}, \Pi\boldsymbol{Y})$ has a hypergeometric distribution with parameters N, $H_{k,+}$ and $H_{+,\ell}$. Thus we obtain the p-values

$$\begin{aligned} \pi_\ell &= F_{N,H_{k,+},H_{+,\ell}}(H_{k,\ell}), \\ \pi_r &= 1 - F_{N,H_{k,+},H_{+,\ell}}(H_{k,\ell} - 1). \end{aligned}$$

In case of binary variables ($K = L = 2$) it is sufficient to consider just one index pair (k, ℓ).

In Chapter 1 we saw already various applications of Fisher's exact test. Now we discuss an erroneous application of this method:

Example 8.10 (A devastating application of Fisher's exact test). In a sensation causing trial, the dutch nurse Lucia de Berk had been convicted of murder in several cases and sentenced to lifelong imprisonment. This trial was triggered by several cases of unexplained deaths in a hospital for children. Some staff members realised that Lucia da Berk was present in all corresponding shifts. During the trial a self-proclaimed expert in statistics presented a two-by-two table summarizing all $N = 1029$ shifts of the hospital in a certain time interval:

	Death case	No death case	
L. de Berk present	9	133	142
L. de Berk not present	0	887	887
	9	1020	1029

(In addition he presented two analogous two-by-two tables of different hospitals in which Lucia de Berk worked previously, but the case numbers had been very low.) Applying Fisher's exact test to this two-by-two table yields the extremely small two-sided p-value

$$\pi_z = 2\pi_r = 2(1 - F_{1029,142,9}(8)) \approx 2.9024 \cdot 10^{-8},$$

indicating an association between Lucia de Berk's presence and the occurence of deaths. One should, however, take into account that in this time period 26 nurses worked in the same ward, and one could have produced and tested a two-by-two table for each of them. Thus the present p-value should be adjusted with a factor of 26, which results in the p-value $52(1 - F_{1029,142,9}(8)) \approx 7.5462 \cdot 10^{-7}$, see also Section 6.6.

The expert made clear that this extremely significant association does *not* prove murder. It could have been possible that due to her experience, Lucia de Berk has been put into particularly challenging shifts. Or maybe she is just a bad nurse, though without bad intentions. All these explanations had been dismissed by her and her principals.

This example shows what damage can be caused by 'hobby statistics'. As soon as the data structure is okay, statistical procedures and software produce output, no matter whether the procedures are justified or not. In particular, Fisher's exact test is often sold as a procedure to possibly verify a 'true association' between two binary variables. But 'no true association' is only an incomplete description of the null hypothesis we are testing. Strictly speaking, Fisher's exact test concerns the null hypothesis that, conditional on the row and colums sums, the entry $H_{k,\ell}$ follows the hypergeometric distribution $\mathrm{Hyp}(N, H_{k,+}, H_{+,\ell})$. A sufficient condition for the latter condition would be exchangeability of the Y-values versus the X-values or vice versa. But in the present case this is questionable. Personal shift schedules in hospitals follow certain regular patterns in order to obey various constraints. There is also empirical evidence that the time points of deaths in hospitals are not distributed randomly. Presumably the present two-by-two table shows only that both the shift schedule of Lucia de Berk and the time points of deaths were somehow correlated with the third variable 'time', but without any causal relationship.

A second weakness of the analysis above is the somewhat arbitrary choice of time interval. Before it there was a longer period without any death cases. Moreover it turned out retrospectively that the underlying raw data were not correct. A careful re-investigation led to the table

	Death case	No death case	
L. de Berk present	7	135	142
L. de Berk not present	4	883	887
	11	1018	1029

which yields a two-sided p-value of $2(1 - F_{1029,142,11}(6)) \approx 2.515 \cdot 10^{-5}$ and, after Bonferroni adjustment, $52(1 - F_{1029,142,11}(6)) \approx 6.5411 \cdot 10^{-4}$. These values are still rather small but not as spectacular as the ones presented during the trial.

Richard Gill and other Dutch and European scientists got involved with this case and succeeded in getting it re-opened. Medical experts made clear again that in *none* of the death cases an external influence had been verified. The new trial ended with a complete acquittal of Lucia de Berk.

The Chi-squared test. Under the null hypothesis H_o we expect the entry $H_{k,\ell}$ to be close to

$$\bar{H}_{k,\ell} \;:=\; \frac{H_{k,+}H_{+,\ell}}{n}.$$

Precisely,

$$\bar{H}_{k,\ell} \;=\; \mathbb{E}\big(H_{k,\ell}(\boldsymbol{X}, \Pi\boldsymbol{Y}) \,\big|\, \boldsymbol{X}, \boldsymbol{Y}\big),$$

where $\mathbb{E}(\cdot \,|\, \boldsymbol{X}, \boldsymbol{Y})$ stands for conditional expectations, given the data $\boldsymbol{X}, \boldsymbol{Y}$. That means, we consider $\boldsymbol{X}, \boldsymbol{Y}$ as fixed objects, and only Π is chosen completely at random. This formula for $\bar{H}_{k,\ell}$ is a consequence of Exercise 8.5. The following chi-squared test statistic due to Karl Pearson quantifies the deviations of the entries $H_{k,\ell}$ from these idealised values $\bar{H}_{k,\ell}$:

$$T(\boldsymbol{X}, \boldsymbol{Y}) \;:=\; \sum_{k=1}^{K}\sum_{\ell=1}^{L} \frac{(H_{k,\ell} - \bar{H}_{k,\ell})^2}{\bar{H}_{k,\ell}} \;=\; \sum_{k=1}^{K}\sum_{\ell=1}^{L} \frac{H_{k,\ell}^2}{\bar{H}_{k,\ell}} - N.$$

The latter equation follows from the fact that both $\sum_{k=1}^{K}\sum_{\ell=1}^{L} H_{k,\ell}$ and $\sum_{k=1}^{K}\sum_{\ell=1}^{L} \bar{H}_{k,\ell}$ are equal to N. We refer to $\bar{H}_{k,\ell}$ as 'idealised value', because a contingency table with entries $H_{k,\ell} \approx \bar{H}_{k,\ell}$ would show no association at all:

Lemma 8.11. *For a contingency table $(H_{k,\ell})_{k,\ell}$ the following conditions are equivalent:*

(i) *All rows (resp. columns) are proportional;*

(ii) *$H_{k,\ell} = \bar{H}_{k,\ell}$ for arbitrary index pairs (k, ℓ).*

Under the null hypothesis H_o one may expect $T(\boldsymbol{X}, \boldsymbol{Y})$ to be of size $(K-1)(L-1)$, because (Exercise 8.7)

$$\mathbb{E}\big(T(\boldsymbol{X}, \Pi\boldsymbol{Y}) \,\big|\, \boldsymbol{X}, \boldsymbol{Y}\big) \;=\; \frac{N}{N-1}\,(K-1)(L-1). \tag{8.1}$$

Unfortunately, the conditional distribution function of $T(\boldsymbol{X}, \Pi\boldsymbol{Y})$, given the data $(\boldsymbol{X}, \boldsymbol{Y})$, is not easy to compute. But one can show that this conditional distribution converges weakly to

$$\chi^2_{(K-1)(L-1)}$$

as

$$\min\big\{H_{1,+}, H_{2,+}, \ldots, H_{K,+},\, H_{+,1}, H_{+,2}, \ldots, H_{+,L}\big\} \;\to\; \infty.$$

Hence whenever all row and column sums are at least 5, one often uses the approximation

$$\pi_r \approx 1 - F_{(K-1)(L-1)}(T(\boldsymbol{X}, \boldsymbol{Y})),$$

where $F_{(K-1)(L-1)}$ denotes the distribution function of $\chi^2_{(K-1)(L-1)}$. Otherwise one could compute Monte-Carlo p-values.

Remark 8.12. If one of the tests in this chapter rejects the null hypothesis H_o, one may conclude that there is a true association between X- and Y-values. But this does not imply a causal relationship. For instance, it could be the case that both variables depend on one or more additional variables. This effect is called "confounding", and such latent variables causing associations are called 'confounders'.

Remark 8.13. If the chi-squared test rejects the null hypothesis H_o, one may claim with a certain confidence that there is a true association between X- and Y-values. But this does not specify *how* the two variables are correlated. Sometimes it is thus more informative to extract from the original contingency table a two-by-two table by merging or deleting certain values in $\{x_1, \ldots, x_K\}$ or $\{y_1, \ldots, y_L\}$. This two-by-two table may then be analyzed with (a two-sided version of) Fisher's exact test. Or one computes a confidence interval for the underlying odds ratio, if the latter is well-defined.

Example 8.14 (Snoring and heart diseases). In a medical cross-sectional study about the potential association between snoring and heart deseases, $N = 2484$ men have been investigated. On the one hand it was checked whether they suffered from a heart desease or not. This corresponds to a binary variable X with possible values 'diseased' and 'healthy'. Moreover the men's spouses were asked to categorise their partners with respect to snoring. This yielded an ordinal variable Y with possible values 'never', 'sometimes', 'often' (at least every second night) and 'always' (every night). Here is the corresponding contingency table:

	never	sometimes	often	always	
diseased	24	35	21	30	110
healthy	1355	603	192	224	2374
	1379	638	213	254	2484

The group of diseased men is (fortunately) much smaller than the group of healthy men, and the group of never snoring men is much larger than the three other groups of snoring men. Hence it is difficult to judge from this contingency table whether there is empirical evidence for an association between snoring and heart diseases. To get a clearer picture we consider the same table with normalised rows (to three decimal digits):

	never	sometimes	often	always
diseased	0.218	0.318	0.191	0.273
healthy	0.571	0.254	0.081	0.094
	0.555	0.257	0.086	0.102

Now it is clear that the relative proportions of often or always snoring men is much higher among the diseased men than among the healthy men. Normalising the columns shows a similar phenomenon:

	never	sometimes	often	always	
diseased	0.017	0.055	0.099	0.118	0.044
healthy	0.983	0.945	0.901	0.882	0.956

The relative proportion of diseased men increases with the ordinal variable Y.

Now we test the null hypothesis H_o that there is no true association between the two variables at level $\alpha = 1\%$: Here is the contingency table, augmented by the idealised values $\bar{H}_{j,k}$:

	never	sometimes	often	always	
diseased	24 (61.1)	35 (28.3)	21 (9.4)	30 (11.2)	110
healthy	1355 (1317.9)	603 (609.7)	192 (203.6)	224 (242.8)	2374
	1379	638	213	254	2484

The chi-squared statistic equals $T(\boldsymbol{X}, \boldsymbol{Y}) = 72.782$, which is substantially larger than the expected value $(K-1)(L-1) = 3$ under H_o. Indeed, the corresponding approximate p-value equals

$$1 - F_3(72.782) \approx 1.1102 \cdot 10^{-15},$$

and the corresponding Monte-Carlo p-values were extremely small, too.

As already mentioned, this does not prove a causal relationship between snoring and heart diseases. It could be the case that (i) snoring leads to heart diseases, (ii) heart diseases cause snoring or (iii) both snoring and heart diseases are influenced by common genetic or environmental factors. Furthermore, the χ^2 test statistics does not indicate any 'direction' of the association.

To get a clear statement including a 'direction', we merge the first two categories 'never' and 'sometimes' of Y to a new category 'rarely', and the two categories 'often' and 'always' are combined to 'frequently'. This yields the two-by-two table

	rarely	frequently	
diseased	59	51	110
healthy	1958	416	2374
	2017	467	2484

The underlying odds ratio ρ may be interpreted in two ways: One could consider the odds of finding a rarely snoring man, among diseased men and among healthy men. Alternatively one could consider the odds of finding a man with a heart disease, among rare snorers and among frequent snorers. The empirical odds ratio equals $\widehat{\rho} = 0.2458$, and a 99%-confidence interval for ρ is given by $[0.1448, 0.4201]$. Since the upper bound is smaller than one, we may conclude with confidence 99% that there is a positive association between snoring and heart diseases.

8.5 Numerical Variables: Sample Comparisons and Correlations

Sample comparisons. Suppose that X is a categorical variable with values in $\{x_1, x_2, \ldots, x_K\}$ while Y is a numerical variable. In this case we could analyze the data with procedures introduced in Chapter 6. (There we talked about (G, X) instead of (X, Y).) With an arbitrary test statistic $T(\boldsymbol{X}, \boldsymbol{Y})$ which quantifies differences between the subsamples $\boldsymbol{Y}_k := (Y_i)_{i\,:\,X_i = x_k}$ for $k = 1, 2, \ldots, K$ we may perform a permutation test of the null hypothesis H_o.

In the special case of $K = 2$, for instance, one could use Wilcoxon's rank sum test: We determine the ranks $R_{Y,1}, R_{Y,2}, \ldots, R_{Y,N}$ of $Y_1, Y_2, \ldots, Y_N$, and then we compute

$$T_W(\boldsymbol{X}, \boldsymbol{Y}) := \sum_{i\,:\,X_i = x_1} R_{Y,i}.$$

With this test statistic one could compute a p-value via a permutation test. That means, we don't have to assume continuous distribution functions of the variables Y_i as in Section 6.5.

In case of $K \geq 3$ one may combine several tests for the comparison of two tests as described in Section 6.6 at the end of Chapter 6. Again it is advisable to apply permutation tests for the $K(K-1)$ one-sided comparisons.

Simple Linear Regression and Correlation

Now we consider the case of two numerical variables. The original question of a true association between X- and Y-values is modified as follows: We want to investigate whether the Y-values may be approximated by a linear function of the X-values. Before introducing explicit tests, let us start with some theoretical considerations.

Linear prediction. Let X and Y be random variables with known joint distribution. Suppose we want to predict the value of Y by a linear function of X. More precisely, we want to determine real parameters a, b such that the mean squared prediction error

$$\mathbb{E}((Y - a - bX)^2)$$

is minimal. Here we assume that $0 < \mathrm{Std}(X), \mathrm{Std}(Y) < \infty$.

