Statistical Data Analysis

M.C.M. de Gunst

Department of Mathematics
Faculty of Science
Vrije Universiteit Amsterdam

version February 18, 2020

with adjustments by F. Bijma, B. T. Knapik, and D. Dobler

These lecture notes are based on the lecture notes *Statistische Data Analyse* (in Dutch) by M.C.M. de Gunst and A.W. van der Vaart.

Contents

Chapter 1

Introduction

Statistics is the science of collecting, analyzing and interpreting data. In the ideal situation a statistician is involved in all stages of a study: in formulation of the question of interest, in determining the experimental procedure, the data collection, the data analysis, and in the interpretation and presentation of the results of the analysis. It will be clear that a statistician not only needs to know the statistical theory, but he/she also should be able to translate a practical problem into statistical terms, to give advice about the experimental design, to judge the quality of the data, to choose a proper statistical model and appropriate analysis tools, to perform the statistical analysis, and to translate the results of the analysis back to the practical problem for which the data were collected. Unfortunately, frequently the help of a statistician is asked after the data are collected. In these situations the data are often not optimal, which makes the statistical analysis more difficult. But also in such case the statistician needs to be aware of what the practical problem is and what this means in statistical terms.

In these lecture notes, the first stages of a study—the formulation of the problem, experimental design, and the data collection—are not considered. The aim is to give practical insight in the analysis of the data. Therefore, it is assumed that the data are already there, and that it is known why they are there and how they were obtained.

The first thing that needs to be done for a data analysis is to get an impression of the data and to summarize them in an appropriate way. This is discussed in Chapter 2. Next, when one wants to choose a model, it is useful to investigate the underlying distribution of the data. This is treated in Chapter 3. Analyzing peculiarities of probability density functions by means of kernel density estimators is done in Chapter 4. A way to prevent making unrealistic model assumptions is to assume as little as possible and use one of the nonparametric methods. These are the topic of Chapter 6. In the other chapters some frequently used techniques are discussed, like the bootstrap (Chapter 5), multivariate linear regression (Chapter 8), and methods for the analysis of categorical data (Chapter 7).

For all methods that are presented, not only the theory is important, but also when and how to use them in practice. These lecture notes contain many examples that illustrate these aspects. In addition, students should gain experience in applying the methods by means of exercises in which (simple) data sets are analyzed with the aid of the computer.

Chapter 2

Summarizing data

The title of this chapter consists of the words 'summarizing' and 'data'. Also the title of the course contains the word 'data'. The first part of this chapter explains what is meant with the term 'data' and what types of data can occur. When the data come in larger quantities, the first step of their analysis is to make an appropriate summary. In the second part of the chapter guidelines for making a good data summary and several techniques for obtaining summaries are discussed.

2.1 Data

The term 'data' is abundantly used in scientific research. Data are the quantified results of a study. A collection of data generally consists of characteristics of individuals or experimental units.

Example 2.1 For every subscriber of a certain magazine the following characteristics are given:

- sex (M, F);
- highest educational level (primary school, secondary school, higher education);
- number of months that the person is a subscriber of the magazine (1, 2, ...).

To quantify these measurements we could use the following code: male = 0, female = 1; primary school = 0, secondary school = 1, higher education = 2. Then the measurement (1,0,23) represents a female subscriber whose highest educational level is primary school and who subscribed 23 months ago.
*****

The data set in the example above contains three different types of data. These different types of data are measured on different *measurement scales*. The sex is an example of a measurement on a so-called *nominal scale* or *nominal level*. With measurements on this

scale the individuals are divided into categories. The different categories are identified by a different code (a number, letter, word, or name). The structure of a nominal scale is not changed by a 1-1 transformation: the coding 'male = 0', 'female = 1' would have been essentially the same as 'male = M, 'female = F'. Note that there is no ordering for this kind of data. Nominal scales are *qualitative* scales. Arithmetic operations cannot be performed on data measured on a qualitative scale. For data measured on a nominal scale the mode can be used as a measure of location (see next section). However, location concepts like median or mean, as well as measures of spread, have no meaning.

The highest educational level is an example of a measurement on an *ordinal scale*. The categories are not only identified, but they can also be ordered. The distances between the categories have no meaning. We cannot say that the difference between the primary and secondary school education is the same as the difference between the secondary and higher school education, or that higher education is three times as good as primary school: there is just only an ordering of the categories. A frequently used ordering is that of a 5-point scale: the patient indicates that she is feeling much worse, worse, as good, better, much better; a test panel judges the taste of chocolate of brand A much less tasty, less tasty, as tasty, more tasty, much more tasty than that of brand B, and so on. Ordinal scales are also qualitative scales: there is only an ordering without measurable distances. For measurements on an ordinary scale, the median and the mode can be useful, but the mean and measures of spread have no meaning.

The number of months that a person is a subscriber is an example of a measurement on a *quantitative scale*. The measurements have more meaning than just falling into a category or indicating an ordering. Someone who subscribed 26 months ago is a subscriber for twice as long as someone who subscribed 13 months ago. For quantitative data intervals are meaningful, and on these data arithmetic operations can be performed. The location measures mean, mode and median can all be used. Moreover, for quantitative data measures of spread can be used too. A quantitative scale for which intervals are meaningful but ratios not is called an *interval scale*; a quantitative scale for which both intervals and ratios are meaningful is a *ratio scale*. An interval scale has an arbitrary zero point, a ratio scale has a unique, true zero point.

The characteristics or properties of individuals or experimental units will be indicated by the term *variable*. Apart from the above mentioned partition of variables according to their measurement level, there are several other partitions:

- *discrete* and *continuous* variables. A discrete variable can only take a finite (or countable) number of values; continuous variables take values in a continuum, i.e. a "full" part of the real line. Nominal and ordinal variables are discrete by definition.

- *univariate*, *bivariate*, and *multivariate* variables. For this partition the dimension of the variable is the determining factor: on one subject one, two or more variables are measured. In Example 2.1 the measurement (1,0,23) is a three-dimensional or trivariate measurement.

- *independent* and *dependent* variables, sometimes called *structural* variables and variables *to be measured*. Dependent variables are the object of the study, whereas the independent or structural variables are quantities that may have an influence on the variables under study.

Example 2.2 In a study about the dependence of political opinions on variables like age, sex, or religion, the political opinion is the dependent variable and answers to a question about political opinion are the values of the dependent variable. Age, sex, religion, and so on, are the independent variables.
***

Example 2.3 When one is interested to know the effect of the composition of a fibre on the strength of the fibre, one could measure the strength of the fibre for different compositions of the fibre. The strength of the fibre is the dependent variable, the composition of the fibre is the independent variable.
***

Whether a variable is a dependent or an independent variable, depends on the context. In example 2.3 the composition of a fibre was an independent variable, but it can also be a dependent variable with, for instance, temperature and humidity as independent variables.

An independent variable that needs special attention is the variable 'time'. The concept 'time' needs to be broadly interpreted. For example, the electricity consumption per month, the number of births per year, and, more generally, every sequence of data points for which the order in time is essential, have 'time' as independent variable. Such ordered data sets are called *time series*.

2.2 Summarizing data

The first step in a data analysis is to perform an exploratory analysis and make an appropriate summary of the data. Although such exploratory analysis should use *known* facts about the data, implicit or explicit *assumptions* on the data should not be used. Data summaries can be just descriptive or they can be intended to investigate the suitability of certain statistical assumptions that are needed for a further analysis.

Example 2.4 In classical statistics it is often assumed that the data are a sample from a normal distribution. If this assumption would be exactly correct, then an appropriate summary would consist of the sample mean and the sample variance only. This is because then these two quantities are in the technical (statistical) sense sufficient. In practice exact normality is of course almost never the case, but it is often sufficient when normality holds only approximately. However, at the start of an analysis, it is usually not known whether the assumption of (approximate) normality is appropriate. Therefore in summarizing a data set one should not let oneself be guided by this assumption. Only after (approximate) normality has been established, a summary consisting of sample mean and sample variance suffices.
***

An appropriate summary should answer a number of basic questions, like
1) what is the area that covers most of the data and how are the data distributed over this area (location, scale)?
2) what are the smallest and largest data points; are there any extreme values?
3) are there any values that occur frequently in the data set; is the distribution unimodal or multi-modal; is there an area without any data values, i.e., is there a 'hole' in the data range?
4) are the values rounded?
5) are the values symmetrically distributed around a certain value; if so, does the distribution look like one of the well-known symmetric distributions: the normal, Cauchy, uniform, etc.; if not, does the distribution resemble one of the well-known asymmetric distributions: chisquare, F-, exponential, etc.?
6) do the data need to be divided into groups for separate analysis?
7) is there influence of other variables, like time?
8) which relationship exists between different variables?
A good summary often needs to answer more specific, context-related questions as well. From the above it will be clear that with a summary as little information as possible should be lost. In the following sections several methods for summarizing data will be discussed. Data summaries can be *graphical* or *numerical*. Examples of both types will be given.

2.2.1 Summarizing univariate data

In this section we consider the situation that we have data consisting of one sequence of numbers. In this case the data can be thought of as a sample of univariate variables.

One way to summarize a data set is to present them as a *stem-and-leaf plot*.

The decimal point is 1 *digit(s) to the right of the* |

5	|	2
5	|	6
6	|	0000004444444
6	|	8888
7	|	222222224
7	|	6666
8	|	004
8	|	888
9	|	2

Figure 2.1: Stem-and-leaf plot for the pulsation data.

Example 2.5 The pulsation of 39 Peruvian Indians was measured:
88, 64, 68, 52, 72, 72, 64, 80, 76, 60, 68, 72, 88, 60, 60, 72, 84, 64, 72, 64, 80, 76, 60, 64, 64, 68, 76, 60, 76, 88, 72, 68, 60, 74, 72, 56, 64, 72, 92.
These data are summarized with a stem-and-leaf plot in Figure 2.1.
***

From the figure the idea behind the plot will be clear: the 2 and the 6 behind the 5s indicate that both 52 and 56 occur once in the data set; 6 zeros behind the 6 means that 60 occurs six times in the data set. The numbers 5, 5, 6, 6, 7, etc. in the first column are the *stems* of the plot; the sequences of numbers behind them in the second column are the *leaves*. Stems without leaves (no occurrence in the data) may also occur in a stem-and-leaf plot. In most statistical packages one can choose the number of stems and the place of the |. The stem-and-leaf plot in the figure is an example of a so-called *split* stem-and-leaf plot, because the natural stems, the tens, are split into two. With stem-and-leaf plots little information about the data gets lost. Stem-and-leaf plots give an impression about the shape of the data distribution while retaining most of the numerical information.

A frequently used graphical summary of univariate data is the *histogram*. A histogram is a graph in which a scale distribution for the measured variable is given along the horizontal axis. Above each interval or *bin* that is induced by this scaling distribution, a bar is drawn such that the *area* of the bar is proportional to the frequency of the data in that bin. When the widths of the bins are all equal, then the *heights* of the bars are also proportional to the frequency of the data in that bin. When the bin widths are not equal,

this does not hold. Obviously, it can be very misleading to plot bins of unequal width with the same size. It can result in a completely wrong impression of the spread of the data. The choice of the bin sizes is somewhat arbitrary. Different histograms of the same data set can give a slightly different impression. Too few or too many intervals always gives a bad result. Figure 2.2 shows two histograms of the pulsation data of Example 2.5. The bin sizes of the histogram on the left were chosen to be all equal to 10, for the histogram on the right the bin sizes were chosen to be 5. Although for the histogram on the right less data information got lost, the histogram on the left gives a better impression of the global spread of the data. Note that several of the bars in the histogram on the right only represent 1 data point. A histogram gives an idea of the distribution of the data. When the histogram is scaled such that the total area of all bars is equal to 1, it can be used as a rough estimate of the probability density function of the distribution from which the data originate. Although with histograms generally more information about the data gets lost than with stem-and-leaf plots, histograms are used much more often.

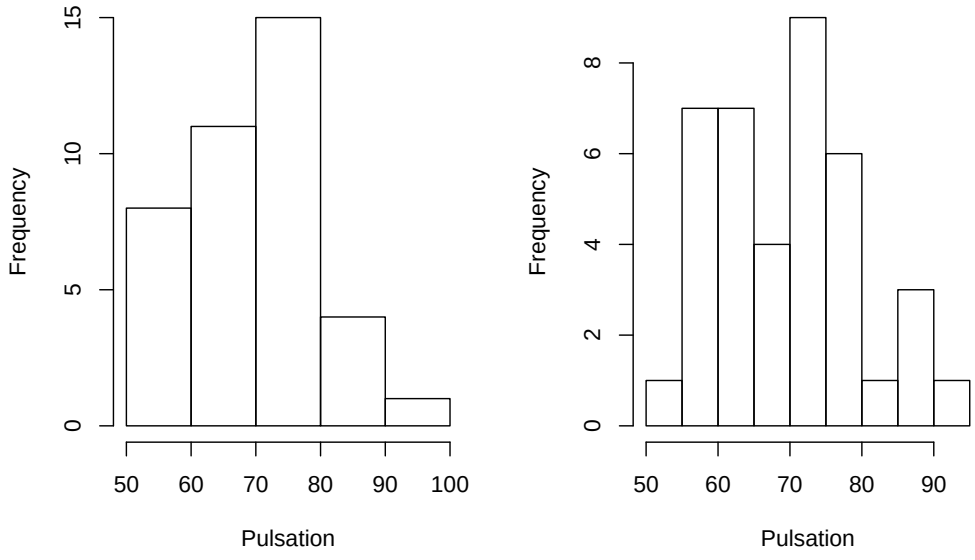


Figure 2.2: Histograms for the pulsation data of Example 2.5.

While a histogram can be used as an estimate of the probability density function, the *empirical (cumulative) distribution function* gives an impression of the (cumulative) distribution function of the underlying distribution. For a sample $\{x_1, \ldots, x_n\}$ the empirical distribution function is defined as

$$\hat{F}_n(x) = \frac{1}{n}(\#(x_j \leq x)) = \frac{1}{n}\sum_{j=1}^{n} 1_{\{x_j \leq x\}}.$$

(Here the *indicator* $1_{\{x_j \leq x\}}$ equals 0 or 1 depending on whether $x_j > x$ or $x_j \leq x$.) Denote the ordered set of sample values, arranged in order from smallest to largest, by $(x_{(1)}, \ldots, x_{(n)})$; $x_{(i)}$ is called *i-th order statistic* of the sample. If $x < x_{(1)}$, $F_n(x) = 0$, if $x_{(1)} \leq x < x_{(2)}$, $F_n(x) = 1/n$, if $x_{(2)} \leq x < x_{(3)}$, $F_n(x) = 2/n$, and so on. If there is a single observation with value x, then F_n has a jump of height $1/n$ at x; if there are k observations with the same value x, then F_n has a jump of height k/n at x. The empirical distribution function is the sample analogue of the distribution function F of a random variable X: $F(x)$ gives the probability that $X \leq x$, $F_n(x)$ gives the proportion of the sample that is less than or equal to x. In Figure 2.3 the empirical distribution function of the pulsation data of Example 2.5 is depicted.

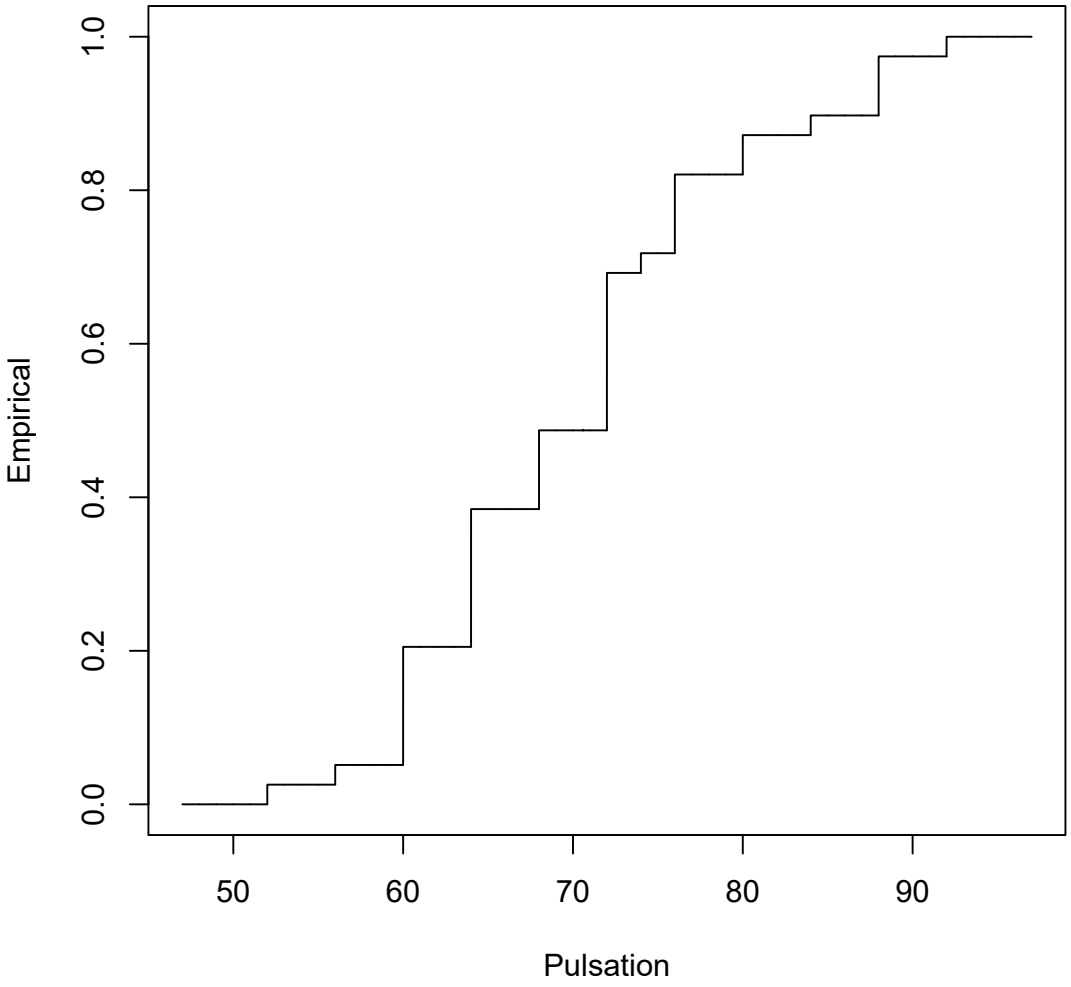


Figure 2.3: Empirical distribution function of the pulsation data of Example 2.5.

Besides graphical summaries, numerical summaries can be given: with one or more

sample size		n
location	*mean*	$\bar{x} = n^{-1}\sum_{i=1}^{n} x_i$
	α-trimmed mean	$\frac{1}{n-2[\alpha n]} \sum_{j=[n\alpha]+1}^{n-[\alpha n]} x_{(j)}, \qquad 0 \le \alpha < \frac{1}{2}$
	median	$\mathrm{med}(x) = \begin{cases} x_{((n+1)/2)}, & \text{if } n \text{ odd} \\ \frac{1}{2}(x_{(n/2)} + x_{(n/2+1)}), & \text{if } n \text{ even} \end{cases}$
scale	*variance*	$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
	standard deviation	$s = \sqrt{s^2}$
	coefficient of variation	$cv = s/\bar{x}$
	median absolute deviation	$\frac{1}{\Phi^{-1}(3/4)} \mathrm{med}(\lvert x_i - \mathrm{med}(x_1, \ldots, x_n)\rvert)$
	range	$(x_{(1)}, x_{(n)})$
	quartiles	$\mathrm{quart}(x)$, $3\mathrm{quart}(x)$
	interquartile range	$3\mathrm{quart}(x) - \mathrm{quart}(x)$
skewness	*skewness*	$b_1 = \frac{\sqrt{n}\sum_{j=1}^{n}(x_j - \bar{x})^3}{\{\sum_{j=1}^{n}(x_j - \bar{x})^2\}^{3/2}}$
size of tails	*kurtosis*	$b_2 = \frac{n\sum_{j=1}^{n}(x_j - \bar{x})^4}{\{\sum_{j=1}^{n}(x_j - \bar{x})^2\}^2}$

Table 2.1: Some numerical summaries of a sample $x_1, \ldots, x_n$.

numbers an impression is given of the whole data set. For a sample $x_1, \ldots, x_n$ one can, for instance, give the *sample size* n, the *sample mean* $\bar{x}$ and the *sample variance* s^2. The latter two quantities give an impression of the location and scale, respectively, of the distribution from which the data were generated. In particular, the sample mean corresponds to the population mean or, in statistical terms, to the expectation of the underlying distribution. Similarly, the sample variance corresponds to the population variance or to the variance of the underlying distribution. There are, however, also other quantities that give information about location and scale of the distribution. Table 2.1 gives an overview of the most commonly used quantities in this context.

The (*sample*) *quartiles* mentioned in Table 2.1 are roughly those values of the sample such that one quarter, and three quarters, respectively, of the sample are smaller than these values. Often the following more precise definition is used:

$$\mathrm{quartile}(x) = \begin{cases} \frac{1}{2}\left(x_{(n/4)} + x_{(n/4+1)}\right), & \text{if } n \text{ is a quadruple;} \\ x_{((n+3)/4)}, & \text{if } n \text{ is a quadruple+1;} \\ x_{((n+2)/4)}, & \text{if } n \text{ is a quadruple+2;} \\ \frac{1}{2}\left(x_{((n+1)/4)} + x_{((n+5)/4)}\right), & \text{if } n \text{ is a quadruple+3.} \end{cases}$$

The third quartile is defined analogously. These sample quartiles are special cases of the *sample quantiles*. For $\alpha \in (0, 1)$ the α-quantile of a sample is a value such that approximately a fraction α of the sample is smaller than this value. Since αn is in general not a round number, the exact definition is again a matter of consensus. Note that the median is the 0.5-quantile. The median and the interquartile range have the advantage that, compared to the mean and the variance, they are less sensitive to outliers or extreme

values. The so-called *five-number summary* consists of the smallest data value, the largest data value, the sample median, and the two quartiles. The 'population version' of the corresponding sample quantile is defined as follows. When for a random variable X there exists a unique number x_α such that $P(X \leq x_\alpha) = \alpha$, then x_α is called the α-*quantile* of the probability distribution of X.

Next to mean and median also the *mode* is frequently used as a location measure. The mode of a distribution is the location of the maximum of the probability density of the distribution. A precise definition of the mode of a sample does not exist. For a data set one could take as the *sample mode* the data value that occurs most frequently in the data set, or the location of the highest bar of a histogram, or the mode of another density estimator (as in Chapter 4).

Another, robust location measure is the α-*trimmed mean*. It is the average of the $n - 2[\alpha n]$ central observations $x_{([\alpha n]+1)}, \ldots, x_{(n-[\alpha n])}$, that is, the $[\alpha n]$ smallest and largest observations are ignored; the bracket indicates the closest integer number. By *robust* we mean that single extremely small or large observations, that might be considered as outliers, have only little influence on the estimate. The usual sample mean and sample median can be seen as special cases of α-trimmed means for $\alpha = 0$ and α tending to 0.5, respectively.

A robust estimator of scale is the *median absolute deviation*. As its name suggests, it is the median deviation of all observations (in absolute value) from the sample median. Dividing by the 75%-quantile of the standard normal distribution, $\Phi^{-1}(3/4)$, makes this a consistent estimator of the standard deviation σ if the independent observations follow any normal distribution with the same mean μ and standard deviation $\sigma > 0$.

The *skewness* and the *kurtosis* are two other characteristics of a distribution. They give an idea of the asymmetry and the size of the tails, respectively, of the distribution. The sample quantities b_1 and b_2 given in Table 2.1 correspond with the population quantities

$$\beta_1 = \frac{\mathrm{E}(X - \mathrm{E}\,X)^3}{\{\mathrm{E}(X - \mathrm{E}\,X)^2\}^{\frac{3}{2}}},$$

and

$$\beta_2 = \frac{\mathrm{E}(X - \mathrm{E}\,X)^4}{\{\mathrm{E}(X - \mathrm{E}\,X)^2\}^2}.$$

Both quantities are location and scale invariant. In Table 2.2 the values of β_1 and β_2 are given for some frequently occurring distributions. We remark that for symmetric distributions $\beta_1 = 0$.

Finally, we mention a combination of a graphical and a numerical summary: the *boxplot*. A scale distribution for the measured variable is given along the vertical axis, next to which a rectangle, 'the box', is drawn. The top and the bottom of the box are situated at the largest and smallest quartile, respectively. The width of the box may vary. A horizontal line in the box gives the position of the sample median. From the middle

distribution	notation	μ	σ^2	β_1	β_2
Bernoulli	$Bern(p)$	p	$p(1-p)$	$\frac{1-2p}{\sqrt{p(1-p)}}$	$3+\frac{1-6p(1-p)}{p(1-p)}$
Binomial	$Bin(n,p)$	np	$np(1-p)$	$\frac{1-2p}{\sqrt{np(1-p)}}$	$3+\frac{1-6p(1-p)}{np(1-p)}$
Hypergeom.	$Hyp(m,r,N)$	$\frac{mr}{N}$	$\frac{mr(N-r)(N-m)}{N^2(N-1)}$	$\frac{(N-2r)(N-2m)(N-1)^{\frac{1}{2}}}{(N-2)(mr(N-r)(N-m))^{\frac{1}{2}}}$	$\frac{N^2(N-1)}{mr(N-r)(N-m)(N-2)(N-3)}$
Poisson	$P(\mu)$	μ	μ	$\mu^{-1/2}$	$3+\mu^{-1}$
Normal	$N(\mu,\sigma^2)$	μ	σ^2	0	3
Uniform	$Unif(a,b)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	0	$\frac{9}{5}$
t	t_ν	0	$\frac{\nu}{\nu-2}, \quad (\nu>2)$	$0, \quad (\nu>3)$	$3+\frac{6}{\nu-4}, \quad (\nu>4)$
Chisquare	χ^2_ν	ν	2ν	$2(2/\nu)^{1/2}$	$3+\frac{12}{\nu}$
Exponential	$exp(\lambda)$	λ^{-1}	λ^{-2}	2	9

Table 2.2: Mean, variance, skewness and kurtosis of some distributions.

of the top and bottom of the box the so-called *whiskers* extend. They connect the box to the most extreme data points that lie at most 1.5 times the interquartile range from the edges of the box. More extreme data points are considered as extreme values and are depicted by separate symbols like $\circ$ or $*$. The factor 1.5 that determines the maximum length of the whiskers is sometimes replaced by another value. Figure 2.3 shows boxplots for the data of Example 2.5 with factors 1.5 and 1.

2.2.2 Data transformation

It may happen that the shape of a data set makes it problematic to summarize or analyze this set. In such situations a suitable transformation of the data may help. The data are then adapted to the available statistical tools. In most cases this means that a transformation is chosen so that the transformed data satisfy certain model assumptions, such as symmetry, normality, mean zero and variance 1, additivity of effects, and so on.

Example 2.6 The picture on the left in Figure 2.5 shows a histogram of the numbers of insects counted on a specific type of bushes in the dunes. This is a skewed distribution. Application of the square-root-transformation on each of the data points yields a data set that looks more symmetric; see the histogram in Figure 2.5 on the right. By application of a simple transformation a more symmetric picture is obtained. Moreover, in this picture two 'tops' can be seen that are less visible in the picture on the left. The meaning of these tops should be further investigated, for instance it may be the case that the number of insects depends on the location

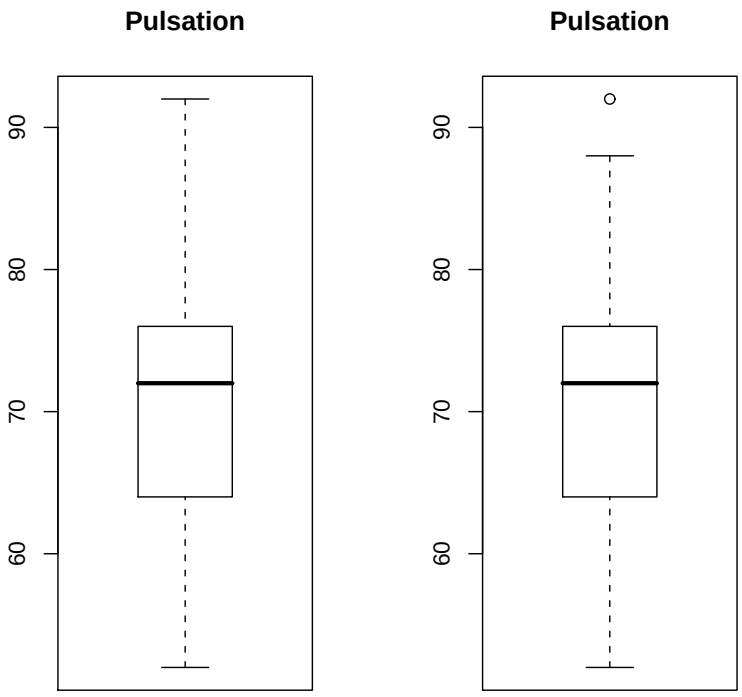


Figure 2.4: Boxplots for the pulsation data of Example 2.5 with factors 1.5 (left) and 1 (right).

of the bush, or that the inspected bushes are located in two very different areas.
***

Frequently used transformations are linear transformations of the form $y = a + bx$, or power transformations of the form $y = x^\lambda$. The values for λ that are usually chosen equal to $\frac{1}{2}$, -1, 2. Another transformation, which is generally applied to make multiplicative effects additive, is $y = \log x$.

It is a point of discussion whether it is acceptable to perform an analysis based on other than the original data. When the transformation is reversible, the conclusions from the results of the analysis of the transformed data can usually be translated to conclusions about the original data. However, caution is needed: a transformation should only be applied when there is a real advantage to do so. Remember that with the present-day computing power even very complex models can be fitted directly to the data.

2.2.3 Summarizing bivariate data

Let us now consider the situation in which for every subject in the study the values of two variables are measured. We then have *paired* data or a *bivariate* data set. Like in

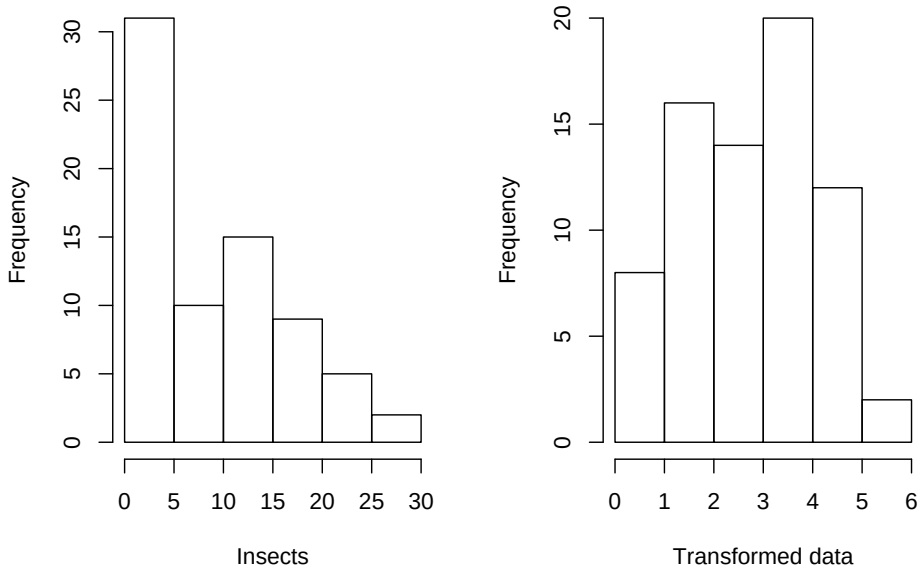


Figure 2.5: Histogram of numbers of insects on bushes (left) and histogram of the square root of the same numbers (right).

the case of univariate data, we can summarize bivariate data graphically or numerically. A frequently used graphical summary for paired observations is a *scatter plot*. Scatter plots are used to plot the values of one quantitative variable against the corresponding values of the other quantitative variable. They can help us to detect relationships between the variables, like linear or quadratic relations, to find extreme values, or to determine clusters of observations. Figure 2.6 shows an example of a scatter plot. The analysis of a time series usually starts by making a scatter plot of the data against time. This type of scatter plot is called a *time plot*. An example of a time plot is given in Figure 2.7.

Another convenient way to summarize bivariate data is by means of a $k \times r$-*frequency table*, also called *contingency table*. To construct such a table the possible values of the first and second variable are partitioned into k and r disjunct categories or intervals, respectively. Then a table of k rows and r columns is made, such that the frequency of the number of subjects with their value of the first variable in the i-th category and their value of the second variable in the j-th category is presented in the cell on the i-th row and in the j-th column.

Example 2.7 In a study on the relationship between blood group and two different diseases a sample of 8766 people consisting of 2697 patients with a stomach ulcer or stomach cancer and a control group of 6087 people without these diseases was divided based on blood group. The result is given in Table 2.3. There seems to be a certain relation between the variables blood group and disease. In Chapter 7 this

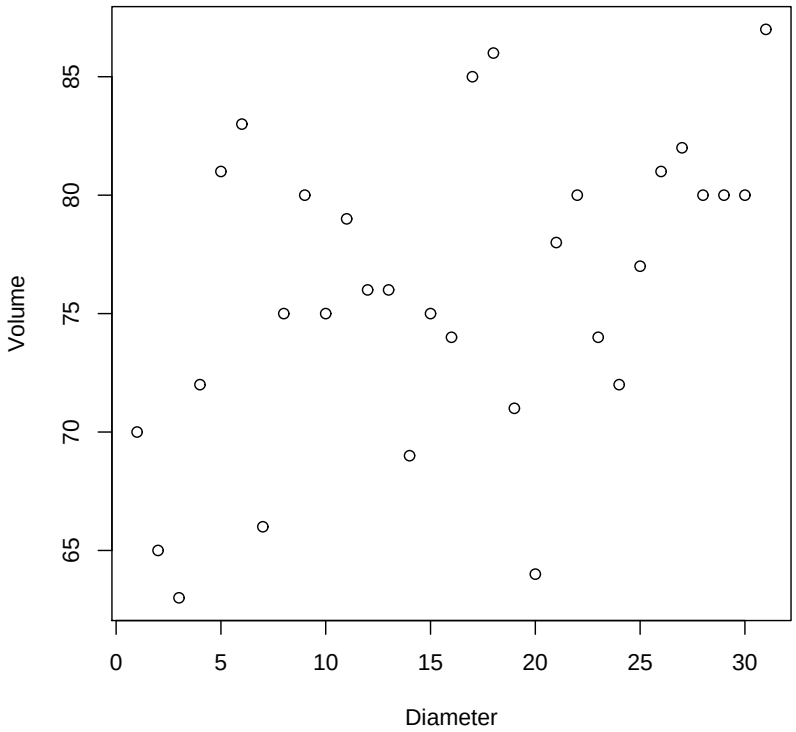


Figure 2.6: Scatter plot of volume of trees against diameter.

table will be further investigated.
***

blood group	stomach ulcer	stomach cancer	control	total
O	983	383	2892	4258
A	679	416	2625	3720
B	134	84	570	788
total	1796	883	6087	8766

Table 2.3: 3×3 contingency table of blood group against disease of 8766 persons.

Whereas in a scatter plot the individual data values can still be recognized, in a contingency table this information may get lost (when the categories consist of more than one value). As is illustrated in the example above, the advantage of contingency tables is that they can be used not only to summarize data that are measured on a quantitative scale, but also to summarize data that are measured on a nominal or ordinal scale. Contingency tables will be discussed further in Chapter 7.

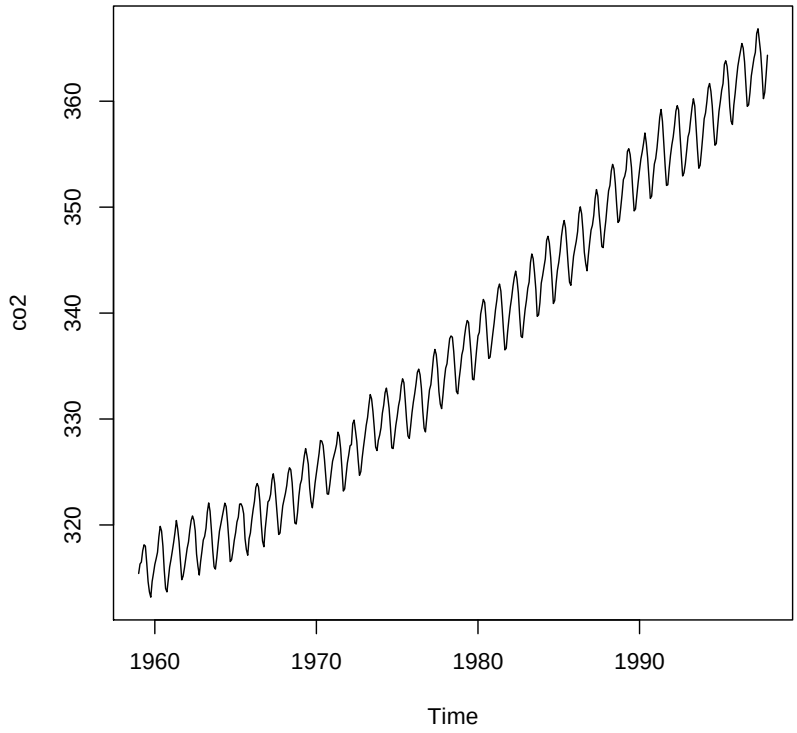


Figure 2.7: Time plot of CO_2 from 1959 to 1997.

The numerical quantities that are most frequently used to summarize bivariate data are given in Table 2.4. Of the quantities mentioned in the table the correlation coefficients need some explanation. The sample correlation coefficient r_{xy} is a measure of the strength of the *linear* relationship between x and y. It can take values from -1 to 1. A r_{xy}-value close to -1 means that there is a strong negative linear relation between x and y. Equality to -1 or 1 means that the relationship is exactly linear.

Rank correlation coefficients, however, are only measures of the *rank* correlation between x and y, i.e. measures for the interdependence between the ranks of the x- and y-values in the ordered samples. Values of both rank correlation coefficients may range from -1 to 1. With a large positive value of the rank correlation coefficient a larger value of x in the sample will generally correspond with a larger value of y, whereas when the rank correlation coefficient has a negative value, then larger values of x will generally correspond with smaller values of y. To obtain Kendall's τ all x_i are compared with all y_j; for Spearman's r_s the x_i are only compared with the corresponding y_i. Note that the correlation coefficient r_{xy} can only be used for measurements on a quantitative scale, whereas rank correlation coefficients can also be used for measurements on an ordinal scale. Beside the three quantities that are mentioned here, there are many other measures for dependence between variables. Correlation measures will reappear in Chapter 6.

mean	$(\bar{x}, \bar{y})$
covariance	$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$
correlation coefficient	$r_{xy} = \frac{s_{xy}}{s_x s_y}$
covariance matrix	$\Sigma = \begin{pmatrix} s_x^2 & s_{xy} \\ s_{xy} & s_y^2 \end{pmatrix}$
Spearman's rank correlation coefficient	$r_s = \frac{\sum_{i=1}^{n}(r_i - \frac{1}{2}(n+1))(t_i - \frac{1}{2}(n+1))}{\sqrt{\sum_{i=1}^{n}(r_i - \frac{1}{2}(n+1))^2 \sum_{i=1}^{n}(t_i - \frac{1}{2}(n+1))^2}}$
Kendall's rank correlation coefficient	$\tau = \frac{\sum\sum_{i \neq j} \operatorname{sgn}(r_i - r_j)\operatorname{sgn}(t_i - t_j)}{n(n-1)} = \frac{4N_\tau}{n(n-1)} - 1$

Table 2.4: Numerical summaries of bivariate data $(x_1, y_1), \ldots, (x_n, y_n)$. The vectors $(r_1, \ldots, r_n)$ and $(t_1, \ldots, t_n)$ are the ranks of, $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_n)$, respectively, in the ordered samples x and y. The quantity N_τ is the number of pairs (i, j) with $i < j$ for which either $x_i < x_j$ ánd $y_i < y_j$, or $x_i > x_j$ ánd $y_i > y_j$. The *sign function* $\operatorname{sgn}(x)$ is defined as: 1 if $x > 0$, 0 if $x = 0$ and -1 if $x < 0$.

2.2.4 Summarizing multivariate data

For a multivariate data set it is often useful to make scatter plots or contingency tables for all (relevant) pairs of variables. With 7 variables this yields already 21 graphs to study. But it can be misleading to project higher-dimensional data into two dimensions. For a general introduction to multivariate statistical analysis see for instance Anderson (2003), Mardia, Kent and Bibby (1980) or Chatfield and Collins (1981). Statistical packages have several possibilities for graphical summaries of higher-dimensional data sets.

Chapter 3

Exploring distributions

In this chapter we discuss several methods to investigate whether data stem from a certain probability distribution. The question whether a data set comes from a 'known' distribution is important for several reasons:

- Fitting a known distribution to the data is an efficient way of summarizing the data. For example, it would be very convenient if the statement 'the data follow a normal distribution with mean μ and variance σ^2' would suffice as a summary of the data.

- It is desirable, if possible, to find a model that somehow 'explains' the data. Investigating whether the data come from a certain distribution is not only helpful for testing an existing model, it can also give an indication which kind of model we need to look for.

- After having conducted an explorative data analysis, one often wants to apply more formal statistical methods, like statistical tests. Such methods are usually based on certain assumptions about the underlying probability mechanism. These assumptions should be plausible, and therefore it is of great importance to test them, at least visually, on the data. For instance, you can think of checking the normality of the measurement errors in a regression analysis before applying a t- or F-test.

Unless stated otherwise, we assume that $x_1, \ldots, x_n$ are realizations from independent random variables $X_1, \ldots, X_n$ which all have the same (unknown) probability distribution with distribution function F. We often shortly speak of 'the distribution F', and say that $x_1, \ldots, x_n$ are independent realizations from F. The goal is to get to know more about F. The independence and identical distribution are not tested! First, we discuss graphical methods to investigate whether univariate data stem from a certain probability distribution, whether the underlying distribution F is symmetric, and whether two samples originate from the same distribution. Some important techniques for obtaining a first impression of the shape of a distribution, like drawing histograms or box plots, have already been treated in the foregoing chapter, and will not be discussed further. Next, we discuss several tests for verifying whether univariate data stem from a certain probability distribution.

3.1 The quantile function and location-scale families

Let F be a distribution function of a probability distribution defined on $\mathbb{R}$. We already saw in Chapter 2 that if for a given $\alpha \in (0,1)$ there exists exactly one $x_\alpha \in \mathbb{R}$ such that $F(x_\alpha) = \alpha$, then x_α is called the *α-quantile* of F, (also the (left) *α-point*). The α-quantile is denoted by $F^{-1}(\alpha)$. As this notation suggests, the quantile function is the function $\alpha \mapsto F^{-1}(\alpha)$, the inverse of F, if this inverse is well-defined. This is the case when F is a strictly increasing function.

Apart from strictly increasing pieces, a cumulative distribution function can have jumps as well as constant pieces. Therefore, for fixed α the equation $F(x) = \alpha$ has exactly one, no or infinitely many solutions (see Figure 3.1). In order to be able to define the α-quantile in the latter two cases, the *quantile function* of F is in general defined by

$$F^{-1}(\alpha) = \inf\{x : F(x) \geq \alpha\}, \qquad \alpha \in (0,1).$$

In words: $F^{-1}(\alpha)$ equals the smallest x with $F(x) \geq \alpha$.

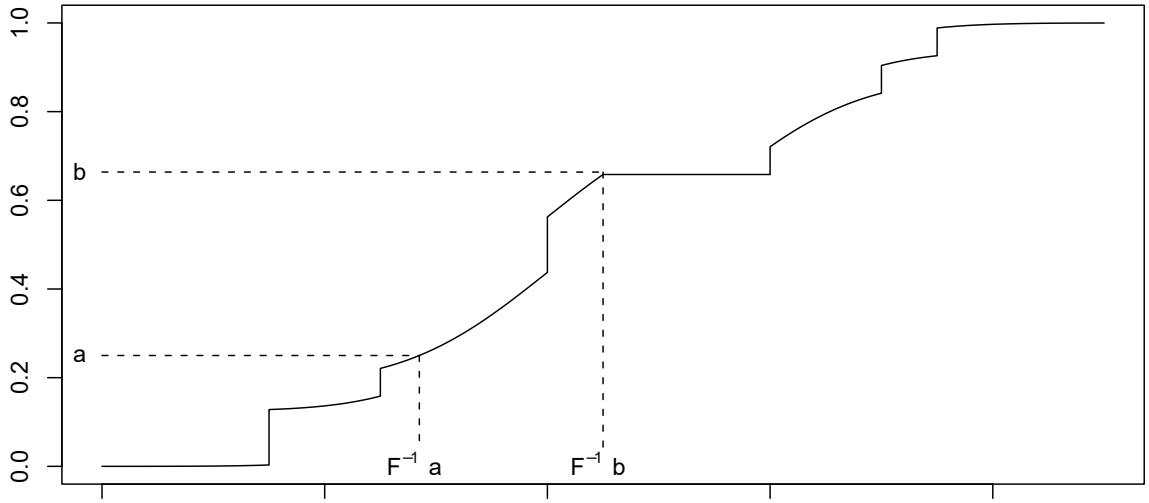


Figure 3.1: A distribution function and two of its quantiles.

When the random variable X has distribution function F, the distribution function of $a + bX$ is $F_{a,b}$, given by

$$F_{a,b}(x) = F(b^{-1}(x-a)), \qquad a \in \mathbb{R},\ b > 0.$$

The family of distributions $\{F_{a,b} : a \in \mathbb{R}, b > 0\}$ is called the *location-scale family* corresponding to F. When X has expectation $\mathrm{E}\, X = 0$ and variance $\mathrm{Var}(X) = 1$, then the expectation and variance of the distribution $F_{a,b}$ are given by a and b^2. It is not difficult to verify the following relation between the quantile functions

$$F_{a,b}^{-1}(\alpha) = a + bF^{-1}(\alpha).$$

In other words: the points $\{(F^{-1}(\alpha), F_{a,b}^{-1}(\alpha)) : \alpha \in (0,1)\}$ are on the straight line $y = a + bx$. Figure 3.2 illustrates that two normal distributions belong to the same location-scale family.

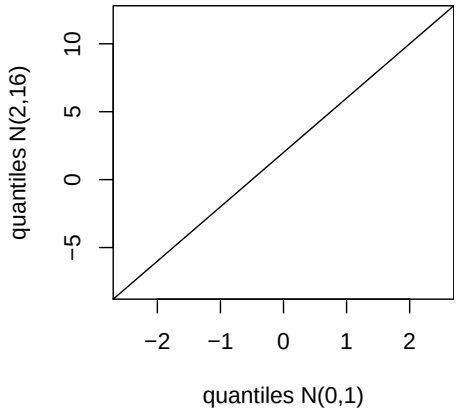


Figure 3.2: The quantiles of N(2,16) plotted against those of N(0,1).

3.2 QQ-plots

Let $x_1, \ldots, x_n$ be independent realizations from a distribution F. Then approximately a fraction $i/(n+1)$ of the data will be below the i-th order statistic $x_{(i)}$. Hence, $x_{(i)}$ approximately equals the $i/(n+1)$-quantile of the data. Therefore, the points

$$\left\{ \left(F^{-1}(\frac{i}{n+1}), x_{(i)} \right) : i = 1, \ldots, n \right\}$$

are expected to lie approximately on a straight line[1]. A *QQ-plot* is a graph of these n points. The Qs in the name are abbreviations of the word 'quantile'. The choice of $i/(n+1)$ is in fact rather arbitrary. One can also plot the points $\{(F^{-1}(\frac{i-\varepsilon_1}{n+\varepsilon_2}), x_{(i)}) : i = 1, \ldots, n\}$ for fixed (small) values of ε_1 and ε_2. In particular in the case that F is a normal distribution, the points

$$\left\{ \left(F^{-1}(\frac{i-3/8}{n+2/8}), x_{(i)} \right) : i = 1, \ldots, n \right\}$$

follow a straight line slightly better than the points defined above.

[1] A more formal argument for this reasoning is the well known result $\mathrm{E}\, F(X_{(i)}) = i/(n+1)$ for measurements from a continuous distribution function F. Therefore, one can expect $\mathrm{E}\, X_{(i)} \approx F^{-1}(i/(n+1))$. The reasoning would be stronger if the last relation would hold with exact equality. However, this is almost never the case. For the normal distribution function Φ one has for example $\mathrm{E}\, X_{(i)} < \Phi^{-1}(i/(n+1))$ if $i < \frac{1}{2}(n+1)$, equality if $i = \frac{1}{2}(n+1)$, and the reverse inequality if $i > \frac{1}{2}(n+1)$.

Example 3.1 Samples consisting of 50 independent measurements from the $N(2, 16)$-distribution were simulated by using a random number generator. In Figure 3.3 the QQ-plots of these samples are shown.
***

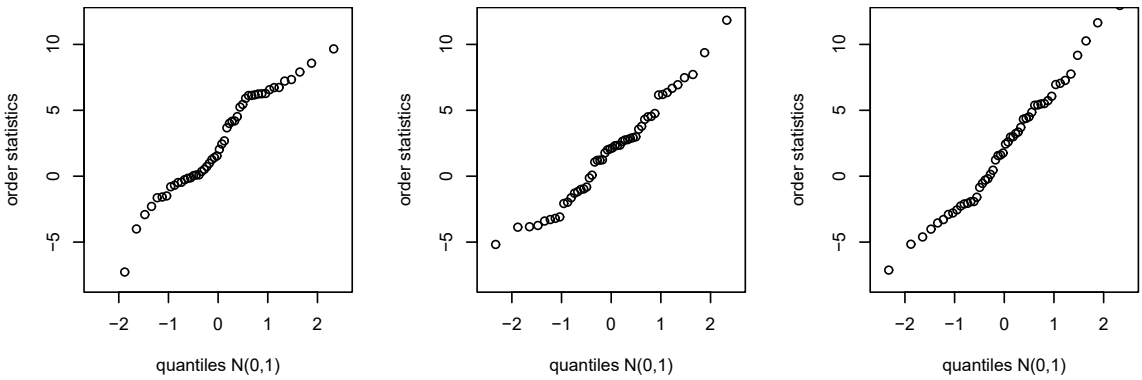


Figure 3.3: Three QQ-plots of samples consisting of 50 data points from N(2,16) against quantiles of N(0,1).

A QQ-plot yields a method to judge whether the data come from a certain distribution by only looking at the plot. When the plot yields approximately the line $y = x$, this is an indication that the data come from the distribution F. Deviations from the line $y = x$ indicate differences between the true distribution of the data and F. The kind of deviation from $y = x$ suggests the kind of difference between the true distribution and F. The simplest case of such a deviation is when the QQ-plot is a straight line but not the line $y = x$, as in Figure 3.3. This is an indication that the data do not originate from F, but come from another member of the location-scale family of F. Interpreting a bent curve is more complicated. Such QQ-plots mainly yield information about the relative thickness of the tails of the distribution of the data with respect to those of F. To give an impression of possible deviations from straight lines in QQ-plots a number of QQ-plots of 'true' quantile functions are plotted in Figure 3.4. In these plots the points $\{(F^{-1}(\alpha), G^{-1}(\alpha)) : \alpha \in (0, 1)\}$ are drawn for different distribution functions F and G.

For the assessment of a QQ-plot one just looks whether the points are more or less on a straight line. This illustrates the way QQ-plots are judged: informally, using prior experience.

Example 3.2 A characteristic application of QQ-plots is the examination of residuals in linear regression analysis. The next data set consists of diameters (in inches)

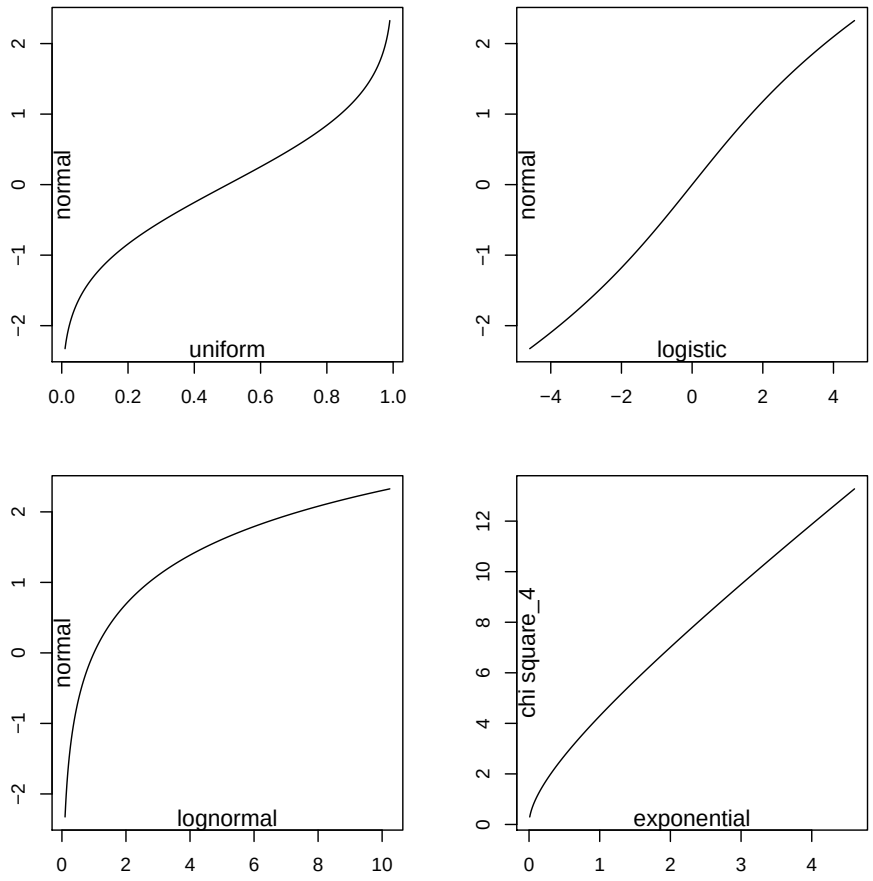


Figure 3.4: Plots of pairs of 'true' quantile functions: uniform-normal, logistic-normal, lognormal-normal, exponential-χ^2_4.

of cherry trees, measured at 4 foot and 6 inches, and their volumes (in cubic feet).

diameter	8.3	8.6	8.8	10.5	10.7	10.8	11.0	11.0	11.1	11.2	11.3
volume	10.3	10.3	10.2	16.4	18.8	19.7	15.6	18.2	22.6	19.9	24.2
diameter	11.4	11.4	11.7	12.0	12.9	12.9	13.3	13.7	13.8	14.0	
volume	21.0	21.4	21.3	19.1	22.2	33.8	27.4	25.7	24.9	34.5	
diameter	14.2	14.5	16.0	16.3	17.3	17.5	17.9	18.0	18.0	20.6	
volume	31.7	36.3	38.3	42.6	55.4	55.7	58.3	51.5	51.0	77.0	

These data were collected from cut cherry trees in order to make predictions of the timber yield (in volume) of a cherry tree from the easy to measure diameter of the uncut tree. For this purpose a linear regression model was used with volume as response variable (Y_i) and diameter as independent variable (x_i). The results are shown in Figure 3.5. The figure on the left shows a scatter plot of the data, together

with the estimated regression line $y = \hat{\alpha} + \hat{\beta}x$.

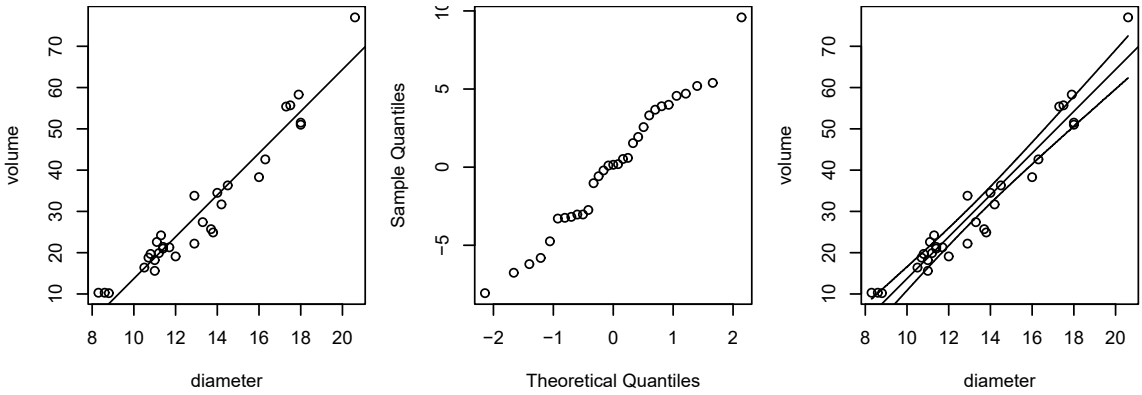


Figure 3.5: Linear regression of volume on diameter of 31 cherry trees and the QQ-plot of the residuals of the regression analysis against the normal distribution.

Often it is assumed that the measurement errors $Y_i - \alpha - \beta x_i$ are independent and $N(0, \sigma^2)$ distributed. Under this assumption it is possible to give confidence regions for the parameters and the estimated regression line. A $(1-2\alpha_0)$-confidence interval for the slope is for example given by

$$\hat{\beta} \pm t_{n-2,\alpha_0}\, s \frac{1}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}},$$

where $s^2 = (n-2)^{-1}\sum_{i=1}^n (y_i - \hat{\alpha} - \hat{\beta}x_i)^2$ is the common estimate of the variance of the measurement errors and t_{n-2,α_0} is the right α_0-point of the t-distribution with $n-2$ degrees of freedom. For the given data set the 90% confidence interval for the slope is equal to $(3.257, 4.098)$.

Under the normality assumption it is also possible to give simultaneous confidence curves for the entire regression line. The formula for these curves is

$$y = \hat{\alpha} + \hat{\beta}x \pm \sqrt{2F_{2,n-2,1-\alpha_0}}\, s \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}.$$

In this expression $F_{2,n-2,1-\alpha_0}$ is the $(1-\alpha_0)$-quantile (or the right α_0-point) of the F-distribution with 2 and $n-2$ degrees of freedom. In the graph on the right in Figure 3.5 the 90% confidence curves are drawn. When the usual assumptions about the linear regression model are valid, including the normality assumption, then the true regression line lies between the two curves for each x with a confidence of 90%.

What is the value of these confidence regions? In order to answer this question, first of all we should inspect the method of data collection. (Unfortunately, nothing is known about the collection of the cherry tree data.) Next, we should investigate the different assumptions that were made to construct the confidence regions. To

judge whether normality of the measurement errors is a plausible assumption, a QQ-plot of the *residuals* $Y_i - \hat{\alpha} - \hat{\beta}x_i$ (i.e., the vertical distances from the data points to the regression curve) was made. It is shown in the middle figure of Figure 3.5. When the measurement errors are independent and normally distributed, the residuals will be normally distributed as well. Based on the QQ-plot there is no reason to worry about this assumption.

We note that the residuals in a linear regression analysis are *not* independent and identically distributed. However, these are not strict conditions for the QQ-plot technique to be valid. A necessary condition is that the data have approximately identical marginal distributions and that the so-called quantile function of the data is a good estimator of the 'true' marginal quantile function. This condition is fulfilled in the case of the cherry tree data.

We will come back to this data set later.

***

With respect to the use of QQ-plots a word of caution is in place. One should be careful not to draw too strong conclusions based on QQ-plots. The notion of 'straight line' is rather subjective and in the case of less than 30 data points it is very difficult to distinguish two distributions anyway.

Example 3.3 The normal and logistic distributions are very difficult to distinguish when only a few measurements are available. Figure 3.6 shows QQ-plots of a sample of 20 data points from a normal distribution against the normal, the logistic and the Cauchy distribution.

***

3.3 Symplots

A random variable X is symmetrically distributed around θ if $X - \theta$ and $\theta - X$ follow the same distribution. If X has a continuous distribution, then X is symmetrically distributed around θ if its probability density is symmetric around θ. A symmetric probability distribution looks simpler than an asymmetric distribution. Moreover, the notion of 'center of the distribution' makes more sense in the case of a symmetric distribution. Therefore, one sometimes tries to find an adequate transformation of the data to obtain symmetry.

To judge whether or not a sample originates from a symmetric distribution, a histogram or a stem-and-leaf plot can be used. Naturally, the skewness parameter gives information about symmetry too, although in spite of its name one should not overestimate its usefulness: a value of zero for the skewness parameter does not necessarily mean that the distribution looks symmetric. Also other parameters can give an indication about

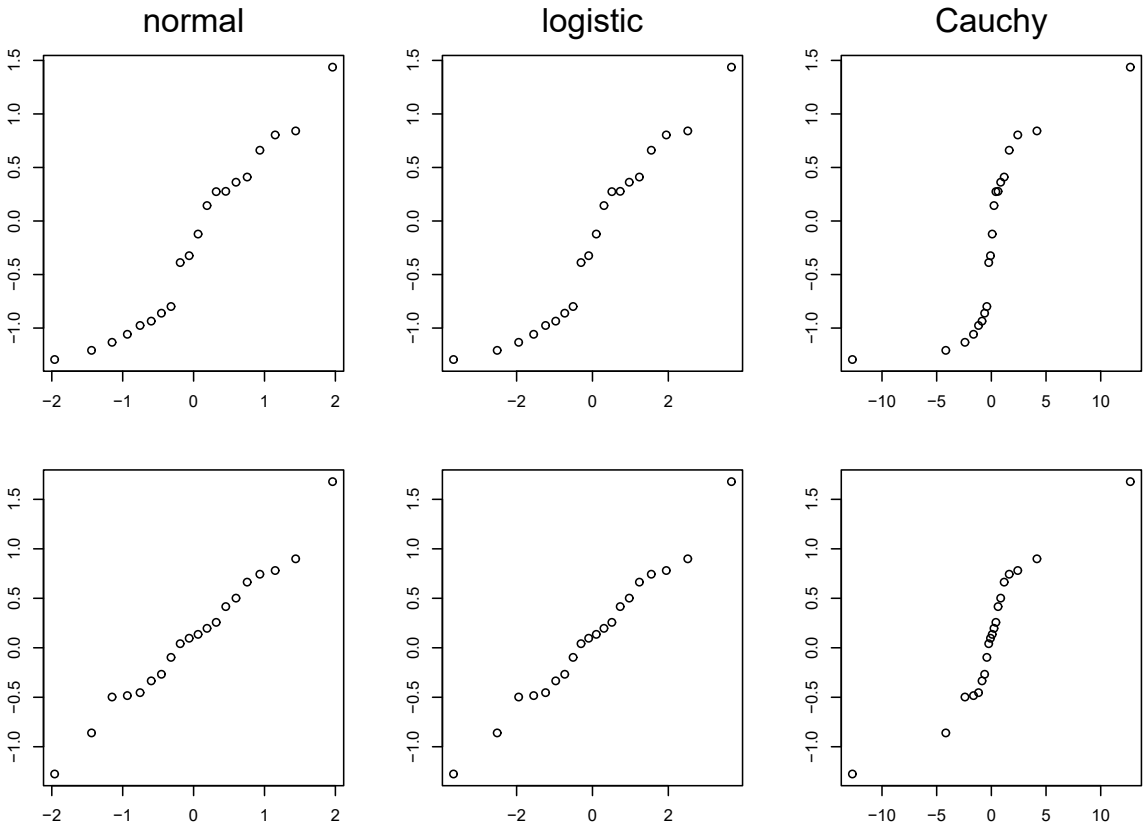


Figure 3.6: QQ-plots of two samples consisting of 20 independent realizations from the N(0,1) distribution against the quantiles of N(0,1) (left), logistic (middle) and Cauchy (right) distribution. The first (second) row shows the three QQ-plots of the first (second) sample.

the amount of symmetry of a distribution. For instance, a large difference between mean and median indicates a skewed distribution.

Skewness can also be assessed with the quantile function. From Figure 3.7 it can easily be derived that the corresponding quantile function satisfies

$$F^{-1}(1-\alpha) - \theta = \theta - F^{-1}(\alpha), \qquad \alpha \in (0,1).$$

This equation holds for each symmetric distribution F. It means that for a symmetric distribution the points $\{(\theta - F^{-1}(\alpha), F^{-1}(1-\alpha) - \theta) : \alpha \in (0,1)\}$ lie on a straight line. For data $x_1, \ldots, x_n$ from a symmetric distribution we expect that the points $\{(\text{med}(x) - x_{(i)}, x_{(n-i+1)} - \text{med}(x)) : i = 1, \ldots, [\frac{1}{2}n]\}$ also lie on a straight line. A plot of these points is called a *symmetry plot* or, briefly, a *symplot*.

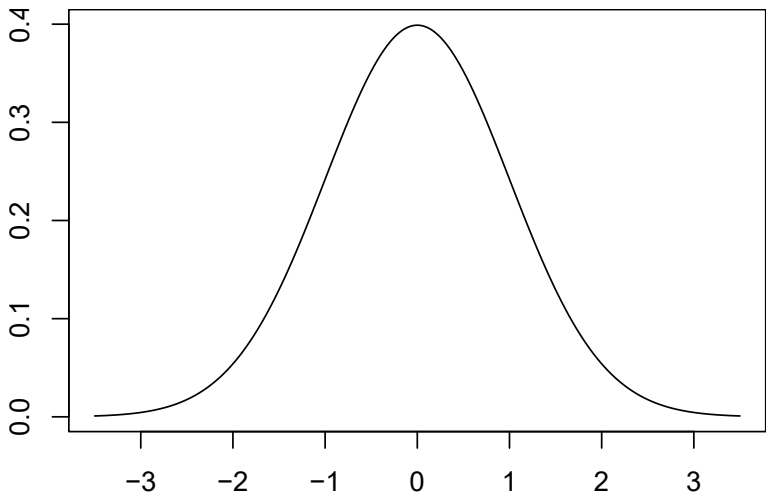


Figure 3.7: A symmetric probability density.

Example 3.4 The phenomenon of Raynaud is the coloring in three phases (white-purple/blue-red) of the fingers and/or toes when somebody is exposed to cold. Below the β-thromboglubine values of three groups of patients are given:

- 32 patients without organ impediments (PRRP)
- 41 patients with scleroderma (SDRP)
- 23 patients with certain other auto-immune illnesses (CTRP).

PRRP: 22, 25, 27, 30.5, 32.5, 34, 41, 41, 43.5, 43.5, 44.5, 44.5, 44.5, 48.5, 50.5, 53, 55.5, 58.5, 58.5, 63.5, 68.5, 68.5, 68.5, 73.5, 80, 89.5, 92, 101.5, 104, 119, 119, 124.5

SDRP: 18, 22, 23.5, 25.5, 27.5, 29.5, 29.5, 31.5, 31.5, 33, 35.5, 35.5, 36.5, 39, 43, 43, 43, 45, 46, 48.5, 49.5, 49.5, 52, 54, 56.5, 62, 65, 67.5, 69.5, 71.5, 72.5, 72.5, 76, 81.5, 95, 105, 106.5, 108.5, 130, 190, 218

CTRP: 20, 23.5, 27, 32, 41, 44, 51, 53, 55.5, 58.5, 62.5, 62.5, 65, 67, 69.5, 72, 80, 88.5, 91, 138, 146.5, 160.5, 219.

Figure 3.8 shows symplots and histograms of these three variables, which clearly have a skewed distribution. A logarithmic transformation yields a more symmetric picture (Figure 3.9), facilitating a comparison between the three data sets. The table below contains the mean values of the three variables as well as the mean values of the logarithms of the variables. Between brackets the standard deviations are given. The difference between the mean of CTRP on the one hand and the means of PRRP and SDRP on the other hand is striking, since this difference vanishes after taking the logarithms. The explanation lies partly in the large standard deviations of the data before the logarithmic transformation. Taking these large standard deviations

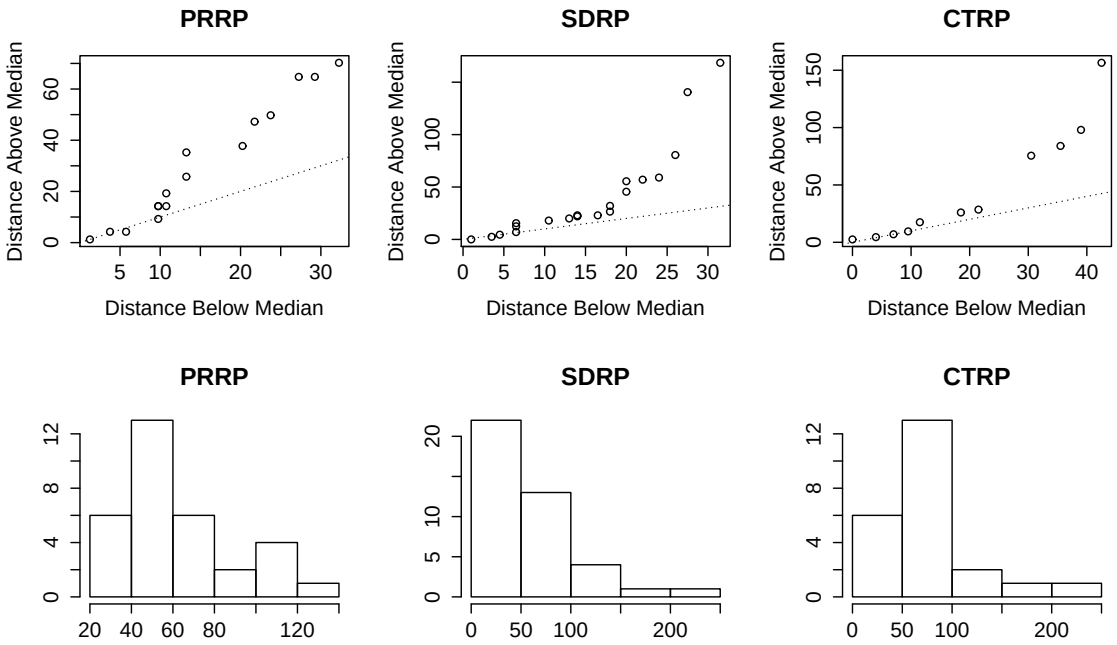


Figure 3.8: Symplots and histograms of the variables PRRP, SDRP and CTRP. The dotted line in the symplots is the line $y = x$.

into account, the initial difference is not that large. Moreover, the distribution of CTRP is rather skewed to the right. The large observations in the right tail may be the reason for the high value of the mean. By the transformation the three distributions become more comparable. The absence of an observable difference in means after transformation, suggests that the β-thromboglubine values are not substantially different in the three groups of patients. Further study is necessary in this case, for example by performing a statistical test.
***

	PRRP	SDRP	CTRP
mean	61.6(28.8)	61.9(42.0)	75.1(48.5)
mean log	4.0(0.5)	4.0(0.6)	4.1(0.6)

Table 3.1: The mean of the variables PRRP, SDRP and CTRP (top row) and the mean of the logarithms of these variables (bottom row). Standard deviations are given between brackets.

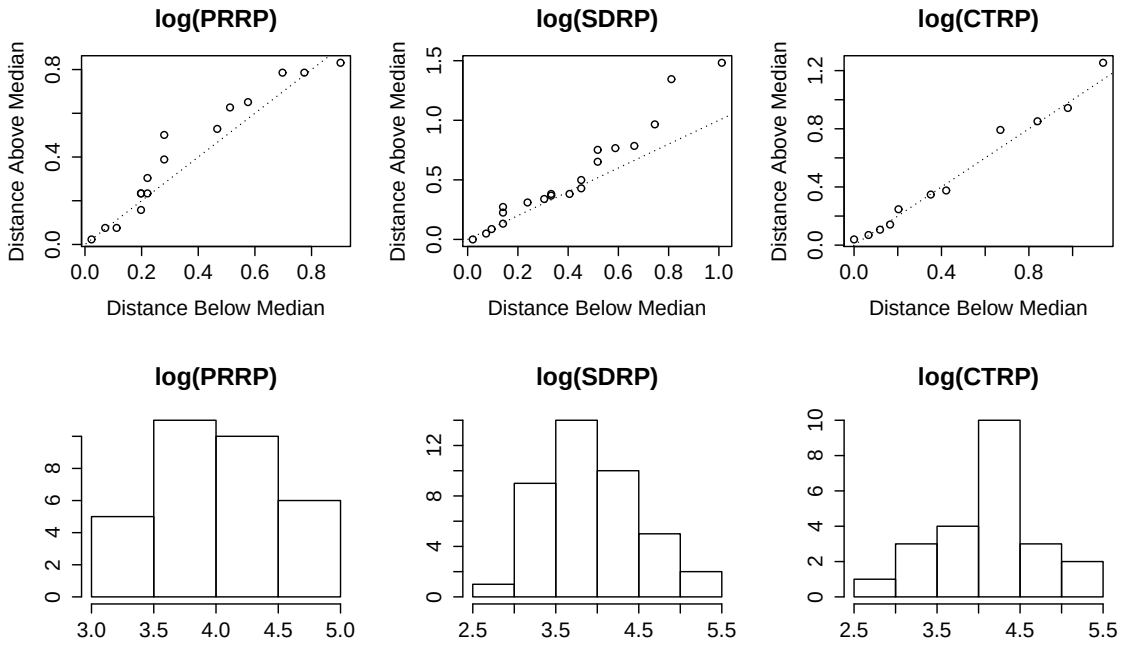


Figure 3.9: Symplots and histograms of the variables log(PRRP), log(SDRP) and log(CTRP). The dotted line in the symplots is the line $y = x$.

3.4 Two-sample QQ-plots

Suppose that $x_1, \ldots, x_m$ and $y_1, \ldots, y_n$ are two independent univariate samples from (possibly) different distributions. A graphical method to check whether the two samples originate from distributions in the same location-scale family is the *two-sample QQ-plot*, also called an *empirical QQ-plot*. When the sample sizes are equal ($m = n$), then this is simply a plot of the points $\{(x_{(i)}, y_{(i)}) : i = 1, 2, \ldots, n\}$. When $m < n$ it is a plot of the points $\{(x_{(i)}, y^*_{(i)}) : i = 1, 2, \ldots, m\}$, where

$$y^*_{(i)} = \tfrac{1}{2}\left(y_{([i\frac{n+1}{m+1}])} + y_{([i\frac{n+1}{m+1}+\frac{m}{m+1}])}\right).$$

The underlying idea is to match $x_{(i)}$ with $y_{(j)}$ for which $i/(m+1) \approx j/(n+1)$, that is, $j \approx i(n+1)/(m+1)$. Assessing a two sample QQ-plot is similar to assessing a standard QQ-plot: points approximately on a straight line indicate that the underlying distributions of the two samples belong to the same location-scale family.

Example 3.5 To compare two methods, A and B, for the production of textile, in a textile factory the number of flaws in pieces of woven fabric were counted. Samples of size 31 and 34 of pieces of fabric produced with method A and method B, respectively, yielded the following data:

A: 7, 9, 8, 10, 7, 9, 8, 6, 2, 11, 14, 8, 9, 12, 4, 6, 9, 10, 8, 11, 10, 9, 11, 8, 9, 8, 12, 10, 7, 9, 11

B: 5, 8, 7, 22, 6, 7, 2, 6, 6, 20, 7, 6, 9, 13, 4, 3, 6, 7, 7, 4, 6, 8, 6, 6, 8, 4, 11, 9, 7, 6, 17, 7, 4, 6.

The two sample QQ-plot of these data is shown in Figure 3.10. We may conclude that these two samples originate from different distributions, since the plot does not show a straight line.
***

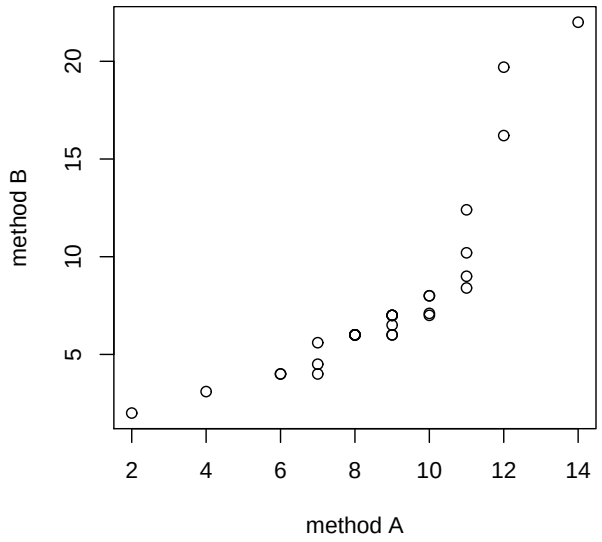


Figure 3.10: Two sample QQ-plot of the number of flaws in weaving in fabric produced by the two methods.