Lemma 8.15. *For arbitrary real numbers a and b,*

$$\mathbb{E}((Y - a - bX)^2) \;\geq\; \mathrm{Var}(Y) - \mathrm{Cov}(X,Y)^2/\,\mathrm{Var}(X).$$

Equality holds if and only if

$$b \;=\; b_* := \mathrm{Cov}(X,Y)/\,\mathrm{Var}(X) \quad \textit{and} \quad a \;=\; a_* := \mathbb{E}(Y) - b_*\,\mathbb{E}(X).$$

The optimal parameters involve only the expectated values of X and Y, the variance of X and the covariance

$$\mathrm{Cov}(X,Y) \;:=\; \mathbb{E}\big((X - \mathbb{E}(X))(Y - \mathbb{E}(Y))\big) \;=\; \mathbb{E}(XY) - \mathbb{E}(X)\,\mathbb{E}(Y)$$

of X and Y. With their correlation

$$\mathrm{Corr}(X,Y) \;:=\; \frac{\mathrm{Cov}(X,Y)}{\mathrm{Std}(X)\,\mathrm{Std}(Y)}$$

one may also write

$$b_* \;=\; \frac{\mathrm{Std}(Y)}{\mathrm{Std}(X)}\,\mathrm{Corr}(X,Y),$$

and the mean squared prediction error equals

$$\mathbb{E}((Y - a_* - b_*X)^2) \;=\; \mathrm{Var}(Y)(1 - \mathrm{Corr}(X,Y)^2).$$

The factor $\mathrm{Var}(Y)$ is the mean squared prediction error if we ignore X and predict Y by the constant $\mathbb{E}(Y)$. It is reduced by the factor $1 - \mathrm{Corr}(X,Y)^2$ if we predict Y by

$$a_* + b_*X \;=\; \mathbb{E}(Y) + b_*(X - \mathbb{E}(X)).$$

Proof of Lemma 8.15. If we fix an arbitrary value b, then $V := Y - bX$ satisfies the equation

$$\mathbb{E}((Y - a - bX)^2) \;=\; \mathbb{E}((V - a)^2) \;=\; \mathrm{Var}(V) + (\mathbb{E}(V) - a)^2.$$

As a function of $a \in \mathbb{R}$ it has the unique minimiser $a_*(b) = \mathbb{E}(V) = \mathbb{E}(Y) - b\,\mathbb{E}(V)$. When plugging in this value $a_*(b)$ for a, we obtain the equation

$$\begin{aligned}\mathbb{E}((Y - a_*(b) - bX)^2) &= \mathbb{E}\big[\big((Y - \mathbb{E}(Y)) - b(X - \mathbb{E}(X))\big)^2\big] \\ &= \operatorname{Var}(Y) - 2b\operatorname{Cov}(X,Y) + b^2\operatorname{Var}(X).\end{aligned}$$

With $b_* = \operatorname{Cov}(X,Y)/\operatorname{Var}(X)$ the right hand side equals

$$\operatorname{Var}(Y) - \operatorname{Cov}(X,Y)^2/\operatorname{Var}(X) + \operatorname{Var}(X)(b - b_*)^2.$$

This shows that b_* is the unique optimal value of b. □

Regression lines. Now we consider a data set with observation vectors $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{R}^N$. We exclude trivial situations and assume that the corresponding sample standard deviations S_X and S_Y are strictly positive. Now we aim for real parameters a and b such that the sum of squares

$$\sum_{i=1}^{N}(Y_i - a - bX_i)^2 = \|\boldsymbol{Y} - a\mathbf{1} - b\boldsymbol{X}\|^2$$

is minimal. Here $\mathbf{1}$ denotes the vector $(1,1,\ldots,1)^\top \in \mathbb{R}^N$, and $\|\cdot\|$ is the standard Euclidean norm on $\mathbb{R}^N$, $\|\boldsymbol{w}\| := \sqrt{\langle \boldsymbol{w}, \boldsymbol{w}\rangle}$ with the standard inner product $\langle\cdot,\cdot\rangle$. Our considerations about linear prediction imply the following conclusions:

Lemma 8.16. *With the centered vectors $\tilde{\boldsymbol{X}} := (X_i - \overline{X})_{i=1}^N$ and $\tilde{\boldsymbol{Y}} := (Y_i - \overline{Y})_{i=1}^N$, for arbitrary real numbers a and b,*

$$\|\boldsymbol{Y} - a - b\boldsymbol{X}\|^2 \geq \|\tilde{\boldsymbol{Y}}\|^2(1 - \widehat{\rho}^2)$$

with the sample correlation coefficient

$$\widehat{\rho} = \widehat{\rho}(\boldsymbol{X}, \boldsymbol{Y}) := \frac{\langle \tilde{\boldsymbol{X}}, \tilde{\boldsymbol{Y}}\rangle}{\|\tilde{\boldsymbol{X}}\|\|\tilde{\boldsymbol{Y}}\|}.$$

Equality holds if and only if

$$b = \widehat{b} := \frac{\langle \tilde{\boldsymbol{X}}, \tilde{\boldsymbol{Y}}\rangle}{\|\tilde{\boldsymbol{X}}\|^2} = \frac{S_Y}{S_X}\,\widehat{\rho} \quad \textit{and} \quad a = \widehat{a} := \overline{Y} - \widehat{b}\overline{X}.$$

Proof. We consider $\boldsymbol{X}$ and $\boldsymbol{Y}$ as fixed vectors. With a random variable J with uniform distribution on $\{1,2,\ldots,N\}$ we define the random pair $(X,Y) := (X_J, Y_J)$. Then

$$\|\boldsymbol{Y} - a\mathbf{1} - b\boldsymbol{X}\|^2 = N\,\mathbb{E}((Y - a - bX)^2).$$

Now the claims follow essentially from Lemma 8.15 and the following identities: $\mathbb{E}(X) = \overline{X}$, $\operatorname{Var}(X) = \|\tilde{\boldsymbol{X}}\|^2/N$, $\mathbb{E}(Y) = \overline{Y}$, $\operatorname{Var}(Y) = \|\tilde{\boldsymbol{Y}}\|^2/N$ and $\operatorname{Cov}(X,Y) = \langle \tilde{\boldsymbol{X}}, \tilde{\boldsymbol{Y}}\rangle/N$. □

The regression line consists of all pairs (x,y) satisfying the equation

$$y = \widehat{a} + \widehat{b}x = \overline{Y} + \frac{S_Y}{S_X}\,\widehat{\rho}\,(x - \overline{X}).$$

One can also write

$$\frac{y - \overline{Y}}{S_Y} = \widehat{\rho}\,\frac{x - \overline{X}}{S_X}.$$

In particular, the regression line passes through the barycentre $(\overline{X}, \overline{Y})$ of all observations (X_i, Y_i).

The sample correlation coefficient $\widehat{\rho}$ is the cosine of the angle between the centered data vectors $\tilde{\boldsymbol{X}}$ and $\tilde{\boldsymbol{Y}}$. The Cauchy–Schwarz inequality implies that $|\widehat{\rho}| \leq 1$. Equality holds if and only if $\tilde{\boldsymbol{Y}} = \widehat{b}\tilde{\boldsymbol{X}}$ with $\widehat{b} \neq 0$. This is equivalent to all observations (X_i, Y_i) sitting on the regression line, where $\operatorname{sign}(\widehat{b}) = \operatorname{sign}(\widehat{\rho})$. In general the slope parameter $\widehat{b} = \widehat{\rho}\, S_Y/S_X$ is always contained in the interval $[-S_Y/S_X, S_Y/S_X]$.

The square $\widehat{\rho}^2$ is sometimes called 'measure of determination'. It is a descriptive measure of how well the Y-values may be approximated by a linear function of the X-values.

Remark 8.17 ($\widehat{\rho}$ as an estimator). Suppose the observations (X_1, Y_1), (X_2, Y_2), …, (X_N, Y_N) are stochastically independent and identically distributed. Then the regression quantities just introduced may be interpreted as point estimators of theoretical quantities related to the distribution of $(X, Y) := (X_1, Y_1)$. On the one hand, $\overline{X}$, $\overline{Y}$, $S_X = \|\tilde{\boldsymbol{X}}\|/\sqrt{N-1}$ and $S_Y = \|\tilde{\boldsymbol{Y}}\|/\sqrt{N-1}$ are estimators of $\mathbb{E}(X)$, $\mathbb{E}(Y)$, $\operatorname{Std}(X)$ and $\operatorname{Std}(Y)$, respectively. Furthermore, $\langle \tilde{\boldsymbol{X}}, \tilde{\boldsymbol{Y}} \rangle/(N-1)$ and $\widehat{\rho}$ are estimators of $\operatorname{Cov}(X, Y)$ and $\operatorname{Corr}(X, Y)$, respectively.

Permutation tests. To verify a true association between X- and Y-values, one may conduct a permutation test with the test statistic $T(\boldsymbol{X}, \boldsymbol{Y}) := \langle \boldsymbol{X}, \boldsymbol{Y} \rangle$ or $T(\boldsymbol{X}, \boldsymbol{Y}) := \langle \tilde{\boldsymbol{X}}, \tilde{\boldsymbol{Y}} \rangle = \langle \boldsymbol{X}, \boldsymbol{Y} \rangle - N\,\overline{X}\,\overline{Y}$. Since $\overline{Y}$ is invariant under permutations of $\boldsymbol{Y}$, the resulting p-values are identical in both cases. Moreover, $\widetilde{\Pi\boldsymbol{Y}} = \Pi\tilde{\boldsymbol{Y}}$, and Exercise 6.9 implies the equations

$$\mathbb{E}\big(\langle \tilde{\boldsymbol{X}}, \Pi\tilde{\boldsymbol{Y}} \rangle \,\big|\, \boldsymbol{X}, \boldsymbol{Y}\big) \;=\; 0 \quad \text{and} \quad \operatorname{Var}\big(\langle \tilde{\boldsymbol{X}}, \Pi\tilde{\boldsymbol{Y}} \rangle \,\big|\, \boldsymbol{X}, \boldsymbol{Y}\big) \;=\; \frac{\|\tilde{\boldsymbol{X}}\|^2 \|\tilde{\boldsymbol{Y}}\|^2}{N-1}.$$

Hence one could also use the standardised test statistic $T(\boldsymbol{X}, \boldsymbol{Y}) := \sqrt{N-1}\,\widehat{\rho}$. Indeed, Theorem A.13 in Section A.9 of the appendix justifies the following approximations:

$$\pi_\ell \;\approx\; \Phi\big(\sqrt{N-1}\,\widehat{\rho}\big) \quad \text{and} \quad \pi_r \;\approx\; \Phi\big(-\sqrt{N-1}\,\widehat{\rho}\big)$$

whenever $\max_{i=1,\ldots,N} |X_i - \overline{X}|/\|\tilde{\boldsymbol{X}}\|$ and $\max_{i=1,\ldots,N} |Y_i - \overline{Y}|/\|\tilde{\boldsymbol{Y}}\|$ are sufficiently small. This is admittedly somewhat vague, but the approximations are a good first step before computing exact (Monte-Carlo) p-values.

Remark 8.18 ($\widehat{a}$ and $\widehat{b}$ as estimators and a classical test). Suppose, $X_1, X_2, \ldots, X_N$ are fixed numbers, for instance, doses or concentrations of a certain substance. Further let

$$Y_i \;=\; a_* + b_* X_i + \epsilon_i \quad \text{for } 1 \leq i \leq N$$

with unknown parameters $a_*, b_* \in \mathbb{R}$ and random errors $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ such that $\mathbb{E}(\epsilon_i) = 0$ for $1 \leq i \leq n$. Under these assumptions, $\widehat{a}$ and $\widehat{b}$ are unbiased estimators of a_* and b_*:

$$\mathbb{E}(\widehat{a}) \;=\; a_* \quad \text{and} \quad \mathbb{E}(\widehat{b}) \;=\; b_*.$$

To prove this we write $\boldsymbol{Y} = a_*\mathbf{1} + b_*\boldsymbol{X} + \boldsymbol{\epsilon}$, $\overline{Y} = a_* + b_*\overline{X} + \overline{\epsilon}$ and $\tilde{\boldsymbol{Y}} = b_*\tilde{\boldsymbol{X}} + \boldsymbol{\epsilon} - \overline{\epsilon}\mathbf{1}$. In particular,

$$\widehat{b} = \frac{\langle \tilde{\boldsymbol{X}}, \tilde{\boldsymbol{Y}} \rangle}{\|\tilde{\boldsymbol{X}}\|^2} \;=\; b_* + \frac{\langle \tilde{\boldsymbol{X}}, \boldsymbol{\epsilon} - \overline{\epsilon}\mathbf{1} \rangle}{\|\tilde{\boldsymbol{X}}\|^2} \;=\; b_* + \frac{\langle \tilde{\boldsymbol{X}}, \boldsymbol{\epsilon} \rangle}{\|\tilde{\boldsymbol{X}}\|^2},$$

because $\langle \tilde{\boldsymbol{X}}, \mathbf{1} \rangle = 0$. Consequently $\mathbb{E}(\widehat{b}) = b_*$, since $\mathbb{E}\langle \tilde{\boldsymbol{X}}, \boldsymbol{\epsilon} \rangle = \sum_{i=1}^N \tilde{X}_i\, \mathbb{E}(\epsilon_i) = 0$. Furthermore,

$$\widehat{a} = \overline{Y} - \widehat{b}\overline{X} \;=\; a_* + (b_* - \widehat{b})\overline{X} + \overline{\epsilon},$$

so $\mathbb{E}(\widehat{a}) = a_*$, since $\mathbb{E}(b_* - \widehat{b}) = \mathbb{E}(\overline{\epsilon}) = 0$.

Under the stronger assumption that $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ are stochastically independent with distribution $\mathcal{N}(0, \sigma^2)$, $\sigma > 0$ unknown,

$$\frac{\sqrt{n-1}\,\widehat{\rho}}{\sqrt{1-\widehat{\rho}^2}} \;\sim\; t_{n-2} \quad \text{if } b_* = 0.$$

This result may be proved with similar arguments as in the proof of Theorem 4.3. It implies a classical test of R. A. Fisher: The null hypothesis that $b_* = 0$ is rejected at level α if

$$\frac{\sqrt{n-1}\,|\widehat{\rho}|}{\sqrt{1-\widehat{\rho}^2}} \;\geq\; t_{n-2;1-\alpha/2}.$$

These considerations are just a glimpse into the important and wide field of linear models and regression methods.

Rank Correlation

The sample correlation coefficient $\widehat{\rho}$ quantifies the empirical *linear* association between X- and Y-values. This is sometimes too restrictive. It could happen, for instance, that the Y-values may be very well approximated by a monotone increasing or monotone decreasing function of the X-values, but this monotonic function is non-linear. In such situations one could replace the original data vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$ with their rank vectors $\boldsymbol{R}_X$ and $\boldsymbol{R}_Y$, respectively. Since the sample mean of a rank vector for N observations is always $(N+1)/2$ (see Exercise 3.6), we obtain *Spearman's rank correlation coefficient*

$$\widehat{\rho}^{(\mathrm{Sp})} = \widehat{\rho}^{(\mathrm{Sp})}(\boldsymbol{X}, \boldsymbol{Y}) \;:=\; \frac{\langle \boldsymbol{R}_X, \boldsymbol{R}_Y \rangle - N(N+1)^2/4}{\sqrt{\|\boldsymbol{R}_X\|^2 - N(N+1)^2/4}\sqrt{\|\boldsymbol{R}_Y\|^2 - N(N+1)^2/4}}.$$

If both the X-values and the Y-values are pairwise different, then $\|\boldsymbol{R}_W\|^2 - N(N+1)^2/4 = \sum_{i=1}^N i^2 - N(N+1)^2/4 = N(N^2-1)/12$ for $W = X, Y$, whence

$$\widehat{\rho}^{(\mathrm{Sp})} \;=\; \frac{\langle \boldsymbol{R}_X, \boldsymbol{R}_Y \rangle - N(N+1)^2/4}{N(N^2-1)/12}.$$

In general we know that $\|\boldsymbol{R}_W\|^2 \leq \sum_{i=1}^N i^2$ (see Exercise 4.20), whence

$$|\widehat{\rho}^{(\mathrm{Sp})}| \;\geq\; \frac{\big|\langle \boldsymbol{R}_X, \boldsymbol{R}_Y \rangle - N(N+1)^2/4\big|}{N(N^2-1)/12}.$$

Again one could conduct permutation tests based on the test statistic $T(\boldsymbol{X}, \boldsymbol{Y}) := \langle \boldsymbol{R}_X, \boldsymbol{R}_Y \rangle$. Approximations for the corresponding p-values are given by

$$\pi_\ell \;\approx\; \Phi\big(\sqrt{N-1} \cdot \widehat{\rho}^{(\mathrm{Sp})}\big) \quad \text{and} \quad \pi_r \;\approx\; \Phi\big(-\sqrt{N-1} \cdot \widehat{\rho}^{(\mathrm{Sp})}\big),$$

provided that $\max_{i=1,\ldots,N} |R_{W,i} - (N+1)/2| / \|\boldsymbol{R}_W\|$ is sufficiently small for $W = X, Y$.