3.5 Goodness of fit tests

Apart from data analytical techniques, there are more formal ways to check whether data stem from a 'known' distribution. In this section a number of so-called *goodness of fit tests* are discussed. With these tests, the null hypothesis that data come from a certain distribution F, or from a member of a certain family of distributions, can be tested against the alternative hypothesis that this is not the case:

$$H_0 : F \in \mathcal{F}_0$$

$$H_1 : F \notin \mathcal{F}_0.$$

Usually the null hypothesis consists of one distribution ($\mathcal{F}_0 = \{F_0\}$) or a small family of distributions (e.g., $\mathcal{F}_0$ is a location-scale family), and, hence, the alternative hypothesis

is extensive. Unfortunately, it is not possible to find a test that has high statistical power against all possible alternatives. Therefore, we will look for a test that has reasonably high statistical power against a large number of alternatives. Such a test is called an 'omnibus test'. When such a test does not reject the null hypothesis, this is considered as an indication that the null hypothesis may be correct. This formulation is on purpose rather vague: strong conclusions can only be drawn when the amount of data is large. For example, distinguishing the logistic from the normal distribution using the tests discussed below—with standard significance levels—requires around 1000 measurements, a number that is often not available in practice. Nevertheless, goodness of fit tests are a meaningful supplement to the graphical methods discussed earlier. The results of these tests can either confirm the impression of a QQ-plot, or give an additional reason for being careful with drawing conclusions.

As all statistical tests, goodness of fit tests only yield qualitative conclusions; the conclusion of a test is about the presence of a certain effect and says little about the size of the effect. In the case of goodness of fit tests, the null hypothesis may be rejected while the true distribution is for all practical purposes very close to the null hypothesis. This problem plays a bigger role as the sample size is larger: for very large data sets the power of a test also becomes very high and as a result even the smallest deviation from the null hypothesis will almost certainly be detected.

3.5.1 Shapiro-Wilk Test

The Shapiro-Wilk test is meant for testing the null hypothesis that the observations are independent and originate from a normal distribution. The test statistic is

$$W = \frac{\left(\sum_{i=1}^{n} a_i X_{(i)}\right)^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

where $a_1, \ldots, a_n$ are certain constants. If the null hypothesis that $X_1, \ldots, X_n$ is a sample from a normal distribution with mean μ and variance σ^2 holds, then the numerator of the statistic W equals a constant times an estimator for σ^2, whereas the denominator contains an estimator for $(n-1)\sigma^2$. For non-normal distributions these two quantities are unrelated, and the value of W will change on account of that. Therefore, extreme values of W indicate a deviation from normality, and W is a useful test statistic. The word 'extreme' is relative in this case; W only takes on values in the interval (0,1]. From simulation studies it is known that for non-normal distributions the distribution of W is shifted to the left, so that it is reasonable to reject the null hypothesis for small values of W. The critical values can be found in tables or by using standard statistical software.

To understand why the test statistic is, under the null hypothesis of normality, a fraction of two estimators of the variance σ^2, we first note that the denominator contains the usual estimator for $(n-1)\sigma^2$. Next, we need to know more about the constants $a_1, \ldots, a_n$

in the numerator. For this we remark that for the order statistics $X_{(1)}, \ldots, X_{(n)}$ a linear regression model holds with $X_{(1)}, \ldots, X_{(n)}$ as the values of the response variable, with the expectations of the order statistics of a sample of size n from a *standard* normal distribution as the corresponding values of the independent variable, and with the unknown values of μ and σ as the intercept and slope coefficients, respectively. The numerator of W turns out to be a constant times the square of the least squares estimator of σ, and hence a constant times an estimator for σ^2.

To see this, let us first introduce some notation. Consider a sample $Z_1, \ldots, Z_n$ from the standard normal distribution. Write $c = (c_1, \ldots, c_n)' = (\mathrm{E}\, Z_{(1)}, \ldots, \mathrm{E}\, Z_{(n)})'$ and define Σ to be the covariance matrix of $(Z_{(1)}, \ldots, Z_{(n)})'$. Now assume that the null hypothesis is true: $X_1, \ldots, X_n$ is a sample from a normal distribution with mean μ and variance σ^2. Since X_i has the same distribution as $\sigma Z_i + \mu$ we have that

$$\mathrm{E}\, X_{(i)} = \mu + \sigma c_i, \qquad i = 1, \ldots, n.$$

This means that for $(Y_1, \ldots, Y_n) = (X_{(1)}, \ldots, X_{(n)})$ the linear regression model holds. The common α and β are now given by μ and σ, while the usual x_i are given by the c_i. There is a difference, however, between this linear regression problem and the usual linear regression setting, because the observations $(Y_1, \ldots, Y_n) = (X_{(1)}, \ldots, X_{(n)})$ are not uncorrelated in this case. One can easily derive that the covariance matrix of $(X_{(1)}, \ldots, X_{(n)})$ is equal to $\sigma^2\Sigma$. In order to find the regression line a *weighted* least squares approach is common: determine estimators $\hat{\mu}$ and $\hat{\sigma}$ by minimizing the following expression for μ and σ

$$\|\Sigma^{-\frac{1}{2}}(Y - \mu 1 - \sigma c)\|^2 = \sum_{i=1}^{n}\sum_{j=1}^{n}(\Sigma^{-1})_{i,j}(Y_i - \mu - \sigma c_i)(Y_j - \mu - \sigma c_j).$$

In this expression $\mu 1$ equals the vector with all entries equal to μ. The so obtained least squares estimator $\hat{\sigma}$ is a linear combination of the components of the observation vector Y, namely $\hat{\sigma} = c'\Sigma^{-1}Y / c'\Sigma^{-1}c$.

The coefficients a_i in the test statistic of the Shapiro-Wilk test are given by

$$a' = \frac{c'\Sigma^{-1}}{\sqrt{c'\Sigma^{-1}\Sigma^{-1}c}}.$$

Obviously, apart from a constant, the a_i equal the coefficients in $\hat{\sigma}$. This constant is chosen such that $\sum_{i=1}^{n} a_i^2 = 1$. Furthermore, it holds that $a_{n-i+1} = -a_i$, which implies $\sum_{i=1}^{n} a_i = 0$. Now it becomes clear that when the null hypothesis holds, also the numerator of the statistic W equals a constant times an estimator for σ^2.

Example 3.6 The Shapiro-Wilk test was applied to the residuals of Example 3.2. The value of the test statistic is equal to $w = 0.9789$. This corresponds to a left

p-value of approximately 0.78, and, hence, the null hypothesis is not rejected. This means that the residuals cannot be distinguished from a sample from a normal distribution, which confirms the impression of the QQ-plot in Figure 3.5.

It should be noted, that in principle this test *cannot* be used for testing normality of the residuals, because the residuals are dependent. However, for large sample sizes and a reasonable choice of design points x_i this dependence vanishes and the critical value is at least approximately correct. A correct way of applying the Shapiro-Wilk test in the present situation is to apply the test on transformed residuals for which independence does hold, as follows. Under the linear regression model the vector R of residuals has a multivariate normal distribution with covariance matrix proportional to $\Sigma = (I - X(X^TX)^{-1}X^T)$ (where $X = (1, (x_1, \ldots, x_n)^T)$, cf. Chapter 8). This matrix has rank $n-2$. Since this matrix is independent of the parameters, it is possible to find a vector of $n-2$ independent components each with the same normal distribution by a linear transformation of R. Computing the Shapiro-Wilk test statistic for this new vector yields a value of $w = 0.974$. It follows that in this case the corrected test yields only a marginally different result.
***

The Shapiro-Wilk test is to a certain extent related to a normal QQ-plot. Since $c_i = \mathrm{E}\, Z_{(i)} \approx \Phi^{-1}((i - 3/8)/(n + 2/8))$ the regression problem above can approximately be regarded as drawing a straight line in a normal QQ-plot. This formalizes the idea of finding a straight line in the QQ-plot when the observations truly come from a normal distribution. The Shapiro-Wilk test statistic is approximately based on an estimator for σ^2, obtained from the best fitting straight line (in least squares sense) in the QQ-plot.

Several other tests for normality, which relate to normal QQ-plots, exist. One example is the following test statistic:

$$W_2 = 1 - \frac{(Y - \hat{\mu}1 - \hat{\sigma}c)'\Sigma^{-1}(Y - \hat{\mu}1 - \hat{\sigma}c)}{(Y - \bar{Y}1)'\Sigma^{-1}(Y - \bar{Y}1)} = \frac{(c'\Sigma^{-1}Y)^2}{c'\Sigma^{-1}c\,(Y - \bar{Y}1)'\Sigma^{-1}(Y - \bar{Y}1)}.$$

This statistic equals the amount of explained variance in the weighted regression problem described above (and it also equals the weighted correlation coefficient between c and Y.) Small values of W_2 indicate that the points in the QQ-plot are not on a straight line, and, therefore, indicate non-normality. Other possibilities for test statistics are the amount of explained variance in an unweighted regression, or the expression above with c_i replaced by $\Phi^{-1}((i - 3/8)/(n + 2/8))$. The distribution of these statistics is more complicated. Apart from the Shapiro-Wilk type tests, that relate to QQ-plots, there exist many other tests for testing normality. In case one is specifically interested in deviations in skewness or kurtosis (which equal 0 and 3, respectively, for all normal distributions), it is better to directly base a test statistic on these statistical quantities.

3.5.2 Kolmogorov-Smirnov test

Next, we discuss two tests for the situation:

$$H_0 : F = F_0$$

$$H_1 : F \neq F_0,$$

this is, we now consider $\mathcal{F}_0 = \{F_0\}$ for a specific distribution F_0. First we consider the *Kolmogorov-Smirnov test*. In Chapter 2, the (sample) empirical distribution function of an observed sample $x_1, \ldots, x_n$ was introduced. Generally, a different sample will give a different empirical distribution function, even if the two samples were drawn from the same distribution. So it is natural to consider the corresponding stochastic version of the (sample) empirical distribution function. Let $x_1, \ldots, x_n$ be realizations from random variables $X_1, \ldots, X_n$ that are independent and identically distributed. The *empirical distribution function* of $X_1, \ldots, X_n$ is the stochastic function

$$\hat{F}_n(x) = \frac{1}{n}\sum_{j=1}^{n} 1_{\{X_j \leq x\}}.$$

The *indicator* $1_{\{X_j \leq x\}}$ equals 0 or 1 when $X_j > x$ or $X_j \leq x$, respectively. Hence, the random variable $n\hat{F}_n(x)$ equals the number $\#(X_j \leq x)$. Note that the notation for the stochastic empirical distribution function and the sample empirical distribution function are the same: both $\hat{F}_n$. When F is the true distribution function of $X_1, \ldots, X_n$, $n\hat{F}_n(x)$ is binomially distributed with parameters n and $F(x)$. Therefore, $\hat{F}_n(x)$ is an unbiased estimator of $F(x)$. Moreover, by the law of large numbers we have $\hat{F}_n(x) \to F(x)$. We may conclude that $\hat{F}_n$ is a reasonable estimator for F. For this conclusion, no assumptions on F are needed.

A test for the null hypothesis that the true distribution F of $X_1, \ldots, X_n$ is equal to F_0 can be based on a distance measure between $\hat{F}_n$ and F_0. The null hypothesis is rejected when the distance between these two functions is large. The Kolmogorov-Smirnov test is based on the maximum vertical distance between $\hat{F}_n$ and F_0: its test statistic is

$$D_n = \sup_{-\infty < x < \infty} \left| \hat{F}_n(x) - F_0(x) \right|.$$

The null hypothesis is rejected for large values of D_n. However, the random variable D_n does not have one of the well-known distributions. Its critical values can be found from tables or from statistical software[2]. One of the properties of the Kolmogorov-Smirnov test that makes this test valuable, is that the distribution of D_n under the null hypothesis that the true distribution equals F_0, is the same for each *continuous* distribution, and

[2]See for example the book *Empirical Processes with Applications to Statistics* by Shorack and Wellner.

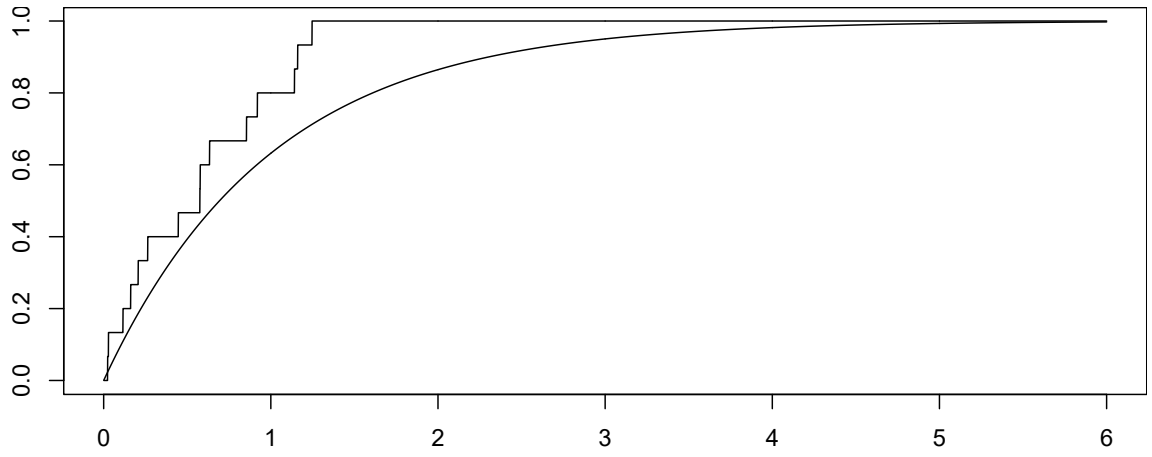


Figure 3.11: Graph of the empirical distribution function (step function) of 15 observations from an exponential distribution, together with its true distribution function (smooth curve).

hence does not depend on F_0. This is expressed as: D_n is *distribution free* over the class of continuous distribution functions, and one says that the Kolmogorov-Smirnov test belongs to the class of distribution free, or *nonparametric*, tests. This means that the critical value of the test is also independent of F_0 and that one table suffices to perform the test. This property of D_n is formulated as a theorem. The proof of the theorem shows a simple way of computing D_n.

Theorem 3.1 *The statistic D_n is distribution free over the class of continuous distribution functions.*

Proof. From Figure 3.11 it can easily be seen that the empirical distribution function $\hat{F}_n(x)$ equals i/n in $X_{(i)}$ while it equals $(i-1)/n$ just left of $X_{(i)}$. Therefore,

$$D_n = \max_{1 \le i \le n} \max \left\{ \left| \frac{i}{n} - F_0(X_{(i)}) \right|, \left| \frac{i-1}{n} - F_0(X_{(i)}) \right| \right\}.$$

Since under H_0, $X_1, \ldots, X_n$ is a sample from F_0, we have that $F_0(X_1), \ldots, F_0(X_n)$ is a sample from the uniform distribution on [0,1] (the so-called 'integral transformation'). By the monotonicity of F_0, the distribution of the vector $(F_0(X_{(1)}), \ldots, F_0(X_{(n)}))$ is equal to the distribution of the order statistics of a sample from the uniform distribution. Since this distribution is independent of F_0, the distribution of D_n is also independent of F_0. ∎

Note that the Kolmogorov-Smirnov test is, in the given form, only applicable for testing simple hypotheses, because the postulated null-distribution F_0 occurs in the definition of the test statistic D_n. This is why in this simple form the test cannot be used to test

'normality', that is, to test the composite null hypothesis that the true distribution is a member of the location-scale family of normal distributions. However, the test can be adapted for such purposes. For example, a test for normality can be based on the adapted statistic

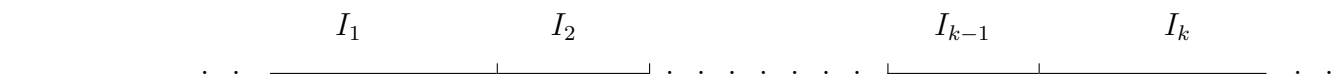
$$\tilde{D}_n = \sup_{-\infty<x<\infty} \left| \hat{F}_n(x) - \Phi(S^{-1}(x - \bar{X})) \right|,$$

where $\bar{X}$ and S^2 are the sample mean and the sample variance, respectively. Finding critical values for such a test is not straightforward. Theorem 3.1 does not hold anymore. It would be incorrect to use the critical values of the Kolmogorov-Smirnov test statistic for this adapted test statistic. In Chapter 5 a method is presented to obtain approximate values for the correct critical values for the adapted test statistic. We refer to the literature for more information about the distribution of the test statistic based on the estimated parameters (e.g., Shorack and Wellner (1986)) as well as for related tests, like the *Cramér-von Mises test*, which is based on the distance measure

$$CM_n = \int_{-\infty}^{\infty} (\hat{F}_n(x) - F_0(x))^2 \, dF_0(x).$$

3.5.3 Chi-Square tests

Alternative tests to the Kolmogorov-Smirnov test, for the same set of hypotheses, are found in the class of *chi-square tests*. For a chi-square test, the real line is divided into adjacent intervals $I_1, \ldots, I_k$:



The number of observations in interval I_i is defined as N_i and the test statistic is given by

$$X^2 = \sum_{i=1}^{k} \frac{[N_i - np_i]^2}{np_i},$$

where $p_i = F_0\{I_i\}$ is the probability under F_0 that an observation lies in I_i. The number np_i is the expectation of N_i when the true distribution of each of the observations $X_1, \ldots, X_n$ is equal to F_0. The statistic X^2 thus measures, in a weighted way, the deviation of the observed frequencies N_i from their expected values under the null hypothesis. The null hypothesis is rejected for large values of X^2. The exact distribution of the statistic under the null hypothesis, and hence the critical value, depends on F_0 and the choice of the intervals I_i. However, for large values of n and for fixed k this distribution is approximately equal to the chi-square distribution with $k - 1$ degrees of freedom, which we denote as χ^2_{k-1}-distribution. As a rule of thumb, this approximation is reliable when

the expected number of observations under the null hypothesis equals at least 5 in each interval.

The theoretical justification of using the χ^2 distribution lies in the following (see also Chapter 7). The vector of cell frequencies $N(n) = (N_1, \ldots, N_k)$ has a multinomial distribution with parameters n and $p_1, \ldots, p_k$. It can be shown that for a fixed number of cells k, the variable $X^2 = X^2(n)$ converges, for $n \to \infty$, in distribution to a χ^2_{k-1}-distribution, that is,

$$P(X^2 \leq x) \to P(\chi^2_{k-1} \leq x), \qquad \text{for all } x,$$

where χ^2_{k-1} now also represents a random variable with the χ^2_{k-1}-distribution. This result is closely related to the Central Limit Theorem and the fact that a sum of squares of independent random variables with the standard normal distribution follows a chi-square distribution.

Like the Kolmogorov-Smirnov test, the chi-square test is in its simplest form only applicable to test simple null hypotheses. For testing a composite null hypothesis, consisting of more than one distribution function, the statistic and the critical values have to be adapted. Usually, the following statistic is used in such a case:

$$\tilde{X}^2 = \sum_{i=1}^{k} \frac{[N_i - n\hat{p}_i]^2}{n\hat{p}_i},$$

where $\hat{p}_i$ is an estimator of p_i, $i = 1, \ldots, k$. The distribution of this adapted statistic depends on which estimators $\hat{p}_i$ are used. When the maximum likelihood estimators based on the distribution of $(N_1, \ldots, N_k)$ under the null hypothesis are chosen (this is a multinomial distribution), and when the null hypothesis consists of an m-dimensional parametric model, then in most cases the distribution of the adapted statistic approximately equals the χ^2-distribution with $(k-m-1)$ degrees of freedom. It is said that for each parameter to be estimated, one degree of freedom gets lost. For example, when the null hypothesis consists of a location-scale family, which is a 2-dimensional model, then the distribution of the test statistic is the χ^2_{k-3}-distribution.

It should be emphasized that the approximation by the chi-square distribution with $(k - m - 1)$ degrees of freedom does not hold when other estimators are used. In those cases the approximation is generally bad. For example, for testing the null hypothesis of normality it is natural to estimate the p_i with the probability of finding an observation in I_i under the $N(\bar{X}, S^2)$-distribution. This $\hat{p}_i$ depends on $\bar{X}$ and S^2 and is not the maximum likelihood estimator based on the vector of cell frequencies $(N_1, \ldots, N_k)$. A reasonable approximation using a chi-square distribution is not possible in this situation. To assume the distribution to be χ^2_{k-3} would lead to incorrect p-values. The difference with the true values is substantial. The maximum likelihood estimators for the parameters $p_1, \ldots, p_k$ based on the distribution of $(N_1, \ldots, N_k)$ under the null hypothesis are in

this case $\hat{p}_1, \ldots, \hat{p}_k$ where $\hat{p}_i = \Phi((I_i - \hat{\mu})/\hat{\sigma})$ for the values $\hat{\mu}$ and $\hat{\sigma}$ that maximize the multinomial probability

$$\binom{n}{N_1, \ldots, N_k} \prod_{i=1}^{k} \Phi((I_i - \mu)/\sigma)^{N_i}.$$

Here we write $\Phi((I - \mu)/\sigma)$ for the probability of finding an observation in I under the $N(\mu, \sigma^2)$-distribution. The values of μ and σ^2 that maximize this multinomial likelihood are in general not equal to $\bar{X}$ and S^2.

Chapter 4

Density estimation

A random variable X is said to have a *probability density function* or, shorter, *density* f if

$$P(a < X < b) = \int_a^b f(t)dt \quad \text{for all } -\infty < a \leq b < \infty.$$

Assume that a random sample $X_1, \ldots, X_n$ originates from a probability distribution with an unknown density. *Density estimation* is the reconstruction of the 'true' density based on the data. A *density estimator* is a stochastic function $\hat{f}(t) = \hat{f}(t, X_1, \ldots, X_n)$ that, in every point t, should approximate the true density $f(t)$ for all realizations of the random sample. As usual, let us omit the notion of the observations: just the hat of $\hat{f}(t)$ reminds us that this concerns a stochastic function. For every t, $\hat{f}(t)$ is an estimator of $f(t)$; however, we prefer to interpret a density estimator as a stochastic function $t \mapsto \hat{f}(t)$. A density in a single point has little meaning – it is the density as a whole which is of interest.

If the observations originate from a probability distribution f_θ in which only the parameter θ is unknown, then the obvious density estimator is $f_{\hat{\theta}}$ where $\hat{\theta}$ is an estimator of the unknown parameter θ. This is called the *parametric method*. The estimator $\hat{\theta}$ can be obtained by any of the methods from classical statistics, such as the maximum likelihood method, the least squares method, or the method of moments.

In many situations, however, a *nonparametric* method is more attractive. By this we understand a method for which it is not necessary to make any strong assumptions on the form of the true density. Instead, density estimators are often used for exploratory analyses and summaries of the data rather than for drawing formal conclusions. It is better to let the data 'speak for themselves'.

Example 4.1 Figure 4.1 displays estimates of two densities of (transformed) mast cell counts[1] in a study about the causes of sudden infant death. The curve A gives

[1]Mast cells belong to the blood and immune system. Due to injuries, local infections, or immune reactions, mast cells eject 'histamine' which widens the blood vessels and it makes them permeable such that serum proteins and leukocytes can reach the injured location. Apart from that, mast cells play a role in the termination of the reaction.

an estimate of the density of these variables based on observations of 95 infants who suddenly died without any obvious reason (sudden infant death). In contrast, the curve B gives a density estimate based on the same transformed measurements from a group of 76 infants who died from a known cause. Based on the form of the estimates, scientists cautiously concluded that the 'true' density of the mast cell counts could be a mixture of the density B and a second density with a somewhat bigger average. A closer inspection of the data revealed that the mast cell numbers of a third of the sudden infant death children seemed to be extremely large. This example illustrates how density estimates can give rise to ideas for further research. However, one should not be misled to drawing strong conclusions from Figure 4.1. Unfortunately, there is still little known about the causes of sudden infant death.
***

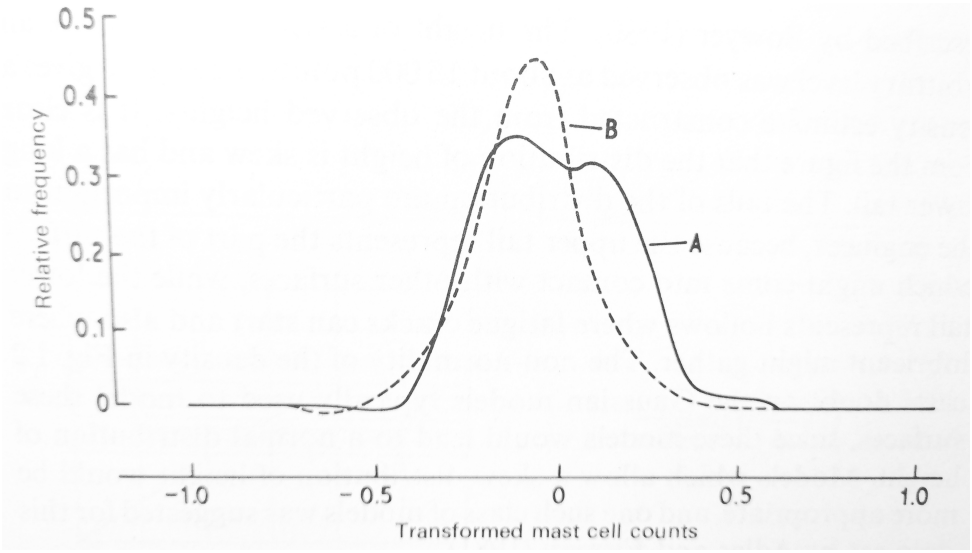


Figure 4.1: Density estimates for transformed mast cell counts in a studie about the causes of sudden infant death. A: sudden infant death; B: control group. Graphic from Silverman (1986).

It is not a simple task to estimate densities. Many observations are necessary for precise estimation. Anyhow, even based on a small sample and if one uses critical reasoning, density estimators give a quite useful insight in aspects of the data such as unimodality, skewness, heaviness of the tails, etc. For analyses in connection with distributions, they are a useful addition to QQ-plots, but cannot replace them.

Out of different methods for nonparametric density estimation, we treat in particular kernel density estimators. We conclude this introduction with a method that we already discussed earlier: the histogram. The histogram, if it is correctly normalized, is indeed an example of a density estimator, perhaps the simplest possible. Simplicity is often an

advantage but in this case there are several drawbacks. The most important problem of the histogram is its discontinuity. This is not only in contrast to our expectation that the true density should often be continuous. It also leads to a certain degree of arbitrariness in the obtained estimation. This is illustrated in Example 4.2.

Example 4.2 *Old Faithful* is a famous geyser in Yellowstone National Park in Wyoming, USA. Tourists enjoy that, at random times, the geyser ejects water into the air. Two histograms of observations of 107 eruption durations are given in Figure 4.2. The histograms display the same observations and use the same bin widths. The only difference is that, relative to the first histogram, the axis partition of the second histogram is shifted. Even though a statistically trained eye would draw the same conclusions from both histograms, both histograms give a quite different impression of the true distribution of the eruption lengths.
***

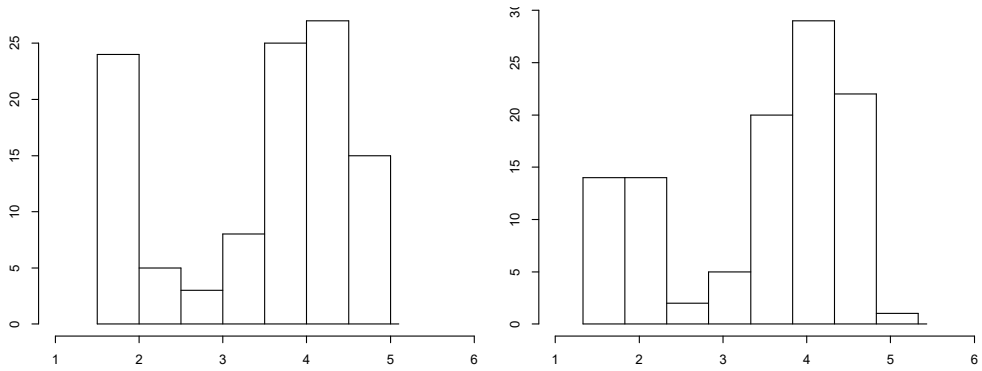


Figure 4.2: Histograms of 107 eruption durations of Old Faithful (in minutes).

4.1 Kernel density estimators

Let K be a given probability density with expectation 0 and variance 1, for example, the standard normal density. A *kernel density estimator* with *kernel* or *window* K is defined by

$$\hat{f}(t) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\Big(\frac{t-X_i}{h}\Big).$$

Here, h is a positive number that is to be specified. It is called the *window width* or *bandwidth*. Any choice of K and h leads to a kernel density estimator. We will later see that the choice of the kernel K is not crucial but that the quality of the estimator will strongly depend on the bandwidth.

Example 4.3 For a sample of size 15 from the normal distribution, kernel density estimators are derived for different bandwidths but with the same standard normal kernel. Figure 4.3 displays the results; each of the three graphs show a kernel estimation together with the true density. It seems as if small as well as large choices of the bandwidths lead to bad estimates. In case of a small bandwidth, the density estimation has many peaks, whereas a large bandwidth results in a too smooth density. The intermediate value leads to a quite precise estimate, despite the small sample size.
*****

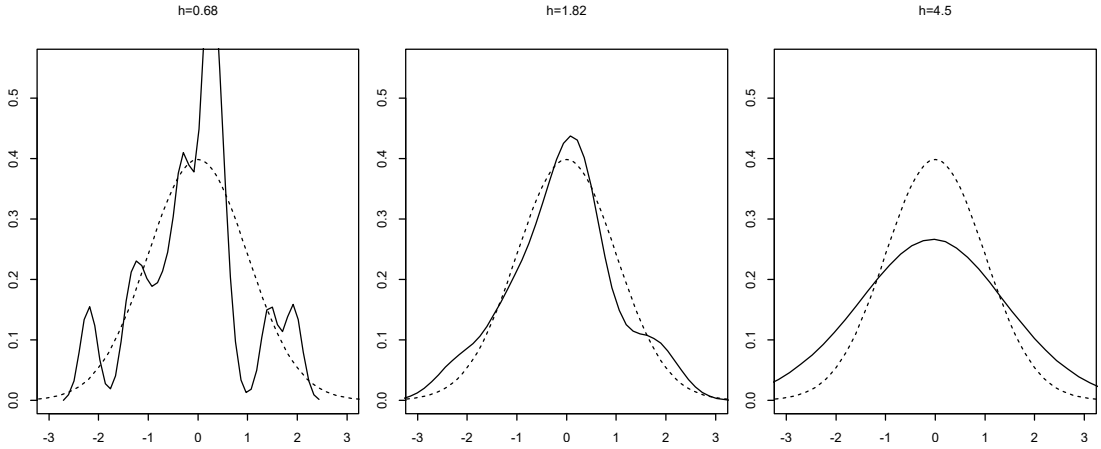


Figure 4.3: Kernel estimates of the density of a sample of size 15 from a standard normal distribution for different bandwidths but with the same normal kernel. The dashed line displays the true density.

A kernel density estimator is an example of a *smoothing method*. One can think of the construction of a density estimation as spreading the total probability mass of 1 around the observations. In the case of independent and identically distributed observations, it is reasonable to first allocate masses of size $1/n$ to each of the n observations. A kernel density estimator then spreads each of these masses around the different X_i, not homogeneously but in accordance with the kernel function and the bandwidth.

More formally, we could consider a kernel density estimator $\hat{f}(t)$ as a sum of the n small hills which are given by the functions

$$t \longmapsto \frac{1}{nh} K\Big(\frac{t - X_i}{h}\Big).$$

Each such hill is centered around the observations X_i and has an area of $1/n$ under the curve, irrespective of the value of the bandwidth h. For smaller bandwidths, the hill is

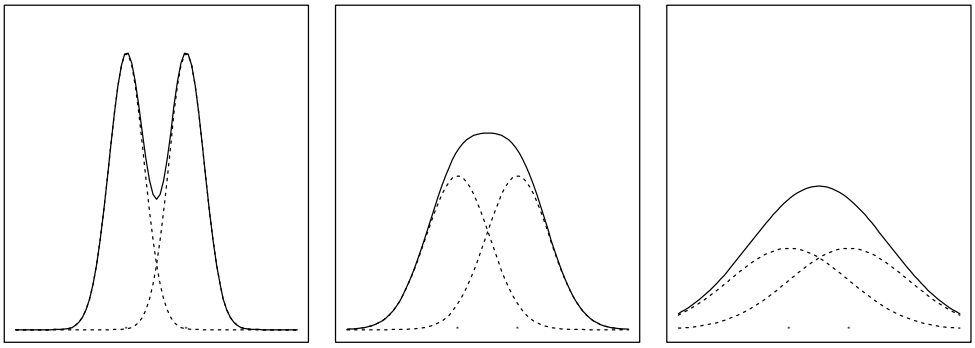


Figure 4.4: The kernel density estimator with the normal kernel based on two observations for three different bandwidths: small (left), medium (middle), and large (right). The figures show the separate contributions of both observations (dashed lines) and the resulting kernel estimator that equals the sum of the dashed lines.

strongly concentrated whereas the hill is very flat for large bandwidths; see Figure 4.4 for an illustration.

This explains the results in Figure 4.3. For a small bandwidth, there is little overlap among the $n = 15$ hills. Consequently, the different peaks are clearly visible in the sum of the n hills. For a large bandwidth, all hills completely overlap which means that the probability mass is spread over a too large region. For intermediate values of the bandwidth, the hills add up to a reasonable result. In the following section, we discuss methods for choosing a suitable bandwidth.

4.2 Choice of kernel and bandwidth

One possibility to choose a suitable kernel and bandwidth is to create multiple density estimates and to choose out of these the ones which 'seem the best'. For an exploratory use of kernel density estimators this 'subjective method' works quite well. Often, the different estimates could even point out different aspects of the data and hence could give rise to different hypotheses about the true underlying distribution. These can then be tested on the basis of newly collected data sets.

For presenting the data it is in general better to choose a somewhat too small bandwidth rather than a too large one. Using her/his imagination, a beholder could still flatten flatten a peaky density but, on the other hand, it is impossible to recreate possible peaks from a too smooth density estimate.

Anyway, there is a need of a more objective method for choosing a suitable bandwidth. For this it is first necessary to find a criterion that can evaluate the quality of a density estimator. In a next step, we choose a kernel and a bandwidth that, according to the criterion, will result in the highest quality. An often used criterion is the mean inte-

grated squared error. The *mean squared error* (MSE) is defined as the expected quadratic deviation of the estimator $\hat{f}(t)$ from the estimand $f(t)$, for a fixed value of t:

$$\mathrm{MSE}(\hat{f}(t)) = E(\hat{f}(t) - f(t))^2.$$

As always, this equals the sum of the variance and the squared bias of the estimator:

$$\mathrm{MSE}(\hat{f}(t)) = \mathrm{var}(\hat{f}(t)) + (E\hat{f}(t) - f(t))^2.$$

The mean squared error is a measure of the quality of $\hat{f}(t)$ for every single t. One manner to derive from this the quality of the estimator as a function is to integrate over all values of t. This brings us to the *mean integrated squared error* (MISE):

$$\begin{aligned} \mathrm{MISE}(\hat{f}(t)) &= \int E(\hat{f}(t) - f(t))^2 dt \\ &= \int \mathrm{var}(\hat{f}(t)) dt + \int (E\hat{f}(t) - f(t))^2 dt. \end{aligned} \tag{4.1}$$

We are in search for a density estimator $\hat{f}$ with a small mean integrated squared error.

For each fixed t, a kernel density estimator is an average of n random variables $h^{-1}K((t - X_i)/h)$. That is why its expectation and variance are easily derived. In particular,

$$E\hat{f}(t) = \int \frac{1}{h} K\Big(\frac{t-x}{h}\Big) f(x) dx \qquad \text{and}$$

$$\mathrm{var}\hat{f}(t) = \frac{1}{n}\Big[\int \frac{1}{h^2} K\Big(\frac{t-x}{h}\Big)^2 f(x) dx - \Big\{ \int \frac{1}{h} K\Big(\frac{t-x}{h}\Big) f(x) dx \Big\}^2 \Big].$$

Inserting these formulas into MISE, we find exact expressions for the mean integrated squared error. However, the obtained formula is rather involved and it seems more useful to consider an approximation of the mean integrated squared error.

By a change of variable $x' = (t - x)/h$ in the formula for $E\hat{f}(t)$, we can express the bias as

$$E\hat{f}(t) - f(t) = \int K(x') f(t - hx') dx' - f(t).$$

In the case that the true density is twice continuously differentiable, we obtain by a Taylor expansion for small values of hx that

$$f(t - hx) = f(t) - hx f'(t) + \frac{1}{2} h^2 x^2 f''(t) + \dots \ .$$

The kernel K is a probability density with expectation 0 and variance 1. Using this, we find that

$$E\hat{f}(t) - f(t) = \int K(x)(f(t) - hx f'(t) + \frac{1}{2} h^2 x^2 f''(t) + \dots) dx - f(t) = \frac{1}{2} h^2 f''(t) + \dots \ .$$

Appropriate conditions on f ensure that the remainder term (the dots) is negligible for small values of the bandwidth h.

The integral of the square of the bias is the second term in the sum (4.1). Using the lastly obtained formula, we see that this integral is equal to

$$\int (E\hat{f}(t) - f(t))^2 dt \approx \frac{1}{4}h^4 \int (f''(t))^2 dt.$$

The variance term in the mean integrated squared error can be bounded with the help of a change of variable and interchanging the order of integration:

$$\begin{aligned}
\int \mathrm{var}(\hat{f}(t))dt &\leq \int \frac{1}{nh^2} \int K\Big(\frac{t-x}{h}\Big)^2 f(x)dxdt \\
&= \frac{1}{nh} \int \int K(x')^2 f(t-hx')dx'dt \\
&= \frac{1}{nh} \int \int f(t-hx')dtK(x')^2 dx' = \frac{1}{nh} \int K(x')^2 dx'.
\end{aligned}$$

Hence, an upper bound of the MISE of a kernel density estimator approximately equals the sum of these two terms:

$$\mathrm{MISE}(\hat{f}(t)) \lesssim \frac{1}{nh} \int K(x)^2 dx + \frac{1}{4}h^4 \int (f''(t))^2 dt. \tag{4.2}$$

This formula illustrates the bias-variance trade-off: the bias of the kernel density estimator can be minimized by choosing a small bandwidth but exactly this would lead to a large variance. The bias term grows proportionally to h^4 whereas the variance term grows inversely proportionally to h. The smallest MISE is obtained by balancing the bias and the variance.

For a given kernel K, we choose a bandwidth that minimizes the mean integrated squared error. When we, instead, minimize the just obtained approximation of the MISE, we find the optimal bandwidth as

$$h_{opt} = \Big\{ \int K(x)^2 dx \Big\}^{1/5} \Big\{ \int (f''(t))^2 dt \Big\}^{-1/5} n^{-1/5}.$$

For finding the approximation above we used some imprecise, heuristic arguments. However, one can show theoretically that the above derivation is asymptotically correct. This means that, if the number of observations grows to infinity, then h_{opt} indeed specifies the optimal bandwidth in the sense of minimizing the MISE for a density for which it is assumed that the first two derivatives exist.[2]

[2]If it can be assumed that the true density has more than two derivatives, better results can be obtained by means of so-called higher order kernels. The formula for h_{opt} is in this case incorrect. However, such assumptions do not align well with our aim of studying nonparametric density estimators.

At first glance, the formula for the optimal bandwidth seems disappointing because h_{opt} does not only depend on the kernel but also on the unknown density through the factor $\int f''(t)^2 dt$. Nevertheless, we can draw a number of important conclusions. First, the optimal bandwidth converges to zero when the number of observations increases; this happens slowly though, at a rate proportionally to $n^{-1/5}$. Thus, if a larger sample is available, a smaller bandwidth should be chosen. Furthermore, there is a dependency between the bandwidth and the (unknown) density trough the 'smoothness' of f. In the case of a strongly fluctuating density, the second derivative f'' would strongly fluctuate as well and $\int f''(t)^2 dt$ would have a large value. For such an f, the formula for h_{opt} suggests that a relatively small bandwidth should be chosen. This qualitative conclusion is also intuitively clear: a large bandwidth leads to a smooth density estimation which is a good result if the density is indeed smooth. However, such an estimate would be suboptimal if the true density is irregular. The precise dependency of the correct bandwidth on $\int f''(t)^2$ is of course not intuitively clear. We can see from this the dependency of h_{opt} on the scale parameter in a location-scale family of densities $f_{\mu,\sigma}(t) = \sigma^{-1} f((t-\mu)/\sigma)$. Because of

$$\int f''_{\mu,\sigma}(t)^2 dt = \frac{1}{\sigma^5} \int f''(t)^2 dt,$$

the optimal bandwidth increases proportionally to σ. The qualitative conclusion is that, for unimodal, smooth densities such as a normal density, a large bandwidth relatively to the standard deviation is favorable.

Finally, the formula for the optimal bandwidth has the following implications for the choice of the kernel. Inserting h_{opt} in the approximation (4.2) for the MISE, we find that

$$\mathrm{MISE}(\hat{f}_{opt}) \lesssim \frac{5}{4} \Big\{ \int K^2(x) dx \Big\}^{4/5} \Big\{ \int f''(t)^2 dt \Big\}^{1/5} n^{-4/5}.$$

Except for the true density, this formula only depends on the kernel and hence expresses the influence of the kernel on the quality of the density estimator (if indeed the optimal bandwidth is being used). The best kernel is the one that minimizes the constant $\int K^2(x) dx$, and hence the mean integrated squared error. Remember that the kernel has to be a probability density with mean zero and variance 1; this is what we used during the derivation of the optimal bandwidth. The optimal kernel is the so-called *Epanechnikov kernel* which is given by

$$K_e(x) = \frac{3}{4\sqrt{5}} \Big(1 - \frac{1}{5} x^2 \Big) \quad \text{for } -\sqrt{5} \leq x \leq \sqrt{5},$$

and is constantly zero otherwise. However, this kernel is rarely used. The reason for this is that other kernels produce bigger values $\int K^2(x) dx$ that are still reasonably close to that of the Epanechnikov kernel. For example, the normal kernel yields a MISE that is just 1.05130 times as large as the MISE for the Epanechnikov kernel. The conclusion is that

the use of the correct bandwidth makes the choice of the kernel of secondary importance. Often, the normal kernel is chosen for the simple reason that the normal density looks 'nicely'. This choice has nothing to do with the question whether the data are normally distributed.

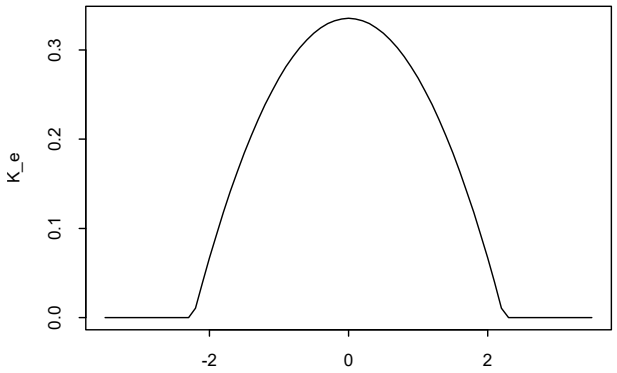


Figure 4.5: The Epanechnikov kernel.

How to choose an appropriate bandwidth? Since the optimal bandwidth depends on the unknown, true density, we should be content with an approximation of h_{opt}. We will now discuss two methods which work reasonably well.

For the first method, we replace $\int f''(t)^2 dt$ by an estimate in the formula for h_{opt} based on the assumption that f belongs to a certain parametric class of densities. For example, we assume that f is a normal density with unknown mean μ and variance σ^2. In this case, it can be calculated that

$$\int f''(t)^2 dt = 3/8\pi^{-1/2}\sigma^{-5} \approx 0.212\sigma^{-5}.$$

In this formula, the only unknown part is the value of σ. We can replace it by an estimate, for example, the sample standard deviation. The resulting estimate of $\int f''(t)^2 dt$ can be used in the formula for h_{opt}. If we use for K the normal kernel, we obtain the bandwidth

$$\hat{h}_{opt} = (4\pi)^{-1/10}\Big(\frac{3}{8}\pi^{-1/2}\Big)^{-1/5}\hat{\sigma}n^{-1/5} \approx 1.06\hat{\sigma}n^{-1/5}. \tag{4.3}$$

If the true underlying density is indeed close to a normal density, then $\hat{h}_{opt}$ will be close to h_{opt}. For most of the remaining cases, this formula works reasonably well, even if the true density is non-normal. Only in cases of multimodal or strongly fluctuating densities, the just obtained rule seems to result in a too large bandwidth: for such distributions, the true value of $\int f''(t)^2 dt$ is, relative to the standard deviation, larger than it would be in the case of a normal distribution. This problem can partially be attenuated by using for $\hat{\sigma}$, instead of the sample standard deviation,

$$\hat{\sigma} = \min(\text{sample standard deviation, interquartile range}/1.34).$$

Example 4.4 The kernel estimate in the middle plot in Figure 4.3 was found by using the formula (4.3). The data were generated according to a normal distribution which is why h_{opt} in this case indeed provides the optimal bandwidth (when using the normal kernel). The result in Figure 4.3 is not surprising.
***

Example 4.5 Figure 4.6 shows kernel density estimates of the eruption durations of Old Faithful. For the first five estimates a normal kernel has been used, together with the following bandwidths (from left to right and top to bottom): $\hat{h}_{opt}/4, \hat{h}_{opt}/3, \hat{h}_{opt}/2, \hat{h}_{opt}, \frac{3}{2} \cdot \hat{h}_{opt}$, where $\hat{h}_{opt}$ is given by (4.3). It is apparent that the use of $\hat{h}_{opt}$ in this case results in an 'over-smoothing': the gap between the two peaks, which is so characteristic in the histograms and also clearly visible in the first three kernel estimates, is obscured in the fourth kernel estimate. This is in accordance with the theory. The sixth plot in Figure 4.6 is a density estimate based on the uniform kernel. Disregarding the crudeness of the graph, the difference from e.g. the second kernel estimate is rather small.
***

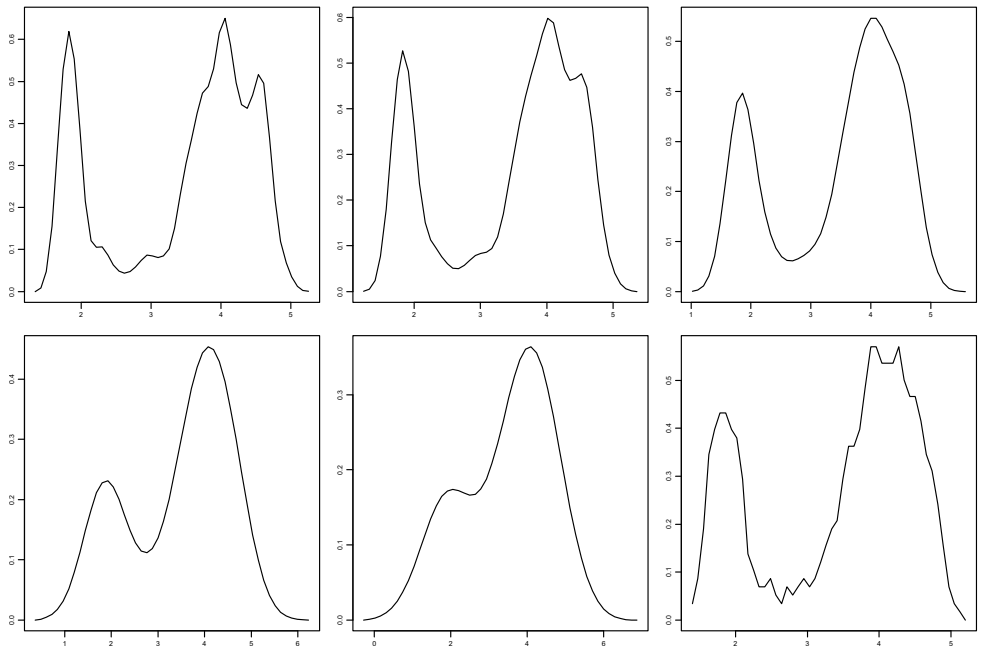


Figure 4.6: Kernel estimates of the density of the eruption durations of the Old Faithful (in minutes).

4.3 Cross-validation

Cross-validation is a method for finding an appropriate bandwidth without further assumptions on the data distribution. There are different regimes for cross-validation but we will only discuss the one based on the integrated squared error.

For a given density estimator, the *integrated squared error* (ISE) is given by

$$\mathrm{ISE}(\hat{f}) = \int (\hat{f}(t) - f(t))^2 dt = \int \hat{f}(t)^2 dt - 2\int \hat{f}(t) f(t) dt + \int f(t)^2 dt.$$

The 'best' density estimator $\hat{f}$ is closest to f and has hence (on average) the smallest value for the integrated squared error. The term $\int f(t)^2 dt$ in the sum on the right-hand side of the previous display only depends on the true density hence cannot be influenced by the choice of the estimator. Thus, the best density estimator also minimizes the expression

$$R(\hat{f}) = \int \hat{f}(t)^2 dt - 2\int \hat{f}(t) f(t) dt.$$

Here, the unknown density is involved one more time. The idea is to replace $R(\hat{f})$ by an estimator $\hat{R}(\hat{f})$ (that does not depend on f anymore) and then to choose a density estimator $\hat{f}$ that minimizes $\hat{R}(\hat{f})$.

The cross-validation estimator of $R(\hat{f})$ is found as follows. If the random variable Y has the same density f as $X_1, \ldots, X_n$ and is independent of those, it holds that

$$E(\hat{f}(Y) \mid X_1, \ldots, X_n) = \int \hat{f}(y) f(y) dy.$$

Here, the expectation on the left-hand side is the conditional expectation, given the values of $X_1, \ldots, X_n$ (the presence of which is expressed by the hats on both sides of the equation). Thus, the random variable $\hat{f}(Y)$ has as a conditional expectation exactly $(-\frac{1}{2})$ times the value of the second term in $R(\hat{f})$. If we had the additional measurements $Y_1, \ldots, Y_m$ at our disposal, we would be able to estimate $R(\hat{f})$ by

$$\int \hat{f}(t)^2 dt - 2\frac{1}{m}\sum_{i=1}^{m} \hat{f}(Y_i).$$

Considering that we, unfortunately, do not have additional measurements at our disposal, we are going to pursue the 'cross-validation principle'. For a fixed i let $\hat{f}_{-i}$ be the density estimator that is based on the observations $X_1, \ldots, X_n$, *except for the i-th.* In the following, X_i is going to play the role of Y_i and $\hat{f}_{-i}$ the role of $\hat{f}$. The situation is exactly as before, just that we are working with $n-1$ instead of n observations which does not

make a big difference for reasonably large values of n. The random variable $\hat{f}_{-i}(X_i)$ can take over the role of $\hat{f}(Y_i)$. This leads us to the *cross-validation criterion*

$$\hat{R}(\hat{f}) = \int \hat{f}(t)^2 - 2\frac{1}{n}\sum_{i=1}^{n} \hat{f}_{-i}(X_i). \tag{4.4}$$

This expression only depends on the data and can be minimized for $\hat{f}$.

In the context of the present chapter we use cross-validation as a method for finding a suitable bandwidth. If we restrict ourselves to kernel density estimators with a fixed, chosen kernel, then the cross-validation criterion $\hat{R}(\hat{f})$ is a function of the bandwidth and the data. Given the data, we choose the value of the bandwidth that minimizes this criterion. It should be noted that this procedure is quite computer-intensive and was no option in the past for a long time. However, the progress in computing power let this difficulty vanish if the method is somewhat efficiently implemented in a programming language. This makes cross-validation a quite attractive method, also in other contexts.

As usual, theoretical analyses consider the behaviour of the thus obtained bandwidth in the situation that the number of observations goes to infinity. In this case, the bandwidth found through cross-validation converges to the optimal bandwidth in the sense of minimizing the integrated squared error. Unfortunately, it seems that the cross-validation method sometimes behaves unfavorably for small bandwidths in the case of small samples and, in particular, if there are many ties, i.e. equal values, in the sample. In some situations, one even ends up with an 'optimal' cross-validation bandwidth of zero. As a consequence, it is reckless to apply the method of cross-validated kernel density estimators without additional reasoning.

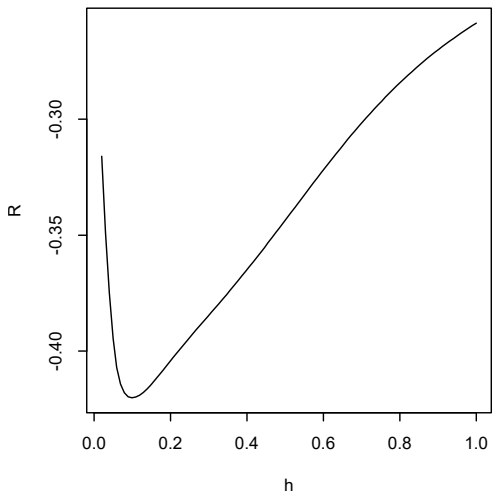


Figure 4.7: The cross-validation criterion (4.4) on the y-axis as a function of the bandwidth for the Old Faithful data when using the standard normal kernel.

Example 4.6 For the eruption durations of Old Faithful an optimal bandwidth was selected with the help of cross-validation. Figure 4.7 illustrates the cross-validation criterion as a function of the bandwidth, while the underlying kernel is standard normal. The criterion reaches its minimum in the neighborhood of $h = 0.1$. This bandwidth corresponds to the first plot in Figure 4.6.
***

4.4 Other density estimators

In many cases kernel density estimators produce decent results, but they also suffer from multiple deficiencies. One obvious deficiency is that kernel estimators with smooth kernels always yield smooth estimates. Possible discontinuities in the 'true' density are obscured by the use of a kernel estimator.

Another undesired 'smoothing' effect is that kernel estimators might assign probability mass to regions which are impossibly attained by any observation. For example, if one uses the normal density kernel, the obtained kernel estimate is positive at each point $-\infty < t < \infty$. For estimating the density of a positive random variable such an effect is undesirable and this sometimes makes it necessary to adjust the estimation. One possibility is to first transform the data, for example logarithmically, $y = \log x$, then do the kernel density estimation for the density of the transformed random variable, and finally transform back. When a transformation is used, one has to make sure that the correct reverse transformation is done to obtain an estimate for the density of X instead of $Y = \log X$.

Another possibility in the case of estimating the density of a positive random variable is to derive a kernel density estimate $\hat{f}_s$ based on the 'symmetrized sample' $X_1, -X_1, \ldots, X_n, -X_n$, and then to use as an estimate the function $t \mapsto 2\hat{f}_s(t)$ for $t > 0$ and the constant zero function on the lower half of the real line. A less appropriate method is to simply set to zero the usual kernel estimate to the left of $t = 0$. Then the result is an estimate with an area less than 1 under the curve. Rescaling this estimate also does not produce a good estimate because in this case probability mass, that is obviously supposed to lie close to $t = 0$, is distributed along the whole upper half of the real line.

A serious deficiency of kernel density estimators such as the ones treated up to now is that the smoothing takes place uniformly over the whole domain. This is particularly undesirable for multimodal distributions with heavy tails.

Example 4.7 Figure 4.8 illustrates kernel estimates of the density of a mixture of three normal distributions. The dashed curve displays the (true) density of the mixture distribution $0.45 * N(0, 1) + 0.3 * N(2, 0.25) + 0.25 * N(5, 16)$. This

density is bimodal and has relatively heavy tails. The other curves in the figure are kernel estimates based on a sample of size 100, both using the normal kernel but different bandwidths. The kernel estimate with a small bandwidth shows a reasonable approximation of the two peaks but it has a sine-like upper tail in which the locations of the largest data points are clearly visible. A larger bandwidth leads to smoother tails but now the bimodality of the distribution is obscured.
***

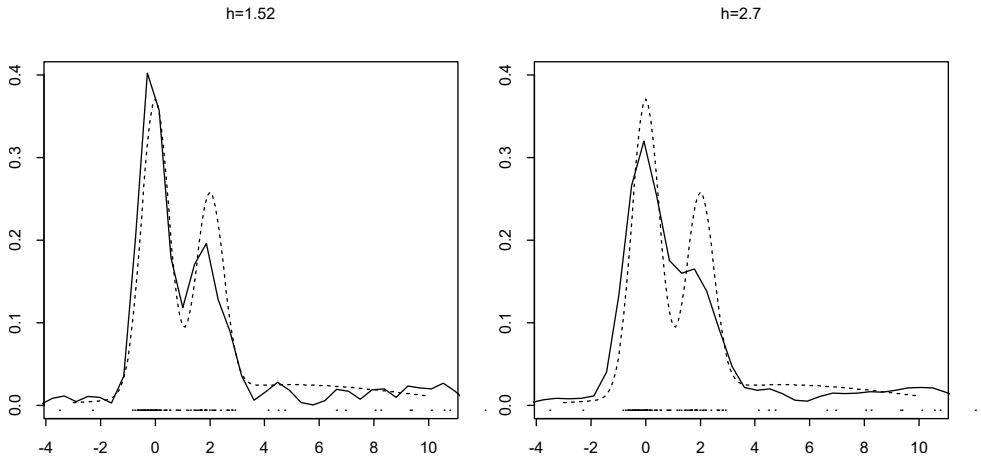


Figure 4.8: Kernel density estimates based on a sample of size 100 from a mixture of three normal distributions. The dashed line is the true density.

This problem can be attacked by allowing for variable bandwidths. Compared to data points which lie accumulated in a certain region, the probability mass of isolated observations should be spread along a wider region. The *variable kernel density estimator* is defined as

$$\hat{f}(t) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{hd_i} K\Big(\frac{t - X_i}{hd_i}\Big).$$

Here, d_i is a measure of the degree of isolation of X_i, for example, the *k-th nearest neighbor* distance: the distance of X_i to a data point such that $k-1$ observations lie closer to X_i and $n-k-1$ observations are farther away. In this connection, k is a number that is chosen beforehand.

4.5 Multivariate density estimation

Obviously, estimation of densities gets even more involved when the data $X_1, \ldots, X_n$ are multivariate. Let us again denote their unknown density by f. Typically, kernel density estimators are based on a spherically symmetric multivariate density K with their center

at the origin as the kernel function. Often, the multivariate standard normal density is used as the kernel. Instead of using a real-valued bandwidth $h > 0$, one has to use a $(d \times d)$ positive-definite and symmetric bandwidth matrix H in the d-variate case. As in the univariate case, the choice of the bandwidth matrix H is crucial. It is involved in the estimation through

$$\hat{f}(x) = \frac{1}{n\sqrt{\det H}} \sum_{i=1}^{n} K(H^{-1/2}(x - X_i)).$$

We will not discuss the peculiarities of the choice of the bandwidth matrix in more detail. Instead, we conclude this chapter with an example.

Example 4.8 The data in this example originate from the following mixture of bivariate normal distributions:

$$\begin{aligned} 0.45 * N\Big(\begin{pmatrix}1.5\\1.5\end{pmatrix}, \begin{pmatrix}1 & -0.5\\-0.5 & 1\end{pmatrix}\Big) &+ 0.3 * N\Big(\begin{pmatrix}-1\\0.5\end{pmatrix}, \begin{pmatrix}0.5 & 0.25\\0.25 & 0.5\end{pmatrix}\Big) \\ &+ 0.25 * N\Big(\begin{pmatrix}-0.5\\-1\end{pmatrix}, \begin{pmatrix}0.75 & -0.6\\-0.6 & 0.75\end{pmatrix}\Big). \end{aligned}$$

A contour plot of the corresponding density f is given on the right in Figure 4.9. The left plot in Figure 4.9 illustrates contour lines of a kernel density estimate based on 400 independent realizations according to f. The estimate was found by applying the function `kde` of the R package `ks`, using the default preferences of the function. The numbers in the left plot are the percentage points of the probability mass that are included in the different curves, the numbers in the right plot are the values of the mixture density along the curves.

The overall picture of the estimate seems to correctly reflect many characteristics of the true density and all components of the mixture distribution are approximated reasonably well. However, the estimated density has three instead of two peaks. Furthermore, recall the different correlation coefficients of the normal components of the mixture distribution. Because of this, it again seems preferable to use varying bandwidth matrices which are more strongly influenced by neighboring observations.

It should be stressed that one should generally not just apply implemented functions without further thinking of the consequences of the choices that are made (by default). Nevertheless, in some situations like in this example, the results could still be satisfactory.

***

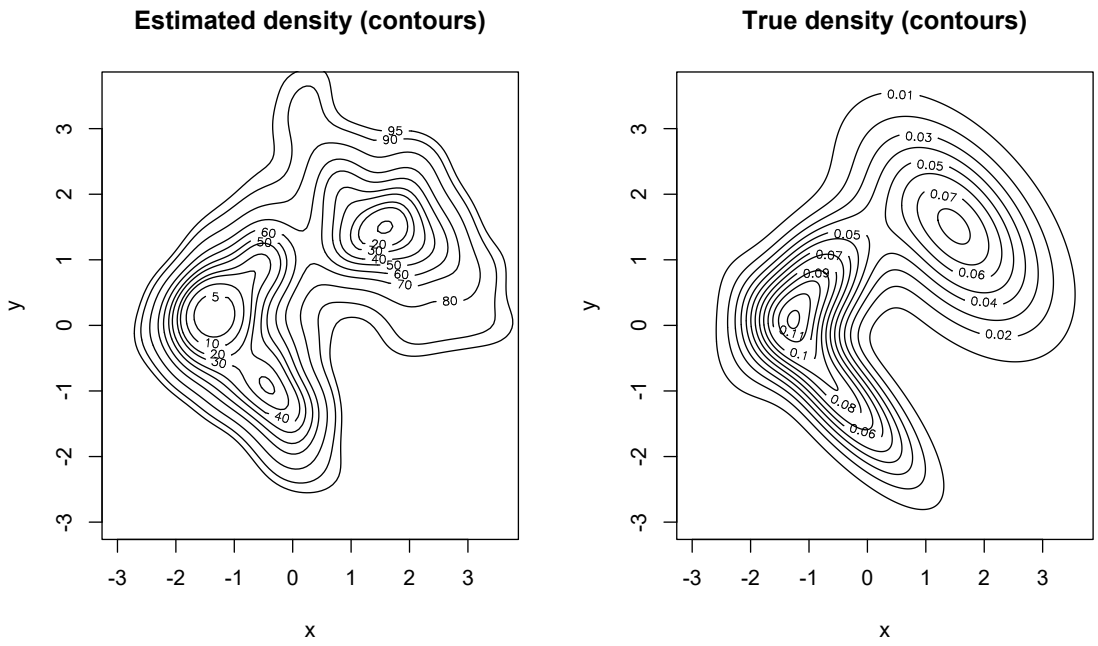


Figure 4.9: Contour plots: kernel density estimate (left) based on a sample of size 400 from a mixture of three bivariate normal distributions and the true density (right).

Chapter 5

The bootstrap

The *bootstrap* [1] is a technique which can be used for

- investigating the variance of an estimator
- computing confidence intervals
- determining critical values of test statistics.

Without modern computers the bootstrap would not be useful, because with this technique the theoretical derivation of a probability distribution generally is replaced by computer simulation. This is why we first discuss the general idea of simulation.

5.1 Simulation

Let $Z_1, Z_2, \ldots, Z_B$ be independent replications from a probability distribution G. Then according to the Law of Large Numbers we have that

$$\frac{1}{B} \cdot \#(Z_i \in A, 1 \leq i \leq B) \to G(A), \qquad \text{with probability 1, if } B \to \infty.$$

For continuously distributed, real valued random variables this principle can be expressed in a graphical way. In this case a properly scaled histogram of $Z_1, Z_2, \ldots, Z_B$ will for large values of B closely look like the density of G. This is illustrated in Figure 5.1, which shows histograms for $B = 20$, 50, 100 and 1000 observations from a χ^2_4 distribution, each time together with the true density.

The above mentioned concept 'independent replications' is precisely defined in probability theory. With a computer it is possible to generate *pseudo random numbers* from most standard probability distributions. A sequence of pseudo random numbers is generated according to a fixed formula and is therefore not a real sequence of independent

[1] This somewhat strange name comes from B. Efron. A bootstrap is just a long shoe lace. The connection between statistics and shoe laces goes via the Baron von Münchhausen, who 'pulled himself up by his own bootstraps'. (Read this chapter and find the connection yourself.)

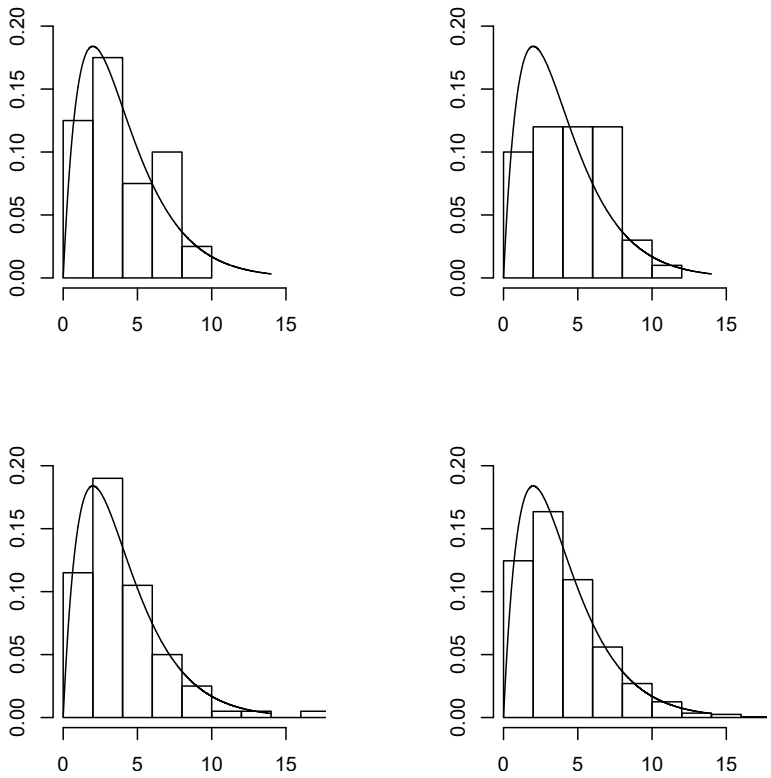


Figure 5.1: Histograms of samples from 20, 50, 100 and 1000 independent observations from a χ^2_4-distribution and the true density.

replications from a given distribution. For most applications the pseudo random numbers cannot be distinguished from real independent observations. This is why the word 'pseudo' is often omitted. Instead of 'generating' one also speaks of *simulating* and, when this principle is used for solving a problem, it is called the *Monte Carlo method*.

Random numbers from more general probability distributions can be generated in two steps.

- Generate random numbers or vectors $U_1, \ldots, U_n$ from a certain (standard) distribution P.

- Next, form with a fixed function ψ the random numbers $\psi(U_1), \ldots, \psi(U_n)$.

A special case of this is the *inverse transformation*. Here P is the uniform distribution on (0,1) and $\psi = G^{-1}$ is the inverse of a distribution function G. In this case $G^{-1}(U_1), \ldots, G^{-1}(U_n)$ form a sample from G, as follows from

$$P(G^{-1}(U) \leq x) = P(U \leq G(x)) = G(x).$$

Evidently, this procedure is only useful if the quantile function G^{-1} is available on the computer.

5.2 Bootstrap estimators for a distribution

Suppose that a set of random variables $X_1, \ldots, X_n$ is available and that one is interested in a function of these random variables, the random variable $T_n = T_n(X_1, \ldots, X_n)$, and in particular in its distribution. This random variable T_n is, for example, an estimator or a test statistic, but it may also depend on an unknown parameter. When T_n is an estimator, then from its distribution a measure for its accuracy, like its variance, can be derived. In the case that T_n is a test statistic, the critical values of the test follow from the distribution of T_n under the null hypothesis. In general, the distribution of T_n is unknown, so that it is an important problem to estimate it from the data. Strictly speaking, the *distribution* of T_n is the set of probabilities $P(T_n \in A)$ of possible events A, but one may as well think of the distribution function or probability density of T_n.

Let the parameter P denote the 'true' probability model for $X_1, \ldots, X_n$. Usually the distribution of T_n depends on P. We use the notation $Q_P = Q_P(T_n)$ for this distribution. The simplest way to find an estimate of Q_P is by replacing in the latter expression the unknown value of P by an estimator $\tilde{P}_n$ of P: estimate Q_P by $Q_{\tilde{P}_n}$. This is the *bootstrap estimator* in its most general, theoretical form. Note that this estimator depends on the data $X_1, \ldots, X_n$ via $\tilde{P}_n$. Once a set of realizations $x_1, \ldots, x_n$ from $X_1, \ldots, X_n$ is available, we have an estimate of Q_P. Often one is specifically interested in the variance $\mathrm{Var}_P(T_n)$. The *bootstrap estimator for the variance* is the variance of the distribution $Q_{\tilde{P}_n}$. In this chapter we shall mostly consider the case where $X_1, \ldots, X_n$ are independent, so that they form a random sample from P.

There are two types of frequently used estimators $\tilde{P}_n$ for the unknown P. The first one is the *empirical distribution* of $X_1, \ldots, X_n$, which is usually denoted by $\hat{P}_n$. The empirical distribution is the discrete probability distribution that distributes the total probability mass uniformly among the n observation points:

$$\hat{P}_n(A) = \frac{1}{n} \cdot \#(X_j \in A).$$

The estimator $Q_{\hat{P}_n}$ for the distribution of T_n is the bootstrap estimator in its original form and is often called the *empirical bootstrap estimator*. The empirical distribution $\hat{P}_n$ is a very simple estimator for P, which always 'works' when the observations are mutually independent replications. In some sense it is the best estimator when nothing is known about the unknown distribution. On the other hand, it is often possible to find better estimators when it is known that the distribution P belongs to a particular class of distributions, like the class of normal distributions or symmetric distributions.

The second type of frequently used estimators for P is the *parametric estimator*, which is appropriate in situations where the unknown distribution P is known to belong to a

parametric family like the normal or exponential family, but its parameter value θ is unknown. In this case it is natural to first find an estimator $\hat{\theta}_n$ of θ, and then estimate P by the distribution in the family which has $\hat{\theta}_n$ as its parameter value. This parametric estimator of P will be denoted by $P_{\hat{\theta}_n}$. The estimator $Q_{P_{\hat{\theta}_n}}$ for the distribution of T_n is a *parametric bootstrap estimator*.

Example 5.1 Consider the case where T_n is the mean of a random sample $X_1, \ldots, X_n$ from an unknown distribution P, i.e. $T_n = \bar{X} = n^{-1}\sum_{j=1}^n X_j$. Then T_n is an unbiased estimator for the expectation μ_P of P. The unbiasedness means that if we would compute T_n for each of a large number of different (independent) samples, the average of these T_n's would be close to μ_P, but this does not say much about the accuracy of T_n for a single sample. One of the measures for the quality of T_n is the variance $\mathrm{Var}_P(T_n)$. It can be expressed in terms of the variance of one observation, $\sigma_P^2 = \mathrm{Var}(X_1)$, according to the formula

$$\mathrm{Var}_P(\bar{X}) = \frac{\sigma_P^2}{n}.$$

Of course, in most cases σ_P^2 is unknown. An unbiased estimator for σ_P^2 is the sample variance

$$S^2 = \frac{1}{n-1}\sum_{j=1}^n (X_j - \bar{X})^2.$$

Hence, the value S^2/n gives an impression of the accuracy of the sample mean $\bar{X} = n^{-1}\sum_{j=1}^n X_j$ as an estimator for μ_P. A rule of thumb which is often used is that the true value of μ_P is equal to $\bar{X}$ plus or minus a couple of times (most often twice) $S/\sqrt{n}$.

This statement can be made precise by means of a confidence interval. To compute this interval more information about the complete distribution of the sample mean is necessary. In classical statistics it is not unusual to assume that the unknown distribution P is a normal distribution. Under this assumption $\bar{X}$ has a $N(\mu_P, \sigma_P^2/n)$ distribution, which can be estimated by the $N(\bar{X}, S^2/n)$ distribution. Here the true distribution of $\bar{X}$, $Q_P = Q_P(\bar{X})$, is estimated by the parametric bootstrap estimator $Q_{P_{\hat{\theta}_n}} = N(\bar{X}, S^2/n)$.

Since frequently the underlying distribution P is not normal, we often do *not* want to assume that P is normal. In case the normality assumption is not made, the (empirical) bootstrap still yields an estimate for the unknown distribution of T_n in this case: the true distribution of $\bar{X}$, $Q_P = Q_P(\bar{X})$, is estimated by $Q_{\hat{P}_n}$. [2]

[2]It should be noted that other estimators than the empirical bootstrap are possible, even if the normality assumption does not hold. For instance, the earlier mentioned $N(\bar{X}, S^2/n)$ estimator. After all, for large n the sample mean will be approximately normally distributed because of the central limit theorem, provided that $\sigma_P^2 < \infty$.

This is the distribution of the mean of a sample of size n from $\hat{P}_n$. Consecutively, if desired, from the estimator of the distribution of $\bar{X}$ estimators for parameters of the distribution can be inferred. In particular, the bootstrap estimator of the variance of $\bar{X}$ is the variance of $Q_{\hat{P}_n}$. In the example this variance turns out to be $(n-1)/n \cdot S^2/n$. [3]
***

The bootstrap estimator can be applied in many more complex situations than that of example 5.1; for example, to estimate the distribution or variance of the median, or of a trimmed mean (see next chapter), for which no simple formulas are available. Moreover, the bootstrap estimator can be used for the assessment of the accuracy of estimators of much more complex parameters, such as regression curves.

Until now, we only considered the theoretical concept of the bootstrap estimator. A practical question is whether a bootstrap estimate can in fact be computed, once the realized sample $x_1, \ldots, x_n$ from $X_1, \ldots, X_n$ is known. In the above the estimator is written as $Q_{\tilde{P}_n}$, but is this, given the observations $x_1, \ldots, x_n$, a useful expression? In most cases the answer is negative. Often $Q_{\tilde{P}_n}$ is a complex function of $\tilde{P}_n$, for which no explicit expression is available. The distribution $Q_{\hat{P}_n}$ in Example 5.1, for example, is a discrete distribution of which the probabilities can in principle be written out, but for which no useful formula is possible. However, when a fast computer is available, it is always possible to give a (stochastic) approximation of the estimate: one can simulate it.

For the bootstrap this goes as follows. Suppose that the original observations $x_1, \ldots, x_n$ were obtained from a distribution P, and that $\tilde{P}_n$ is an estimator for P. Given the values $x_1, \ldots, x_n$, the estimate $\tilde{P}_n = \tilde{P}_n(x_1, \ldots, x_n)$ is fixed. But this means that given $x_1, \ldots, x_n$, the probability mechanism $\tilde{P}_n = \tilde{P}_n(x_1, \ldots, x_n)$ is known, and that a new sample $X_1^*, \ldots, X_n^*$ can be generated from it by computer simulation. Obviously, the new sample $X_1^*, \ldots, X_n^*$ is not a randomly selected sample like the original sample, but a pseudo random, simulated sample. It is called a *bootstrap sample*, and the X_i^* are called the *bootstrap values*. It is important that $X_1^*, \ldots, X_n^*$ are generated from $\tilde{P}_n$ in the same manner as $X_1, \ldots, X_n$ were sampled from P (for instance of the same size and with the same (in)dependence structure). Write $T_n^* = T_n(X_1^*, \ldots, X_n^*)$ for the corresponding *bootstrap value* of T_n. Then the bootstrap estimate for the distribution of T_n can be written as

$$Q_{\tilde{P}_n(x_1,\ldots,x_n)} = Q_{\tilde{P}_n}(T_n^* \mid x_1, \ldots, x_n).$$

[3] In the example the bootstrap estimator is the variance of the mean of a sample of size n from the distribution $\hat{P}_n$. Given the values of the original sample $x_1, \ldots, x_n$, which need to be considered as fixed, the bootstrap estimate can in this example be explicitly determined. This goes as as follows. Like before, the variance of the mean is $1/n$ times the variance of one observation. One observation from $\hat{P}_n$ has expectation $\bar{x} = \sum_{j=1}^n x_j \cdot n^{-1}$ and variance $\sum_{j=1}^n (x_j - \bar{x})^2 \cdot n^{-1}$. (Use the formulas for expectation and variance of a discrete distribution to compute this.)

The vertical bar in the middle stands for 'given the value of': determination of the bootstrap distribution is always done with the original observations $x_1, \ldots, x_n$ being given, thus fixed. A (stochastic) approximation of the estimate $Q_{\tilde{P}_n}(T_n^* \mid x_1, \ldots, x_n)$ can now be obtained by simulating B values $T_{n,1}^*, \ldots, T_{n,B}^*$ from this distribution and then approximating $Q_{\tilde{P}_n}(T_n^* \mid x_1, \ldots, x_n)$ by the empirical distribution of these B bootstrap values. Simulation of the B bootstrap values $T_{n,1}^*, \ldots, T_{n,B}^*$ of T_n can be done according to the the following bootstrap sampling scheme which consists of three steps:
given $x_1, \ldots, x_n$,

1. simulate a large number, B say, bootstrap samples (like $X_1^*, \ldots, X_n^*$) according to $\tilde{P}_n(x_1, \ldots, x_n)$, all of size n, and independent from each other;

2. compute for each of the B bootstrap samples the corresponding bootstrap value T_n^*;

3. approximate the bootstrap estimate $P_{\tilde{P}_n}(T_n^* \in A \mid x_1, \ldots, x_n)$ with the fraction of the B bootstrap values T_n^* that fall in A; in other words, approximate $Q_{\tilde{P}_n}(T_n^* \mid x_1, \ldots, x_n)$ by the empirical distribution of the B bootstrap values.