Remark 8.19 (Properties of $\widehat{\rho}^{(\mathrm{Sp})}$). Spearman's rank correlation coefficient has some properties which distinguish it from the sample correlation coefficient $\widehat{\rho}$:

- $\widehat{\rho}^{(\mathrm{Sp})}$ is invariant under strictly monotone increasing transformations of the X-values or the Y-values.
- $\widehat{\rho}^{(\mathrm{Sp})}$ equals $\xi \in \{-1, 1\}$ if and only if

$$\mathrm{sign}(Y_i - Y_j) \;=\; \xi \cdot \mathrm{sign}(X_i - X_j) \quad \text{for } 1 \leq i < j \leq N.$$

This is equivalent to $Y_i = \xi \cdot u(X_i)$ for a strictly monotone increasing function $u : [X_{(1)}, X_{(N)}] \to \mathbb{R}$.

- $\widehat{\rho}^{(\mathrm{Sp})}$ is robust against a few outliers in the raw data.
- One can compute $\widehat{\rho}^{(\mathrm{Sp})}$ for ordinal rather than numerical variables.

Remark 8.20 ($\widehat{\rho}^{(\mathrm{Sp})}$ as an estimator). Suppose that the observations (X_i, Y_i) are stochastically independent and identically distributed, where $X = X_1$ and $Y = Y_1$ have continuous distribution functions F and G, respectively. Then $R_{X,i}/(N+1)$ and $R_{Y,i}/(N+1)$ may be viewed as proxies for $F(X_i)$ and $G(Y_i)$, respectively, and $\widehat{\rho}^{(\mathrm{Sp})}$ is an estimator for the correlation

$$\rho^{(\mathrm{Sp})} \;:=\; \mathrm{Corr}\big(F(X), G(Y)\big).$$

The transformations $X \mapsto F(X)$ and $Y \mapsto G(Y)$ yield a random variable $\big(F(X), G(Y)\big)$ with values in $[0,1] \times [0,1]$, and both components are uniformly distributed on $[0,1]$. In particular, $\mathbb{E}(F(X)) = \mathbb{E}(G(Y)) = 1/2$ and $\mathrm{Var}(F(X)) = \mathrm{Var}(G(Y)) = 1/12$, so

$$\rho^{(\mathrm{Sp})} \;=\; 12\big(\mathbb{E}\big(F(X)G(Y)\big) - 1/4\big).$$

Example 8.21. We consider once more Example 6.4 with professional baseball players. For a generic player let X be the number of years he is playing in the professional league and Y his annual salary (in kUSD). Both variables are now viewed as numerical.

The $N = 263$ observations yielded the regression parameters $\widehat{a} \approx 260.234$ (unit: kUSD), $\widehat{b} \approx 37.705$ (unit: kUSD/year) and $\widehat{\rho} \approx 0.401$. Here one may interpret $\widehat{b}$ as mean increase of income per year. Figure 8.2 shows a scatter plot of these data plus regression line. The sample means $\overline{X}, \overline{Y}$ and the lines through $(\overline{X}, \overline{Y})$ with slopes $\pm S_Y/S_X$ are indicated as well. As probably expected, the slope $\widehat{b}$ is positive, but the association between X and Y seems to be nonlinear rather than linear. The measure of determination has the relatively low value $\widehat{\rho}^2 \approx 0.161$.

The sample correlation $\widehat{\rho}$ is invariant under monotone increasing linear transformations of the X-values or Y-values. But it may change under monotone increasing nonlinear transformations. For instance, if we replace Y with $\log_{10}(Y)$, then we obtain the larger value $\widehat{\rho} \approx 0.537$ and $\widehat{\rho}^2 \approx 0.289$; see Figure 8.3. Still the scatter plot indicates a monotone but nonlinear relationship between X and $\log_{10}(Y)$.

Now let's consider rank correlation: Neither the X-values nor the Y-values are pairwise different. Here $\|\boldsymbol{R}_X\|^2 = 6089630$, $\|\boldsymbol{R}_Y\|^2 = 6098224$ and $\langle \boldsymbol{R}_X, \boldsymbol{R}_Y \rangle = 5528264$. Moreover, $N(N+1)^2/4 = 263 \cdot 264^2/4 = 4582512$. Consequently,

$$\widehat{\rho}^{(\mathrm{Sp})} \;=\; \frac{(5528264 - 4582512)}{\sqrt{(6089630 - 4582512)(6098224 - 4582512)}} \;\approx\; 0.626$$

and $(\widehat{\rho}^{(\mathrm{Sp})})^2 \approx 0.392$. Interestingly Spearman's rank correlation coefficient is larger than the sample correlation coefficient for the original variables X and Y (or $\log_{10}(Y)$). Figure 8.4 shows a scatter plot of the rank pairs $(R_{X,i}, R_{Y,i})$ plus regression line.

Computing the standardised correlation coefficients indicates already that the empirical positive correlation between X and Y is significant: For the raw data, $\sqrt{N-1}\,\widehat{\rho} \approx 6.4852$, after replacing Y with $\log_{10}(Y)$ we even get $\sqrt{N-1}\,\widehat{\rho} \approx 8.698$, and $\sqrt{N-1}\,\widehat{\rho}^{(\mathrm{Sp})} \approx 10.129$. In all three cases, corresponding two-sided (Monte-Carlo) p-values with $m \geq 9'999$ simulations turned out to be no larger than 10^{-5}.

8.6 Exercises

Exercise 8.1 (Groups). Let $(\mathcal{G}, *)$ be an arbitrary group and h any element of $\mathcal{G}$. Show that the mappings $g \mapsto g * h$, $g \mapsto h * g$ and $g \mapsto g^{-1}$ are bijective from $\mathcal{G}$ to $\mathcal{G}$.

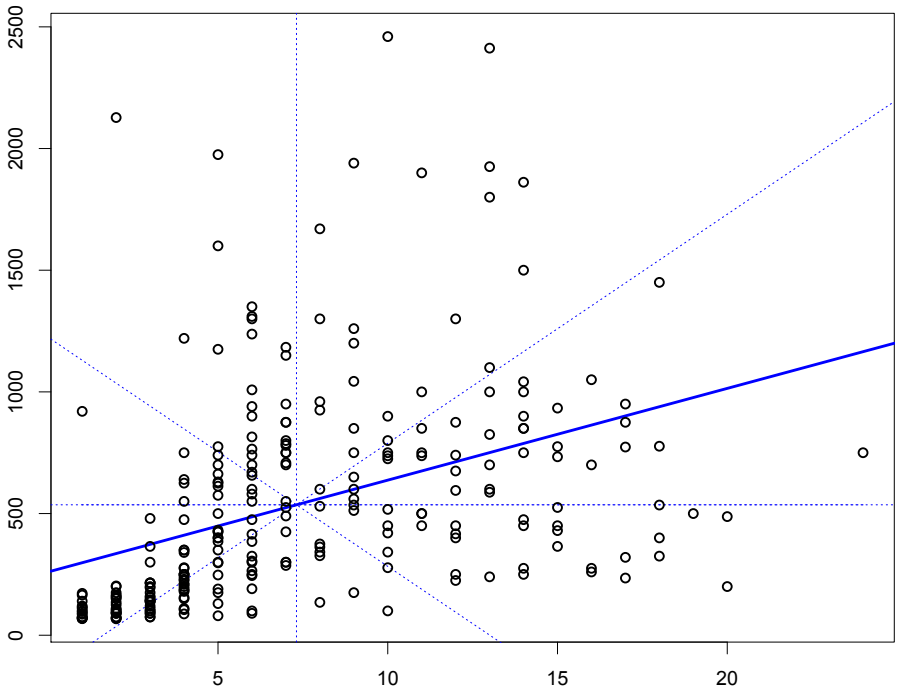

Figure 8.2: Salary versus years of employment for baseball players.

Now let $\#\mathcal{G} < \infty$, and let G be uniformly distributed on $\mathcal{G}$. Explain why each of the three random variables $h * G$, $G * h$ and G^{-1} is uniformly distributed on $\mathcal{G}$, too.

Exercise 8.2. Each vector $\boldsymbol{y} \in \{0,1\}^N$ is uniquely determined by the set $\big\{i \in \{1,2,\ldots,n\} : y_i = 1\big\}$. Suppose that $1 \le y_+ = \sum_{i=1}^N y_i < N$. Let Π be uniformly distributed on the set $\mathcal{S}_N$. Show that $\big\{i \in \{1,2,\ldots,N\} : y_{\Pi(i)} = 1\big\}$ and $\big\{\Pi(1),\ldots,\Pi(y_+)\big\}$ are identically distributed, namely, uniformly on the set of all $\binom{N}{y_+}$ subsets of $\{1,2,\ldots,N\}$ with exactly y_+ elements.

Exercise 8.3 (Moments of the runs test statistic). Prove the following (in)equalities for the runs test statistic $T(\boldsymbol{y}) = \sum_{i=1}^{n-1} 1_{[y_i \ne y_{i+1}]}$, $\boldsymbol{y} \in \{0,1\}^N$:

$$\begin{aligned}
\mathbb{E}(T(\Pi\boldsymbol{y})) &= 2y_+(N-y_+)/N \le N/2,\\
\mathrm{Var}(T(\Pi\boldsymbol{y})) &= \mathbb{E}(T(\Pi\boldsymbol{y}))\big(\mathbb{E}(T(\Pi\boldsymbol{y})) - 1\big)/(N-1),\\
\mathrm{Std}(T(\Pi\boldsymbol{y})) &\le \mathbb{E}(T(\Pi\boldsymbol{y}))/\sqrt{N} \le \sqrt{N}\,/2.
\end{aligned}$$

Exercise 8.4 (Good vintages and exchangeability). A vintage is considered to be good if it is better than both its predecessor and its successor. Among wine experts it is well-known that about every third vintage is a good one. At first glance this indicates a mysterious three-year rhythm. But there is also a rather simple explanation:

Let $Y_0, Y_1, \ldots, Y_N, Y_{N+1}$ be random variables which are almost surely pairwise different. Suppose that the tuple of these $N+2$ random variables is exchangeable (in distribution). Show that the random variable $Z := \sum_{i=1}^N 1_{[Y_i > \max(Y_{i-1}, Y_{i+1})]}$ satisfies

$$\mathbb{E}(Z/N) = 1/3$$

and

$$\mathrm{Std}(Z/N) = O(N^{-1/2}).$$

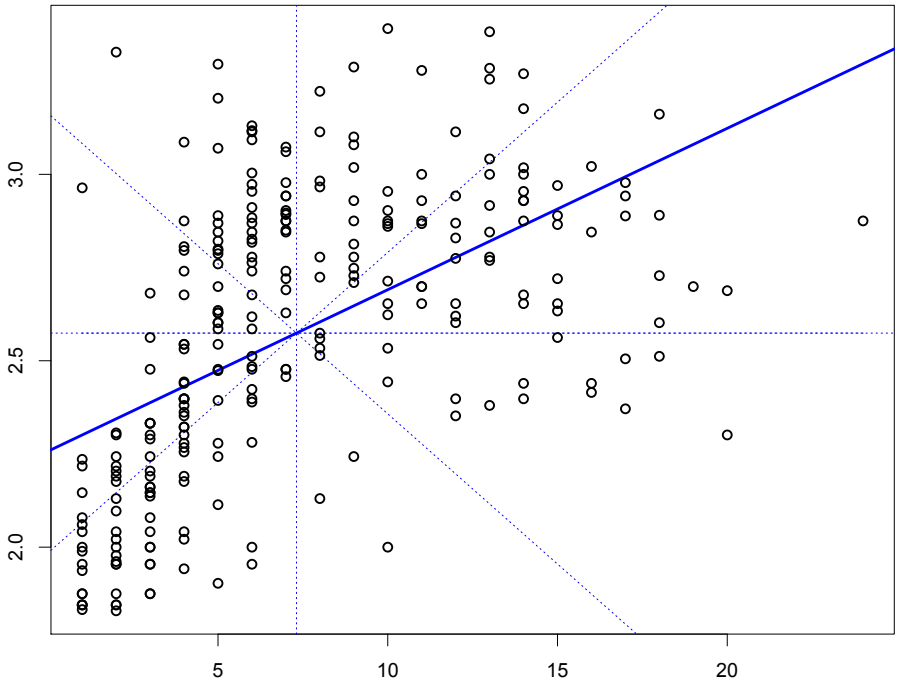

Figure 8.3: $\log_{10}(\text{Salary})$ versus years of employment for baseball players.

Exercise 8.5 (Linear permutation statistics, II). As in Exercise 6.9 let $T := \sum_{i=1}^{N} a_i b_{\Pi(i)}$ with fixed vectors $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^N$ and a random permutation Π with uniform distribution on $\mathcal{S}_N$.

(i) Show that the distribution of T remains unchanged if we interchange the vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ or permute the components of $\boldsymbol{a}$ or $\boldsymbol{b}$.

(ii) Suppose that $\boldsymbol{a}, \boldsymbol{b} \in \{0,1\}^N$. Show that T has a hypergeometric distribution with parameters N, $a_+ = \sum_{i=1}^{N} a_i$ and $b_+ = \sum_{i=1}^{N} b_i$. Now deduce from Exercise 6.9 that

$$\mathbb{E}(T) \;=\; \frac{a_+ b_+}{N} \quad \text{and} \quad \text{Var}(T) \;=\; \frac{a_+ b_+ (N - a_+)(N - b_+)}{N^2 (N-1)}.$$

Exercise 8.6. Prove Lemma 8.11.

Exercise 8.7. Prove Equation (8.1) by means of Exercise 8.5 (ii).

Exercise 8.8 (Order of siblings and personality). It is a well-known stereotype that within families with several children the youngest kids are often the most comic ones. To test this working hypothesis one could obtain the data of n families of comedians with at least one sibling. Then the data set would consist of pairs (G_1, K_1), (G_2, K_2), …, (G_n, K_n), where $G_i \geq 2$ would be the total number of kids in the family of the i-th comedian, and $K_i \in \{1, \ldots, G_i\}$ would specify its order with respect to age. How could one test the working hypothesis?

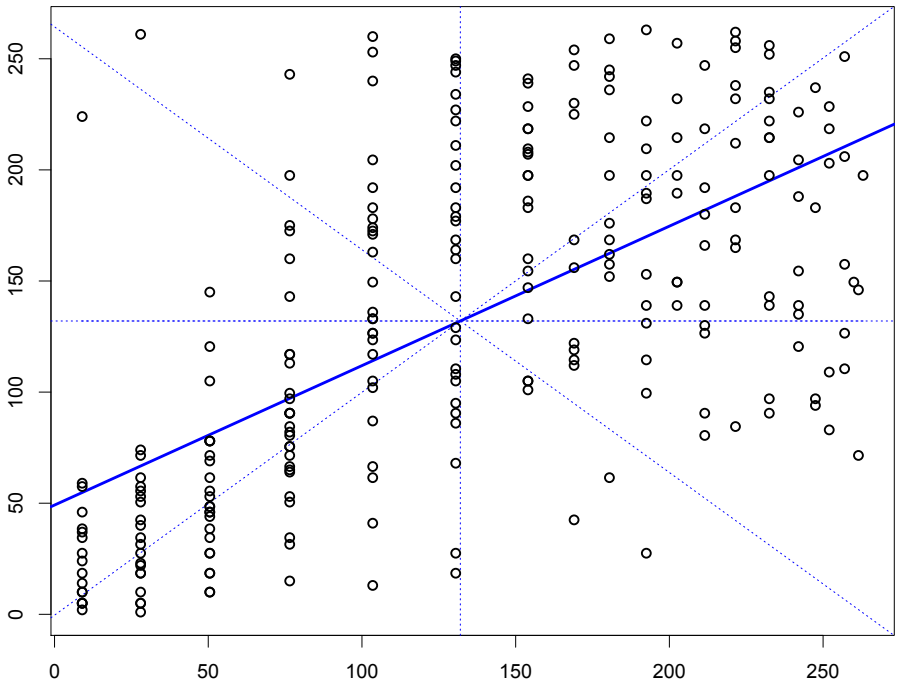

Figure 8.4: Rank(Salary) versus Rank(Years) for baseball players.

Appendix A

Complements

This appendix contains background information on some topics covered in the lecture notes. Moreover it contains some material which is of interest for students who are planning to delve further into statistics.

A.1 Hints for R

For statistical analyses and simulations as well as for the implementation of new procedures the software and programming environment R [22] is very suitable. This is an open-source software based on the programming language S which is available for all major operating systems.

Chapter I. The most important distributions are implemented in R, in each case by means of four functions:

- dfamily(x, θ): Weight function (for discrete distributions) or density function (for absolutely continuous distributions) at point $x \in \mathbb{R}$;
- pfamily(x, θ): Distribution function at point $x \in \mathbb{R}$;
- qfamily(u, θ): Quantile function at point $u \in [0, 1]$;
- rfamily(n, θ): Simulation of n independent random variables.

Here family is a place holder for the explicit distribution family, and θ denotes the parameter(s):

- hyper (hypergeometric distributions): $\mathrm{Hyp}(N, \ell, n)$ corresponds to hyper$(\cdot, \ell, N-\ell, n)$ or hyper$(\cdot, \mathsf{m} = \ell, \mathsf{n} = N-\ell, \mathsf{k} = n)$!
- binom (binomial distributions): $\mathrm{Bin}(n, p)$ corresponds to binom$(\cdot, n, p)$ or binom$(\cdot, \mathsf{size} = n, \mathsf{prob} = p)$.
- norm (Normal distributions): $\mathcal{N}(\mu, \sigma^2)$ corresponds to norm$(\cdot, \mu, \sigma)$ or norm$(\cdot, \mathsf{mean} = \mu, \mathsf{sd} = \sigma)$!
- t (Student distributions): t_k corresponds to t$(\cdot, k)$ or t$(\cdot, \mathsf{df} = k)$.
- gamma (Gamma distributions): $\mathrm{Gamma}(a, b)$ corresponds to gamma$(\cdot, \mathsf{shape} = a, \mathsf{scale} = b)$ oder gamma$(\cdot, \mathsf{shape} = a, \mathsf{rate} = 1/b)$!

Concerning Fisher's exact test we refer to the hints for Chapter VII.

Chapter II. The Clopper–Pearson confidence bounds for a binomial parameter p may be obtained with the built-in function $\mathsf{binom.test}(\cdot)$. Precisely,

$$\mathsf{binom.test}(\mathsf{x} = H, \mathsf{n} = n, \mathsf{conf.level} = 1 - \alpha)$$

yields (among other things) the $(1-\alpha)$-confidence interval $\big[a_{\alpha/2}(H), b_{\alpha/2}(H)\big]$ for p, based on the observation $H \sim \mathrm{Bin}(n, p)$. The argument $\mathsf{conf.level}$ is optional with default 95%. The one-sided confidence bounds may be obtained as follows:

$$\mathsf{binom.test}(\mathsf{x} = H, \mathsf{n} = n, \mathsf{conf.level} = 1 - \alpha, \mathsf{alternative} = \mathsf{'greater'})$$

yields the interval $\big[a_\alpha(H), 1\big]$ and

$$\mathsf{binom.test}(\mathsf{x} = H, \mathsf{n} = n, \mathsf{conf.level} = 1 - \alpha, \mathsf{alternative} = \mathsf{'less'})$$

the interval $\big[0, b_\alpha(H)\big]$ for p.

The function $\mathsf{binom.test}(\cdot)$ has another optional parameter p with default 0.5. This is a *hypothetical* value of p which is to be tested. Precisely, for an arbitrary value $p_o \in [0, 1]$ one may determine p-values for the following testing problems:

Null hypothesis "$p = p_o$", working hypothesis "$p \neq p_o$":

$$\mathsf{binom.test}(\mathsf{x} = H, \mathsf{n} = n, \mathsf{p} = p_o).$$

Null hypothesis "$p \geq p_o$", working hypothesis "$p < p_o$":

$$\mathsf{binom.test}(\mathsf{x} = H, \mathsf{n} = n, \mathsf{p} = p_o, \mathsf{alternative} = \mathsf{'less'}).$$

Null hypothesis "$p \leq p_o$", working hypothesis "$p > p_o$":

$$\mathsf{binom.test}(\mathsf{x} = H, \mathsf{n} = n, \mathsf{p} = p_o, \mathsf{alternative} = \mathsf{'greater'}).$$

The chi-squared goodness-of-fit test may be carried out with the commands

$$\mathsf{chisq.test}(\mathsf{x} = \boldsymbol{X}, \mathsf{p} = \boldsymbol{p}^o) \quad \text{or} \quad \mathsf{chisq.test}(\mathsf{x} = \boldsymbol{H}, \mathsf{p} = \boldsymbol{p}^o).$$

Thus one may use the raw data vector $\boldsymbol{X} = (X_i)_{i=1}^n$ or the vector $\boldsymbol{H} = (H_k)_{k=1}^K$ of absolute frequencies as first argument. By the way, the latter vector $\boldsymbol{H}$ may be determined with the command $\mathsf{table}(\boldsymbol{X})$. The argument p is optional with default $(1/K)_{k=1}^K$. To compute a Monte-Carlo p-value by means of m simulations, one can write

$$\mathsf{chisq.test}(\mathsf{x} = \boldsymbol{X}, \mathsf{p} = \boldsymbol{p}^o, \mathsf{simulate.p.value} = \mathsf{TRUE}, \mathsf{B} = m).$$

Chapter III. For a vector $\boldsymbol{X} = (X_i)_{i=1}^n$ of real-valued observations, $\mathsf{sort}(\boldsymbol{X})$ yields the vector $(X_{(i)})_{i=1}^n$ of its order statistics. With $\mathsf{range}(\boldsymbol{X})$ one obtains the pair $(X_{(1)}, X_{(n)})$. The empirical distribution function $\widehat{F}$ may be plotted, for instance, with

$$\mathsf{plot.ecdf}(\boldsymbol{X}) \quad \text{or} \quad \mathsf{plot.ecdf}(\boldsymbol{X}, \mathsf{verticals} = \mathsf{TRUE}).$$

The sample γ-quantile $\widehat{q}_\gamma$ is obtained with

$$\mathsf{quantile}(\boldsymbol{X}, \mathsf{probs} = \gamma, \mathsf{type} = 2).$$

Here $\mathsf{type} = 2$ indicates our convention to use the arithmetic mean of the smallest and largest sample γ-quantile.

To draw confidence bands, the functions $\mathsf{stepfun}(\cdot)$ and $\mathsf{plot}(\mathsf{stepfun}(\cdot))$ might be useful. For Monte-Carlo simulations in connection with confidence bands and elsewhere the function $\mathsf{runif}(\cdot)$ is essential. Precisely, $\mathsf{runif}(n)$ simulates a vector of n independent random variables with uniform distribution on $[0, 1]$.

Chapter IV. Sample mean and sample standard deviation of a data vector $\boldsymbol{X}$ are obtained with $\mathsf{mean}(\boldsymbol{X})$ and $\mathsf{sd}(\boldsymbol{X})$. The function

$$\mathsf{t.test}(\boldsymbol{X}, \mathsf{conf.level} = 1 - \alpha)$$

yields the $(1-\alpha)$-confidence interval $\left[\bar{X} \pm t_{n-1;1-\alpha/2} S_X/\sqrt{n}\right]$ for the underlying theoretical mean. Similarly as with $\mathsf{binom.test}(\cdot)$ one may specify the additional parameter $\mathsf{alternative} = \mathsf{'greater'}$ or $\mathsf{alternative} = \mathsf{'less'}$ to get one-sided bounds.

We mentioned already the function $\mathsf{quantile}(\cdot)$ for sample quantiles; the sample median is also implemented separately as $\mathsf{median}(\cdot)$. The trimmed mean $\overline{X}_\tau$ of $\boldsymbol{X}$ can be determined with $\mathsf{mean}(\boldsymbol{X}, \mathsf{trim} = \tau)$.

The median of absolute deviations (from the median) is implemented as $\mathsf{mad}(\cdot)$, the inter quartile range as $\mathsf{IQR}(\cdot)$. We defined the range of $\boldsymbol{X}$ to be the difference $X_{(n)} - X_{(1)}$, but $\mathsf{range}(\boldsymbol{X})$ provides the pair$(X_{(1)}, X_{(n)})$.

For a vector $\boldsymbol{X} = \boldsymbol{Y} - \boldsymbol{Z}$ of observation differences, Wilcoxon's signed-rank test may be executed with the command $\mathsf{wilcox.test}(\boldsymbol{X})$ or $\mathsf{wilcox.test}(\mathsf{x} = \boldsymbol{Y}, \mathsf{y} = \boldsymbol{Z}, \mathsf{paired} = \mathsf{TRUE})$. This function yields the test statistic $T_o(\boldsymbol{X})$ (!) and an exact p-value $\pi_z(\boldsymbol{X})$. However, all moduli $|X_i|$ have to be pairwise different and non-zero. Otherwise a warning is issued, and R uses a normal approximation. The confidence bounds for the centre of a symmetric distribution may be computed in principle with $\mathsf{wilcox.test}(\mathsf{x} = \boldsymbol{X}, \mathsf{conf.int} = \mathsf{TRUE})$, but for sample sizes $n \geq 50$ certain approximations are used.

Chapter V. For a data vector $\boldsymbol{X}$ and a vector $\boldsymbol{a} = (a_k)_{k=0}^K$ of break points $a_0 < a_1 < \cdots < a_K$, the corresponding histogram is generated by means of

$$\begin{cases} \mathsf{hist}(\boldsymbol{X}, \mathsf{breaks} = \boldsymbol{a}, \mathsf{freq} = \mathsf{TRUE}) & \text{(Convention 1)}, \\ \mathsf{hist}(\boldsymbol{X}, \mathsf{breaks} = \boldsymbol{a}, \mathsf{freq} = \mathsf{FALSE}) & \text{(Convention 2)}. \end{cases}$$

The kernel density estimator $\widehat{f}_h$ with Gaussian kernel $K = \phi$ may be depicted on the interval $[a, b]$ with

$$\mathsf{density}(\boldsymbol{X}, \mathsf{bw} = h, \mathsf{from} = a, \mathsf{to} = b).$$

Different kernel functions K may be specified with the optional argument kernel. They are all standardised such that $\int_{-\infty}^{\infty} K(y) y^2 \, dy = 1$.

Q-Q-Plots are easily implemented. In particular for Gaussian distributions one can also use the function $\mathsf{qqnorm}(\cdot)$.

Chapter VI. The multiple box-and-whiskers plot of $K \geq 2$ data vectors $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_K$ can be generated via

$$\mathsf{boxplot}(\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_K).$$

The single box plots are denoted with the numbers $1, 2, \ldots, K$. With the optional parameter names one may specify alternative annotations:

$$\mathsf{boxplot}(\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_K, \mathsf{names} = \boldsymbol{g}).$$

Here $\boldsymbol{g}$ is a vector of K numbers or character strings (in the latter case with quotation marks), for instance $\boldsymbol{g} = \mathsf{c}(\mathsf{'Basel'}, \mathsf{' Bern'}, \mathsf{' Chur'}, \ldots)$. If the data are provided as a numerical vector $\boldsymbol{X}$ and a categorical vector $\boldsymbol{G}$ with entries $G_i \in \{g_1, g_2, \ldots, g_K\}$, then

$$\mathsf{boxplot}(\boldsymbol{X} \sim \boldsymbol{G})$$

provides a multiple box-and-whiskers plot for the corresponding subvectors $\boldsymbol{X}_k = (X_i)_{i:G_i=g_k}$, $1 \leq k \leq K$.

Suppose one wants to compute confidence bounds for the difference $\mu_1 - \mu_2$ of the means μ_1 and μ_2 underlying two data vectors $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$, respectively. Again the function t.test is applicable. If one assumes identical standard deviations $\sigma_1 = \sigma_2$, then

$$\mathsf{t.test}(\mathsf{x} = \boldsymbol{X}_1, \mathsf{y} = \boldsymbol{X}_2, \mathsf{alternative} = \ldots, \mathsf{conf.level} = 1 - \alpha, \mathsf{var.equal} = \mathsf{TRUE})$$

yields corresponding student confidence bounds. Welch's method for arbitrary standard deviations σ_1, σ_2 is applied when typing

$$\mathsf{t.test}(\mathsf{x} = \boldsymbol{X}_1, \mathsf{y} = \boldsymbol{X}_2, \mathsf{alternative} = \ldots, \mathsf{conf.level} = 1 - \alpha),$$

or one replaces $\mathsf{var.equal} = \mathsf{TRUE}$ with $\mathsf{var.equal} = \mathsf{FALSE}$.

Chapter VII und VIII. With the command $\mathsf{table}(\boldsymbol{X}, \boldsymbol{Y})$ one can generate a contingency table from categorical data vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$. To ensure that the potential values of X_i and Y_i are listed completely and in a specific order, one can replace the data vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$ with $\mathsf{factor}(\boldsymbol{X}, \mathsf{levels} = \mathsf{c}(x_1, x_2, \ldots, x_K))$ and $\mathsf{factor}(\boldsymbol{Y}, \mathsf{levels} = \mathsf{c}(y_1, y_2, \ldots, y_L))$, respectively.

Especially a two-by-two table $\boldsymbol{H}$ may be analyzed with the commands $\mathsf{fisher.test}(\boldsymbol{H})$ or

$$\mathsf{fisher.test}(\boldsymbol{H}, \mathsf{alternative} = \ldots, \mathsf{conf.level} = 1 - \alpha).$$

Alternatively one could write $\mathsf{fisher.test}(\mathsf{x} = \boldsymbol{X}, \mathsf{y} = \boldsymbol{Y})$ or

$$\mathsf{fisher.test}(\mathsf{x} = \boldsymbol{X}, \mathsf{y} = \boldsymbol{Y}, \mathsf{alternative} = \ldots, \mathsf{conf.level} = 1 - \alpha).$$

This yields a two-sided p-value with Fisher's exact test for the null hypothesis that $\rho = 1$, combined with a $(1 - \alpha)$-confidence interval for ρ.

Analogously, the chi-squared test for association may be applied with the command

$$\mathsf{chisq.test}(\boldsymbol{H}) \quad \text{or} \quad \mathsf{chisq.test}(\mathsf{x} = \boldsymbol{X}, \mathsf{y} = \boldsymbol{Y}).$$

These commands yield the numerical value of the chi-squared test statistic and an approximate p-value via the chi-squared distribution with $(K - 1)(L - 1)$ degrees of freedom. If some row or column numbers are rather small, a corresponding warning is issued. The variant

$$\mathsf{chisq.test}(\ldots, \mathsf{simulate.p.value} = \mathsf{TRUE}, \mathsf{B} = m)$$

leads to a Monte-Carlo p-value for a permutation test with m pseudo-random permutations.