Expressed in formulas: step 2 gives B values $T_{n,1}^*, \ldots, T_{n,B}^*$ and the probability in step 3 is approximated by $B^{-1} \cdot \#(T_{n,i}^* \in A)$. In practice the third step is often replaced by drawing a histogram of the realizations of the B bootstrap values $T_{n,i}^*$. This histogram then is an estimate of the true distribution of the estimator T_n, and it can give a reasonable insight into the accuracy of T_n. If the histogram is broadly spread, then T_n is not to be trusted as a very good estimator. When the obtained estimate $T_n(x_1, \ldots, x_n)$ lies in the tail of the histogram, this is a reason to be cautious too. The obtained estimate $T_n(x_1, \ldots, x_n)$ then seems to be an outlier.

When one is interested in the bootstrap estimator for the variance, then this is approximated by the sample variance of the B bootstrap values:

$$\frac{1}{B-1} \sum_{i=1}^{B} \left(T_{n,i}^* - \bar{T}_n^*\right)^2,$$

where $\bar{T}_n^* = B^{-1} \sum_{i=1}^{B} T_{n,i}^*$. In general, these stochastic approximations of the bootstrap estimates are also called bootstrap estimators and their realizations bootstrap estimates.

Example 5.2 Consider again the case where $X_1, \ldots, X_n$ is a random sample from a distribution P and where for $\tilde{P}_n$ the empirical estimator $\hat{P}_n$ is used. Then the bootstrap values $X_1^*, \ldots, X_n^*$ form a sample from the discrete distribution $\hat{P}_n$. Now, sampling from a discrete distribution where all mass points have equal probability is the same as randomly selecting one of the mass points. Therefore, in this case the values $X_1^*, \ldots, X_n^*$ can be obtained by drawing a random sample with replacement

of size n from the set $\{x_1, \ldots, x_n\}$. With replacement, because the different X_i^* need to be independent, like the values X_i in the original sample. Hence, each of the B bootstrap samples in step 1 is obtained by 'resampling' n times with replacement from the original observations $\{x_1, \ldots, x_n\}$. This technique is therefore also called a *resampling scheme*.

Random sampling from a set of n points can be implemented on a computer in a reasonably efficient way. This is why the simple bootstrap which is based on the empirical estimator, is frequently used. In principle it is possible to sample from any probability distribution, so that instead of the empirical estimator any other estimator for $\tilde{P}_n$ can be used. With respect to the approximation for $Q_{\tilde{P}_n}(T_n^* \mid x_1, \ldots, x_n)$, the choice of $\tilde{P}_n$ only makes a difference for the first step of the bootstrap scheme, the second and third steps are the same.
***

Example 5.3 The statistic $T_n = (n-1)/\sum_{i=1}^n X_i$ is an unbiased estimator for the parameter λ when a sample $X_1, \ldots, X_n$ from an exponential distribution with parameter λ is available. The observed values:

> 2.24, 0.029, 0.155, 0.551, 0.495, 0.15, 0.64, 1.132, 0.03, 0.062, 1.771, 1.001, 0.478, 0.897, 0.205, 0.254, 0.564, 0.274, 0.517, 0.505, 0.51, 0.565, 1.011, 1.357, 2.76

($n = 25$) yield an estimate $\hat{\lambda} = 1.322$. In order to obtain an impression of the accuracy of this estimate 100 bootstrap values of T_n were generated according to the empirical bootstrap method.

- Generate 25 values $X_1^*, \ldots, X_{25}^*$ by randomly sampling 25 times with replacement from the observed numbers given above.
- Compute the bootstrap value $T_{25}^* = (n-1)/\sum_{i=1}^n X_i^*$.
- Repeat this scheme 100 times.

Figure 5.2 gives a histogram of the 100 computed bootstrap values. The corresponding (approximate) bootstrap estimate for the standard deviation of T_n is 0.275, the (sample) standard deviation of the 100 bootstrap values.

The histogram of the 100 bootstrap values is an estimate for the true distribution of $(n-1)/\sum_{i=1}^n X_i$. It can be derived theoretically that this distribution converges, for $n \to \infty$, to a normal distribution. The here obtained estimate ($n = 25$) looks already more or less symmetric.

An alternative for the empirical bootstrap in this example is the parametric bootstrap. For this, in step 1 of the bootstrap scheme, the bootstrap samples are generated from the exponential distribution with parameter $\hat{\lambda} = 1.322$, instead of from the empirical distribution. This procedure with $B = 100$ bootstrap samples yielded an estimate of 0.274 for the standard deviation of T_n.

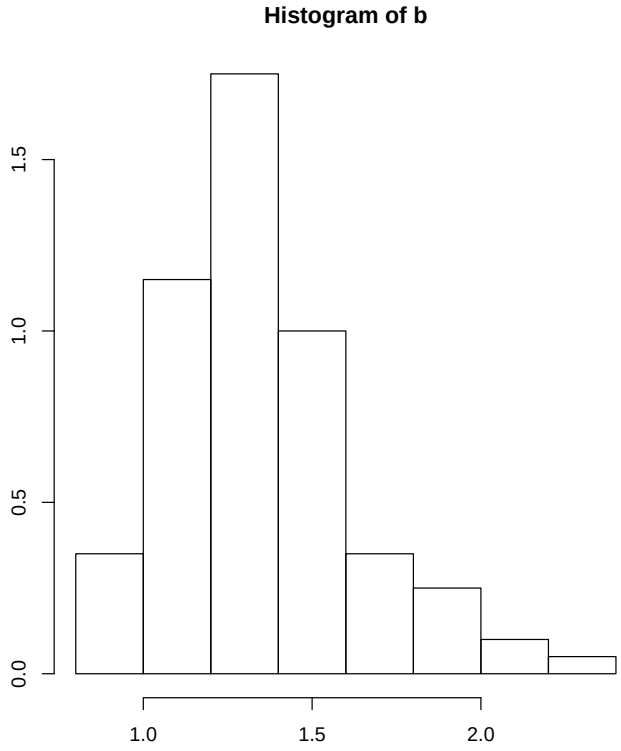


Figure 5.2: Histogram of the 100 bootstrap values of the statistic $(n-1)/\sum_{i=1}^{n} X_i$ of Example 5.3.

The true value of λ was known in this case: $\lambda = 1$. The theoretical value of the standard deviation of T_n is equal to $\sqrt{\lambda^2/(n-2)} = 0.21$. In this example the empirical bootstrap produced a nearly equally good (or bad!) estimate as the parametric bootstrap. This counterintuitive result is explained by the fact that the original sample is rather extreme, in the sense that the estimate $\hat{\lambda} = 1.322$ lies at a distance $1.322 - 1 = 0.322 \approx 1\frac{1}{2} \times \sigma_1(T_n)$ from the true value of λ. On the other hand, the simulation for the parametric bootstrap worked very well: the parametric bootstrap samples were simulated from the exponential distribution with parameter 1.322 and the estimate 0.274 is a very good estimate for the 'true' standard deviation of T_n of $\sqrt{\hat{\lambda}^2/(n-2)} = 0.276$ under $\hat{\lambda} = 1.322$.
***

As mentioned above, the approximation for the bootstrap estimate by computer simulation is a *stochastic* approximation: the bootstrap estimate itself is being estimated. If we are unlucky, the obtained B bootstrap samples are all extreme, and the approximation

very bad. However, the law of large numbers says that for $B \to \infty$ this approximation converges in probability to the bootstrap estimate. Informally, 'when $B = \infty$ then the approximation is exact'. In theory B can be made arbitrarily large, because it does not depend on the original data but is chosen by the statistician. The larger B, the better the approximation. The fact that also the computation time increases with B is a practical limitation. It depends on the complexity of the estimator T_n, how severe this practical limitation is. For instance, some estimators need one day to compute them only once. Then for the computation of the bootstrap values (in step 2) B days, or B computers, are needed!

How large do we take B and how well does the bootstrap work? While using the (approximating) bootstrap estimator in principle two errors are made:

- the difference between $Q_{\tilde{P}_n}(T_n^* \mid X_1, \ldots, X_n)$ and $Q_P(T_n)$.
- the difference between the empirical distribution of the simulated B bootstrap values $T_{n,i}^*$ and $Q_{\tilde{P}_n}(T_n^* \mid X_1, \ldots, X_n)$.

The first error is in some sense an unavoidable error, because on the basis of the original observations the true distribution of T_n cannot be determined with certainty. The magnitude of this error depends on the specific statistic T_n and the chosen estimator $\tilde{P}_n$. When the empirical distribution $\hat{P}_n$ is used, this error will be small, provided that the distribution $Q_P(T_n)$ does not change too much by 'discretizing the underlying distribution P'. As we shall see, this is not always the case.

The second error is of a completely different nature. This one we can control ourselves and it can be made arbitrarily small by making B sufficiently large. Most of the time we will not let the computer run endlessly, but be satisfied with making this second error relatively small with respect to the first one. It has turned out that for simple cases values of B between 20 and 200 yield good results, although sometimes values of 1000 are necessary. For more complex situations values of 10,000 or more are no exception.

5.3 Bootstrap confidence intervals

A correct and precise manner to present the accuracy of an estimate is by means of a confidence interval. Suppose that T_n is used as an estimator of an unknown parameter θ. Write $G_n = Q_P(T_n - \theta)$. Then due to the properties of the quantile function, we have $P\left(T_n - \theta \in [G_n^{-1}(\alpha), G_n^{-1}(1-\alpha)]\right) \geq 1 - 2\alpha$. (When T_n has a continuous distribution, exact equality holds.) In other words,

$$[T_n - G_n^{-1}(1-\alpha), T_n - G_n^{-1}(\alpha)]$$

is a confidence interval with confidence level $1 - 2\alpha$ for θ. The quantile function G_n^{-1} is usually unknown, but by replacing it by an estimator $\tilde{G}_n^{-1}$ a confidence interval

$$[T_n - \tilde{G}_n^{-1}(1-\alpha), T_n - \tilde{G}_n^{-1}(\alpha)] \tag{5.1}$$

	PRRP	CTRP
mean	61.56	75.11
median	54.25	62.5

Table 5.1:

with confidence level *approximately* $1-2\alpha$ is obtained. When for $\tilde{G}_n^{-1}$ the quantile function of the bootstrap estimator $\tilde{G}_n = Q_{\tilde{P}_n}(T_n^* - T_n \mid X_1, \ldots, X_n)$ is used, then this interval is called a *bootstrap confidence interval*. In practice a bootstrap estimator $\tilde{G}_n$ will itself be approximated by means of simulation.

In summary, a bootstrap confidence interval with approximate confidence level $1-2\alpha$ is obtained in 3 steps: given $x_1, \ldots, x_n$,

- simulate a large number, B say, bootstrap samples (like $X_1^*, \ldots, X_n^*$) from $\tilde{P}_n$, all of size n, and independent of each other;

- compute for each of the B bootstrap samples the corresponding bootstrap value (like $T_n^* = T_n(X_1^*, \ldots, X_n^*)$); call the resulting values $T_{n,1}^*, \ldots, T_{n,B}^*$ and write $Z_i^* = T_{n,i}^* - T_n$ $(i = 1, \ldots, B)$

- the bootstrap confidence interval is

$$[T_n - Z^*_{([(1-\alpha)B])}, T_n - Z^*_{([\alpha B]+1)}] = [2T_n - T^*_{n,([(1-\alpha)B])}, 2T_n - T^*_{n,([\alpha B]+1)}]. \tag{5.2}$$

Note, that in step 3 the quantiles $\tilde{G}_n^{-1}(\alpha)$ and $\tilde{G}_n^{-1}(1-\alpha)$ in (5.3) are replaced by the corresponding sample quantiles of $Z_1^*, \ldots, Z_B^*$.

Example 5.4 Example 3.4 gives data concerning the β-thromboglobulin level in the blood of three different groups of patients: PRRP, SDRP and CTRP. The mean and median β-thromboglobulin level of the first and last group are given in Table 5.1. It seems that the β-thromboglobulin level of the group CTRP is much higher than that of the group PRRP. In Chapter 3 we already saw that the difference vanishes after a logarithmic transformation.
The boxplots (Figure 5.3) confirm that the conclusion is not that simple. It is true that the β-thromboglobulin levels of the CTRP group seem to be somewhat larger, but this seems in the first place to be the result of a larger variation.

In order to be able to draw a more clear conclusion with the bootstrap an (approximate) 90% confidence interval for the difference in median between the two groups was derived. For this a vector of 1000 bootstrap values of the difference was generated according to the scheme:

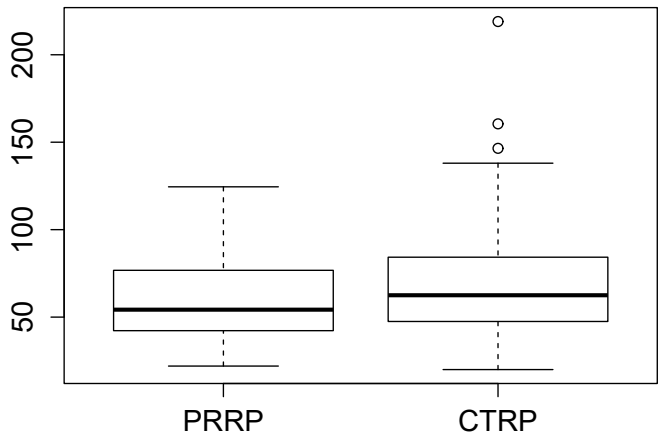


Figure 5.3: Boxplots of the β-thromboglobulin levels of PRRP (left) and CTRP (right).

- from the original β-thromboglobulin levels $a_1, \ldots, a_{32}$ of group PRRP, a sample $a_1^*, \ldots, a_{32}^*$ was formed by randomly sampling 32 times with replacement from $a_1, \ldots, a_{32}$;
- analogously, from the original β-thromboglobulin levels of the patients in group CTRP a sample $b_1^*, \ldots, b_{23}^*$ was formed;
- the value $z^* = \text{med}(a_1^*, \ldots, a_{32}^*) - \text{med}(b_1*, \ldots, b_{23}^*) - (\text{med}(a_1, \ldots, a_{32}) - \text{med}(b_1, \ldots, b_{23}))$ was computed.

This scheme was repeated 1000 times. Figure 5.4 gives a histogram of the obtained values $z_1^*, \ldots, z_{1000}^*$. The 5% and 95% quantile of the distribution of this bootstrap sample were -13.75 and 14.25.
According to formula 5.2 this leads to the confidence interval (-22.5, 5.5) with approximate confidence level 90% for the difference in median. The fact that the value 0 lies is this interval, confirms the doubt about the existence of a systematic difference between the two groups of patients. The observed difference can very well be the result of chance.

The choice of 90% is, as always, rather arbitrary. Figure 5.4 (and the corresponding set of z^*-values) is therefore very helpful, because from this the confidence intervals for all confidence levels can be determined. In particular, we remark that for this set of z^*-values the confidence interval with level 70% still contains the value 0.

The bootstrap scheme here is more complex than in Example 5.3, because we now deal with a two sample problem. Formally, we can describe the statistical model of the observations by the parameter $P = (F, G)$, where F and G are the distributions of the β-thromboglobulin levels in the groups PRRP and CTRP, respectively. This parameter is estimated in the bootstrap scheme above by the pair

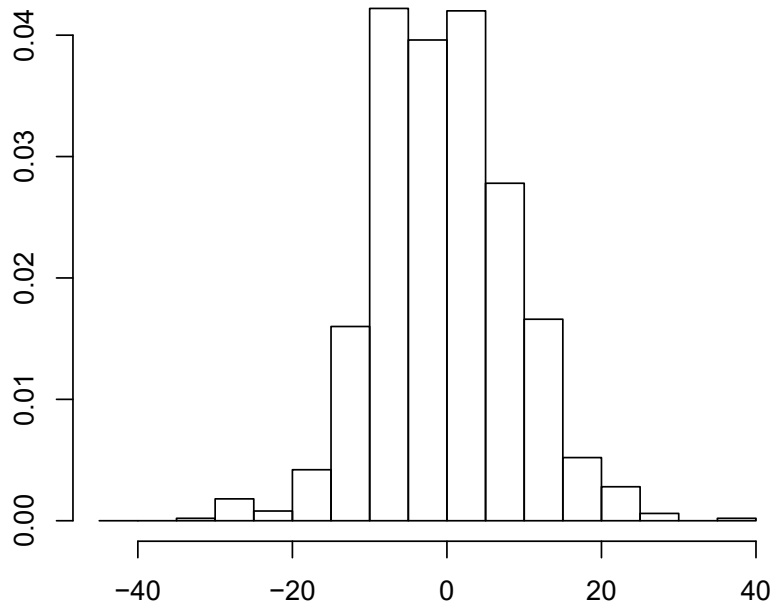


Figure 5.4: Histogram of the 1000 bootstrap values z^* described in Example 5.4.

$\hat{P}_n = (\hat{F}_n, \hat{G}_n)$ of the empirical distributions of the two samples.
$***$

5.4 Bootstrap tests

The bootstrap can also be used for the computation of an approximation of the p-value of a test. Let $\mathcal{P}_0$ be a collection of probability distributions. As before, P is the unknown 'true' distribution of $X_1, \ldots, X_n$. Suppose that the null hypothesis $H_0 : P \in \mathcal{P}_0$ is rejected for large values of a test statistic T_n. First, assume that T_n is *nonparametric* under the null hypothesis; e.g. the distribution $Q_{P_0}(T_n)$ is the same for all $P_0 \in \mathcal{P}_0$. The recipe

"When t is observed, reject H_0 if $P_{H_0}(T_n > t) \leq \alpha$",

gives a test with significance level $\leq \alpha$. For actually performing the test, the p-value $P_{H_0}(T_n > t)$ has to be evaluated. Because T_n is *nonparametric*, this can be done by choosing a suitable distribution $P_0 \in \mathcal{P}_0$ and evaluating $P_{P_0}(T_n > t)$ for this P_0. Sometimes this can be done exactly. Often an asymptotic approximation for $n \to \infty$ is known, for instance a normal approximation. A third possibility is simulation: simulate B bootstrap values of T_n from the distribution $Q_{P_0}(T_n)$ and approximate the p-value with the fraction of bootstrap values that exceed the observed value t of T_n.

In case T_n is not nonparametric under the null hypothesis, one can proceed as follows. Choose, under the assumption that the null hypothesis is correct, an appropriate estimator

$\tilde{P}_n$, i.e. $\tilde{P}_n \in \mathcal{P}_0$, for P. Determine the bootstrap estimate $Q(T_n^* \mid x_1, \ldots, x_n) = Q_{\tilde{P}_n}$ for $Q_{P_0} = Q_{P_0}(T_n)$. Simulate B bootstrap values T_n^* from $Q_{\tilde{P}_n}$ and determine a 'p-value' as the fraction of bootstrap values which exceed the observed value t of T_n. The test "Reject H_0 if $p \leq \alpha$" is at best *approximately* of level $\leq \alpha$. Asymptotically, the probability of an error of the first type under the true distribution is often exactly equal to α. When T_n for large n is approximately nonparametric, then the significance level of the test is also approximately equal to $\leq \alpha$. (Remember, that the significance level is the supremum of the probabilities of an error of the first type.)

Example 5.5 In Example 3.2 a linear regression was performed of the volume y_i on the diameter x_i of 31 cherry trees. The least squares estimates for the intercept and slope of the regression line were $\hat{\alpha} = -36.94$ and $\hat{\beta} = 5.07$. We wish to test whether the normality assumption of the regression model ("the measurement errors $Y_i - \alpha - \beta x_i$ are $N(0, \sigma^2)$ distributed, with σ^2 unknown") is satisfied by means of the adapted test statistic

$$\tilde{D}_n = \sup_{-\infty<x<\infty} |\hat{F}_n(x) - \Phi(x/S)|.$$

Here $\hat{F}_n$ is the empirical distribution function of the residuals $R_i = Y_i - \hat{\alpha} - \hat{\beta} x_i$ and $S^2 = \sum_{i=1}^n R_i^2/(n-2)$. For the cherry tree data the observed value of the latter is $s^2 = 18.079$. The test statistic is nonparametric under the null hypothesis: for every α, β and σ, $\tilde{D}_n$ has the same distribution. This distribution can be approximated while using a simulation scheme.

The observed value of $\tilde{D}_n$ is 0.0953. To approximate the p-value $P_{H_0}(\tilde{D}_n > 0.0953)$ 1000 bootstrap values of $\tilde{D}_n$ were generated according to the following scheme:

- simulate a sample of size 31 from $N(0, s^2)$; denote the values by $e_1^*, \ldots, e_{31}^*$
- compute $y_i^* = \alpha + \beta x_i + e_i^*$, $(i = 1, \ldots, 31)$ with α and β as computed before, that is $\alpha = \hat{\alpha}$ and $\beta = \hat{\beta}$.
- determine the least squares estimates $\hat{\alpha}^*$ and $\hat{\beta}^*$ for linear regression of y_i^* on x_i
- compute $R_i^* = y_i^* - \hat{\alpha}^* - \hat{\beta}^* x_i$, $(i = 1, \ldots, 31)$
- compute $\tilde{D}_n^*$ from $R_1^*, \ldots, R_{31}^*$ in the same manner as $\tilde{D}_n$ from $R_1, \ldots, R_{31}$.

This scheme was executed 1000 times. Figure 5.5 gives a histogram of the 1000 values of $\tilde{D}_n^*$. The observed $\tilde{D}_n = 0.0953$ is the 0.316-quantile of the values $\tilde{D}_n^*$. The p-value of about 68% gives absolutely no reason to doubt the normality of the residuals.

In step one a parametric bootstrap is performed: the bootstrap sample is drawn from a normal distribution and not from the empirical distribution of the residuals.

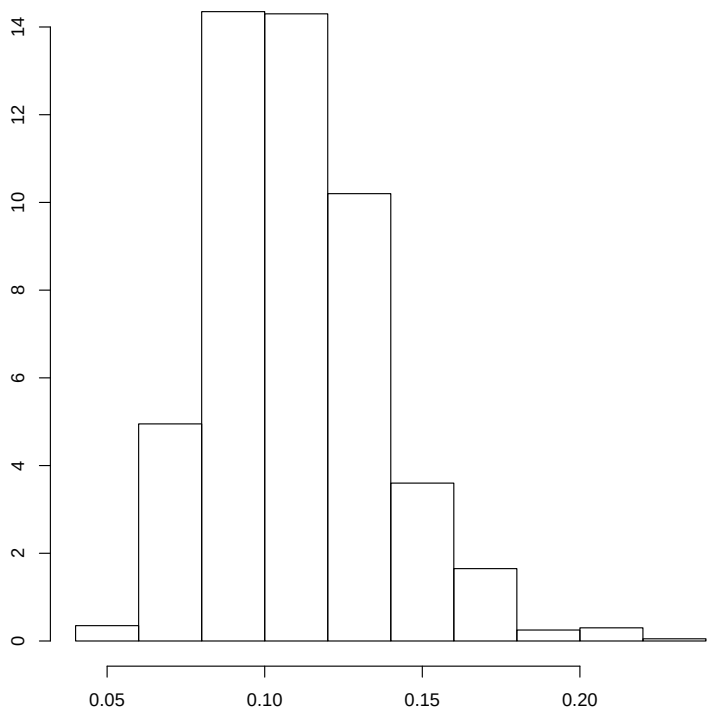


Figure 5.5: Histogram of 1000 bootstrap values of the statistic $\tilde{D}_n$ in Example 5.5.

This is because the goal is to approximate the distribution of $\tilde{D}_n$ under the null hypothesis, which says that the measurement errors are normally distributed. Instead of $\tilde{D}_n$ also another test statistic can be used. One point of criticism on the use of $\tilde{D}_n$ is, for example, that the residuals under the null hypothesis are indeed normally distributed, but not independent and with equal variances. A small adaptation of the test would be appropriate. However, this is usually omitted due to the amount of extra work that is needed for this.
***

5.5 Limitations of the bootstrap

The bootstrap technique is nowadays widely used due to the efficiency of modern computers. Because of its general applicability it is very popular. In general it gives the same results as the classical methods when these can also be used, and it gives reasonable results in many situations where the classical methods cannot be used. However, one should always be cautious. The following example shows that thoughtless use of the bootstrap may yield incorrect results: the bootstrap does not automatically yield the right answer,

and the mathematical theory is needed to investigate when the technique does or does not work.

Example 5.6 In a simulation study the following experiment was performed 12 times in order to estimate the distribution of the mean $\bar{X}$ of 30 observations from a standard Cauchy distribution.

- Simulate a sample of size 30 from a standard Cauchy distribution.
- Determine the (stochastic approximation for the) empirical bootstrap estimator for the distribution of the mean based on $B = 500$ bootstrap replicates.

When we write the sample in step one as $X_1, \ldots, X_{30}$, then in step two the simulation approximation for $Q_{\hat{P}_{30}}(\bar{X}^* \mid X_1, \ldots, X_{30})$ is determined. This scheme was repeated 12 times to get an idea of the variation of the quality of the bootstrap estimator. Figure 5.6 gives the histograms of the 500 bootstrap values. The smooth curve in the graphs is the true density of the mean. (According to the theory the mean itself has a standard Cauchy distribution also.) It is clear that the bootstrap estimator in this case behaves rather badly.

The explanation lies in the heavy tails of the Cauchy distribution. Firstly, they cause the 12 obtained samples to be of a quite different nature. As a consequence, the bootstrap histogram is in some cases very concentrated and in other cases fairly spread out. Secondly, the empirical distribution $\hat{P}_{30}$ of the 30 original observations differs very much from the underlying Cauchy distribution with respect to the tails. After all, the empirical distribution does not really have tails. Because the distribution of the mean strongly depends on the tails of the underlying distribution, the distribution of the mean of 30 observations from $\hat{P}_{30}$ (in Figure 5.6 approximated by a histogram) hardly resembles the distribution of the mean of 30 observations from a Cauchy distribution.

One could object, that $n = 30$ or $B = 500$ is not large enough. This is not correct. Increasing B does not yield any different results. Moreover, it can be proved theoretically that in this case the bootstrap estimator does not converge to the true distribution when $B \to \infty$ and $n \to \infty$, but that it remains of a stochastic nature. The simulation results that are shown in Figure 5.6 therefore give exactly the right impression.

***

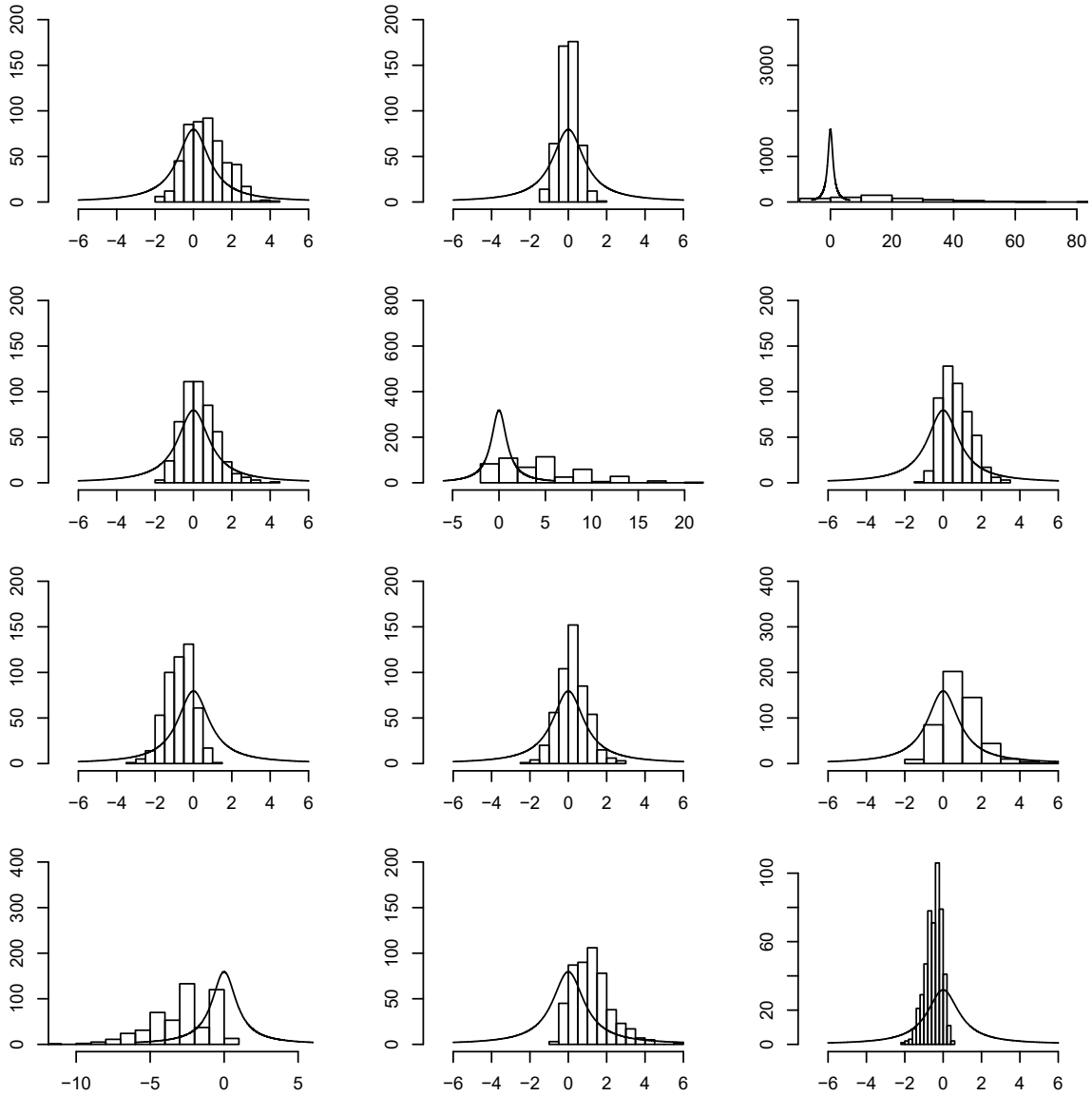


Figure 5.6: Histograms of 12 independent realizations of the empirical bootstrap estimator (B=500) for the distribution of the mean of a sample of size 30 from a Cauchy distribution, together with the density function of the Cauchy distribution.

Chapter 6

Nonparametric methods

When one uses one of the classical tests, such as the t- or F-test, in the situation where the assumptions—like normality—that are necessary for this test do not exactly hold, then the actual level and power of the test will be different from the level and power that are predicted by the theory. In the usual asymmetrical testing framework it is considered particularly undesirable to have the real level of the test larger than the α that was chosen beforehand. The so-called nonparametric (or distribution free) tests accommodate for this problem. For these tests the actual level is always exactly α for a broad class of possible underlying distributions. In the terminology of the foregoing chapter these tests are very robust with respect to the level of the test. Moreover, a number of these tests turns out to have a reasonably high power for a broad set of alternatives. This is why they are good competitors of the classical tests, like the t-test. The Wilcoxon-tests, for example, are a little less efficient than the t-test when the observations are exactly normally distributed, but they result in a large gain in power when the data do not come from a normal distribution.

In this chapter we discuss some nonparametric tests for the one-sample problem, the two-sample problem and the correlation in bivariate data. For nonparametric tests for other problems we refer to the literature. For simplicity we omit in the notation in this chapter, except in Section 6.2, the dependence of the test statistics on n, e.g. a test statistic T_n is denoted as T.

6.1 The one-sample problem

Let $X_1, \ldots, X_n$ be independent and identically distributed random variables. The one-sample problem that we consider concerns testing hypotheses about the location of their common distribution. In the 'classical' model for this problem $X_1, \ldots, X_n$ are normally $N(\mu, \sigma^2)$ distributed and the null hypothesis is $H_0 : \mu = \mu_0$. The most frequently used test, the t-test, is based on the statistic $T = \sqrt{n}(\bar{X} - \mu_0)/S$. In this section we do *not* make the normality assumption.

6.1.1 The sign test

Assume that the true underlying distribution of the observations has a unique median m and that each observation is equal to m with probability 0. In other words, assume that there exists a unique number m with the property

$$P(X_i < m) = P(X_i > m) = \tfrac{1}{2}.$$

Consider the testing problem

$$H_0 : m = m_0$$
$$H_1 : m \neq m_0,$$

for a fixed given number m_0. Note that the null hypothesis is composite. It concerns all probability distributions with median m_0, of which there are infinitely many.

The *sign test* is based on the test statistic

$$T = \#(X_i > m_0) = \sum_{i=1}^{n} 1_{\{X_i > m_0\}}.$$

Under the null hypothesis X_i has median m_0, so that under the null hypothesis T has a binomial distribution with parameters n and $\frac{1}{2}$. Because the distribution of the test statistic turns out to be the same for every possible distribution of the null hypothesis, the test is *distribution free* or *nonparametric*. The statistic T is called *nonparametric under the null hypothesis*.

A relatively large value of T indicates that the true median is larger than m_0, whereas a relatively small value of T is an indication that the true median is smaller than m_0. In the given two-sided problem the null hypothesis is therefore rejected for large and small values of T. In terms of p-values: H_0 is rejected when the observed value t of T satisfies

$$P_{H_0}(T \leq t) \leq \tfrac{1}{2}\alpha, \qquad \text{or} \qquad P_{H_0}(T \geq t) \leq \tfrac{1}{2}\alpha.$$

These probabilities can be found by straightforward computation, from tables of the binomial distribution or with a statistical computer package. For large n one can also apply the normal approximation.

Of course, the test statistic T can also be used for left- or right-sided testing problems. An equivalent test statistic—for which testing a pair of hypotheses for a particular data set gives the same p-value, and hence the same conclusion, as testing the same pair of hypotheses based on T—is

$$\tilde{T} = \sum_{i=1}^{n} \operatorname{sgn}(X_i - m_0) = 2T - n.$$

This explains the name of the sign test.

Example 6.1 To judge the level of an exam a group of 13 randomly chosen students was asked to make the exam beforehand. The results were:

3.7, 5.2, 6.9, 7.2, 6.4, 9.3, 4.3, 8.4, 6.5, 8.1, 7.3, 6.1, 5.8. We test, with level $\alpha = 0.05$, the null hypothesis that the median of the (future) exam results in the population of all students is ≤ 6, against the alternative that the median is > 6. We thus have a right-sided test. The value $t = 9$ yields the p-value $P_{H_0}(T \geq 9) = 0.13$. This is larger than α and hence, the null hypothesis is not rejected. We may conclude that the level of the exam is probably suitable.
***

In the foregoing example we were lucky that there were no grades exactly equal to 6. If there had been grades equal to 6, this would have violated the assumption that the probability of an observation being equal to the median is 0. For less 'lucky' results the sign test can be adjusted in a simple way: leave out the values for which $X_i = m_0$ and perform the test as described above. This yields a test which is nonparametric *conditional on* the number of the X_i that is equal to m_0.

6.1.2 The signed rank test

For applying a sign test one has to make only a very weak assumption about the underlying distribution. This can be done because only little information in the data is used for the test. Only the signs $\text{sgn}(X_i - m_0)$ are of importance for the sign test; the absolute value of the deviation $X_i - m_0$ does not play a role for this test. The symmetry test of Wilcoxon, called the (*Wilcoxon*) *signed rank test*, makes use of more information in the observations, but requires a stronger assumption about the underlying distribution of $X_1, \ldots, X_n$. Instead of assuming a unique median, it is now assumed that the underlying distribution is symmetric.

For convenience we assume for the time being that the underlying distribution F has a continuous distribution function. Let m denote the point of symmetry of F. We consider the testing problem

$$H_0 : m = m_0$$

$$H_1 : m \neq m_0.$$

Note that, for a distribution that has a unique median and is symmetric, the point of symmetry is equal to the median, so that in this case the hypotheses also concern the median. To see how the test statistic is constructed, let $Z_i = X_i - m_0$. Since F has a continuous distribution function, the n values $|Z_1|, \ldots, |Z_n|$ are different from each other with probability 1. Let $(R_1, \ldots, R_n)$ denote the vector of *ranks* of $|Z_1|, \ldots, |Z_n|$ in the

corresponding vector of order statistics. This is, $|Z_i|$ is the R_i-th in size (in increasing order) of the $|Z_1|, \ldots, |Z_n|$. The signed rank test is based on the test statistic

$$V = \sum_{i=1}^{n} R_i \ \operatorname{sgn}(X_i - m_0).$$

Each of the n signs $\operatorname{sgn}(X_i - m_0)$ is equal to 1 or -1. A value of 1 is an indication that the true symmetry point is larger than m_0. This indication is stronger when $|X_i - m_0|$, and hence R_i, is larger. Relatively large values of V thus indicate that the true distribution of $X_1, \ldots, X_n$ has a larger point of symmetry than m_0, whereas relatively small values of V point to the opposite. Critical values and p-values again follow from the distribution of V under the null hypothesis. From the following theorem it follows that the signed rank test is nonparametric.

Theorem 6.1 *Let $Z_1, \ldots, Z_n$ be independent random variables, with a distribution that is symmetric around 0 and with a continuous distribution function. Let $(R_1, \ldots, R_n)$ be the vector of ranks of $|Z_1|, \ldots, |Z_n|$ in the corresponding vector of order statistics $(|Z|_{(1)}, \ldots, |Z|_{(n)})$. Then the following three properties hold.*

(i) The vectors $(R_1, \ldots, R_n)$ and $(\operatorname{sgn}(Z_1), \ldots, \operatorname{sgn}(Z_n))$ are independent.

(ii) $P(R_1 = r_1, \ldots, R_n = r_n) = 1/n!$ for every permutation $(r_1, \ldots, r_n)$ of $\{1, 2, \ldots, n\}$.

(iii) The variables $\operatorname{sgn}(Z_1), \ldots, \operatorname{sgn}(Z_n)$ are independent and identically distributed with $P(\operatorname{sgn}(Z_i) = -1) = P(\operatorname{sgn}(Z_i) = 1) = \frac{1}{2}$.

Proof. Since Z_i is distributed symmetrically around 0, it holds that

$$P(|Z_i| \leq x) = 2P(0 < Z_i \leq x)$$

and

$$P(\operatorname{sgn}(Z_i) = 1) = P(\operatorname{sgn}(Z_i) = -1) = \tfrac{1}{2}.$$

From this last equation property (iii) follows immediately.