If one wants to implement a Monte-Carlo permutation test by oneself, one may use the function $\mathsf{sample}()$. The command $\mathsf{sample}(\boldsymbol{Y})$ simulates a random permutation $\Pi\boldsymbol{Y}$, and with $\mathsf{sample}(n)$ one can simulate a random permutation $\Pi \in \mathcal{S}_n$, represented as a tuple $(\Pi(i))_{i=1}^n$.

The parameters of the regression line for data vectors $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{R}^n$ are computed when typing $\mathsf{lm}(\boldsymbol{Y} \sim \boldsymbol{X})$.[1] The corresponding correlation coefficients are provided by

$$\mathsf{cor}(\mathsf{x} = \boldsymbol{X}, \mathsf{y} = \boldsymbol{Y}) \quad \text{or} \quad \mathsf{cor}(\mathsf{x} = \boldsymbol{X}, \mathsf{y} = \boldsymbol{Y}, \mathsf{method} = \mathsf{'spearman'}).$$

With $\mathsf{cor.test}(\ldots)$ instead of $\mathsf{cor}(\ldots)$ one obtains p-values for the null hypothesis that there is no true association between X- and Y-values. In particular,

$$\mathsf{cor.test}(\mathsf{x} = \boldsymbol{X}, \mathsf{y} = \boldsymbol{Y}, \mathsf{method} = \mathsf{'spearman'}, \mathsf{exact} = \mathsf{TRUE})$$

performs an exact permutation test, provided the components of both $\boldsymbol{X}$ and $\boldsymbol{Y}$ are pairwise different.

[1] The function $\mathsf{lm}(\ldots)$ offers many more methods for so-called linear models.

A.2 Affine Transformations of Random Variables

Let X be a real-valued random variable with distribution function F, and let

$$Y \;=\; a + bX$$

with constants $a \in \mathbb{R}$ and $b > 0$. This is equivalent to saying that

$$X \;=\; \frac{Y - a}{b}.$$

Since $Y \le y$ if and only if $X \le (y-a)/b$, the distribution function G of Y is given by

$$G(y) := \mathbb{P}(Y \le y) \;=\; F\Big(\frac{y-a}{b}\Big).$$

The corresponding quantile functions F^{-1} and G^{-1} satisfy

$$G^{-1}(u) \;=\; a + bF^{-1}(u)$$

for $0 < u < 1$.

If F is given by a density f, i.e. $F(x) = \int_{-\infty}^{x} f(t)\,dt$, then G is given by the density

$$g(y) \;:=\; \frac{1}{b} f\Big(\frac{y-a}{b}\Big).$$

In case of a continuous density f, this follows from computing G' via the chain rule. In general, with $t(v) := (v-a)/b$ for $v \in \mathbb{R}$,

$$G(y) \;=\; \int_{-\infty}^{(y-a)/b} f(t)\,dt \;=\; \int_{-\infty}^{y} f(t(v))t'(v)\,dv \;=\; \int_{-\infty}^{y} f((v-a)/b)/b\,dv.$$

A.3 Weak Convergence of Distributions

For $n = 1, 2, 3, \ldots$ let X_n be a random variable with distribution P_n on $\mathbb{R}^d$ (equipped with its Borel σ-field). Further let X be a random variable with distribution P on $\mathbb{R}^d$.

Definition A.1 (Convergence in distribution; weak convergence). One says that "X_n converges in distribution to X (as $n \to \infty$)" and writes

$$X_n \;\to_{\mathcal{L}}\; X$$

if

$$\lim_{n\to\infty} \mathbb{E}(f(X_n)) \;=\; \mathbb{E}(f(X))$$

for any bounded and continuous function $f : \mathbb{R}^d \to \mathbb{R}$.

This is equivalent to the following statement about the distributions P_n: One says that "*P_n converges weakly to P (as $n \to \infty$)*" and writes

$$P_n \;\to_{w}\; P$$

if

$$\lim_{n\to\infty} \int f(x)\,P_n(dx) \;=\; \int f(x)\,P(dx)$$

for any bounded and continuous function $f : \mathbb{R}^d \to \mathbb{R}$.

In statistics, these facts are often stated as "X_n *has asymptotic distribution* P *(as* $n \to \infty$*)*". Sometimes one also writes "$X_n \to_{\mathcal{L}} P$".

To verify these statements, it suffices to consider functions $f : \mathbb{R}^d \to \mathbb{R}$ which are infinitely often differentiable and have compact support. Here is yet another characterisation:

The sequence $(X_n)_n$ converges in distribution to X if and only if

$$\limsup_{n\to\infty} \mathbb{P}(X_n \in A) \;\leq\; \mathbb{P}(X \in A)$$

for arbitrary closed sets $A \subset \mathbb{R}^d$. This is equivalent to the statement that

$$\liminf_{n\to\infty} \mathbb{P}(X_n \in U) \;\geq\; \mathbb{P}(X \in U)$$

for arbitrary open sets $U \subset \mathbb{R}^d$.

In the special case $d = 1$, convergence in distribution and weak convergence may be characterised by means of the distribution functions F_n and F of X_n and X, respectively:

$$\lim_{n\to\infty} F_n(x) \;=\; F(x) \tag{A.1}$$

for any point $x \in \mathbb{R}$ at which F is continuous. If the distribution function F is continuous, then property (A.1) is even equivalent to

$$\lim_{n\to\infty} \sup_{\text{intervals } B \subset \mathbb{R}} \big|P_n(B) - P(B)\big| \;=\; 0.$$

A.4 Lindeberg's Central Limit Theorem

The univariate case. The Central Limit Theorem specifies the vague statement that a sum of independent random variables has approximately a Gaussian distribution if each single summand has only small influence on the total sum.

Theorem A.2. *Let* $Y_1, Y_2, \ldots, Y_n$ *be stochastically independent random variables with* $\mathbb{E}(Y_i) = 0$ *and*

$$\sum_{i=1}^{n} \mathrm{Var}(Y_i) \;=\; \sum_{i=1}^{n} \mathbb{E}(Y_i^2) \;=\; 1.$$

Further let

$$L \;:=\; \sum_{i=1}^{n} \mathbb{E}\big(Y_i^2 \min(1, |Y_i|)\big).$$

Then

$$\sup_{\text{intervals } B \subset \mathbb{R}} \big|\mathbb{P}(Y \in B) - \mathcal{N}(0,1)(B)\big| \;\to\; 0 \quad \text{as } L \to 0.$$

The quantity L measures the influence of the single summands Y_i on the total sum. For instance, if $|Y_i| \leq \kappa$ almost surely for all indices i and a constant κ, then

$$L \;\leq\; \sum_{i=1}^{n} \mathbb{E}(Y_i^2 \kappa) \;=\; \kappa.$$

The previous formulation of the Central Limit Theorem is similar in spirit to the versions of J.W. Lindeberg[2] and A.M. Ljapunov[3] . Ljapunov considered the critical quantity $\sum_{i=1}^{n} \mathbb{E}(|Y_i|^3) \geq L$.

[2] Jarl W. Lindeberg (1876-1932): Finnish mathematician.

[3] Alexander M. Ljapunov (1857-1918): Russian mathematician and physicist.

Example A.3 (Binomial distributions). If $X \sim \mathrm{Bin}(n,p)$, then the standardised quantity $Y := (X-np)/\sqrt{np(1-p)}$ has approximately a standard Gaussian distribution as $np(1-p) \to \infty$. To verify this, we write $Y = \sum_{i=1}^n Y_i$ with summands $Y_i := (X_i - p)/\sqrt{np(1-p)}$ and stochastically independent, $\{0,1\}$-valued random variables $X_1, X_2, \ldots, X_n$, where $\mathbb{P}(X_i = 1) = \mathbb{E}(X_i) = p$, so $\mathrm{Var}(X_i) = p(1-p)$. Obviously, $|Y_i| \le 1/\sqrt{np(1-p)}$, whence $L \le 1/\sqrt{np(1-p)}$.

Example A.4 (Sample means). Let $X_1, X_2, \ldots, X_n$ be stochastically independent and identically distributed random variables with mean μ and standard deviation $\sigma > 0$. Then

$$Y := n^{1/2}(\bar{X}_n - \mu)/\sigma \;=\; \sum_{i=1}^n Y_i$$

with $Y_i := n^{-1/2}(X_i - \mu)/\sigma$, and

$$L \;=\; \mathbb{E}\Big(\frac{(X_1-\mu)^2}{\sigma^2}\min\Big(1, \frac{|X_1-\mu|}{\sqrt{n}\,\sigma}\Big)\Big).$$

For fixed distribution of X_1, the latter quantity converges to zero as $n \to \infty$.

The multivariate case. For a random vector $\boldsymbol{Y} = (Y_k)_k \in \mathbb{R}^K$ and a random matrix $\boldsymbol{M} = (M_{k\ell})_{k,\ell} \in \mathbb{R}^{K\times L}$, their expectation is defined component-wise, i.e. $\mathbb{E}(\boldsymbol{Y}) := \big(\mathbb{E}(Y_k)\big)_k$ and $\mathbb{E}(\boldsymbol{M}) := \big(\mathbb{E}(M_{k\ell})\big)_{k,\ell}$.

Theorem A.5. *For $n \in \mathbb{N}$ let $\boldsymbol{Y}_n = \sum_{i=1}^n \boldsymbol{Y}_{ni}$ with stochastically independent random vectors $\boldsymbol{Y}_{ni} \in \mathbb{R}^K$ such that $\mathbb{E}(\boldsymbol{Y}_{ni}) = \boldsymbol{0}$ and $\mathbb{E}\big(\|\boldsymbol{Y}_{ni}\|^2\big) < \infty$. Suppose that the following two conditions as satisfied as $n \to \infty$:*

$$\boldsymbol{\Sigma}_n := \sum_{i=1}^n \mathbb{E}(\boldsymbol{Y}_{ni}\boldsymbol{Y}_{ni}^\top) \;\to\; \boldsymbol{\Sigma}$$

for a symmetric, positive semidefinite matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{K\times K}$, and

$$L_n := \sum_{i=1}^n \mathbb{E}\big(\|\boldsymbol{Y}_{ni}\|^2 \min(1, \|\boldsymbol{Y}_{ni}\|)\big) \;\to\; 0.$$

Then $\boldsymbol{Y}_n$ converges in distribution to a Gaussian random vector $\boldsymbol{Y}$ with mean $\boldsymbol{0}$ and covariance matrix $\boldsymbol{\Sigma}$.

The statement that a random vector $\boldsymbol{Y} \in \mathbb{R}^K$ is Gaussian with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ may be circumscribed as follows: If we write $\boldsymbol{\Sigma}$ as a sum $\sum_{k=1}^K \lambda_k \boldsymbol{u}_k \boldsymbol{u}_k^\top$ with eigenvalues $\lambda_1, \ldots, \lambda_K \ge 0$ and orthonormal eigenvectors $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_K$, then $\boldsymbol{Y}$ has the same distribution as $\boldsymbol{\mu} + \sum_{k=1}^K \sqrt{\lambda_k}\, Z_k\, \boldsymbol{u}_k$ with independent, standard Gaussian random variables $Z_1, \ldots, Z_K \in \mathbb{R}$.

Example A.6 (Multinomial distributions). For $n \in \mathbb{N}$ let $X_{n1}, X_{n2}, \ldots, X_{nn}$ be stochastically independent random variables with values in $\{1, 2, \ldots, K\}$, where $\mathbb{P}(X_{ni} = k) = p_{nk} > 0$ for $1 \le k \le K$. Then $\boldsymbol{H}_n = (H_{nk})_{k=1}^K$ with $H_{nk} := \#\{i \le n : X_{ni} = k\}$ follows a multinomial distribution with parameters n and $\boldsymbol{p}_n = (p_{nk})_{k=1}^K \in (0,1)^K$.

Suppose that the sequence $(\boldsymbol{p}_n)_n$ converges to a probability vector $\boldsymbol{p} = (p_k)_{k=1}^K \in [0,1]^K$, where

$$\lim_{n\to\infty} \min_{k=1,2,\ldots,K} np_{nk} \;=\; \infty.$$

Then the random vector

$$\boldsymbol{Y}_n \;:=\; \Big(\frac{H_{nk} - np_{nk}}{\sqrt{np_{nk}}}\Big)_k$$

converges in distribution to a Gaussian random vector $\boldsymbol{Y}$ with mean $\mathbf{0}$ and covariance matrix

$$\boldsymbol{\Sigma} \;:=\; \boldsymbol{I} - \sqrt{\boldsymbol{p}}\sqrt{\boldsymbol{p}}^{\top},$$

where $\sqrt{\boldsymbol{p}} := \big(\sqrt{p_k}\big)_{k=1}^{K}$. This particular limit distribution may be described as follows: If we complement the unit vector $\sqrt{\boldsymbol{p}}$ to an orthonormal basis $\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_{K-1}, \sqrt{\boldsymbol{p}}$ of $\mathbb{R}^K$, then $\boldsymbol{Y}$ has the same distribution as

$$\sum_{j=1}^{K-1} Z_j \boldsymbol{b}_j$$

with independent, standard Gaussian random variables $Z_1, Z_2, \ldots, Z_{K-1}$.

That $\boldsymbol{Y}_n$ converges in distribution to $\boldsymbol{Y}$ follows from the Central Limit Theorem. For $\boldsymbol{Y}_n = \sum_{i=1}^{n} \boldsymbol{Y}_{ni}$ with the independent random vectors

$$\boldsymbol{Y}_{ni} \;:=\; \Big(\frac{1_{[X_{ni}=k]} - p_{nk}}{\sqrt{np_{nk}}}\Big)_k$$

which have the following properties:

$$\mathbb{E}(\boldsymbol{Y}_{ni}) \;=\; \mathbf{0}, \quad \mathbb{E}(\boldsymbol{Y}_{ni}\boldsymbol{Y}_{ni}^{\top}) \;=\; n^{-1}\big(\boldsymbol{I} - \sqrt{\boldsymbol{p}_n}\sqrt{\boldsymbol{p}_n}^{\top}\big),$$

and

$$\|\boldsymbol{Y}_{ni}\| \;\le\; \kappa_n := \sqrt{\Big(\min_{k=1,2,\ldots,K} np_{nk}\Big)^{-1} + n^{-1}}.$$

In particular, $\boldsymbol{\Sigma}_n = \boldsymbol{I} - \sqrt{\boldsymbol{p}_n}\sqrt{\boldsymbol{p}_n}^{\top}$ converges to $\boldsymbol{\Sigma}$, and $L_n \le \mathrm{Spur}(\boldsymbol{\Sigma}_n)\kappa_n$ converges to 0 as $n \to \infty$.

A.5 Fubini's Theorem

Fubini's[4] theorem is a general result from measure theory. Here we restrict ourselves to a version involving independent random variables.