Furthermore, we may conclude that for every $x > 0$

$$P(|Z_i| \leq x \wedge \operatorname{sgn}(Z_i) = 1) = P(0 < Z_i \leq x) = P(|Z_i| \leq x)P(\operatorname{sgn}(Z_i) = 1).$$

Analogously,

$$P(|Z_i| \leq x \wedge \operatorname{sgn}(Z_i) = -1) = P(|Z_i| \leq x)P(\operatorname{sgn}(Z_i) = -1).$$

This proves that $|Z_i|$ and $\operatorname{sgn}(Z_i)$ are independent. But then the same holds for the vectors $(|Z_1|, \ldots, |Z_n|)$ and $(\operatorname{sgn}(Z_1), \ldots, \operatorname{sgn}(Z_n))$. This implies (i), because the ranks $(R_1, \ldots, R_n)$ are a function of the first vector only.

Property (ii) is an immediate consequence of the fact that $|Z_1|, \ldots, |Z_n|$ are independent and identically distributed with a continuous distribution function. ∎

Let $\tilde{R}_1, \ldots, \tilde{R}_n$ be a random permutation of the numbers $\{1, 2, \ldots, n\}$, and $Q_1, \ldots, Q_n$ a sequence of independent random variables with $P(Q_i = -1) = P(Q_i = 1) = \frac{1}{2}$, which is also independent from the permutation $\tilde{R}_1, \ldots, \tilde{R}_n$. Then, according to Theorem 6.1, under the null hypothesis the test statistic V of the signed rank test satisfies

$$V \stackrel{\mathcal{D}}{=} \sum_{i=1}^{n} Q_i \tilde{R}_i.$$

Hence, the signed rank test is nonparametric. Moreover, it easily follows that under H_0

$$V \stackrel{\mathcal{D}}{=} \sum_{i=1}^{n} Q_i \tilde{R}_i \stackrel{\mathcal{D}}{=} \sum_{i=1}^{n} (-Q_i) \tilde{R}_i = -\sum_{i=1}^{n} Q_i \tilde{R}_i \stackrel{\mathcal{D}}{=} -V,$$

from which it follows that under H_0 the statistic V is symmetrically distributed around 0.

Critical values and p-values of the signed rank test can be deduced from Theorem 6.1. They are tabulated and the test is standardly available in statistical software packages. For large n a normal approximation can be applied. Under H_0 it holds that

$$\frac{V}{\sqrt{n(n+1)(2n+1)/6}} \stackrel{\mathcal{D}}{\to} N(0,1), \qquad n \to \infty.^{1}$$

An equivalent test statistic, which is frequently used instead of V, is

$$V_+ = \sum_{i: X_i > m_0} R_i,$$

which is the sum of the ranks of only the *positive* differences $Z_i = X_i - m_0$, where the ranks are as defined above, i.e. the ranks in the ordered sequence of *all* absolute differences $|Z_i|$. A test based on V_+ is also nonparametric. Although the distribution of V_+ under H_0 is different from that of V, for a particular data set testing a pair of hypotheses based on V_+ gives the same p-value, and hence the same conclusion, as testing the same pair of hypotheses based on V. It holds that

$$V_+ = \frac{1}{2} V + \frac{n(n+1)}{4},$$

and under H_0

$$\frac{V_+}{\sqrt{n(n+1)(2n+1)/24}} - \frac{n(n+1)}{4} \stackrel{\mathcal{D}}{\to} N(0,1), \qquad n \to \infty.$$

[1] The arrow with $\mathcal{D}$ on top, denotes convergence in distribution: '$X_n \stackrel{\mathcal{D}}{\to} X, \quad n \to \infty$' means that for every x it holds that $P(X_n \leq x) \to P(X \leq x)$, when $n \to \infty$.

Example 6.2 (Example 6.1 continued.)
We now test the null hypothesis that the point of symmetry $m \leq 6$ against the alternative $m > 6$. Again we have a right-sided test. The ordered sequence of absolute values of the differences of the exam grades with 6 is
0.1, 0.2, 0.4, 0.5, 0.8, 0.9, 1.2, 1.3, 1.7, 2.1, 2.3, 2.4, 3.3.
The vector of ranks is
$(11, 5, 6, 7, 3, 13, 9, 12, 4, 10, 8, 1, 2)$
and the corresponding vector of signs is
$(-1, -1, 1, 1, 1, 1, -1, 1, 1, 1, 1, 1, -1)$.
This yields $v = 37$ and $v_+ = 64$. Because $P_{H_0}(V \geq 37) = P_{H_0}(V_+ \geq 64) = 0.1082$, H_0 is not rejected.
$***$

In practice it may happen that one or more of the observations equal m_0. In that case, the procedure is the same as with the sign test: the observations that equal m_0 are deleted and the test is performed as described above, conditionally on the number of the observations that equal m_0.

It also often occurs that groups of the same values, called *ties*, are present in the sample. The occurrence of ties is contrary to the assumption of the continuity of the distribution function that we made up to now. The signed rank test is adapted to this situation as follows. One starts with deleting all values X_i that equal m_0, or equivalently, all values $Z_i = 0$. Next, one assigns adjusted ranks to the remaining values: every member of a group of equal values gets a 'pseudo rank', namely the mean of the ranks that the group members would have gotten if they all would have been different. For example, the ranks of $(3, 2, 2, 5, 3, 3)$ in the corresponding vector of order statistics $(2, 2, 3, 3, 3, 5)$ become $(4, 1\frac{1}{2}, 1\frac{1}{2}, 6, 4, 4)$. The test statistic V is then computed as before. Under H_0 and the given pattern of ties this statistic still has a fixed distribution, which depends on the pattern of the ties (the number and sizes of the groups of equal observations). Theorem 6.1 is no longer applicable and the critical values of the test need to be adjusted. In many statistical packages for large n a normal approximation with the correct adjustment is used by default.

6.2 Asymptotic efficiency

Nonparametric tests are in the first place important, because they always have the correct significance level. A good test also needs to have a large *power*. It turns out that the presented tests also perform well in this respect. They can be infinitely better than the classical tests and perform reasonably in situations where the classical tests are optimal.

A test that has more power than another test for the same number of observations is said to be more efficient. In other words, a more efficient test needs fewer observations to obtain the same power as the less efficient test. Like in the case of (robust) estimators we will consider the *asymptotic* case for the number of observations n tending to infinity, and we will see that the asymptotic variance (now of the test statistic) plays a role in determining the asymptotic efficiency.

To illustrate this we consider the one-sample problem again. Let $X_1, \ldots, X_n$ be observations from a distribution F. According to H_0 F belongs to a class $\mathcal{F}_0$ (for example, all distributions with median m_0), whereas according to H_1 F belongs to a class $\mathcal{F}_1$. The *power* of a test is the function

$$\pi(F) = P_F(H_0 \text{ is rejected}).$$

For a test to be good $\pi(F)$ should be small when $F \in \mathcal{F}_0$ and large when $F \in \mathcal{F}_1$. Because in this chapter we are specifically interested in the situation that $\mathcal{F}_0$ and $\mathcal{F}_1$ are large classes of probability distributions, it is rather difficult to compare the power of two tests for every possible F. This is why we will only discuss the so-called *shift alternatives*. These are alternatives that can be obtained by shifting a distribution that belongs to $\mathcal{F}_0$ over a certain distance θ. Such alternatives therefore have a location that is shifted over this distance θ, whereas the scale stays the same. To fix thoughts, we limit the discussion to *right-sided* testing problems, i.e. $\theta > 0$; the reasoning is the same for left-sided and two-sided problems.

Choose a fixed $F_0 \in \mathcal{F}_0$, and assume that the distribution that is shifted with respect to F_0 over a positive distance θ belongs to the alternative hypothesis. The shifted distribution is denoted by $F_\theta(\cdot) := F_0(\cdot - \theta)$. If, for example, one would consider hypotheses about the median, F_0 could be a distribution with median m_0, and then F_θ would have median $m_0 + \theta$. The power for the class of shift alternatives F_θ can be written as

$$\pi_n(\theta) = P_\theta(H_0 \text{ is rejected}).$$

For a suitable right-sided test the value $\pi_n(0)$ of the power function under H_0 is small, whereas $\pi_n(\theta)$ is 'large' for $\theta > 0$, i.e. under the alternative hypothesis.

Suppose that H_0 is rejected for a large value of the test statistic T_n. Assume that this test statistic is asymptotically normally distributed in the sense that

$$\sqrt{n}(T_n - \mu(\theta)) \overset{\mathcal{D},\theta}{\to} N(0, \sigma^2(\theta)), \qquad n \to \infty, \tag{6.1}$$

for suitable functions $\mu(\theta)$ and $\sigma^2(\theta)$ (the asymptotic mean and asymptotic variance of T_n). This means: for large n, T_n is approximately distributed as $N(\mu(\theta), \sigma^2(\theta)/n)$, when θ is the true value of the parameter.

Rejecting the null hypothesis for large values of T_n, is equivalent to rejecting the null hypothesis if $\sqrt{n}(T_n - \mu(0))/\sigma(0) > c_n$ for a suitably chosen critical value c_n. Due to the

asymptotic normality of T_n the level of the test is then

$$\pi_n(0) = P_0\left(\frac{\sqrt{n}(T_n - \mu(0))}{\sigma(0)} > c_n\right) \approx 1 - \Phi(c_n),$$

where Φ is the distribution function of the standard normal distribution. From this we see that to make the level of the test equal to a chosen α, the critical value c_n should be equal to $\xi_{1-\alpha}$, the $1-\alpha$-quantile of the $N(0,1)$ distribution. The test thus becomes

$$\text{“ Reject } H_0 \text{ if } \sqrt{n}(T_n - \mu(0))/\sigma(0) > \xi_{1-\alpha} \text{ ”}.$$

The power of the test is then equal to

$$\begin{aligned}\pi_n(\theta) &= P_\theta\left(\frac{\sqrt{n}(T_n - \mu(0))}{\sigma(0)} > \xi_{1-\alpha}\right) \\ &= P_\theta\left(\frac{\sqrt{n}(T_n - \mu(\theta))}{\sigma(\theta)} > \frac{\sigma(0)}{\sigma(\theta)}\,\xi_{1-\alpha} - \frac{\sqrt{n}(\mu(\theta) - \mu(0))}{\sigma(\theta)}\right) \\ &\approx 1 - \Phi\left(\frac{\sigma(0)}{\sigma(\theta)}\,\xi_{1-\alpha} - \frac{\sqrt{n}(\mu(\theta) - \mu(0))}{\sigma(\theta)}\right), \qquad (6.2)\end{aligned}$$

where the last approximation comes from (6.1).

From (6.2) several things can be concluded. First, a sequence of tests $\{T_n\}$ is called *consistent* when for a fixed level α the power tends to 1 for each alternative when $n \to \infty$. Consistency is a very desirable property of a test: with infinitely many observations a test should be able to distinguish the alternative from the null hypothesis with certainty. In the present case the sequence of tests is consistent for the shifted alternatives if $\mu(\theta) > \mu(0)$ for every $\theta > 0$, because then $\sqrt{n}(\mu(\theta) - \mu(0)) \to \infty$, so that $\pi_n(\theta) \to 1 - \Phi(-\infty) = 1$.

Next, since every reasonable test will be consistent, for a comparison of tests it is not sufficient to only consider consistency. To make a useful comparison of the quality of tests we will need to study the performance of the tests, and in particular the power of the tests, in a situation where the testing problem becomes more difficult when more observations become available. This situation is created by choosing the alternatives to come close to the null hypothesis when the number n of observations grows. However, they should not come too close too quickly, because then it becomes impossible to make the distinction between the two hypotheses for any test. This is why we now choose θ_n to be of the form $\theta_n = h/\sqrt{n}$ for a constant h, so that $\theta_n \to 0$, but not too fast, and study the power $\pi_n(\theta_n) = \pi_n(h/\sqrt{n})$ for different h while letting $n \to \infty$. It can be derived from (6.2) that when μ is differentiable in 0, and $\sigma(\theta_n) \to \sigma(0)$, then

$$\pi_n(\theta_n) = \pi_n(h/\sqrt{n}) \approx 1 - \Phi\left(\xi_{1-\alpha} - \frac{\mu'(0)}{\sigma(0)}h\right) \qquad (6.3)$$

for large n. [2] The value $\mu'(0)/\sigma(0)$ is called the *slope* of the test. From (6.3) it can be seen that when h increases, the power of the test increases faster when the slope is larger. This is illustrated in Figure 6.1. Hence: the larger the slope, the better the test.

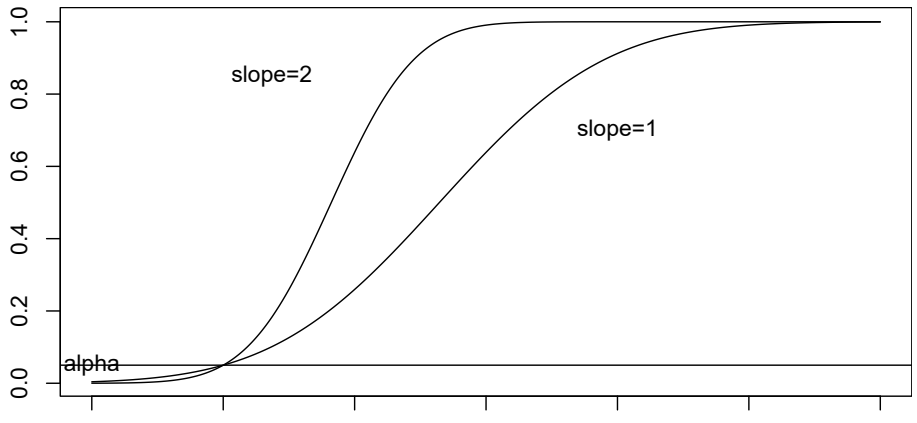


Figure 6.1: The asymptotic power of two tests with different slope (y-axis) as function of h (x-axis).

Let a second test based on statistic $\tilde{T}_n$ have power $\tilde{\pi}_n$ which satisfies

$$\tilde{\pi}_n(\theta_n) = \tilde{\pi}_n(h/\sqrt{n}) \approx 1 - \Phi\left(\xi_{1-\alpha} - \frac{\tilde{\mu}'(0)}{\tilde{\sigma}(0)}h\right).$$

Then the *asymptotic relative efficiency* of the test T_n with respect to $\tilde{T}_n$ is defined as the square of the quotient of the slopes of the tests:

$$\text{are}(T_n, \tilde{T}_n) = \left(\frac{\mu'(0)/\sigma(0)}{\tilde{\mu}'(0)/\tilde{\sigma}(0)}\right)^2. \tag{6.4}$$

If $\text{are}(T_n, \tilde{T}_n) > 1$, then T_n is more efficient than $\tilde{T}_n$; if $\text{are}(T_n, \tilde{T}_n) < 1$, then $\tilde{T}_n$ is more efficient.

Note that, because $\sigma(0)$ in (6.4) is the asymptotic standard deviation of the test statistic under the null hypothesis, the asymptotic variance of the test statistic T_n plays a role in determining the asymptotic efficiency, as was already indicated in the beginning of this section. The smaller the asymptotic variance, the more efficient the test.

[2] When μ is differentiable in 0,

$$\sqrt{n}\big(\mu(h/\sqrt{n}) - \mu(0)\big) = h\,\frac{\mu(h/\sqrt{n}) - \mu(0)}{(h/\sqrt{n})} \to h\,\mu'(0), \qquad n \to \infty.$$

When also $\sigma(\theta_n) \to \sigma(0)$, (6.3) follows from (6.2).

Like in the case of estimators, where the asymptotic relative efficiency could be explained in terms of the number of observations needed for two estimators to have approximately equal asymptotic variances, it is possible to interpret the asymptotic relative efficiency of two tests in terms of sample sizes. Namely, it is the ratio of the numbers of observations needed for the tests to have approximately equal asymptotic power, for a given level α. To see this, we note that for a sequence of alternatives $\theta_n = h/\sqrt{n}$ it follows from (6.3) that the first test T_n with n observations reaches a power

$$\pi_n(\theta_n) \approx 1 - \Phi\left(\xi_{1-\alpha} - \frac{\mu'(0)}{\sigma(0)}\sqrt{n}\,\theta_n\right),$$

and that for the second test $\tilde{T}_n$ with $\tilde{n}$ observations the power is

$$\tilde{\pi}_{\tilde{n}}(\theta_n) \approx 1 - \Phi\left(\xi_{1-\alpha} - \frac{\tilde{\mu}'(0)}{\tilde{\sigma}(0)}\sqrt{\tilde{n}}\,\theta_n\right).$$

Therefore, the powers of the two tests are approximately equal when

$$\frac{\mu'(0)}{\sigma(0)}\sqrt{n} = \frac{\tilde{\mu}'(0)}{\tilde{\sigma}(0)}\sqrt{\tilde{n}},$$

or when

$$\frac{\tilde{n}}{n} = \left(\frac{\mu'(0)/\sigma(0)}{\tilde{\mu}'(0)/\tilde{\sigma}(0)}\right)^2 = \text{are}(T_n, \tilde{T}_n).$$

In words: for the second test there are $\text{are}(T_n, \tilde{T}_n)$ as many observations needed as for the first test to obtain approximately the same power as the first test. If $\text{are}(T_n, \tilde{T}_n) > 1$, then the first test is to be preferred.

What has been neglected so far, is the fact that $\text{are}(T_n, \tilde{T}_n)$ in fact depends on the type of shift alternatives F_θ that is considered. Indeed, we see from (6.1) and (6.4) that $\text{are}(T_n, \tilde{T}_n)$ depends on the asymptotic distribution of the two test statistics for $\theta = 0$, which in turn depends on F_0. Table 6.1 gives the relative efficiencies of the one-sample tests that we have discussed for a couple of shift alternatives. From the table we see that none of the three tests is optimal in all four cases. The sign test, which is based on a very simple principle, turns out to be the best against Laplace shift alternatives. The Wilcoxon signed rank test takes a middle position in between the t-test and the sign test. For the signed rank test the loss of efficiency with respect to the t-test in case the true underlying distribution is exactly normal, is small ($3/\pi = 0.955 \approx 1$), whereas the gain with respect to the t-test for Laplace shift alternatives is considerable (3/2=1.5). This makes the signed rank test a serious competitor of the t-test: besides the fact that this test has the correct level of significance for a large class of null-distributions, it also has good efficiency properties!

An advantage of the sign test which does not show from the table is that this test is nonparametric against all alternatives, whereas for the Wilcoxon signed rank test and the

	t	s	w
t	1		
s	$\frac{2}{\pi}$	1	
w	$\frac{3}{\pi}$	$\frac{3}{2}$	1

N(0,1)

	t	s	w
t	1		
s	$\frac{\pi^2}{12}$	1	
w	$\frac{\pi^2}{9}$	$\frac{4}{3}$	1

logistic

	t	s	w
t	1		
s	$\frac{1}{3}$	1	
w	1	3	1

uniform

	t	s	w
t	1		
s	2	1	
w	$\frac{3}{2}$	$\frac{3}{4}$	1

Laplace

Table 6.1: Asymptotic relative efficiencies (row-variable with respect to column-variable) of the t-test (t), sign test (s) and Wilcoxon signed rank test (w) for shift alternatives of different F_0.

t test symmetry and normality, respectively, are required. This makes the sign test the best choice in many situations.

We have restricted our comparison of the different tests to shift alternatives of a couple of distributions. Although these are in most cases the alternatives of interest, in some contexts it may be necessary to know how the tests compare for other alternatives.

6.3 Two-sample problems

It frequently happens that one wishes to compare two univariate samples with each other. The samples can be paired or unpaired. In the *paired*-sample problems the data consist of a sequence of independent bivariate random variables $(X_1, Y_1), \ldots, (X_n, Y_n)$. For example, each of these variables may have been obtained by taking two measurements on one and the same experimental unit. In the *unpaired* two-sample problem one has two independent groups of univariate variables $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$. In this case the number of variables in each group need not be the same.

Example 6.3 One wants to investigate whether there is a difference in weight between people who go to work by bike and people who go by car. Taking samples of sizes m and n from these categories of people and determining the weights of the people in the samples leads to unpaired data.
***

Example 6.4 From a group of patients the concentration of a chemical substance in the blood is measured one week before and one week after application of a medicine. This leads to paired data.
***

Mostly X_i and Y_i within a pair (X_i, Y_i) of a paired data set are not independent, although the differences $Z_i = Y_i - X_i$ for the different i often are independent. When the latter holds, one could test whether one sample is stochastically larger[3] than another one by applying a one-sample test to $Z_1, \ldots, Z_n$. For instance, with the sign test one could test whether the median of the distribution of the differences is significantly different from 0. When X_i and Y_i are independent, then under the null hypothesis that the two samples have the same distribution, the Z_i are automatically symmetrically distributed around zero. Wilcoxon's symmetry test, the signed rank test, is in this case the obvious test.

From now on we shall limit ourselves in this section to two unpaired, independent samples $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$. Suppose that $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ have true distributions F and G, respectively. Consider the testing problem

$$H_0 : F = G$$
$$H_1 : F \neq G.$$

In the 'classical model' for this problem the two samples are both normally distributed with expectations μ and ν and the testing problem is $H_0 : \mu = \nu$ against $H_1 : \mu \neq \nu$. If, in addition, it is assumed that the two samples have equal variances, then the test is based on the statistic

$$T = \frac{\bar{X} - \bar{Y}}{S_{X,Y}\sqrt{1/m + 1/n}},$$

where $S_{X,Y} = \sqrt{S^2_{X,Y}}$ and $S^2_{X,Y} = \frac{1}{m+n-2}\{\sum_{i=1}^m (X_i - \bar{X})^2 + \sum_{i=1}^n (Y_i - \bar{Y})^2\}$. This statistic has a t_{m+n-2} distribution. When the variances are not assumed to be equal, then a test can be based on

$$\tilde{T} = \frac{\bar{X} - \bar{Y}}{\sqrt{S^2_X/m + S^2_Y/n}}.$$

This statistic does not have a t-distribution, but approximations for the critical values are available.

In the sequel we do *not* make the normality assumption. Unless stated otherwise, we assume for convenience that F and G have continuous distribution functions. When this assumption is not fulfilled, the presented tests can still be used, but need to be adapted. In practice this only needs to be done when there is a relatively large number of ties.

6.3.1 The median test

Combine the two samples into one sample $X_1, \ldots, X_m, Y_1, \ldots, Y_n$ of size $m + n$, and determine the median of this new sample. The *median test* has as test statistic

$$Z = \#\Big(X_i \leq \text{med}(X_1, \ldots, X_m, Y_1, \ldots, Y_n)\Big).$$

[3]When X has distribution function F_X, and Y has distribution function F_Y, then (the distribution of) Y is stochastically larger than (the distribution of) X if $F_Y(x) \leq F_X(x)$ for all x.

Under H_0 the observations $X_1, \ldots, X_m, Y_1, \ldots, Y_n$ are independent and identically distributed. Every ordering of these $m+n$ variables then has the same probability, so that the distribution of Z under H_0 is the same as in the following experiment. Randomly select without replacement one half, or, more precisely, $p = [\frac{1}{2}(m+n+1)]$ variables from the combined sample $X_1, \ldots, X_m, Y_1, \ldots, Y_n$. [4] Next, count the number Z of elements in the selected set of size p that belongs to $X_1, \ldots, X_m$. The distribution of Z under H_0 is hypergeometric:

$$P_{H_0}(Z = z) = \frac{\binom{m}{z}\binom{n}{p-z}}{\binom{m+n}{p}}.$$

The median test is thus nonparametric. In the two-sided problem H_0 is rejected for both large and small values of Z. Large values of Z are more likely when $P(Y > X) > \frac{1}{2}$, and small values are more likely when $P(Y > X) < \frac{1}{2}$, whereas under H_0 it holds that $P(Y > X) = \frac{1}{2}$. Of course, the test can also be applied right- or left-sided.

6.3.2 The Wilcoxon two-sample test

Again combine the two samples into one sample $X_1, \ldots, X_m, Y_1, \ldots, Y_n$ of size $N = m+n$ and let $R_1, \ldots, R_n$ be the ranks of $Y_1, \ldots, Y_n$ in the combined sample. (Thus $R_1, \ldots, R_n$ form a subset of $\{1, 2, \ldots, N\}$). The *Wilcoxon two-sample test*, also called the *Wilcoxon rank sum test* or the *Mann-Whitney test*, is based on the test statistic

$$W = \sum_{i=1}^{n} R_i.$$

The null hypothesis is rejected for large and small values of W.

Under the null hypothesis each of the $N!$ possible orderings of the combined sample has the same probability. Therefore, the ordered ranks $R_{(1)}, \ldots, R_{(n)}$ of $Y_1, \ldots, Y_n$ satisfy

$$P_{H_0}((R_{(1)}, \ldots, R_{(n)}) = (r_1, \ldots, r_n)) = \frac{1}{\binom{N}{n}},$$

for every subset $r_1 < r_2 < \cdots < r_n$ of $\{1, 2, \ldots, N\}$. The distribution of the test statistic is given by

$$P_{H_0}(W = w) = \frac{\#(r_1 < r_2 < \cdots < r_n \text{ with } \sum_{i=1}^{n} r_i = w)}{\binom{N}{n}}.$$

Hence, the test is nonparametric. Critical values and p-values can be found by straightforward computation, from tables or with statistical software. For large m and n a normal approximation can be used, because

$$\frac{W - \frac{1}{2}n(N+1)}{\sqrt{mn(N+1)/12}} \xrightarrow{\mathcal{D}} N(0,1), \qquad m, n \to \infty,$$

[4] The notation $[x]$ stands for the *entier* or *floor* of x, which is the largest integer not greater than x.

(provided $0 < P(X_i < Y_j) < 1$). An equivalent test statistic, which is often used instead of W and which gives different critical values but exactly the same p-values as W, is

$$U = \sum_{i=1}^{m}\sum_{j=1}^{n} 1_{\{X_i<Y_j\}} = W - \tfrac{1}{2}n(n+1).$$

Under H_0 the statistic U is symmetrically distributed around $\frac{1}{2}mn$, and

$$\frac{U - \frac{1}{2}mn}{\sqrt{mn(N+1)/12}} \xrightarrow{\mathcal{D}} N(0,1), \qquad m,n \to \infty.$$

It should be noted that there is no unanimity about the definition of the Mann-Whitney test. Not only are both the above-defined equivalent test statistics W and U used, but also the test statistics $\tilde{W}$ and, equivalently to this, $\tilde{U}$, where in their definitions the roles of the first and second sample are reversed with respect to the definitions of W and U. This means that for two-sided testing problems testing with $\tilde{W}$ or $\tilde{U}$ yields the same p-value as testing with W or U. Note that for one-sided testing problems testing with $\tilde{W}$ or $\tilde{U}$ instead of W or U results in the same p-values, but the critical regions lie on the other side.

When there are groups of equal values, ties, in the combined sample $X_1, \ldots, X_m, Y_1, \ldots, \ldots, Y_n$, the test can be adjusted as follows. First, pseudo ranks are defined in the following way. Each element of a tie gets a pseudo rank which is the mean of the ranks the elements of the tie would have gotten if they would have been different from each other. Observations that have no equal get their normal rank. Let $R_1, \ldots, R_n$ be the pseudo ranks of $Y_1, \ldots, Y_n$ in the combined samples $X_1, \ldots, X_m, Y_1, \ldots, Y_n$. The test statistic is $W = \sum_{i=1}^{n} R_i$. It can be proved that, given the pattern of ties, W has a fixed distribution under H_0, so that the test is nonparametric given the pattern of ties.

To see what is meant with the 'pattern of ties', suppose that K different values are present in the sample $X_1, \ldots, X_m, Y_1, \ldots, Y_n$, and that the smallest value occurs T_1 times, the one but smallest T_2 times,..., and the largest T_K times. The pattern of ties is then described by the vector $(K, T_1, \ldots, T_K)$. The conditional distribution of $(R_1, R_2, \ldots, R_n)$ given $(K, T_1, \ldots, T_K)$ is, under the null hypothesis $H_0 : F = G$, the same as the distribution of n randomly and without-replacement selected numbers from the set

B:	5	8	7	22	6	7	2	6	6	20	7	6
rank:	9	36	26.5	65	15.5	26.5	1.5	15.5	15.5	64	26.5	15.5
B:	9	13	4	3	6	7	7	4	6	8	6	6
rank:	45	61	6	3	15.5	26.5	26.5	6	15.5	36	15.5	15.5
B:	8	4	11	9	7	6	17	7	4	6		
rank:	36	6	56	45	26.5	15.5	63	26.5	6	15.5		

Table 6.2: Number of flaws in sample B with its ranks.

T_1 times the smallest pseudo rank,
T_2 times the one but smallest pseudo rank,
$\vdots$
T_K times the largest pseudo rank.

This is why the conditional distribution of $W = \sum_{i=1}^{n} R_i$ given $(K, T_1, \ldots, T_K)$ is fixed under H_0. Its form is somewhat complicated, but the corresponding p-values can be easily computed with a computer. For large m and n an adjusted normal approximation can be used: conditionally on $K = k, T_1 = t_1, \ldots, T_k = t_k$,

$$\frac{W - \frac{1}{2}n(N+1)}{\sqrt{mn(N^3 - \sum_{i=1}^{k} t_i^3)/(12N(N-1))}} \xrightarrow{\mathcal{D}} N(0,1), \qquad m, n \to \infty.$$

Example 6.5 In Example 3.6 data were given concerning flaws in samples of sizes 31 and 34 of pieces of textile woven with methods A and B. Between the samples there is a mean difference of about 2 flaws, method B yielding the fewer flaws. Whether the difference is systematic or occurred by chance (namely, obtaining just these perhaps not representative samples), can be further investigated by means of a test. Because the data are clearly not normally distributed, the use of a t-test will be misleading. We chose the Mann-Whitney (Wilcoxon two-sample) test. The ordered combined sample is:

2, 2, 3, 4, 4, 4, 4, 4, 5, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 7, 7, 7, 7, 7, 7, 7, 7, 7, 8, 8, 8, 8, 8, 8, 8, 8, 8, 9, 9, 9, 9, 9, 9, 9, 9, 9, 10, 10, 10, 10, 11, 11, 11, 11, 11, 12, 12, 13, 14, 17, 20, 22.

The second sample together with its ranks in the combined sample are given in Table 6.2.

We test H_0: "The location of the second sample produced with method B is larger than the location of the first sample produced with method A" against H_1: "The location of the second sample produced with method B is smaller than the

location of the first sample produced with method A". The test statistic W has value $w = 885$. The normal approximation for the left p-value $P_{H_0}(W \leq 885)$ is equal to 0.0009. Hence, H_0 is rejected for every reasonable value of α. This indicates that the location of the second sample produced with method B is considerably smaller than that of the first sample.

Use of the test statistic $\tilde{U}$ with H_0 and H_1 reversed, would have given $\tilde{u} = 764$, and the same (but now right instead of left) p-value 0.0009, and, hence, the same conclusion.

Some doubt about the normal approximation is always justified. Although in principle it is possible to compute the p-value exactly, as a control the bootstrap was applied. From the set of pseudo ranks of the combined sample 6000 times a sample of size $n = 34$ was taken without replacement. Of each sample the sum w^* was computed. The value $w = 885$ was the 0.0005-quantile of the empirical distribution of the set of 6000 values of w^*. The bootstrap approximation for the left p-value is therefore equal to 0.0005, which confirms the normal approximation.
***

6.3.3 The Kolmogorov-Smirnov test

Let $\hat{F}_m$ and $\hat{G}_n$ be the empirical distribution function of $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$, respectively. The *Kolmogorov-Smirnov* two-sample test is based on the statistic

$$D = \sup_{-\infty<x<\infty} |\hat{F}_m(x) - \hat{G}_n(x)|.$$

An easy way to compute the test statistic when there are no ties follows from

$$\begin{aligned} D &= \max_{1\leq i\leq n} \max\Big\{ |\hat{F}_m(Y_{(i)}) - \hat{G}_n(Y_{(i)})|, |\hat{F}_m(Y_{(i)}) - (\hat{G}_n(Y_{(i)}) - \frac{1}{n})| \Big\} \\ &= \max_{1\leq i\leq n} \max\Big\{ |\frac{1}{m}(R_{(i)} - i) - \frac{i}{n}|, |\frac{1}{m}(R_{(i)} - i) - \frac{i-1}{n}| \Big\}. \end{aligned} \tag{6.5}$$

Here $R_{(1)}, \ldots, R_{(n)}$ are the ordered ranks of $Y_1, \ldots, Y_n$ in the combined sample $X_1, \ldots, X_m$, $Y_1, \ldots, Y_n$. From this formula it follows that the value of D is completely determined by the positions taken by $Y_1, \ldots, Y_n$ in the ordered combined sample $X_1, \ldots, X_m, Y_1, \ldots, Y_n$. There are $\binom{m+n}{n}$ possible groups of n positions. Under the null hypothesis these are all equally likely. This is why the test is nonparametric. The null hypothesis is rejected for large values of D. The p-values again can be found in tables, from computer packages or by direct computation. Figure 6.2 illustrates the idea of the Kolmogorov-Smirnov test, although for the data set used in this picture, the Kolmogorov-Smirnov test should be adapted, because it contains relatively many ties.

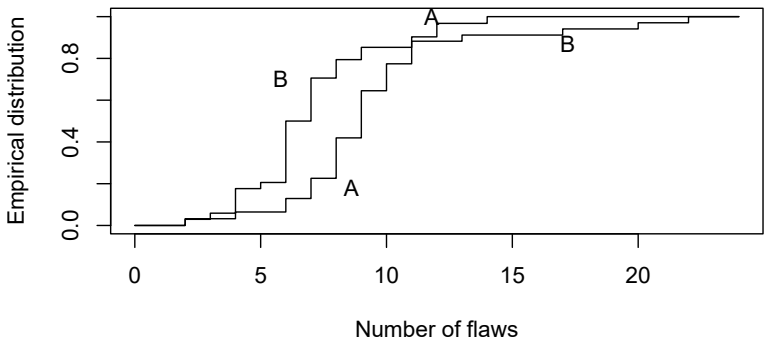


Figure 6.2: The empirical distribution function of the number of flaws in textile produced by methods A and B from Example 3.6.

6.3.4 Permutation tests

Write the ordered combined sample as $Z_{(1)}, \ldots, Z_{(m+n)}$. Under the null hypothesis that the two samples have the same distribution, each of the $(m+n)!$ possible permutations of these quantities leads with the same probability back to the original observations $X_1, \ldots, X_m, Y_1, \ldots, Y_n$. This is why it holds that

$$P_{H_0}((X_1, \ldots, X_m, Y_1, \ldots, Y_n) = (Z_{(\pi_1)}, \ldots, Z_{(\pi_{m+n})}) | Z_{(1)}, \ldots, Z_{(m+n)}) = \frac{1}{(m+n)!},$$

for every permutation $(\pi_1, \ldots, \pi_{m+n})$ of $\{1, 2, \ldots, m+n\}$.

Now let $T = T(X_1, \ldots, X_m, Y_1, \ldots, Y_n)$ be a statistic for which a large (or small) value expresses an indication of a difference between the two samples, for example $T = \bar{X} - \bar{Y}$. When t is the observed value of T, the right-sided *permutation test* based on T rejects the null hypothesis if

$$P_{H_0}(T \geq t | Z_{(1)}, \ldots, Z_{(m+n)})$$
$$= \frac{\#(\text{permutations } \pi \text{ with } T(Z_{(\pi_1)}, \ldots, Z_{(\pi_m)}, Z_{(\pi_{m+1})}, \ldots, Z_{(\pi_{m+n})}) \geq t)}{(m+n)!}$$

is smaller than or equal to the level α. We thus see that given the values $Z_{(1)}, \ldots, Z_{(m+n)}$, a permutation test is nonparametric. A left-sided or two-sided test is defined analogously. Of course, the distribution of $Z_{(1)}, \ldots, Z_{(m+n)}$ does depend on the underlying distribution, so that the unconditional distribution is not nonparametric. The unconditional level of the test, however, is smaller than or equal to α. Indeed,

$$P_{F=G}(H_0 \text{ is rejected } | \ Z_{(1)}, \ldots, Z_{(m+n)}) \leq \alpha$$

for all $Z_{(1)}, \ldots, Z_{(m+n)}$. From this it easily follows that

$$\sup_{F=G} P_{F,G}(H_0 \text{ is rejected }) \leq \alpha.$$

In principle a permutation test can always be performed by complete enumeration of all $(m+n)!$ permutations. Often this time consuming approach can be replaced by a more efficient procedure. For instance, for $T = \bar{X} - \bar{Y}$ it turns out that $m!n!$ permutations give the same value of T, so that it is sufficient to compute the value of T for only $(m+n)!/(m!n!)$ permutations. In case the number of values T to be computed is too large for practical purposes, one could *approximate* the p-value $P_{H_0}(T \geq t | Z_{(1)}, \ldots, Z_{(m+n)})$ by the fraction

$$\frac{\#(\text{permutations } \pi \text{ with } T(Z_{(\pi_1)}, \ldots, Z_{(\pi_m)}, Z_{(\pi_{m+1})}, \ldots, Z_{(\pi_{m+n})}) \geq t)}{B}$$

for a large number B of *randomly* chosen permutations π of $\{1, 2, \ldots, m+n\}$. Note that if we perform a permutation test in this way, we actually perform a bootstrap test, because we *simulate* B values $T(Z_{(\pi_1)}, \ldots, Z_{(\pi_m)}, Z_{(\pi_{m+1})}, \ldots, Z_{(\pi_{m+n})})$ from the distribution of the test statistic under the null hypothesis, and use the empirical distribution of these B values as an estimator of the distribution of the test statistic under the null hypothesis.

We remark that permutation tests are used much more generally and in many other settings than the two-sample one. The above test is just one example of a permutation test.

6.3.5 Power and asymptotic efficiency

The median test and the Wilcoxon two-sample test are particularly sensitive, that is have large power, for alternatives where Y is stochastically larger or smaller than X. For example, these alternatives can be used when one is more interested in discovering a difference in location than a difference in scale. Shift-alternatives in which G is shifted with respect to F are an example of such alternatives. In most cases the median test has smaller power than Wilcoxon's two-sample test. This is because the median test only considers the position of each observation relative to the overall median, whereas the Wilcoxon test takes the ranks of each observation into account. For this reason the median test is rarely-used nowadays. The Kolmogorov-Smirnov two-sample test is well-suited for alternatives in which G has a different shape than F.

For the asymptotic efficiency of tests for two-sample problems, similar results hold as for the one-sample tests.

6.4 Tests for correlation

Suppose that $(X_1, Y_1), \ldots, (X_n, Y_n)$ are independent random vectors from a bivariate distribution. A frequently asked question is whether the X_i and Y_i variables are independent.

Example 6.6 In a study of the consequences of the development from a 'primitive' to a 'modern' society, systolic blood pressure was measured of 39 randomly selected Peruvian men who live in a city. Figure 6.3 gives a scatter plot of the blood pressure against the number of years since migration to the city. At first sight there is no relationship between the two variables. The sample correlation turns out to be -0.09, which confirms the conclusion of no relationship.
***

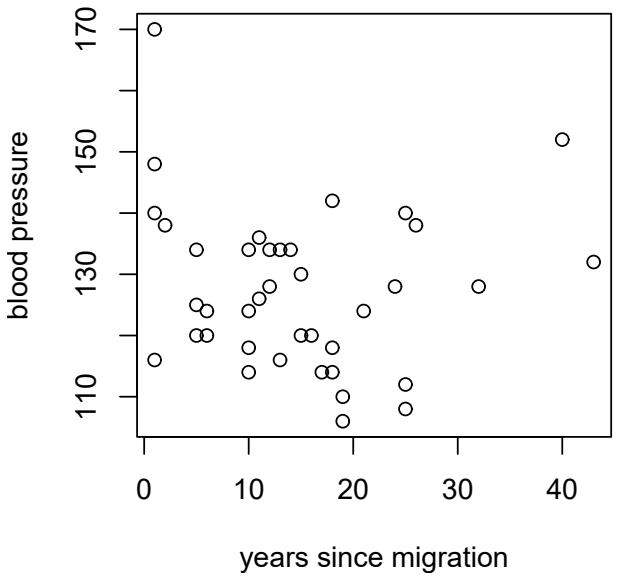


Figure 6.3: Scatter plot of systolic blood pressure against years since migration of 39 Peruvian men.

In the classical model for this situation it is assumed that the vectors (X_i, Y_i) originate from a bivariate normal distribution. This distribution is completely determined by five parameters: the two expectations, the two variances and the correlation coefficient ρ. Only the last parameter concerns the dependence of the X_i and Y_i variables. Therefore, to investigate dependencies among these variables, it is sufficient to estimate ρ or test hypotheses about ρ. The natural statistic for this is the sample correlation coefficient r_{xy}, which was introduced in Chapter 2. Because of the normality of the (X_i, Y_i) it has under the null hypothesis of independence a t_{n-2}-distribution. In the following we do not make the normality assumption.

Let $S_1, \ldots, S_n$ be the ranks of $X_1, \ldots, X_n$ in the ordered sequence $X_{(1)}, \ldots, X_{(n)}$, and define $R_1, \ldots, R_n$ analogously for $Y_1, \ldots, Y_n$. If $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ are independent, then the two groups of ranks $S_1, S_2, \ldots, S_n$ and $R_1, \ldots, R_n$ are also independent. Given that there are no ties, the groups of ranks can be seen as random permutations of the numbers $\{1, 2, \ldots, n\}$. If, however, the two samples are positively dependent, then

we expect that the two sequences of ranks will run in parallel, whereas under negative dependence the sequences of ranks will run in the opposite direction. We discuss three tests for the problem

$$H_0: \ X_i \text{ and } Y_i \text{ are independent}, \qquad i = 1, \ldots, n$$

$$H_1: \ X_i \text{ and } Y_i \text{ are dependent}, \qquad i = 1, \ldots, n.$$

When there are no ties, then every test which is based only on the vectors $R_1, \ldots, R_n$ and $S_1, \ldots, S_n$ is nonparametric under the null hypothesis. In the situation that ties do occur, such a test is nonparametric conditionally on the pattern of the ties.

6.4.1 The rank correlation test of Spearman

This test is based on the (sample) correlation coefficient of the two groups of ranks. The null hypothesis is rejected for values close to 1 or -1 of the statistic

$$r_s = \frac{\sum_{i=1}^n (R_i - \bar{R})(S_i - \bar{S})}{\left[\sum_{i=1}^n (R_i - \bar{R})^2 \sum_{i=1}^n (S_i - \bar{S})^2\right]^{\frac{1}{2}}}.$$

Note that r_s is just the sample correlation coefficient of the *ranks* of the two samples. Here it holds that $\bar{R} = \bar{S} = (n+1)/2$. If there are no ties in the samples, so that $\{R_1, \ldots, R_n\} = \{S_1, \ldots, S_n\} = \{1, \ldots, n\}$, we also have

$$\sum_{i=1}^n R_i^2 = \frac{n(2n+1)(n+1)}{6}; \qquad \sum_{i=1}^n \left(R_i - \bar{R}\right)^2 = \frac{n^3 - n}{12}.$$

A little computation shows that r_s then can be written as

$$r_s = 1 - \frac{6 \sum_{i=1}^n (R_i - S_i)^2}{n^3 - n}.$$

Therefore the test may as well be based on the statistic $L = \sum_{i=1}^n (R_i - S_i)^2$. For this quantity under the null hypothesis a normal approximation holds, irrespectively of the existence of ties. More precisely, given the pattern of ties, L is asymptotically normally distributed, with the asymptotic expectation and variance depending on this pattern:

$$\frac{L - \mu_n}{\sigma_n} \xrightarrow{\mathcal{D}} N(0,1), \qquad n \to \infty$$

where

$$\mu_n = \mathrm{E}_{H_0} L = (n^3 - n)/6 - \sum (U_i^3 - U_i)/12 - \sum (V_i^3 - V_i)/12$$

and

$$\sigma_n^2 = \sigma_{H_0}^2(L) = \frac{(n-1)n^2(n+1)^2}{36}\left[1 - \frac{\sum(U_i^3 - U_i)}{n^3 - n}\right]\left[1 - \frac{\sum(V_i^3 - V_i)}{n^3 - n}\right].$$

In these formulas the U_i and the V_i give the patterns of ties for the two sequences of ranks; U_1 is the number of times that the smallest rank occurs in the sequence $R_1, \ldots, R_n$, U_2 is is the number of times that the one but smallest rank occurs, etc.

Of course, a p-value can also be approximated by means of simulation instead of with a normal approximation.

6.4.2 The rank correlation test of Kendall

With this test H_0 is rejected for values close to 1 or -1 of the statistic

$$\tau = \frac{\sum\sum_{i \neq j} \operatorname{sgn}(R_i - R_j)\operatorname{sgn}(S_i - S_j)}{n(n-1)} = \frac{4N_\tau}{n(n-1)} - 1,$$

where the statistic N_τ is equal to the number of pairs (i, j) with $i < j$ for which either $X_i < X_j$ *and* $Y_i < Y_j$, or $X_i > X_j$ *and* $Y_i > Y_j$.

Sometimes the statistic N_τ itself is used as test statistic. This statistic lies between 0 and $n(n-1)/2$, and the test based on N_τ is equivalent to the test based on τ.

6.4.3 Permutation tests

Write $Y_{(1)}, \ldots, Y_{(n)}$ for the ordered second sample. For a given permutation $(Y_{(\pi_1)}, \ldots, Y_{(\pi_n)})$ of these values, the pairs $(X_1, Y_{(\pi_1)}), \ldots, (X_n, Y_{(\pi_n)})$ can be formed. Under the null hypothesis of independence, each of the $n!$ possible permutations leads with the same probability back to the original pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$. Suppose that large values of the statistic $T = T(X_1, \ldots, X_n, Y_1, \ldots, Y_n)$ express (positive) dependence between X_i and Y_i. For example, $T = r_{xy}$, the sample correlation coefficient, or $T = r_s$, Spearman's rank correla-

tion. A permutation test based on T rejects the null hypothesis for an observed value t of T if

$$P_{H_0}(T \geq t | X_1, \ldots, X_n, Y_{(1)}, \ldots, Y_{(n)})$$
$$= \frac{\#(\text{permutations } \pi \text{ with } T(X_1, \ldots, X_n, Y_{(\pi_1)}, \ldots, Y_{(\pi_n)}) \geq t)}{n!}$$

is smaller than or equal to α.

In case $n!$ is too large to compute the probability $P_{H_0}(T \geq t | X_1, \ldots, X_n, Y_{(1)}, \ldots, Y_{(n)})$ in practice, this probability can be approximated by the fraction

$$\frac{\#(\text{permutations } \pi \text{ with } T(X_1, \ldots, X_n, Y_{(\pi_1)}, \ldots, Y_{(\pi_n)}) \geq t)}{B}$$

for a large number B of randomly chosen permutations π of $1, \ldots, n$. Like for the permutation tests in the two-sample case, this is nothing else than applying a bootstrap test.

Example 6.7 (Continuation of Example 6.6.) The rank correlation between the variables systolic blood pressure and migration is equal to -0.17. The rank correlation as well as the ordinary correlation is small, so that there seems to be no relationship between the two variables. A bootstrap approximation (with $B = 1000$) for the left p-value of the permutation test based on Spearman's rank correlation coefficient is equal to 0.151. This confirms that there is no significant correlation between systolic blood pressure and years since migration.
$***$

Chapter 7

Analysis of categorical data

In this chapter we discuss the analysis of categorical data that can be summarized in a so-called *contingency table* which was already mentioned in Chapter 1. A contingency table consisting of r rows and c columns generally looks like Table 7.1. In this table a number of n studied objects is divided into categories of two variables. The quantity N_{ij} is the number of objects that are grouped in category i of the *row* variable and in category j of the *column* variable. We say that there are N_{ij} objects in *cell* (i, j), and N_{ij} is called the *cell frequency* of cell (i, j). A dot instead of an index means that the sum is taken over the quantities with all possible values of that index. For example, $N_{\cdot j}$ is the sum of the frequencies in the j-th column, that is: the total number of objects in the j-th category of the column variable. The $N_{i\cdot}$ and $N_{\cdot j}$ are called the *marginal frequencies* or, shortly, *marginals*. The goal is to investigate whether there is a relationship between the row and column variable, and if so, which categories are involved in this relationship.

Example 7.1 In a study on the relationship between blood group and certain diseases a large sample of 8766 people consisting of patients with an ulcer, patients with stomach cancer and a control group of people without these diseases was divided into blood groups O, A or B. The division is given in Table 7.2.
***

7.1 Fisher's exact test for 2×2 tables

The smallest contingency table possible is a 2×2 contingency table. If we assume that all marginals are fixed, the entire table is determined by N_{11}, as shown in Table 7.3. Because of the fixed row marginal $N_{1\cdot}$, the value of N_{12} is found te be $N_{1\cdot} - N_{11}$. Similarly, N_{21} equals $N_{\cdot 1} - N_{11}$, and finally N_{21} equals $n - N_{11} - N_{12} - N_{21}$. Note that the entire table can be determined in this way by either of the 4 entries.

To investigate the relationship between row variable and column variable in a 2×2 contingency table with fixed marginals, the null hypothesis of independence between these

	B_1	$\cdots\cdots\cdots$	B_j	$\cdots$	B_c	total
A_1	N_{11}	$\cdots\cdots\cdots$	N_{1j}	$\cdots$	N_{1c}	$N_{1\cdot}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
A_i	N_{i1}	$\cdots\cdots\cdots$	N_{ij}	$\cdots$	N_{ic}	$N_{i\cdot}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
A_r	N_{r1}	$\cdots\cdots\cdots$	N_{rj}	$\cdots$	N_{rc}	$N_{r\cdot}$
total	$N_{\cdot 1}$	$\cdots\cdots\cdots$	$N_{\cdot j}$	$\cdots$	$N_{\cdot c}$	$n = N_{\cdot\cdot}$

Table 7.1: A $r \times c$ contingency table.

blood group	ulcer	cancer	control	total
O	983	383	2892	4258
A	679	416	2625	3720
B	134	84	570	788
total	1796	883	6087	8766

Table 7.2: 3×3 contingency table of blood group against disease of 8766 people.

	B_1	B_2	total
A_1	k	$N_{1\cdot} - k$	$N_{1\cdot}$
A_2	$N_{\cdot 1} - k$	$N_{2\cdot} - N_{\cdot 1} + k$	$N_{2\cdot}$
total	$N_{\cdot 1}$	$N_{\cdot 2}$	n

Table 7.3: The general 2×2 contingency table with fixed marginals, determined by the value k in cell $(1, 1)$.

study type	exact	arts	total
male	5	3	8
female	11	16	27
total	16	19	35

Table 7.4: 2×2 contingency table of 35 students scored against gender and study type.

two variables is tested. The test statistic is the variable N_{11}, which has under the null hypothesis a hypergeometric distribution with parameters $(n, N_{1\cdot}, N_{\cdot 1})$ with probability mass function

$$P(N_{11} = k) = \frac{\binom{N_{1\cdot}}{k}\binom{N_{2\cdot}}{N_{\cdot 1}-k}}{\binom{n}{N_{\cdot 1}}}.$$

This test is called *Fisher's exact test*, since the distribution of the test statistic under the null hypothes is exact. For testing independence between row and column variables the test is performed two sided.

Example 7.2 In Table 7.4 a 2×2 contingency table is shown, in which 60 students are scored against gender and type of study. The marginals are assumed to be fixed. We test the null hypothesis of no dependence between gender and type of study. The distribution of the test statistic N_{11} under the null hypothesis is hypergeometric with parameters $(35, 8, 16)$, and its value is $k = 5$. The p-value of the two sided test equals 0.42, implying that the null hypothesis of independence cannot be rejected based on these data.
***

7.2 The chi-square test for contingency tables

For contingency tables larger than 2×2 Fisher's exact test does not apply. Instead, an approximate chi-square test can be used. This approximate test can be used for large samples from multinomial distributions. A contingency table leads to a multinomial distribution in a natural way, so that we can base estimation and testing procedures for contingency tables on a statistical model that assumes a multinomial distribution for certain random variables. However, there are different experimental procedures that may lead to Table 7.2, or more generally to Table 7.1. Obviously, a statistical model should reflect the way in which the data are collected. This is why we need to use different models with different assumptions on the parameters for different experimental procedures.

Let us discuss the different possibilities by considering the data of Example 7.1.

A. One single random sample of size $n = 8766$ is classified in two ways: according to blood group and according to disease. In the model that belongs to this situation the rc-vector of all cell frequencies N_{ij} has a rc-nomial (a 9-nomial) distribution with parameters $n, p_{11}, \ldots, p_{rc}$ where

$$\sum_{i=1}^{r}\sum_{j=1}^{c} p_{ij} = 1. \tag{7.1}$$

All marginals are stochastic and satisfy

$$\sum_{i=1}^{r} N_{i\cdot} = \sum_{j=1}^{c} N_{\cdot j} = n. \tag{7.2}$$

In the example: $4258 + 3720 + 788 = 1796 + 883 + 6087 = 8766$.

B. The table can also be the result of r independent, random samples, one per row. In the example we have $r = 3$ samples: one of size 4258 from people with blood group O, one of size 3720 from people with blood group A and one of size 788 from people with blood group B). The objects in the r samples are classified according to the c (also 3 in this example) categories of the column variable (disease). In the model that corresponds to this experimental set-up the contingency table contains data from r independent samples each from a c-nomial distribution. For the i-th sample the c-vector $N_{i1}, \ldots, N_{ic}$ has a c-nomial distribution with parameters $N_{i\cdot}, p_{i1}, \ldots, p_{ic}$, $i = 1, \ldots, r$ where

$$p_{i\cdot} = \sum_{j=1}^{c} p_{ij} = 1, \qquad i = 1, \ldots, r. \tag{7.3}$$

The relation (7.2) still holds, but in this case only the marginals $N_{\cdot 1}, \ldots, N_{\cdot c}$, the column totals in the table, are random variables, whereas the marginals $N_{1\cdot}, \ldots, N_{r\cdot}$, the row marginals in the table, are not. This is because the latter are the fixed sample sizes of the r samples that were determined by the investigator.

C. In this model the role of the variables are interchanged with respect to Model B: choose c independent random (column) samples (in the example one of size 1796 from people with an ulcer, one of size 883 from stomach cancer patients and one of size 6087 from people who do not have an ulcer nor suffer from stomach cancer), and classify the objects according to the row variable (blood group). The data in the table then originate from c independent samples from a r-nomial distribution with parameters $N_{\cdot j}, p_{1j}, \ldots, p_{rj}$, $j = 1, \ldots, c$, for the j-th sample, where

$$p_{\cdot j} = \sum_{i=1}^{r} p_{ij} = 1, \qquad j = 1, \ldots, r. \tag{7.4}$$

In this case the $N_{\cdot j}$ are the fixed sample sizes and thus not random.

Each of the three models for a $r \times c$-contingency table is a set of multinomially distributed random variables. As mentioned above, one is most often interested in investigating whether there is a relationship between the row and column variable. Hypotheses about such relationship can be translated into hypotheses about the cell probabilities p_{ij}. The chi-square test for a contingency table is in a way a generalization of the chi-square test for goodness of fit as discussed in Chapter 3. In Chapter 3 only one k-nomial sample was present, and hence $r = 1$. The theory is the same though. These approximate chi-square tests are based on the following two theorems, which we will not prove.

Theorem 7.1 *Let for $m = 2, 3, \ldots$, the ℓ-vector $N^m = (N_1, \ldots, N_\ell)$ with $\sum_{j=1}^{\ell} N_j = m$ be multinomially distributed with parameters $m, p_1, \ldots, p_\ell$ which satisfy $p_j > 0$ for all j and $\sum_{j=1}^{\ell} p_j = 1$. Then it holds that*

$$\sum_{j=1}^{\ell} \frac{(N_j - mp_j)^2}{mp_j} \xrightarrow{\mathcal{D}} \chi^2_{\ell-1}, \qquad m \to \infty,$$

where χ^2_ν denotes a chi-square distribution with ν degrees of freedom.

Theorem 7.2 *The sum of s independent χ^2_ν distributed random variables has a $\chi^2_{s\nu}$ distribution.*

When the sample sizes are sufficiently large Theorems 7.1 and 7.2 can be used to design approximate chi-square tests for each of the three cases A, B and C. However, in general the probabilities p_{ij} are unknown and need to be estimated. When these parameters are estimated by means of maximum likelihood the number of degrees of freedom of the limit distribution is decreased by the number of estimated parameters.

For Model A the hypothesis of *independence* of the row and column variables is tested; in terms of cell probabilities this becomes:

$$H_{0A} : p_{ij} = p_{i\cdot} p_{\cdot j}, \qquad i = 1, \ldots, r,\ j = 1, \ldots, c. \tag{7.5}$$

Under the null hypothesis H_{0A} the statistic

$$X^2_A = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(N_{ij} - np_{ij})^2}{np_{ij}}, \tag{7.6}$$

with $p_{ij} = p_{i\cdot} p_{\cdot j}$ has approximately a χ^2_{rc-1}-distribution when n is large. The maximum likelihood estimators for the unknown probabilities $p_{i\cdot} p_{\cdot j}$ are equal to

$$\text{under } H_{0A}: \qquad \hat{p}_{i\cdot} \hat{p}_{\cdot j} = \frac{N_{i\cdot}}{n} \frac{N_{\cdot j}}{n}, \qquad i = 1, \ldots, k,\ j = 1, \ldots, r; \tag{7.7}$$

The number of estimated parameters is $(r-1)+(c-1)$. Inserting the estimated probabilities of (7.7) in (7.6) yields the test statistic

$$X_A^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(N_{ij} - n\hat{p}_{ij})^2}{n\hat{p}_{ij}}, \tag{7.8}$$

where $\hat{p}_{ij}$ is defined for $i = 1, \ldots, r$ and $j = 1, \ldots, c$ by

$$\hat{p}_{ij} = \frac{N_{i\cdot}N_{\cdot j}}{n^2}. \tag{7.9}$$

If n is large this test statistic has approximately a chi-square distribution with $(rc-1) - ((r-1)+(c-1)) = (r-1)(c-1)$ degrees of freedom under H_{0A}.

For Model B the hypothesis is tested that the r samples come from r *identical c-nomial distributions*, or:

$$H_{0B} : p_{1j} = p_{2j} = \cdots = p_{rj} \equiv p_j, \qquad j = 1, \ldots, c. \tag{7.10}$$

Under H_{0B} the statistic

$$X_B^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(N_{ij} - N_{i\cdot}p_{ij})^2}{N_{i\cdot}p_{ij}}, \tag{7.11}$$

with $p_{ij} = p_j$, has approximately a $\chi^2_{r(c-1)}$-distribution when $N_{i\cdot}$ is large for $i = 1, \ldots, r$. The maximum likelihood estimators for the unknown probabilities p_j are equal to

$$\text{under } H_{0B}: \qquad \hat{p}_j = \frac{N_{\cdot j}}{n}, \qquad j = 1, \ldots, c; \tag{7.12}$$

The number of estimated parameters is $c-1$. Substituting the estimated $\hat{p}_j$ of (7.12) in (7.11) yields the test statistic

$$X_B^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(N_{ij} - n\hat{p}_{ij})^2}{n\hat{p}_{ij}}, \tag{7.13}$$

where $\hat{p}_{ij}$ is defined as in (7.9). If n is large this test statistic has approximately a chi-square distribution with $r(c-1) - (c-1) = (r-1)(c-1)$ degrees of freedom under H_{0B}.

In model C the roles of row and column variable interchange with respect to model B and one tests the hypothesis that the c samples originate from c *identical r-nomial distributions*, or:

$$H_{0C} : p_{i1} = p_{i2} = \cdots = p_{ic} \equiv p_i, \qquad i = 1, \ldots, r. \tag{7.14}$$

Under H_{0C} the statistic

$$X_C^2 = \sum_{j=1}^{c} \sum_{i=1}^{r} \frac{(N_{ij} - N_{\cdot j} p_{ij})^2}{N_{\cdot j} p_{ij}}, \tag{7.15}$$

with $p_{ij} = p_i$, has approximately a $\chi^2_{c(r-1)}$-distribution when $N_{\cdot j}$ is large for $j = 1, \ldots, c$. The maximum likelihood estimators for the unknown probabilities p_i are equal to

$$\text{under } H_{0C}: \qquad \hat{p}_i = \frac{N_{i\cdot}}{n}, \qquad i = 1, \ldots, r. \tag{7.16}$$

The number of estimated parameters is $r - 1$. Substituting the estimated $\hat{p}_i$ of (7.16) in (7.15) yields the test statistic

$$X_C^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(N_{ij} - n\hat{p}_{ij})^2}{n\hat{p}_{ij}}, \tag{7.17}$$

where $\hat{p}_{ij}$ is defined as in (7.9). If n is large this test statistic has approximately a chi-square distribution with $c(r-1) - (r-1) = (r-1)(c-1)$ degrees of freedom under H_{0C}.

It turns out that the three test statistics in (7.8), (7.13) and (7.17) are identical and all have approximately a chi-square distribution with $(r-1)(c-1)$ degrees of freedom under the corresponding null hypothesis.

Theorem 7.3 *Under the null hypotheses H_{0A}, H_{0B}, H_{0C} in the Models A, B and C, and for n, the row totals, and the column totals, respectively, sufficiently large, the test statistic*

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(N_{ij} - n\hat{p}_{ij})^2}{n\hat{p}_{ij}}, \tag{7.18}$$

with $\hat{p}_{ij}$ defined as in (7.9)*, approximately has a χ^2-distribution with $(r-1)(c-1)$ degrees of freedom.*

Example 7.3 For the data in Example 7.1 it is not known whether they originate from one sample, from three samples from the three blood groups, or from three samples from the different diseases. Let us assume that they come from one sample of size 8766. We then assume Model A, and we wish to test whether the variable blood group and the variable disease are independent. We test the hypothesis H_{0A} against the alternative that $p_{ij} \neq p_{i\cdot} p_{\cdot j}$ for at least one pair (i, j). We use the test statistic X^2. The null hypothesis is rejected for large values of X^2. The value of X^2 is in this case equal to 40.54; the number of degrees of freedom is 4. The (right) p-value is smaller than 0.001, so that for the usual significance levels H_{0A} will be rejected.

Under the Models B and C for testing the null hypothesis H_{0B} and H_{0C}, respectively, the same test statistic would have been used as under Model A. H_{0B} and H_{0C} would have been rejected also.
***

The methods that were discussed in this section rely on a chi-square *limit* distribution. In order for the approximation to be reasonable for a given data set, the following rule of thumb is used. If $\mathrm{E}_{H_0} N_{ij} > 1$ for all (i, j) and at least 80% of the $\mathrm{E}_{H_0} N_{ij} > 5$, then the approximation will be reasonable. Here $\mathrm{E}_{H_0} N_{ij}$, the expected number of observations in cell (i, j) under H_0, equals $np_{i\cdot}p_{\cdot j}$, $N_{i\cdot}p_j$, and $N_{\cdot j}p_i$ for Model A, B and C, respectively. In practice the expectations $\mathrm{E}_{H_0} N_{ij}$ are not known and are to be estimated. For this the probabilities in the three expressions for $\mathrm{E}_{H_0} N_{ij}$ are replaced by their estimators (7.7), (7.12) and (7.16), so that for all models $\mathrm{E}_{H_0} N_{ij}$ is estimated by $\frac{N_{i\cdot}N_{\cdot j}}{n} = n\hat{p}_{ij}$ with $\hat{p}_{ij}$ given by (7.9).

7.3 Identification of cells with extreme values

When, like in the example above, the null hypothesis of independence (Model A) or homogeneity of samples (Models B and C) is rejected, usually further investigation is needed to establish the type of dependence between the variables (Model A) or the type of inhomogeneity between the samples (Models B and C). In other words, one wants to know which categories of the two variables are involved in the relationship between the two variables. For this the cell frequencies in the separate cells can be considered. Cells with extreme values, i.e., extreme under the null hypothesis, need to be detected. A cell with an extremely large value indicates a positive relation between the corresponding two categories of the row and column variable, whereas a cell with an extremely small value indicates that the corresponding categories exclude each other.

A first step in the further analysis is to compare for the different models the cell probabilities as estimated *not* under the null hypothesis, i.e. N_{ij}/n under Model A, $N_{ij}/N_{i\cdot}$ under Model B and $N_{ij}/N_{\cdot j}$ under Model C, with those as estimated under the null hypothesis, i.e. with $\hat{p}_{i\cdot}\hat{p}_{\cdot j}$ under H_{0A}, $\hat{p}_j$ under H_{0B} and $\hat{p}_i$ under H_{0C}, as given by (7.7), (7.12) and (7.16). This comparison gives an indication of a possible deviance from the null hypothesis and of the size of this deviance.

7.3.1 Determining outliers based on the residuals

As a second step in the further analysis, the cell frequencies N_{ij} are compared with their estimated expectations $n\hat{p}_{ij}$. One often studies the residuals $N_{ij} - n\hat{p}_{ij}$. However, the sizes of these residuals do not mean much, because a large value of $N_{ij} - n\hat{p}_{ij}$ generally goes hand in hand with a large variance of N_{ij}. Therefore the residuals need to be normalized.

One way to normalize the residuals is already used in the chi-square test statistic in (7.8). A table of the square roots of the so-called *contributions* of this test statistic (often also called contributions), defined by

$$C_{ij} = \frac{N_{ij} - n\hat{p}_{ij}}{\sqrt{n\hat{p}_{ij}}}, \qquad i = 1, \ldots, r,\ j = 1, \ldots, c \tag{7.19}$$

gives insight into which cells contribute relatively much to the value of the test statistic. For cells with a considerably large contribution it should be checked whether the observation in the cell is correct. Note that an incorrect observation not only leads to an incorrect value in the corresponding cell, but also in the other cells in the same row and/or column: the contributions are dependent. This is why it is not easy to decide which cells are extreme solely on a table with contributions.

Often other normalizations of the residuals are used. We mention

$$V_{ij} = \frac{N_{ij} - n\hat{p}_{ij}}{\sqrt{\dfrac{N_{i\cdot}(n - N_{i\cdot})N_{\cdot j}(n - N_{\cdot j})}{n^2(n-1)}}}, \tag{7.20}$$

$$U_{ij} = \frac{N_{ij} - n\hat{p}_{ij}}{\sqrt{n\hat{p}_{ij}(1 - \dfrac{N_{i\cdot}}{n})(1 - \dfrac{N_{\cdot j}}{n})}} = \sqrt{\frac{n}{n-1}}V_{ij}, \tag{7.21}$$

and

$$\tilde{V}_{ij} = \frac{1}{g_{ij}}V_{ij}, \quad \text{with} \quad g_{ij} = 1 + e^{-n\hat{p}_{ij}}, \tag{7.22}$$

for $i = 1, \ldots, r,\ j = 1, \ldots, c$. Since given the row and column totals, N_{ij} has under the three models and under the corresponding null hypotheses (exactly) a hypergeometric distribution with expectation $n\hat{p}_{ij}$ and variance $N_{i\cdot}(n - N_{i\cdot})N_{\cdot j}(n - N_{\cdot j})/(n^2(n-1))$, the statistic V_{ij} has conditionally on the row and column totals under the null hypothesis expectation 0 and variance 1 under all three models. Under the three models, for large n V_{ij} is approximately $N(0,1)$ distributed under the null hypothesis, so that p-values for the V_{ij}, based on the standard normal distribution, preferably computed with a continuity correction, give an impression of the extremeness of each of the cell frequencies separately: small and large p-values indicate that the value in the corresponding cell is an outlier under the null hypothesis. For large n U_{ij} almost does not differ from V_{ij}, and each U_{ij} is for large n also approximately $N(0,1)$ distributed under the null hypothesis. The same holds for $\tilde{V}_{ij}$. The additional factor $1/g_{ij}$ in (7.22) serves as a correction for the fact that V_{ij} is not exactly standard normally distributed. The p-values of $\tilde{V}_{ij}$ based on the $N(0,1)$-distribution in some cases give a better approximation of the exact p-values per cell, than those of the other statistics.

Example 7.4 Let us consider the data of Example 7.1 again. Because for these data the null hypothesis was rejected (for all models), the next step is to investigate the nature of the relationship, i.e. of the dependence in Model A or of the inhomogeneity of the samples in Models B and C. Below a couple of tables are presented that give insight in this.

First we compare for the different models the cell probabilities as estimated under the respective null hypotheses (Tables 7.6–7.8), with those as estimated not under the null hypothesis (Table 7.5). We see that there are differences between the two types of estimates, but that they are not very large. In particular, the Tables 7.5 and 7.6 are almost identical. This illustrates the fact that it makes sense to apply a chi-square test: based on the comparison of the cell probabilities estimated under and not under the null hypothesis it is difficult to conclude that the null hypothesis does not hold.

On the other hand, the Tables 7.5 and 7.6 are not irrelevant, for they indicate how big the deviance from independence is. The difference between the estimated probabilities in the two tables is about 0.01. At first sight this is quite small. Whether or not this is of practical relevance, depends on the context and goal of the analysis. This cannot be inferred from the fact that the deviance from independence or homogeneity is *statistically significant*. The statistical significance only says that the observed relationship (dependence/inhomogeneity) is not due to an artifact of taking a sample (which is large in size in this example). It does not mean that the relationship is strong, i.e. that the dependence or inhomogeneity is large, but only that the relationship almost surely exists.

Next, we investigate the nature of the relationship by studying the residuals (Table 7.9). The residuals in the cells (1,1), (1,3) and (2,1) seem to be relatively large. However, in Tables 7.9 and 7.10 it is illustrated that it is good to *normalize* the residuals: the residuals: 41.28 and 41.88 yield rather different contributions. Vice versa, the contributions -2.16 and -2.22 originate from two very different residuals. From Table 7.10 with the contributions we see that cell (1,1) and cell (2,1) do make a large contribution to the chi-square test statistic indeed, but that the contribution of cell (1,3) is not very large after all. Also in Table 7.11 with the normalized residuals U_{ij} the cells (1,1) and (2,1) stand out. We therefore may conclude that there seems to be a positive relationship between having blood group O and getting an ulcer, and a negative relationship between getting an ulcer and having blood group A. In this example the values of U_{ij}, V_{ij} and $\tilde{V}_{ij}$ are the same, due to the fact that the sample size is very large.
***

7.3.2 Determining outliers based on the empirical distribution

In a large table, that is when rc is large, there are almost always some outliers. A simple alternative for studying tables of standardized residuals in that case is to use the

blood group	ulcer	cancer	control	total
O	0.10	0.05	0.34	0.49
A	0.09	0.04	0.29	0.42
B	0.02	0.01	0.06	0.09
total	0.20	0.10	0.69	

Table 7.5: Table of estimated cell probabilities. The inner part of the table gives the values $\hat{p}_{i\cdot}\hat{p}_{\cdot j}$ under H_{0A}; the row totals are the values $\hat{p}_i$ under H_{0C}; the column totals are the values $\hat{p}_j$ under H_{0B}.

blood group	ulcer	cancer	control
O	0.11	0.04	0.33
A	0.08	0.05	0.30
B	0.02	0.01	0.07

Table 7.6: Table of estimated cell probabilities p_{ij} under Model A, not under the null hypothesis; the table gives the values N_{ij}/n.

blood group	ulcer	cancer	control
O	0.23	0.09	0.68
A	0.18	0.11	0.71
B	0.17	0.11	0.72

Table 7.7: Table of estimated cell probabilities p_j under Model B, not under the null hypothesis; the table gives the values $N_{ij}/N_{i\cdot}$.

blood group	ulcer	cancer	control
O	0.55	0.43	0.48
A	0.38	0.47	0.43
B	0.07	0.10	0.09

Table 7.8: Table of estimated cell probabilities p_i under Model C, not under the null hypothesis; the table gives the values $N_{ij}/N_{\cdot j}$.

blood group	ulcer	cancer	control
O	110.61	-45.91	-64.70
A	-83.16	41.28	41.88
B	-27.45	4.62	22.82

Table 7.9: Table of the residuals $N_{ij} - n\hat{p}_{ij}$.

blood group	ulcer	cancer	control
O	3.74	-2.22	-1.19
A	-3.01	2.13	0.82
B	-2.16	0.52	0.98

Table 7.10: Table of the contributions C_{ij}.

blood group	ulcer	cancer	control
O	5.86	-3.26	-3.00
A	-4.45	2.96	1.96
B	-2.54	0.57	1.86

Table 7.11: Table of the standardized residuals V_{ij}; U_{ij} and $\tilde{V}_{ij}$ are the same as V_{ij} in this case.

empirical distribution of these standardized residuals. Values that lie more than 1.5 times the interquartile range below the first quartile or above the third quartile, are considered as extreme values. This procedure can be executed for each of the above defined normalized residuals. Note that this method gives the same outliers as the (standard) boxplot. For small tables this approach generally does not make much sense.

Example 7.5 For the data of Example 7.1 it turns out that the C_{ij} and the U_{ij} (and hence the V_{ij} and $\tilde{V}_{ij}$) do not have any outliers based on their empirical distribution, whereas in fact there are a couple of large contributions. In view of the fact that the chi-square test rejects the null hypothesis, this illustrates the limitations of this method for smaller tables.
*****

7.4 The bootstrap for contingency tables

Analogously to earlier described bootstrap procedures, the exact distribution under the different null hypotheses of the statistics defined in the foregoing sections can be estimated by means of a re-sampling procedure. In this procedure, which is executed conditionally on the marginals of the contingency table of the data, B bootstrap contingency tables are generated with the same row and column totals as those of the table of the data and under the null hypothesis. For each of the B bootstrap tables the value of the statistic of interest can then be computed, which gives B bootstrap values of this statistic. The empirical distribution of the B bootstrap values of the statistic are an estimator of the conditional distribution of the statistic under the null hypothesis. In particular,

comparison of the value of the chi-square test statistic for the original data with the exact conditional distribution of the chi-square test statistic can replace the chi-square test described in Section 7.1. This is especially useful in situations where n is small, for then the chi-square test is not applicable. When n is large, the chi-square test should be preferred, because the bootstrap procedure takes a lot of computation time in that case.

The bootstrap is, however, more generally applicable than just for determining a bootstrap estimate of the exact, conditional, distribution of the chi-square test statistic. With the bootstrap also estimates of the conditional distribution of any other function of the cell frequencies can be derived. Hence, the bootstrap can be employed for finding cells with extreme values too. For instance, estimation of the conditional distribution—under one of the three null hypotheses— of the largest, one but largest, ect. cell frequency, allows us to compare the largest, one but largest, ect. cell frequency in the contingency table of the data with the corresponding estimated distribution. If, for example, the largest value in the table lies in the tail of the estimated distribution, then this value is extreme under the null hypothesis.

We now describe the general procedure for deriving a bootstrap estimate of the distribution of a function of the cell frequencies under one of the null hypotheses and given the row and column totals. Because the distribution of N_{ij} *given the marginals* is the same under all three hypotheses, namely hypergeometric with expectation $n\hat{p}_{ij}$ and variance $N_{i\cdot}(n - N_{i\cdot})N_{\cdot j}(n - N_{\cdot j})/(n^2(n-1))$, it does not matter whether we work under Model A, B or C. Let $N^n = (N_{11}, \ldots, N_{1c}, N_{21}, \ldots, N_{2c}, \ldots, N_{r1}, \ldots, N_{rc})^T$, and let F_0 be the conditional distribution of N^n under the null hypothesis given the marginals. Let the function of the cell frequencies $T = T(N^n)$ be the statistic of interest. The bootstrap procedure for estimation of its distribution is as follows:

- Generate B new rc-vectors (or B new $r \times c$-contingency tables) by simulating B random samples of size 1 from F_0. Denote these bootstrap vectors by $N_1^{n*}, \ldots, N_B^{n*}$.
- Compute the B bootstrap values $T_{n,1}^* = T(N_1^{n*}), \ldots, T_{n,B}^* = T(N_B^{n*})$.
- The empirical distribution of the B bootstrap values $T_{n,1}^*, \ldots, T_{n,B}^*$ is a bootstrap estimate of the conditional distribution of T under the null hypothesis given the marginals of the original data set.