Let Z_1 and Z_2 be stochastically independent random variables with values in measurable spaces $(\mathcal{Z}_1, \mathcal{B}_1)$ and $(\mathcal{Z}_2, \mathcal{B}_2)$, respectively. Further let $H = h(Z_1, Z_2)$ with a measurable function $h : \mathcal{Z}_1 \times \mathcal{Z}_2 \to \mathbb{R}$ such that $h \ge 0$ or $\mathbb{E}(|H|) < \infty$. For fixed points $z_j \in \mathcal{Z}_j$ let

$$h_1(z_1) \;:=\; \mathbb{E}(h(z_1, Z_2)) \quad \text{and} \quad h_2(z_2) \;:=\; \mathbb{E}(h(Z_1, z_2)).$$

Fubini's theorem says that for $j = 1, 2$, the set B_j of all $z_j \in \mathcal{Z}_j$ such that $h_j(z_j)$ is well-defined, satisfies $\mathbb{P}(Z_j \in B_j) = 1$, and

$$\mathbb{E}(H) \;=\; \mathbb{E}(h_1(Z_1)) \;=\; \mathbb{E}(h_2(Z_2)).$$

Instead of $h_j(z_j)$ one often writes $\mathbb{E}(H \,|\, Z_j = z_j)$, and $h_j(Z_j)$ is often written as $\mathbb{E}(H \,|\, Z_j)$. Thus

$$\mathbb{E}(H) \;=\; \mathbb{E}\big(\mathbb{E}(H \,|\, Z_j)\big).$$

In particular for events A which can be defined in terms of Z_1 and Z_2,

$$\mathbb{P}(A) \;=\; \mathbb{E}\big(\mathbb{P}(A \,|\, Z_1)\big) \;=\; \mathbb{E}\big(\mathbb{P}(A \,|\, Z_2)\big).$$

[4] Guido Fubini (1879-1943): Italian mathematician.

A.6 Jensen's Inequality

Jensen's[5] Inequality is one of the most important inequalities in probability theory and analysis. We consider a random variable X with values in an interval $J \subset \mathbb{R}$ and finite expectation $\mathbb{E}(|X|)$. Then $\mathbb{E}(X)$ is necessarily a number in J, and for any convex function $\psi : J \to \mathbb{R}$,

$$\mathbb{E}(\psi(X)) \;\geq\; \psi(\mathbb{E}(X)).$$

If ψ is even strictly convex, then

$$\mathbb{E}(\psi(X)) \;>\; \psi(\mathbb{E}(X)) \quad \text{or} \quad \mathbb{P}(X = \mathbb{E}(X)) \;=\; 1.$$

Proof of Jensen's inequality. A function $\psi : J \to \mathbb{R}$ is called convex if for arbitrary $x_0, x_1 \in J$ and $\lambda \in [0,1]$,

$$\psi((1-\lambda)x_0 + \lambda x_1) \;\leq\; (1-\lambda)\psi(x_0) + \lambda\psi(x_1).$$

One calls ψ strictly convex if the preceding inequality is strict whenever $x_0 \neq x_1$ and $0 < \lambda < 1$. Suppose that $\mu := \mathbb{E}(X)$ equals $a := \inf(J)$ or $b := \sup(J)$. This implies that $\mathbb{P}(X = \mu) = 1$, and the stated inequality is trivial. Hence let $a < \mu < b$. One can derive from convexity of ψ that the function

$$J \setminus \{\mu\} \ni x \;\mapsto\; \frac{\psi(x) - \psi(\mu)}{x - \mu}$$

is monotone increasing. In particular the two limits

$$\psi'(\mu-) \;:=\; \lim_{x \uparrow \mu} \frac{\psi(x) - \psi(\mu)}{x - \mu} \quad \text{and} \quad \psi'(\mu+) \;:=\; \lim_{x \downarrow \mu} \frac{\psi(x) - \psi(\mu)}{x - \mu}$$

exist in $\mathbb{R}$, where $\psi'(\mu-) \leq \psi'(\mu+)$. For any number $\psi'(\mu) \in \big[\psi'(\mu-), \psi'(\mu+)\big]$ and $x \in J$ this implies the inequality

$$\psi(x) \;\geq\; \psi(\mu) + \psi'(\mu)(x - \mu).$$

Consequently,

$$\mathbb{E}(\psi(X)) \;\geq\; \mathbb{E}\big(\psi(\mu) + \psi'(\mu)(X - \mu)\big) \;=\; \psi(\mu).$$

In case of a strictly convex function ψ, the difference ratio $(\psi(x) - \psi(\mu))/(x - \mu)$ is even strictly monotone increasing in $x \in J \setminus \{\mu\}$, whence $\psi(x) > \psi(\mu) + \gamma(x - \mu)$ for $x \in J \setminus \{\mu\}$. Then $\mathbb{E}(\psi(X)) > \psi(\mu)$, unless $\mathbb{P}(X \neq \mu) = 0$. □

A.7 Technical Details about Student Distributions

We first note that chi-squared distributions are special gamma distributions:

Theorem A.7. *For any natural number k,*

$$\chi^2_k \;=\; \mathrm{Gamma}(k/2, 2).$$

The student distribution with $k \in \mathbb{N}$ degrees of freedom, t_k, was defined as the distribution of

$$Z_0 \Big/ \sqrt{\frac{1}{k} \sum_{i=1}^{k} Z_i^2}$$

with independent, standard Gaussian random variables Z_0, Z_1, …, Z_k. With Theorem A.7 in mind, one can extend this definition to non-integer parameters k:

[5] Johan Jensen (1859-1925): Danish mathematician and engineer.

Definition A.8 (Student distributions in general). *Student's t distribution with $k > 0$ degrees of freedom*, denoted by t_k, is defined as the distribution of

$$Z/\sqrt{G_k/k}$$

with stochastically independent random variables $Z \sim \mathcal{N}(0,1)$ and $G_k \sim \mathrm{Gamma}(k/2, 2)$. Its β-quantile is denoted with $t_{k;\beta}$.

Now we derive and investigate the density functions and quantiles of these distributions t_k:

Theorem A.9. *For each $k > 0$, the student distribution t_k has a density function f_k given by*

$$f_k(x) \;=\; \frac{\Gamma((k+1)/2)}{\sqrt{k\pi}\Gamma(k/2)}\Big(1 + \frac{x^2}{k}\Big)^{-(k+1)/2}.$$

For each $x \in \mathbb{R}$,

$$\lim_{k\to\infty} f_k(x) \;=\; \phi(x),$$

and

$$\int_{-\infty}^{\infty} \big|f_k(x) - \phi(x)\big|\,dx \;=\; O(k^{-1/2}).$$

Since student distributions are obviously symmetric around zero, they satisfy $t_{k;1/2} = 0$ and

$$t_{k;1-\beta} \;=\; -t_{k;\beta}.$$

From the last statement of Theorem A.9 and strict monotonicity of Φ one can easily deduce that

$$\lim_{k\to\infty} t_{k;\beta} \;=\; \Phi^{-1}(\beta).$$

Furthermore, student quantiles and densities satisfy certain monotonicity properties with respect to the parameter $k > 0$:

Theorem A.10. *For $1/2 < \beta < 1$, the quantile $t_{k;\beta}$ is strictly monotone decreasing in $k > 0$, while $f_k(0)$ is strictly monotone increasing in $k > 0$.*

Proof of Theorem A.7. Essentially we have to show that the random sum $Y := \sum_{i=1}^k Z_i^2/2$ with independent, standard Gaussian random variables Z_i has a gamma distribution with parameters $k/2$ and 1. That means,

$$\mathbb{P}(Y \le y) \;=\; \Gamma(k/2)^{-1} \int_0^y x^{k/2-1} e^{-x}\,dx \quad \text{for } y > 0. \tag{A.2}$$

The random vector $\boldsymbol{Z} := (Z_i)_{i=1}^k$ has density function

$$\phi_k(\boldsymbol{z}) \;:=\; C_k e^{-\|\boldsymbol{z}\|^2/2}$$

on $\mathbb{R}^k$, where $C_k = (2\pi)^{-k/2}$. Now we use polar coordinates, i.e. we write $\boldsymbol{z} \in \mathbb{R}^k \setminus \{\boldsymbol{0}\}$ as $\boldsymbol{z} = r\boldsymbol{u}$ with radius $r = \|\boldsymbol{z}\|$ and directional vector $\boldsymbol{u} = r^{-1}\boldsymbol{z}$. For functions $h : [0,\infty) \to [0,\infty)$ it is well-known that $\int_{\mathbb{R}^k} h(\|\boldsymbol{z}\|)\,d\boldsymbol{z} = C_k' \int_0^\infty r^{k-1} h(r)\,dr$ with a certain constant $C_k' > 0$. Hence

for arbitrary $y > 0$,

$$\begin{aligned}
\mathbb{P}(Y \le y) &= C_k \int_{\mathbb{R}^k} 1_{[\|\boldsymbol{z}\|^2/2 \le y]} e^{-\|\boldsymbol{z}\|^2/2}\, d\boldsymbol{z} \\
&= C_k C_k' \int_0^\infty r^{k-1} 1_{[r^2/2 \le y]} e^{-r^2/2}\, dr \\
&= 2^{(k-1)/2} C_k C_k' \int_0^\infty (r^2/2)^{(k-1)/2} 1_{[r^2/2 \le y]} e^{-r^2/2}\, dr \\
&= 2^{(k-1)/2} C_k C_k' \int_0^\infty x^{(k-1)/2} 1_{[x \le y]} e^{-x}\, (2x)^{-1/2} dx \\
&= 2^{k/2-1} C_k C_k' \int_0^y x^{k/2-1} e^{-x}\, dx.
\end{aligned}$$

Here in the second last step we employed the transformation $x = r^2/2$, so $r = (2x)^{1/2}$ and $dr = (2x)^{-1/2} dx$. As $y \to \infty$, the probability $\mathbb{P}(Y \le y)$ converges to 1, and we obtain the formula $2^{k/2-1} C_k C_k' = \Gamma(k/2)^{-1}$. This implies for fixed $y > 0$ the asserted equation (A.2). □

Proof of Theorem A.9. We consider the random variable $Z/\sqrt{Y_a/a}$ with independent random variables $Z \sim \mathcal{N}(0,1)$ and $Y_a \sim \mathrm{Gamma}(a,1)$, where $a := k/2$. Stochastic independence of Z and Y_a and Fubini's theorem imply that the distribution function F_k of $Z/\sqrt{Y_a/a}$ has the following form:

$$\begin{aligned}
F_k(x) := \mathbb{P}\big(Z/\sqrt{Y_a/a} \le x\big) &= \mathbb{P}\big(Z \le x\sqrt{Y_a/a}\big) \\
&= \int_0^\infty \mathbb{P}\big(Z \le x\sqrt{y/a}\big) g_a(y)\, dy \\
&= \int_0^\infty \Phi\big(x\sqrt{y/a}\big) g_a(y)\, dy.
\end{aligned}$$

Here g_a denotes the density function of $\mathrm{Gamma}(a,1)$. But the derivative of $\Phi(x\sqrt{y/a})$ with respect to x equals $\phi(x\sqrt{y/a})\sqrt{y/a}$, so $\Phi(x\sqrt{y/a}) = \int_{-\infty}^x \phi(t\sqrt{y/a})\sqrt{y/a}\, dt$. Hence a further application of Fubini's theorem yields

$$\begin{aligned}
F_k(x) &= \int_0^\infty \int_{-\infty}^x \phi(t\sqrt{y/a})\sqrt{y/a}\, g_a(y)\, dt\, dy \\
&= C_k \int_{-\infty}^x \int_0^\infty y^{(k+1)/2-1} e^{-(1+t^2/k)y}\, dy\, dt \\
&= C_k \int_{-\infty}^x (1+t^2/k)^{-(k+1)/2} \int_0^\infty \tilde{y}^{(k+1)/2-1} e^{-\tilde{y}}\, d\tilde{y}\, dt \\
&= C_k' \int_{-\infty}^x (1+t^2/k)^{-(k+1)/2}\, dt
\end{aligned}$$

with $C_k := \big(\sqrt{k\pi}\Gamma(k/2)\big)^{-1}$ and $C_k' := \Gamma((k+1)/2) C_k$. Here in the second last step we used the transformation $y \mapsto \tilde{y} := (1+t^2/k)y$, and the last step involves the definition of $\Gamma((k+1)/2)$. Concerning the additional statements about the density functions f_k,

$$\mathbb{E}(Y_a^u) = \frac{\Gamma(a+u)}{\Gamma(a)} \quad \text{for } u > -a.$$

This, together with the well-known identity $\Gamma(b+1) = b\Gamma(b)$, yields the equations $\mathbb{E}(Y_a) = a$ and $E(Y_a^2) = (a+1)a$, whence $\mathrm{Var}(Y_a) = a$. Now we may write

$$f_k(0) = \frac{\Gamma((k+1)/2)}{\sqrt{k\pi}\Gamma(k/2)} = \frac{\mathbb{E}\big(\sqrt{Y_a/a}\big)}{\sqrt{2\pi}},$$

and this converges to $\phi(0)$ as $k \to \infty$. For

$$\big|\mathbb{E}\big(\sqrt{Y_a/a}\big) - 1\big| \;\le\; \mathbb{E}\big(\big|\sqrt{Y_a/a} - 1\big|\big) \;\le\; \mathbb{E}(|Y_a/a - 1|) \;\le\; \mathrm{Std}(Y_a/a) \;=\; 1/\sqrt{a}.$$

Now for an arbitrary fixed number $x \in \mathbb{R}$ one can write

$$\begin{aligned} f_k(x) \;&=\; f_k(0)\exp\big(-(k+1)\log(1+x^2/k)/2\big) \\ &=\; (\phi(0)+o(1))\exp\big(-(k+1)(x^2/k + O(k^{-2}))/2\big) \\ &=\; (\phi(0)+o(1))\exp\big(-x^2/2 + O(k^{-1})\big) \\ &\to\; \phi(x) \quad (k \to \infty). \end{aligned}$$

Finally, to verify that $\int_{-\infty}^{\infty} \big|f_k(x) - \phi(x)\big|\,dx = O(k^{-1/2})$, we note first that

$$\begin{aligned} \int_{-\infty}^{\infty} \big|f_k(x) - \phi(x)\big|\,dx \;&=\; \int_{\{f_k > \phi\}} \big(f_k(x) - \phi(x)\big)\,dx + \int_{\{f_k \le \phi\}} \big(\phi(x) - f_k(x)\big)\,dx \\ &=\; 2\int_{\{f_k \le \phi\}} \big(\phi(x) - f_k(x)\big)\,dx \end{aligned}$$

because $\int_{-\infty}^{\infty} \big(f_k(x) - \phi(x)\big)\,dx = 0$, and

$$\int_{\{f_k \le \phi\}} \big(\phi(x) - f_k(x)\big)\,dx \;=\; \int_{\{f_k \le \phi\}} f_k(x)\Big(\frac{\phi(x)}{f_k(x)} - 1\Big)\,dx \;\le\; \sup_{x\in\mathbb{R}} \frac{\phi(x)}{f_k(x)} - 1.$$

But

$$\begin{aligned} \sup_{x\in\mathbb{R}} \frac{\phi(x)}{f_k(x)} \;&=\; \sup_{x\in\mathbb{R}} \exp\big((k+1)\log(1+x^2/k)/2 - x^2/2\big)\,\frac{\phi(0)}{f_k(0)} \\ &=\; \exp\Big(\sup_{y\ge 0}\big((k+1)\log(1+y/k) - y\big)/2\Big) \big/ \mathbb{E}\big(\sqrt{Y_a/a}\big) \\ &=\; \exp\Big(\big((k+1)\log(1+1/k) - 1\big)/2\Big) \big/ \big(1 + O(a^{-1/2})\big) \\ &\le\; \exp((2k)^{-1}) \big/ \big(1 + O(k^{-1/2})\big) \\ &=\; 1 + O(k^{-1/2}). \end{aligned}$$

□

The proof of Theorem A.10 is essentially an application of the following lemma about gamma random variables.

Lemma A.11. *For $a > 0$ let Y_a be a random variable with distribution $\mathrm{Gamma}(a,1)$. For any convex, non-linear function $\psi : (0,\infty) \to \mathbb{R}$, the expectation $\mathbb{E}(\psi(Y_a/a))$ is strictly monotone decreasing in $a > 0$.*

Proof of Lemma A.11. For $x > 0$ we have $\mathbb{P}(Y_a/a \le x) = \Gamma(a)^{-1}\int_0^{ax} y^{a-1}e^{-y}\,dy$. Differentiating this with respect to x shows that Y_a/a has density function

$$\tilde{g}_a(x) \;:=\; \Gamma(a)^{-1}a^a x^{a-1}e^{-ax}, \quad x > 0.$$

Thus for fixed parameters $0 < a < b$ and arbitrary numbers $x > 0$,

$$\rho(x) := \frac{\tilde{g}_b(x)}{\tilde{g}_a(x)} \;=\; C \cdot (xe^{-x})^{b-a},$$

where $C = C(a,b) > 0$. This likelihood ratio ρ is continuous, and it is strictly monotone increasing on $(0,1]$ and strictly decreasing on $[1,\infty)$ with limit 0 as $x \to 0$ or $x \to \infty$. Moreover,

$\rho(1) > 1$ because otherwise $\int_0^\infty \tilde{g}_b(x)\,dx = \int_0^\infty \rho(x)\tilde{g}_a(x)\,dx < 1$. Hence there exist numbers $0 < x_1 < x_2$ with

$$\rho(x) \begin{cases} > 1 & \text{for } x \in (x_1, x_2), \\ < 1 & \text{for } x \in (0,\infty) \setminus (x_1, x_2). \end{cases}$$

Now we utilise that $\mathbb{E}(Y_a/a) = \mathbb{E}(Y_b/b) = 1$, so $\int_0^\infty x(\rho(x)-1)\tilde{g}_a(x)\,dx = 0 = \int_0^\infty (\rho(x)-1)\tilde{g}_a(x)\,dx$. Consequently,

$$\begin{aligned} \mathbb{E}(\psi(Y_b/b)) - \mathbb{E}(\psi(Y_a/a)) &= \int_0^\infty \psi(x)\tilde{g}_a(x)(\rho(x)-1)\,dx \\ &= \int_0^\infty (\psi(x) - c - dx)(\rho(x)-1)\tilde{g}_a(x)\,dx \end{aligned}$$

for arbitrary $c, d \in \mathbb{R}$. If we choose c and d such that $c + dx_1 = \psi(x_1)$ and $c + dx_2 = \psi(x_2)$, then it follows from convexity of ψ that

$$\psi(x) - c - dx \begin{cases} \le 0 & \text{for } x \in [x_1, x_2], \\ \ge 0 & \text{for } x \in (0, x_1] \cup [x_2, \infty). \end{cases}$$

In particular, $(\psi(x) - c - dx)(\rho(x)-1) \le 0$ for arbitrary $x > 0$, so $\mathbb{E}(\psi(Y_b/b)) - \mathbb{E}(\psi(Y_a/a)) \le 0$. Equality may hold only if $\psi(x) = c + dx$ for almost all $x > 0$. By convexity of ψ this would be equivalent to $\psi(x) = c + dx$ for all $x > 0$. □

Proof of Theorem A.10. Let Z and Y_a be stochastically independent, where $Z \sim \mathcal{N}(0,1)$ and $Y_a \sim \text{Gamma}(a,1)$ with $a = k/2$. Then $Z/\sqrt{Y_a/a}$ has distribution t_k, and in the proof of Theorem A.9 we saw that

$$f_k(0) = \mathbb{E}\big(\sqrt{Y_a/a}\big)/\sqrt{2\pi}.$$

According to Lemma A.11, this is strictly monotone increasing in $k > 0$, because $-\sqrt{x}$ is strictly convex in $x \ge 0$.

Now we consider the distribution function F_k of t_k at a fixed point $t > 0$. In the proof of Theorem A.9 we showed that

$$F_k(t) = 1 - \mathbb{E}\big(\Phi\big(-t\sqrt{Y_a/a}\big)\big).$$

Elementary calculations show that $\Phi(-t\sqrt{y})$ is a strictly convex function of $y \ge 0$. Thus it follows from Lemma A.11 that $F_k(t)$ is strictly monotone increasing in $k > 0$. For $k' > k$ this implies that $\beta = F_k(t_{k;\beta}) < F_{k'}(t_{k;\beta})$, whence $t_{k';\beta} < t_{k;\beta}$. □

A.8 Consistency of Empirical Distribution Functions

The results about $\|\widehat{F} - F\|_\infty$ stated at the end of Chapter 3 are based on the theory of *empirical processes*, a field at the intersection of probability theory and statistics. In particular, it is shown there that for large sample sizes n, the stochastic process (the random function) $\sqrt{n}(\widehat{F} - F)$ behaves essentially like

$$B \circ F$$

with a *Brownian bridge* $B = (B(t))_{t\in[0,1]}$. The latter is a stochastic process with remarkable properties. For instance, B is continuous with $B(0) = B(1) = 0$, but it is nowhere differentiable. Such results are beyond the scope of this book, but we illustrate them by means of a few simulations:

Figure A.1 shows for two samples of size $n = 100$ and $n = 1000$, respectively, from the Gaussian distribution $\mathcal{N}(100, 15^2)$ the functions F and $\widehat{F}$ in the upper panels. In the lower panels one sees

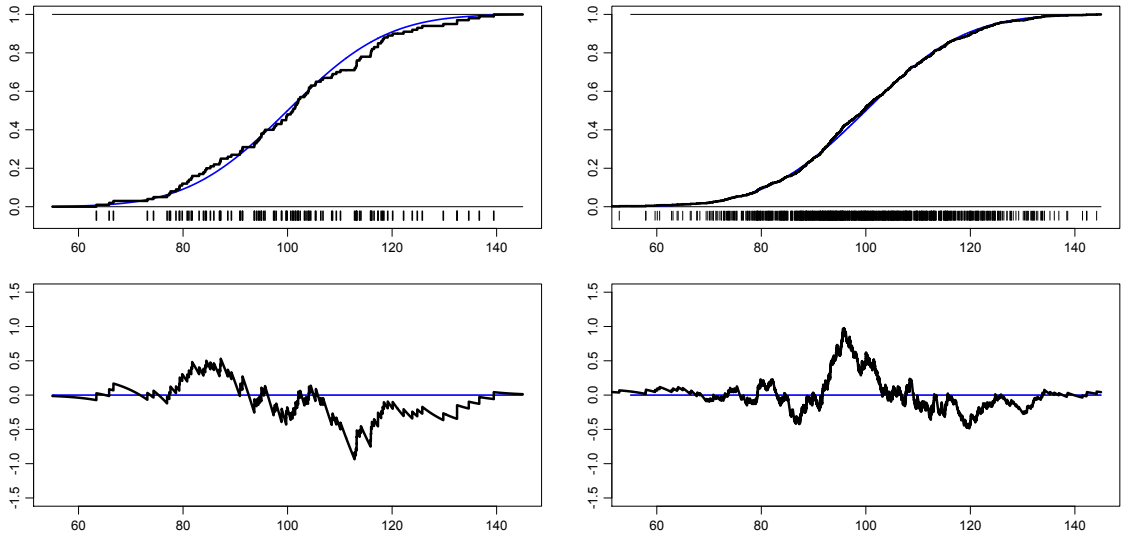

Figure A.1: Distribution functions $F, \widehat{F}$ and empirical processes $\sqrt{n}(\widehat{F} - F)$ for $n = 100$ (left) and $n = 1000$ (right).

$\sqrt{n}(\widehat{F} - F)$. The plots of F and $\widehat{F}$ show clearly the impact of the different sample sizes. But the standardised process $\sqrt{n}(\widehat{F} - F)$ looks rather similar in both situations.

Inequality 3.5 implies that

$$\mathbb{E}\big(\|\widehat{F} - F\|_\infty\big) \;=\; O\big(n^{-1/2}\big).$$

However, the proof of (3.5) is quite involved. Alternatively we shall derive a weaker inequality but of similar type:

Theorem A.12. *For arbitrary sample sizes $n \in \mathbb{N}$, distribution functions F and constants $c \geq 0$,*

$$\mathbb{P}\big(\|\widehat{F} - F\|_\infty \geq c\big) \;\leq\; e^2 \exp(-nc^2).$$

This result and the subsequent proof are presented in the monograph of Shorack and Wellner [27]. In the fourth step of proof we employ a trick of J.H.B. Kemperman.

Proof of Theorem A.12. The proof consists of four steps.

1st step: According to Lemma 3.11 (b) it suffices to consider the case of a continuous distribution function F. Then one easily verifies that $\sup(\widehat{F} - F)$ and $-\inf(\widehat{F} - F)$ are identically distributed; just replace all variables X_i with $-X_i$. Here we use the shorthand notation $\sup(h) := \sup_{x\in\mathbb{R}} h(x)$ and $\inf(h) := \inf_{x\in\mathbb{R}} h(x)$. This leads to the following inequality:

$$\begin{aligned}\mathbb{P}\big(\|\widehat{F} - F\|_\infty \geq c\big) \;&=\; \mathbb{P}\big(\sup(\widehat{F} - F) \geq c \text{ or } -\inf(\widehat{F} - F) \geq c\big)\\ &\leq\; 2\,\mathbb{P}\big(\sup(\widehat{F} - F) \geq c\big).\end{aligned}$$

2nd step: Eventually we want to apply Lemma 6.9. Hence we consider independent random variables $X_1, \ldots, X_n, X_{n+1}, \ldots, X_{2n}$ with continuous distribution function F and define, in addition to $\widehat{F}$, the empirical distribution function $\check{F}$ of the variables $X_{n+1}, \ldots, X_{2n}$. According to Lemma 6.9, for arbitrary $c \geq 0$,

$$\mathbb{P}\big(\sup(\widehat{F} - \check{F}) \geq c\big) \;=\; \binom{2n}{n + \lceil nc \rceil} \Big/ \binom{2n}{n},$$

and now we shall show that the right hand side is no larger than

$$(e/2)\exp(-nc^2).$$

It suffices to consider the case $c \le 1$, so $z := \lceil nc \rceil \in \{1, 2, \dots, n\}$. Then $\binom{2n}{n+z}/\binom{2n}{n}$ is equal to

$$\begin{aligned}
\frac{n!\,n!}{(n+z)!(n-z)!} &= (1+z/n)^{-1} \prod_{i=1}^{z-1} \frac{n-i}{n+i} \\
&= \exp\Big(-\log(1+z/n) + \sum_{i=1}^{z-1} \log\Big(\frac{1-i/n}{1+i/n}\Big)\Big) \\
&\le \exp\Big(-\log(1+z/n) - 2\sum_{i=1}^{z-1} i/n\Big) \\
&= \exp\big(-\log(1+z/n) - (z-1)z/n\big) \\
&= \exp(z/n - \log(1+z/n) - z^2/n) \\
&\le \exp(1 - \log(2) - z^2/n) \\
&\le (e/2)\exp(-nc^2).
\end{aligned}$$

The first inequality is based on the known identity $\log\big((1-x)/(1+x)\big) = -2\sum_{k=0}^{\infty} x^{2k+1}/(2k+1) \le -2x$ for $x \in [0,1)$. In the second inequality we used the fact that $x - \log(1+x)$ is monotone increasing in $x \in [0,1]$.

3rd step: Now we show that $\sup(\widehat{F} - F)$ tends to get larger if F is replaced with $\check{F}$. Precisely, let $\psi : \mathbb{R} \to \mathbb{R}$ be a monotone increasing and convex function, e.g. $\psi(x) := \max(x-b, 0)$ with $b \in \mathbb{R}$. Then

$$\mathbb{E}\big(\psi\big(\sup(\widehat{F} - F)\big)\big) \;\le\; \mathbb{E}\big(\psi\big(\sup(\widehat{F} - \check{F})\big)\big).$$

For $\widehat{F}(x) - F(x)$ may be viewed as conditional expectation

$$\mathbb{E}\big(\widehat{F}(x) - \check{F}(x) \,\big|\, \boldsymbol{X}\big).$$

Here $\mathbb{E}(\cdot \,|\, \boldsymbol{X})$ means that $\boldsymbol{X} = (X_i)_{i=1}^n$ is treated temporarily as a fixed vector, and we average only over the potential values of $(X_i)_{i=n+1}^{2n}$. Thus

$$\begin{aligned}
\mathbb{E}\big(\psi\big(\sup(\widehat{F} - F)\big)\big) &= \mathbb{E}\Big(\psi\Big(\sup_{x\in\mathbb{R}} \mathbb{E}\big(\widehat{F}(x) - \check{F}(x) \,\big|\, \boldsymbol{X}\big)\Big)\Big) \\
&= \mathbb{E}\Big(\sup_{x\in\mathbb{R}} \psi\big(\mathbb{E}\big(\widehat{F}(x) - \check{F}(x) \,\big|\, \boldsymbol{X}\big)\big)\Big) \\
&\le \mathbb{E}\Big(\sup_{x\in\mathbb{R}} \mathbb{E}\big(\psi\big(\widehat{F}(x) - \check{F}(x)\big) \,\big|\, \boldsymbol{X}\big)\Big) \\
&\le \mathbb{E}\Big(\mathbb{E}\Big(\sup_{x\in\mathbb{R}} \psi\big(\widehat{F}(x) - \check{F}(x)\big) \,\Big|\, \boldsymbol{X}\Big)\Big) \\
&= \mathbb{E}\Big(\mathbb{E}\Big(\psi\big(\sup(\widehat{F} - \check{F})\big) \,\Big|\, \boldsymbol{X}\Big)\Big) \\
&= \mathbb{E}\big(\psi\big(\sup(\widehat{F} - \check{F})\big)\big).
\end{aligned}$$

Here the second and second last step rely on the fact that $\psi(\sup(h)) = \sup(\psi \circ h)$ because ψ is non-decreasing and continuous. In the third step we used Jensen's inequality, and the last step is based on Fubini's theorem.

Final step: For fixed $c > 0$, arbitrary numbers $b \in [0, c)$ and $z \in \mathbb{R}$ we have the inequality $1_{[z \geq c]} \leq \max(z - b, 0)/(c - b)$. Consequently

$$\begin{aligned}
\mathbb{P}\big(\sup(\widehat{F} - F) \geq c\big) &\leq \frac{1}{c-b}\,\mathbb{E}\big(\max\big(\sup(\widehat{F} - F) - b, 0\big)\big) \\
&\leq \frac{1}{c-b}\,\mathbb{E}\big(\max\big(\sup(\widehat{F} - \check{F}) - b, 0\big)\big) \\
&= \frac{1}{c-b}\int_0^\infty \mathbb{P}\Big(\max\big(\sup(\widehat{F} - \check{F}) - b, 0\big) \geq t\Big)\,dt \\
&= \frac{1}{c-b}\int_b^\infty \mathbb{P}\big(\sup(\widehat{F} - \check{F}) \geq s\big)\,ds \\
&\leq \frac{e/2}{c-b}\int_b^\infty \exp(-ns^2)\,ds.
\end{aligned}$$

Here we applied the result from the third step with the particular function $\psi(z) := \max(z - b, 0)$. Then we applied the standard formula $\mathbb{E}(Z) = \int_0^\infty \mathbb{P}(Z \geq t)\,dt$ for nonnegative random variables $Z \geq 0$ and finally the exponential inequality from the second step. But $\exp(-ns^2) \leq \exp(-nc^2 - 2nc(s - c))$ for arbitrary $s \in \mathbb{R}$, whence

$$\begin{aligned}
\frac{1}{c-b}\int_b^\infty \exp(-ns^2)\,ds &\leq \frac{1}{c-b}\int_b^\infty \exp(-2nc(s-c))\,ds\ \exp(-nc^2) \\
&\leq \frac{1}{c-b}\int_{b-c}^\infty \exp(-2nct)\,dt\ \exp(-nc^2) \\
&= \frac{\exp(2nc(c-b))}{2nc(c-b)}\ \exp(-nc^2).
\end{aligned}$$

Elementary calculations show that $\exp(x)/x \geq \exp(1)/1 = e$ for all $x > 0$. Hence if $2nc(c-b) = 1$ and $b \geq 0$, that means, $b = c - 1/(2nc) \geq 0$, then

$$\mathbb{P}\big(\|\widehat{F} - F\|_\infty \geq c\big) \leq e^2 \exp(-nc^2).$$

The restriction $b = c - 1/(2nc) \geq 0$ is equivalent to $c^2 \geq 1/(2n)$. But for $c^2 \leq 1/(2n)$, the expression $e^2 \exp(-nc^2) \geq \exp(3/2)$ is greater than one, rendering the asserted inequality trivial. □

A.9 Normal Approximation of Linear Permutation Statistics

For two fixed vectors $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^N$ and a permutation Π of $\{1, 2, \ldots, N\}$ drawn completely at random let

$$T := \sum_{i=1}^N a_i b_{\Pi(i)}.$$

This random variable has the same distribution as $\sum_{i=1}^N a_{\Pi(i)} b_i$. In particular, in Exercise 6.9 it is shown that

$$\mathbb{E}\,T = N\bar{a}\bar{b} \quad \text{and} \quad \mathrm{Var}(T) = \frac{\big(\|\boldsymbol{a}\|^2 - N\bar{a}^2\big)\big(\|\boldsymbol{b}\|^2 - N\bar{b}^2\big)}{N-1}.$$

As shown in the following Theorem, the standardised random quantity

$$\tilde{T} := \frac{T - \mathbb{E}\,T}{\mathrm{Std}(T)}$$

has approximately a standard Gaussian distribution under certain conditions on the vectors $\boldsymbol{a}$ and $\boldsymbol{b}$. Of course we assume that $\mathrm{Std}(T) > 0$, that means, $\boldsymbol{a} \neq (\bar{a})_{i=1}^N$ and $\boldsymbol{b} \neq (\bar{b})_{i=1}^N$.

Theorem A.13 (Hájek[6]).

$$\sup_{\text{intervals } B\subset\mathbb{R}} \big|\mathbb{P}\big(\tilde{T}\in B\big) - \mathcal{N}(0,1)(B)\big| \;\to\; 0$$

as

$$\frac{\max_{i=1,\dots,N}(a_i-\bar{a})^2}{\sum_{j=1}^N (a_j-\bar{a})^2} + \frac{\max_{i=1,\dots,N}(b_i-\bar{b})^2}{\sum_{j=1}^N (b_j-\bar{b})^2} \;\to\; 0.$$

This is a classical result from nonparametric statistics. Its proof is treated in detail in the monograph of Hájek and Šidak [10]. Subsequently we sketch the essential considerations.

Sketch of proof for Theorem A.13. Without loss of generality let $\sum_{i=1}^N a_i = \sum_{i=1}^N b_i = 0$. For

$$T - \mathbb{E}\,T \;=\; \sum_{i=1}^N (a_i-\bar{a}) b_{\Pi(i)} \;=\; \sum_{i=1}^N (a_i-\bar{a})(b_{\Pi(i)}-\bar{b}),$$

so we may replace $\boldsymbol{a}$ and $\boldsymbol{b}$ with $(a_i-\bar{a})_{i=1}^N$ and $(b_i-\bar{b})_{i=1}^N$, respectively. Then the critical quantities mentioned in the theorem are $\|\boldsymbol{a}\|_\infty^2/\|\boldsymbol{a}\|^2$ and $\|\boldsymbol{b}\|_\infty^2/\|\boldsymbol{b}\|^2$ with the maximum norm $\|\cdot\|_\infty$ and the usual Euclidean norm $\|\cdot\|$.

Now we represent the random permutation Π as follows: Let $U_1, U_2, \dots, U_N$ be stochastically independent random variables with uniform distribution on $[0,1]$. Then we define

$$\Pi(i) \;:=\; \sum_{j=1}^N 1_{[U_j\le U_i]} \quad\text{and}\quad \check{\Pi}(i) \;:=\; \lceil NU_i\rceil.$$

In other words, Π contains the ranks of the random variables U_1, U_2, …, U_N. The random variables $\check{\Pi}(1), \check{\Pi}(2), \dots, \check{\Pi}(N)$ are independent and uniformly distributed on $\{1,2,\dots,N\}$.

Now we show that $\check{T} := \sum_{i=1}^N a_i b_{\check{\Pi}(i)}$ and T differ by a small amount only. With elementary calculations, similarly as in Exercise 6.9, one can show that

$$\begin{aligned}\mathbb{E}\big((\check{T}-T)^2\big) \;&= \|\boldsymbol{a}\|^2\,\mathbb{E}\,\big((b_{\check{\Pi}(1)}-b_{\Pi(1)})^2\big)\\ &\quad -\;\|\boldsymbol{a}\|^2\,\mathbb{E}\big((b_{\check{\Pi}(1)}-b_{\Pi(1)})(b_{\check{\Pi}(2)}-b_{\Pi(2)})\big)\\ &\le 2\|\boldsymbol{a}\|^2\,\mathbb{E}\big((b_{\check{\Pi}(1)}-b_{\Pi(1)})^2\big).\end{aligned}$$

The latter inequality is just a consequence of the Cauchy–Schwarz inequality. On the other hand, $\check{T}$ is a sum of independent random variables with

$$\mathbb{E}(\check{T}) \;=\; 0 \quad\text{and}\quad \mathrm{Var}(\check{T}) \;=\; \|\boldsymbol{a}\|^2\|\boldsymbol{b}\|^2/N \;=\; \frac{N-1}{N}\,\mathrm{Var}(T).$$

Without loss of generality one may arrange the components of $\boldsymbol{b}$ such that $b_1\le b_2\le\dots\le b_N$. Then it follows from Lemma A.14 below that

$$\mathbb{E}\big((b_{\check{\Pi}(1)}-b_{\Pi(1)})^2\big) \;\le\; 2^{3/2}\|\boldsymbol{b}\|_\infty\|\boldsymbol{b}\|/N,$$

whence

$$\frac{\mathbb{E}\,\big((\check{T}-T)^2\big)}{\mathrm{Var}(\check{T})} \;\le\; 2^{5/2}\|\boldsymbol{b}\|_\infty/\|\boldsymbol{b}\|.$$

[6] Jaroslav Hájek (1926-1974): Czech mathematician who contributed substantially to mathematical statistics.

All in all these considerations show that

$$\frac{T}{\operatorname{Std}(T)} \;=\; \sqrt{\frac{N}{N-1}}\,\frac{T}{\operatorname{Std}(\check{T})} \;=\; \sqrt{\frac{N}{N-1}}\,\frac{\check{T}}{\operatorname{Std}(\check{T})} + R$$

with $\mathbb{E}(R^2) \;=\; O\big(\|\boldsymbol{b}\|_\infty/\|\boldsymbol{b}\|\big)$. With the Central Limit Theorem one can show that the random variable $\check{T}/\operatorname{Std}(\check{T})$ has asymptotically a standard Gaussian distribution as $\|\boldsymbol{a}\|_\infty/\|\boldsymbol{a}\|$ and $\|\boldsymbol{b}\|_\infty/\|\boldsymbol{b}\|$ tend to 0. Thus the same conclusion is true for $T/\operatorname{Std}(T)$. □

Lemma A.14 (Hájek 1961). *For Π and $\check{\Pi}$ as in the proof of Theorem A.13 and arbitrary vectors $\boldsymbol{b} \in \mathbb{R}^n$ with $b_1 \le b_2 \le \cdots \le b_N$,*

$$\mathbb{E}\big((b_{\check{\Pi}(1)} - b_{\Pi(1)})^2\big) \;\le\; 2^{3/2} \max_{i=1,\ldots,N} |b_i - \bar{b}| \,\Big(\sum_{i=1}^N (b_i - \bar{b})^2\Big)^{1/2} \Big/ N.$$

Proof of Lemma A.14 in a special case. We prove this lemma only in case of $\boldsymbol{b} = (1_{[i>q]})_{i=1}^N$ with a number $q \in \{1, \ldots, N-1\}$. For the general case we refer to the original paper of Hájek (1961) or the monograph of Hájek and Šidak [10].

For symmetry reasons, the N random pairs $\big(\Pi(i), \check{\Pi}(i)\big) \;=\; \big(\Pi(i), \lceil NU_i \rceil\big)$, $1 \le i \le N$, are identically distributed. In case of $\Pi(i) = j$, $b_{\Pi(i)} = 1_{[j>q]}$ and $b_{\check{\Pi}(i)} = 1_{[U_{(j)}>q/N]}$. Consequently

$$\begin{aligned}
\mathbb{E}\big((b_{\check{\Pi}(1)} - b_{\Pi(1)})^2\big) \;&=\; \mathbb{E}\Big(\frac{1}{N}\sum_{i=1}^N (b_{\check{\Pi}(i)} - b_{\Pi(i)})^2\Big) \\
&=\; \mathbb{E}\Big(\frac{1}{N}\sum_{j=1}^N (1_{[U_{(j)}>q/N]} - 1_{[j>q]})^2\Big).
\end{aligned}$$

With elementary considerations one can show that

$$\frac{1}{N}\sum_{j=1}^N (1_{[U_{(j)}>q/N]} - 1_{[j>q]})^2 \;=\; |\widehat{G}(q/N) - q/N|$$

with the empirical distribution function $\widehat{G}(v) := N^{-1}\#\{i : U_i \le v\}$ of the uniform random variables U_i. Consequently,

$$\begin{aligned}
\mathbb{E}\big((b_{\check{\Pi}(1)} - b_{\Pi(1)})^2\big) \;&=\; \mathbb{E}\big(|\widehat{G}(q/N) - q/N|\big) \\
&\le\; \operatorname{Std}\big(\widehat{G}(q/N)\big) \\
&=\; \sqrt{q(1-q/N)}/N \\
&=\; \Big(\sum_{i=1}^N (b_i - \bar{b})^2\Big)^{1/2}/N.
\end{aligned}$$

Moreover,

$$\max_{i=1,\ldots,N} |b_i - \bar{b}| \;=\; \max(q/N, 1-q/N) \;\ge\; 1/2.$$

Thus our special vector $\boldsymbol{b}$ satisfies the asserted inequality with 2 instead of $2^{3/2}$. □

Bibliography

[1] ALAN AGRESTI (2002). *Categorical Data Analysis (2nd edition).* Wiley & Sons.

[2] PETER J. BICKEL, EUGENE A. HAMMEL and J. W. O'CONNELL (1975). Sex Bias in Graduate Admissions: Data from Berkeley. *Science* ***187***, 398-404.

[3] PETER J. BICKEL and ERICH L. LEHMANN (1976). Descriptive Statistics for Nonparametric Models III: Dispersion. *Annals of Statistics* ***4***, 1139-1158.

[4] C. CLOPPER and EGON S. PEARSON (1934). The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial. *Biometrika* ***26***, 404-413.

[5] DAVID L. DONOHO and PETER J. HUBER (1983). The Notion of Breakdown Point. In: *A Festschrift for Erich Lehmann (P.J. Bickel, K. Doksum and J.L. Hodges, Jr., Eds.)*, Wadsworth, Belmont (CA), pp. 157-184.

[6] LUTZ DÜMBGEN (2003). *Stochastik für Informatiker.* Springer-Verlag.

[7] LUTZ DÜMBGEN (2006/2007). *Wahrscheinlichkeitstheorie.* Vorlesungsskriptum, Univ. Bern.

[8] V.A. EPANEČNIKOV (1969). Non-Parametric Estimation of a Multivariate Probability Density. *Theory of Probability and its Applications* ***14***, 153-158.

[9] JAROSLAV HÁJEK (1961). Some Extensions of the Wald-Wolfowitz-Noether Theorem. *Annals of Mathematical Statistics* ***32***, 506–523.

[10] JAROSLAV HÁJEK and ZBYNĚK ŠIDAK (1967). *Theory of Rank Tests.* C.S.A.V., Prague

[11] FRANK R. HAMPEL (1971). A General Qualitative Definition of Robustness. *Annals of Mathematical Statistics* ***42***, 1887-1896.

[12] JOSEPH L. HODGES and ERICH L. LEHMANN (1963). Estimates of Location Based on Rank Tests. *Annals of Mathematical Statistics* ***34***, 598-611.

[13] WASSILY HOEFFDING (1948). A Class of Statistics with Asymptotically Normal Distribution. *Annals of Mathematical Statistics* ***19***, 293-325.

[14] WASSILY HOEFFDING (1963). Probability Inequalities for Sums of Bounded Random Variables. *Journal of the American Statistical Association* ***58***, 13-30.

[15] M. CHRIS JONES, J. STEVEN MARRON and SIMON J. SHEATHER (1996). A Brief Survey of Bandwidth Selection for Density Estimation. *Journal of the American Statistical Association* ***91***, 401-407.

[16] GÖTZ KERSTING und ANTON WAKOLBINGER (2008). *Elementare Stochastik.* Birkhäuser Verlag.

[17] ERICH L. LEHMANN (2006). *Nonparametrics: Statistical Methods Based on Ranks.* Springer Verlag.

[18] HENRY B. MANN and DONALD WHITNEY (1947). On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *Annals of Mathematical Statistics* ***18***, 50-60.

[19] PASCAL MASSART (1990). The Tight Constant in the Dvoretzky-Kiefer-Wolfowitz Inequality. *Annals of Probability* ***18***, 1269-1283.

[20] GOTTFRIED E. NOETHER (1971). Introduction to Statistics - A Fresh Approach. Houghton Mifflin Company.

[21] EMANUEL PARZEN (1962). On Estimation of a Probability Density Function and Mode. *Annals of Mathematical Statistics* ***33***, 1065-1076.

[22] R CORE TEAM (2013). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

[23] MICHAEL L. RADELET and GLENN L. PIERCE (1991). Choosing Those Who will Die: Race and the Death Penalty in Florida. *Florida Law Review* ***43***, 1-34.

[24] JOHN A. RICE (1995). *Mathematical Statistics and Data Analysis.* Wadsworth.

[25] MURRAY ROSENBLATT (1956). Remarks on Some Nonparametric Estimates of a Density Function. *Annals of Mathematical Statistics* ***27***, 832-837.

[26] LOTHAR SACHS (1973). *Angewandte Statistik.* Springer-Verlag.

[27] GALEN R. SHORACK and JON A. WELLNER (1986). *Empirical Processes with Applications to Statistics.* Wiley.

[28] BERNARD W. SILVERMAN (1986). *Density Estimation.* Chapman and Hall.

[29] EDWARD H. SIMPSON (1951). The Interpretation of Interaction in Contingency Tables. *Journal of the Royal Statistical Society, Series B* ***13***, 238-241. Chapman and Hall.

[30] FRANK WILCOXON (1945). Individual Comparisons by Ranking Methods. *Biometrics Bulletin* ***1***, 80-83.

Index