The value t of $T(N^n)$ for the original data set can now be compared with the bootstrap estimate of the distribution of $T(N^n)$. We remark that, although this does not show in the notation, the bootstrap values $T_{n,1}^*, \ldots, T_{n,B}^*$ of course do not only depend on n, but also on the row and column totals. Hoaglin, Mosteller and Tukey (1985) describe how to generate the samples from F_0. Unfortunately, this procedure takes a long computation time even for moderately large n.

	illiterate	literate	total
urban	6	30	36
rural	36	28	64
total	42	58	100

Table 7.12: Table of the number of illiterate people in different areas

Example 7.6 Table 7.12 gives the results of a study into illiteracy in urban and rural areas. As expected, the row and column variables are not independent: we find a value of 14.82 for the chi-square statistic, which has for one degree of freedom a very small p-value. For these data 500 bootstrap 2×2-contingency tables with the same marginals as Table 7.12 were simulated under the null hypothesis of independence. For each of the tables the smallest and largest value in the table was determined. The empirical distributions of these values are bootstrap estimates for the conditional distributions, given the marginals, under the null hypothesis of the minimum and the maximum in a table. Based on these bootstrap estimates the value 6 in the table turns out to be an outlier, (6 was the 0.002-quantile of the empirical distribution of the bootstrapped smallest values), but the value 36 not (36 was the 0.41-quantile of the empirical distribution of the bootstrapped largest values). We may therefore conclude that in particular there are less illiterate people in the urban areas then would be the case under independence of the variables 'area' and 'literacy'.
$***$

Chapter 8

Linear regression analysis

Regression analysis in its simplest form consists of fitting models for one continuous *response* to experimental data. In the models it is assumed that the value of the response depends on the values of a number of other variables, the *independent* or *structural* variables. In the context of a regression model, these variables often are called the *explanatory* variables. The response is assumed to be observed with a measurement error; the corresponding values of the explanatory variables are assumed to be known exactly. The relationship between the two types of variables depends on the unknown values of a set of parameters. These unknown parameter values need to be estimated. Often the estimation is done by the so-called *least squares method*, but other methods are also available. The least squares method tries to find those values of the parameters that minimize the sum of the quadratic deviances of the observations from their expected values under the model. In the case of *linear regression* the expected response depends in a linear way on the vector of unknown parameters. Generally the response is assumed to be normally distributed. Two more general classes of regression models are the class of *nonlinear regression models* in which the expected response depends in a nonlinear way on the parameters, and the class of *generalized linear models* in which the expected response depends on a function of a linear model and the response can have a different distribution than the normal one. In this chapter we restrict ourselves to univariate linear regression models in which the one response variable can depend on more than one explanatory variable, the so-called *multiple* linear regression models; see, for instance, Seber and Lee (2003) or Weisberg (2005). If there is more than one response variable, one speaks of *multivariate* linear regression, see Mardia et al. (1980) for an introduction. We shall not address this case. For nonlinear models we refer to Seber and Wild (2003) or Bates and Watts (2007), and for generalized linear models to McCullagh and Nelder (1989) or Dobson and Barnett (2008).

8.1 The multiple linear regression model

The multiple linear regression model for n observations and p explanatory variables, $p < n$, is formulated as follows:

for $i, k = 1, \ldots, n$,

$$
\begin{aligned}
&Y_i = \beta_0 + x_{i1}\beta_1 + \cdots + x_{ip}\beta_p + e_i, \\
&\mathrm{E}\, e_i = 0, \\
&\mathrm{E}\, e_i e_k = \begin{cases} \sigma^2, & i = k, \\ 0, & i \neq k, \end{cases}
\end{aligned}
\tag{8.1}
$$

where Y_i is the i-th observation of the response, x_{ij} the (known) value of the j-th explanatory variable for this observation, $\beta_0, \beta_1, \ldots, \beta_p$, and σ^2 are unknown constants, and e_i is the unknown stochastic measurement error in the i-th observation. We see that in this model the Y_i are uncorrelated random variables with $\mathrm{E}\, Y_i = \beta_0 + x_{i1}\beta_1 + \cdots + x_{ip}\beta_p$, and $\operatorname{Var} Y_i = \sigma^2$. The constant β_0 is called the *intercept*.

It is practical to use the matrix notation in this situation. Model (8.1) then becomes

$$
\begin{aligned}
&Y = X\beta + e, \\
&\mathrm{E}\, e = 0, \\
&\operatorname{Cov}(e) = \sigma^2 I_{n\times n}.
\end{aligned}
\tag{8.2}
$$

Here $Y = (Y_1, \ldots, Y_n)^T$ is the stochastic vector of observations with $\mathrm{E}\, Y = X\beta$ and $\operatorname{Cov}(Y) = \operatorname{Cov}(e) = \sigma^2 I_{n\times n}$; X the $n \times (p+1)$-matrix with i-th row $x_i^T = (1, x_{i1}, \ldots, x_{ip})$; $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^T$ the unknown parameter vector; $e = (e_1, \ldots, e_n)^T$ the stochastic vector of measurement errors, and $I_{n\times n}$ the $n \times n$-identity matrix. The matrix X is called the *design matrix*. We assume that X has maximal rank, $\operatorname{rank}(X) = p+1$. For simplicity the first column of X, which consists of only ones, is often called the 0-th column of X and we denote it by 1; the other columns are denoted as $X_1, X_2, \ldots, X_p$, so that the coefficient β_j belongs to the explanatory variable X_j. According to this convention, in the following we shall call the column that is associated with β_j, the j-th column–even though in fact this is the $(j+1)$-th column in the design matrix.

A third way to describe the model is in terms of the columns of X:

$$
\begin{aligned}
&Y = \beta_0 1 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + e, \\
&\mathrm{E}\, e = 0, \\
&\operatorname{Cov}(e) = \sigma^2 I_{n\times n}.
\end{aligned}
\tag{8.3}
$$

The first goal is to estimate β and σ^2. Next the quality of the model with the estimated parameter values needs to be investigated. In the sequel the "i-th observation point" means the i-th response y_i combined with the corresponding values $x_{i1}, \ldots, x_{ip}$ of the explanatory variables. Unless stated otherwise, we assume from now on that $e_1, \ldots, e_n$ are independent and normally distributed.

8.1.1 Parameter estimation

One way to estimate β is with the so-called least squares method. The least squares estimator $\hat{\beta}$ of the parameter vector β minimizes the sum of squares

$$S(\beta) = (Y - X\beta)^T(Y - X\beta). \tag{8.4}$$

It can be obtained in the usual way, by differentiation of (8.4) with respect to the different β_j and setting the result equal to 0. Doing this, we see that $\hat{\beta}$ satisfies $X^T(Y - X\hat{\beta}) = 0$, or $X^TX\hat{\beta} = X^TY$, and we find

$$\hat{\beta} = (X^TX)^{-1}X^TY. \tag{8.5}$$

It follows that $\mathrm{E}\,\hat{\beta} = \beta$, so that $\hat{\beta}$ is an unbiased estimator of β. Furthermore, the covariance matrix of $\hat{\beta}$ is $\mathrm{Cov}(\hat{\beta}) = \sigma^2(X^TX)^{-1}$.

Once $\hat{\beta}$ is computed, several other useful quantities can easily be derived. The vector of *predicted* responses $\hat{Y} = X\hat{\beta}$, has i-th element $\hat{Y}_i = x_i^T\hat{\beta}$. The vector of *residuals* $R_Y(X)$, or shortly R, is defined by

$$R_Y(X) = R = Y - \hat{Y},$$

and has i-th element

$$R_i = Y_i - \hat{Y}_i = Y_i - x_i^T\hat{\beta}.$$

The function S of (8.4) at the point $\hat{\beta}$ is called the *residual sum of squares* RSS:

$$RSS = S(\hat{\beta}) = (Y - X\hat{\beta})^T(Y - X\hat{\beta}) = (Y - \hat{Y})^T(Y - \hat{Y}) = R^TR,$$

or

$$RSS = \sum_{i=1}^{n} R_i^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2, \tag{8.6}$$

hence its name.

The estimator

$$\hat{\sigma}^2 = \frac{RSS}{n - p - 1} \tag{8.7}$$

is used to estimate σ^2, so that the covariance matrix of $\hat{\beta}$ can be estimated by $\widehat{\mathrm{Cov}}(\hat{\beta}) = \hat{\sigma}^2(X^TX)^{-1}$. This means that the j-th value on the diagonal of this matrix is an estimate of the variance of the estimator $\hat{\beta}_j$ and thus gives an indication of the accuracy of the resulting estimates. Finally, we note that under the assumption that the vector of measurement errors e is normally distributed, the quantity $(n-p-1)\sigma^{-2}\hat{\sigma}^2 = RSS/\sigma^2$ has a chi-squared distribution with $(n - p - 1)$ degrees of freedom, so that $\mathrm{E}\,\hat{\sigma}^2 = \sigma^2$ and $\hat{\sigma}^2$ is an unbiased estimator of σ^2.

diameter	8.3	8.6	8.8	10.5	10.7	10.8	11.0	11.0	11.1	11.2	11.3
volume	10.3	10.3	10.2	16.4	18.8	19.7	15.6	18.2	22.6	19.9	24.2
length	70.0	65.0	63.0	72.0	81.0	83.0	66.0	75.0	80.0	75.0	79.0
diameter	11.4	11.4	11.7	12.0	12.9	12.9	13.3	13.7	13.8	14.0	
volume	21.0	21.4	21.3	19.1	22.2	33.8	27.4	25.7	24.9	34.5	
length	76.0	76.0	69.0	75.0	74.0	85.0	86.0	71.0	64.0	78.0	
diameter	14.2	14.5	16.0	16.3	17.3	17.5	17.9	18.0	18.0	20.6	
volume	31.7	36.3	38.3	42.6	55.4	55.7	58.3	51.5	51.0	77.0	
length	80.0	74.0	72.0	77.0	81.0	82.0	80.0	80.0	80.0	87.0	

Table 8.1: Tree data.

Example 8.1 In Example 3.2 univariate regression was performed in which the response variable 'volume' was regressed on the explanatory variable 'diameter' for data of 31 cut cherry trees. These data were collected to investigate to which extent the timber yield (volume) of a tree can be predicted by means of the diameter at the height of 4 feet and 6 inches of the uncut tree. The question is whether the diameter on its own is sufficient to predict the volume. Therefore, besides the volume (in cubic feet) and the diameter (in inches), also the length (in feet) of the 31 trees was measured, see Table 8.1.

We now assume the model (8.1) with $p = 2$, where Y_i, x_{i1} and x_{i2} represent the volume, the diameter and the length of the i-th tree. With the least squares method we find

$$\hat{\beta} = (-57.99, 4.71, 0.34)^T;\ RSS = 421.92;\ \hat{\sigma}^2 = 15.07;$$

$$\widehat{\mathrm{Cov}}(\hat{\beta}) = \begin{pmatrix} 74.62 & 0.43 & -1.05 \\ 0.43 & 0.07 & -0.02 \\ -1.05 & -0.02 & 0.02 \end{pmatrix}.$$

The corresponding values for the univariate regression on diameter only were

$$\hat{\beta} = (-36.94, 5.07)^T;\ RSS = 524.30;\ \hat{\sigma}^2 = 18.08;$$

$$\widehat{\mathrm{Cov}}(\hat{\beta}) = \begin{pmatrix} 11.32 & -0.81 \\ -0.81 & 0.06 \end{pmatrix}.$$

We see that addition of the variable 'length' to the model results in a smaller residual sum of squares, but that the value of the intercept is estimated less accurately for the larger model. We will come back to this later.

Before we use the results of the regression on diameter and length in a further analysis, like in the case of the univariate regression, we should check the normality of the measurement errors. This can once again be done by means of a normal QQ-plot of the residuals. This plot, see Figure 8.1, gives us no reason to doubt the normality assumption.
***

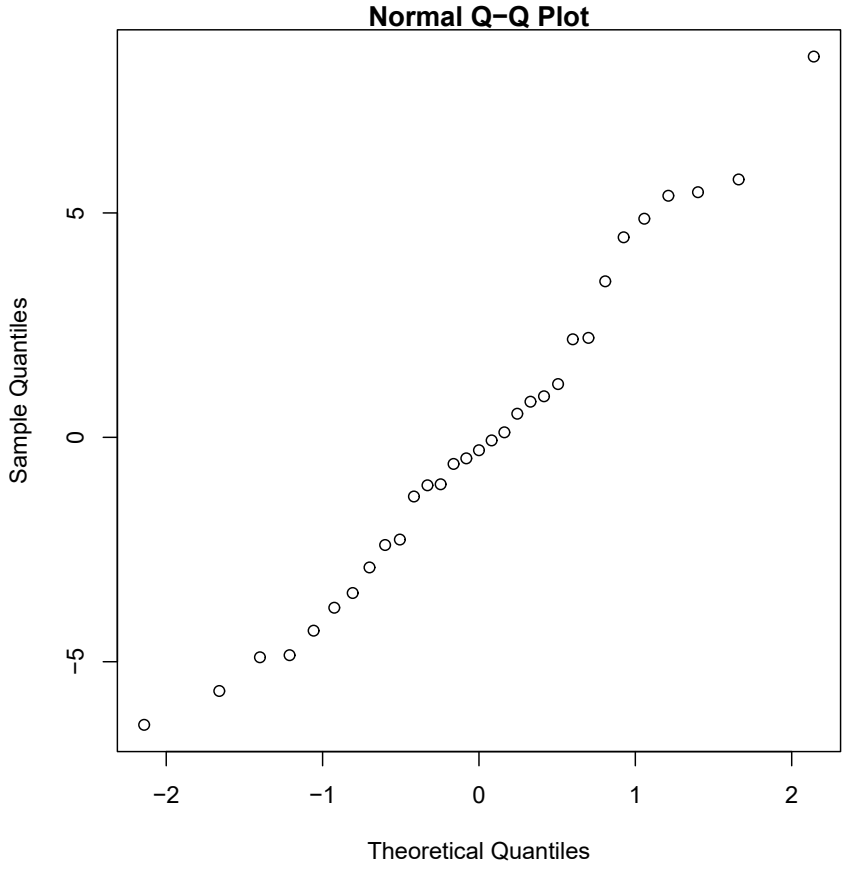


Figure 8.1: Normal QQ-plot of the residuals of the bi-variate linear regression of volume on diameter and length for data of 31 cherry trees.

We remark that under the assumption of normality of the measurement errors the least squares estimates are the same as the maximum likelihood estimates. In case it is unlikely that the measurement errors are normally distributed, for example if a Poisson or gamma distribution or a symmetric distribution with heavier tails seems more appropriate, then it is recommended to use maximum likelihood or a robust method to estimate β, instead of the least squares method.

8.1.2 Selection of explanatory variables

Choosing the explanatory variables is not an easy task. The goal is to obtain the best possible model with the smallest possible number of explanatory variables. These two criteria potentially contradict each other, especially if we think of the 'best possible model' as the one with the best fit to the data. One way to perform the model selection is by building up the model step by step, by adding explanatory variables one by one according to a prescribed rule. Another way is to start with the largest possible model that contains all available variables, and reducing this model step by step by deleting variables one by one according to a prescribed rule. Addition or deletion are stopped when the two above mentioned criteria are in balance. Although these two techniques have the advantage to be sort of systematic, the criteria for judging whether or not a variable should be added (deleted), generally depend on which variables are already in the model (still in the model). A thorough statistical analysis should therefore consider all possible models, with all combinations of explanatory variables, and compare them to each other. In several statistical packages this is standard practice. The statistician's task is to choose proper selection criteria. Below we will discuss several techniques to compare two different linear regression models with each other in a model selection procedure.

8.1.2.1 The determination coefficient

The most frequently reported measure that gives information about the fit of a linear regression model is the *coefficient of determination* $\mathcal{R}^2$, which in fact compares the residual sum of squares of the model under consideration with the residual sum of squares of the model without any explanatory variables and with only the intercept β_0. Let us consider the linear regression model with p explanatory variables

$$Y = \beta_0 1 + \beta_1 X_1 + \cdots + \beta_p X_p + e \tag{8.8}$$

with residual sum of squares RSS given by (8.6), and compare it with the model without explanatory variables,

$$Y = \beta_0 1 + e. \tag{8.9}$$

The residual sum of squares for this special model without explanatory variables is denoted by SSY. Because for this model the design matrix reduces to the n-vector containing only ones, $\hat{\beta}_0 = \bar{Y}$, and $\hat{Y} = X\hat{\beta} = \hat{\beta}_0 1 = \bar{Y}1$. This means that $\hat{Y}_i = \bar{Y}$ for every i, and the residual sum SSY, for this special case thus turns into

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

which is called the *sum of squares for* Y or the *total sum of squares*. The difference

$$SS_{reg} = SSY - RSS$$

corresponds to the part of the sum of squares for Y that is explained by the larger model (8.8), but not by the small model (8.9). Therefore, the size of SS_{reg} is an indicator of the usefulness of regression of Y on the X-variables. The coefficient of determination $\mathcal{R}^2$ is a scaled version of this difference:

$$\mathcal{R}^2 = \frac{SS_{reg}}{SSY} = \frac{SSY - RSS}{SSY} = 1 - \frac{RSS}{SSY} \tag{8.10}$$

It represents the fraction of the variability in Y which is explained by regression of Y on X. The quantity $\mathcal{R}^2$ is also called the *fraction of explained variance*. Its magnitude thus is a measure of the overall fit of the model (8.8). For linear regression models with an intercept, it holds that $0 \leq \mathcal{R}^2 \leq 1$. The closer $\mathcal{R}^2$ is near 1, the better the model. In general, addition of a new variable to a model will make $\mathcal{R}^2$ larger.

For simple linear regression, where there is only one explanatory variable, the coefficient of determination is equal to the square of the correlation coefficient between Y and X_1. In the case of multiple regression it is equal to the square of the so-called *multiple correlation coefficient* of Y and the X-variables. The multiple correlation coefficient is the largest correlation in absolute value between Y and any linear combination of the X-variables, which is by definition equal to $\rho(Y, \hat{Y})$, the correlation coefficient between Y and its expectation under the considered regression model. Although the determination coefficient is very often used, its importance should not be overrated. It only gives a *global* indication of the model fit.

8.1.2.2 F-tests

A different scaling of SS_{reg} than the one used for the definition of the determination coefficient in (8.10), leads to a frequently used test statistic. Indeed, if SS_{reg} exceeds some critical level, we will conclude that exploiting our knowledge of the values of the X's yields a better model than if we would ignore them. More precisely, define the F-statistic

$$\mathcal{F} = \frac{SS_{reg}/(\sigma^2 p)}{RSS/(\sigma^2(n-p-1))} = \frac{(n-p-1)(SSY - RSS)}{pRSS}. \tag{8.11}$$

If $e_1, \ldots, e_n$ are independent and normally distributed, then under the model (8.8) $\mathcal{F}$ is the ratio of two independent chi-square distributed random variables divided by their numbers of degrees of freedom p and $(n-p-1)$, respectively, and $\mathcal{F}$ therefore has an F-distribution with p and $(n-p-1)$ degrees of freedom. Hence, for the testing problem under the model (8.8):

H_0: $\beta_1 = \cdots = \beta_p = 0$,

H_1: $\beta_j \neq 0$ for some j, $1 \leq j \leq p$,

the following F-test may be used: reject H_0 if $\mathcal{F} \geq F_{p,(n-p-1);1-\alpha}$. Here $F_{\mu,\nu;1-\alpha}$ is the $(1-\alpha)$-quantile of the F-distribution with μ and ν degrees of freedom, and α is the level of the test.

Example 8.2 For the multiple regression with two variables of Example 8.1 we find $SSY = 8106.08$, so that $SS_{reg} = 7684.16$, $\mathcal{R}^2 = 0.95$ and $\mathcal{F} = 254.97$. We see from the value of $\mathcal{R}^2$ that a large fraction of the variability in volume is explained by regression on diameter and length. The given $\mathcal{F}$-value relates to the following testing problem:

H_0: $\beta_1 = \beta_2 = 0$,

H_1: $\beta_j \neq 0$ for some j, $1 \leq j \leq 2$.

Because the right p-value of 254.97 for the $F_{2,28}$-distribution is equal to 0, H_0 is rejected for all reasonable significance values. Later on we will see that diameter is a good predictor of volume, whereas length is less useful in this respect.
***

It is not always relevant to compute the *overall* test statistic $\mathcal{F}$: often it is already known that at least some of the variables are correlated with the response variable, so that a large value for $\mathcal{F}$ is to be expected. Instead of testing whether all variables together are of importance, it is then more interesting to investigate which part of the available set of variables should be selected for inclusion in the model. For this purpose another F-test can be used. If one is interested to know whether besides the explanatory variables $X_1, \ldots, X_p$ also $X_{p+1}, \ldots, X_q$ $(q > p)$ should be included in the model, one considers the testing problem

H_0: $\beta_{p+1} = \cdots = \beta_q = 0$; $\beta_0, \beta_1, \ldots, \beta_p$ arbitrary,

H_1: $\beta_j \neq 0$ for some j, $p+1 \leq j \leq q$; $\beta_0, \beta_1, \ldots, \beta_p$ arbitrary.

The procedure is as follows. Fit the model with $X_1, \ldots, X_p$, but without $X_{p+1}, \ldots, X_q$ and determine the residual sum of squares RSS_p for this model. Next, fit the larger model with $X_1, \ldots, X_p$ ànd $X_{p+1}, \ldots, X_q$ and determine the residual sum of squares RSS_q of this larger model. The difference $RSS_p - RSS_q$ of these sums of squares can be written as $RSS_p - RSS_q = (SSY - RSS_q) - (SSY - RSS_p) = SS_{reg,q} - SS_{reg,p}$ and thus can be seen to correspond with the fraction of the sum of squares for Y that is explained by the larger set $X_1, \ldots, X_q$, but not by the smaller subset $X_1, \ldots, X_p$ alone. Define

$$\mathcal{F}^{p,q} = \frac{(RSS_p - RSS_q)/(\sigma^2(q-p))}{RSS_q/(\sigma^2(n-q-1))} = \frac{(n-q-1)(RSS_p - RSS_q)}{(q-p)RSS_q}. \tag{8.12}$$

If $e_1, \ldots, e_n$ are independent and normally distributed, then under H_0 also $\mathcal{F}^{p,q}$ is the ratio of two independent chi-squared distributed random variables divided by their numbers of

degrees of freedom, $(q-p)$ and $(n-q-1)$. Therefore $\mathcal{F}^{p,q}$ has an F-distribution with $(q-p)$ and $(n-q-1)$ degrees of freedom and a test with significance level α becomes: reject H_0 if $\mathcal{F}^{p,q} \geq F_{(q-p),(n-q-1);1-\alpha}$. This test is a *partial F-test*.

8.1.2.3 *t*-tests

Investigating whether or not to include one single variable, X_k say $(0 \leq k \leq p)$, is generally done by means of a *t-test*, rather than by using a partial F-test. In this case the hypotheses are:

H_0: $\beta_k = 0$; β_j arbitrary for $0 \leq j \leq p,\ j \neq k$,

H_1: $\beta_k \neq 0$; β_j arbitrary for $0 \leq j \leq p,\ j \neq k$.

If the model (8.8) with p explanatory variables $X_1, \ldots, X_p$ is fitted, then—again under the assumption of independence and normality of the measurement errors—the statistic T_k obtained by dividing $\hat{\beta}_k$ by its estimated standard deviation, under H_0 has a t-distribution with $(n-p-1)$ degrees of freedom. Hence, for significance level α H_0 is rejected if

$$|T_k| = \frac{|\hat{\beta}_k|}{\sqrt{\widehat{\mathrm{Cov}}(\hat{\beta})_{kk}}} \geq t_{(n-p-1);1-\alpha/2}, \tag{8.13}$$

where $t_{\nu;1-\alpha/2}$ is the $(1-\alpha/2)$-quantile of the t-distribution with ν degrees of freedom and $\widehat{\mathrm{Cov}}(\hat{\beta})_{kk}$ the k-th diagonal element of the matrix $\widehat{\mathrm{Cov}}(\hat{\beta})$. In most statistical computer packages for regression the vector of t-values corresponding to the estimates $\hat{\beta}_k$ belong to the standard output.

We note that the testing problem of this t-test is the same as that of the partial F-test for the case where the question is whether X_k should be included next to the other $(p-1)$ explanatory variables. For this situation $\mathcal{F}^{p-1,p} = T_k^2$. Under the relevant null hypothesis $\mathcal{F}^{p-1,p}$ has an F-distribution with 1 and $(n-p-1)$ degrees of freedom and the two tests are equivalent. This means that the tests in this case yield the same p-value, and, hence, the same conclusion for the same significance level. It is important to keep in mind that both the F- and t-tests should only be used if the independence and normality assumptions for the measurement errors are plausible.

Unfortunately, the partial F-tests and t-tests have the undesirable property that the conclusion about whether or not one or more new variables should be included in the model (that is when these tests are used in a model selection procedure), may depend on which variables were already present in the model. For instance, if a series of tests is performed such that each time one single variable is added if it is tested significant, the final model may depend on the order in which the available variables were tested.

It also often happens that based on a t- or F-test it is concluded that a certain variable should be added to a model whereas the determination coefficient of the new model with

this variable is barely larger than that of the old model without the variable. In that case, addition of the variable hardly improves the fit. Since simpler models are easier to work with, a general guideline for the choice of a model is: choose the model with the largest determination coefficient, but only if the difference of its determination coefficient with the smaller model is relevant. In addition, one should always keep in mind that the modeling problem concerns a concrete, practical situation: it should always be questioned whether addition or deletion of variables makes sense in the context of the situation that the model intends to describe.

8.1.2.4 The partial correlation coefficient

Finally, we introduce the *partial correlation coefficient* as an indicator of linear relationship. As we have seen before, the correlation coefficient $\rho(X_1, Y)$ is a global measure for the linear relationship between X_1 and Y. As such, it gives an indication of the usefulness of inclusion of X_1 as the explanatory variable in a simple linear regression model with response variable Y. Analogously, if p explanatory variables are available, there exists a measure for the linear relationship between one of them, X_k say, and Y corrected for the $(p-1)$ other variables, which gives a global indication of the usefulness of including X_k into a multiple regression model *in addition* to the other $(p-1)$ variables. This measure therefore addresses the same problem as the above described t-test and is called the partial correlation coefficient of X_k corrected for the $(p-1)$ other variables. It is defined as the correlation coefficient between two vectors of residuals, namely between $R_Y(X_{-k})$ which is the vector of residuals from the linear regression model (with intercept) that regresses Y on all explanatory variables except X_k, and $R_{X_k}(X_{-k})$ which is the vector of residuals from the linear regression model that regresses X_k on the same $(p-1)$ other explanatory variables. The vectors $R_Y(X_{-k})$ and $R_{X_k}(X_{-k})$ represent the part of Y and X_k, respectively, that cannot be explained by the $(p-1)$ other variables. This means that if the absolute value of the partial correlation coefficient is close to 1, it makes sense to include X_k in a model for Y in which the other $(p-1)$ explanatory variables are already present.

Example 8.3 Let us consider once more the cherry tree data of the foregoing examples, and let us investigate whether or not the variable 'length' should be added to the model in addition to the variable 'diameter'. To this end, consider the testing problem

H_0: $\beta_2 = 0$; β_0, β_1 arbitrary,

H_1: $\beta_2 \neq 0$; β_0, β_1 arbitrary.

We find $t_2 = 2.607$ with a (two-sided) p-value of 0.0145 for the t-distribution with 28 degrees of freedom, and $\mathcal{F}^{1,2} = 6.794$ with the same p-value, 0.0145 for the

F-distribution with 1 and 28 degrees of freedom. We may conclude that H_0 gets rejected (assuming we selected α to be greater than 0.0145.) Note that $t_2^2 = 6.796$ equals $F^{1,2}$ up to rounding error.

Taking the determination coefficient into account, we see that for simple linear regression of volume on diameter it is $\mathcal{R}^2 = 0.94$, whereas the determination coefficient increases to $\mathcal{R}^2 = 0.95$ due to addition of the variable length (multiple regression on diameter and length). Since this is a minor increase, the values of the determination coefficients suggest that inclusion of the variable 'length' is not very useful, especially since this variable is hard to measure in practice.

This conclusion is to some extent supported by the value 0.44 of the partial correlation coefficient of length corrected for diameter. But how do we explain the results of the $\mathcal{F}$- and t-test, which with the small p-value of around 0.014 seem to indicate that inclusion of length is significantly better? In this respect the general warning for thoughtlessly applying statistical tests is in place. The rejection of the null hypothesis (since in that case the p-value is below the significance level such a test is often called "significant") means that an observed effect, here the increase of $\mathcal{R}^2$ from 0.94 to 0.95, cannot be attributed to chance: if new data would be collected, one would most likely find a similar increase in $\mathcal{R}^2$. However, significance of a test does not mean that the effect is of practical importance. In the case of the cherry tree data one will probably decide that diameter on its own is sufficient to predict the timber yield, not only because the difference of the two determination coefficients is small, but also because the length of an uncut tree is difficult to establish.

Although practically not very relevant, for illustrative purposes we could also investigate whether a simple linear regression model with only length would be appropriate. In other words, we could test whether the variable 'diameter' should be included in the model in addition to the variable 'length'. In that case we would test the hypotheses:

H_0: $\beta_1 = 0$; β_0, β_2 arbitrary,

H_1: $\beta_1 \neq 0$; β_0, β_2 arbitrary.

The resulting values $t_1 = 17.82$ and $\mathcal{F}^{1,2} = 317.41$ for the test statistics indicate that the null hypothesis will be rejected for all reasonable values of the significance level. A simple linear regression with only length is therefore not appropriate. This also follows from the value 0.35 of the determination coefficient for this model and the value 0.96 of the partial correlation coefficient for diameter corrected for length. ***

8.2 Diagnostics

For the estimation and testing procedures we have postulated a certain model including a couple of model assumptions, and we tacitly assumed that the model and the model

assumptions were correct or at least plausible. Of course, in practice this is not always the case. Hence, another important aspect of a statistical analysis is to check the model assumptions and to investigate the quality of the model. This is called *diagnostic regression* or shortly: *diagnostics*. Quantities that give information on the model quality and assumptions are called *diagnostic quantities*. Unlike the global quantities that we saw until now, such as $\hat{\beta}$ and $\mathcal{R}^2$, which give a global picture, a diagnostic quantity generally has a different value for each observation point. The importance of diagnostic regression is illustrated by the following example.

Example 8.4 For four fictional data sets 1–4 the values of the response variable Y and explanatory variable X, so-called Anscombe's quartet, are:

1		2		3		4	
X	Y	X	Y	X	Y	X	Y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

If the simple linear regression model

$$Y = \beta_0 + \beta_1 X + e$$

is fitted to each of the four data sets, the parameter estimates and the determination coefficient have the same values for each of the four cases:

$$\hat{\beta}_0 = 3.0,\ \hat{\beta}_1 = 0.5,\ \hat{\sigma}^2 = 1.5,\ \mathcal{R}^2 = 0.67.$$

From the fact that these global quantities are the same for the four sets, it is tempting to conclude that simple linear regression is equally good in the four situations. The most simple form of diagnostic regression, making scatter plots, shows that this is not true. Figure 8.2 gives scatter plots of the Y-values against the X-values for each of the four cases. In the first case the data are as we would expect if a simple linear regression model would be a good description of reality. The graph for the second case suggests that the analysis based on simple linear regression is incorrect and that a smooth curve, for instance a quadratic polynomial could be fitted to

the data so that much less unexplained variation in Y would remain. The third case shows a picture from which we see that a simple linear regression model seems appropriate for all points, except one. It would perhaps be better not to include this point in the global analysis. This would give: $\hat{y} = 4.0 + 0.34x$, a very different line. Without more information it is impossible to say which of the two lines is 'better'. In the last case, there is not sufficient information to infer something about the quality of the linear regression model. The slope $\hat{\beta}_1$ of the regression line is mainly determined by the value y_8. If the eighth observation point would be deleted, the parameter β_1 could not even be estimated. A global analysis that depends on one point to such large extent should not be trusted!
***

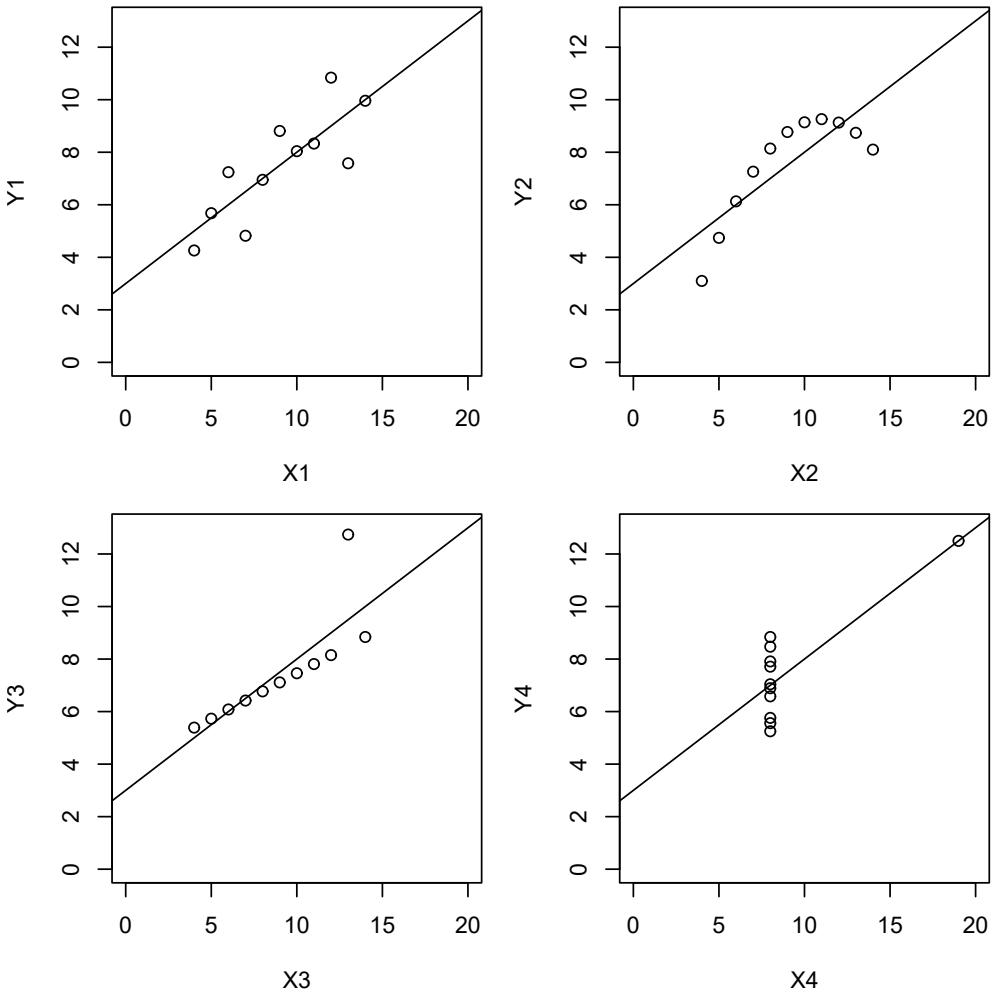


Figure 8.2: Scatter plots of four fictional data sets with their linear regression lines.

8.2.1 Scatter Plots

From Example 8.4 it is clear that making plots is indispensable for model building and assessment of the quality of a model. In general, to investigate how well a model fits the data it is helpful to consider not only the value of $\mathcal{R}^2$, but also a number of graphs.

For a simple linear regression model the relationship between the response Y and the explanatory variable $X = X_1$ can simply be represented by a scatter plot to which the estimated regression line may be added. For multiple regression it is less straightforward. It is generally useful to make scatter plots of the response Y against each of the explanatory variables and, because there may be relationships between the explanatory variables themselves, also of each pair of explanatory variables.

Example 8.5 Figure 8.3 shows several scatter plots for the cherry tree data. The results of our global analysis are confirmed in these plots: there appears to be a strong linear relationship between volume and diameter. Moreover, there is a moderate linear relationship between the two explanatory variables 'diameter' and 'length'. ***

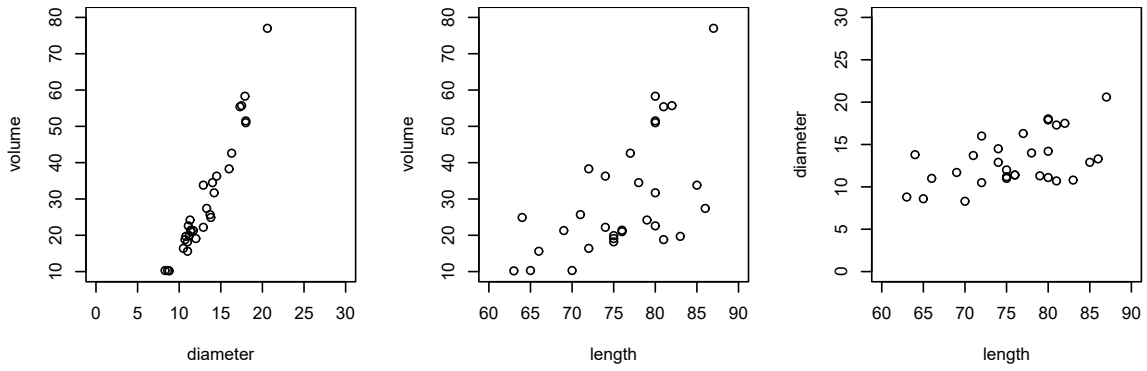


Figure 8.3: Some scatter plots for the cherry tree data.

What a scatter plot of Y against one of the explanatory variables, X_k for instance, does not show, is the relationship between Y and X_k after correction for possible relationships between X_k and the other explanatory variables. Going back to the definition of the partial correlation coefficient, we see that a scatter plot of the vector of residuals $R_Y(X_{-k})$ of the regression of Y on all X_j except X_k, against the vector of residuals $R_{X_k}(X_{-k})$ of regression of X_k on the other X_j represents just this relationship. This graph is called the *added variable plot* for X_k. A strong linear relationship between the plotted residuals corresponds with a strong linear relationship between Y and X_k corrected for the other

variables, and indicates that X_k should be added to the model that already contains the other variables. The added variable plot has the property that if the linear regression model

$$R_Y(X_{-k}) = \alpha_0 + \alpha_1 R_{X_k}(X_{-k}) + e,$$

of $R_Y(X_{-k})$ on $R_{X_k}(X_{-k})$ is fitted, then the least squares estimators of α_0 and α_1 satisfy $\hat{\alpha}_0 = 0$ and $\hat{\alpha}_1 = \hat{\beta}_k$, respectively. Here $\hat{\beta}_k$ is the least squares estimator of β_k in (8.8). The residuals of both regression procedures are the same, so that the residual sums of squares RSS for both models are equal too.

Added variable plots can be interpreted in the same way as scatter plots for simple linear regression. They can give information about nonlinear relationships or outliers. To determine whether or not a variable should be included in the model, the added variable plot is a very informative tool: whereas the partial correlation coefficient and the statistic t_k defined in (8.13) summarize the overall effect of adding the variable X_k, the added variable plot for X_k shows this effect for each observation separately.

Example 8.6 Figure 8.4 displays added variable plots for the cherry tree data. The added variable plot for diameter (left) shows a strong linear relationship between the two plotted variables, which indicates that the variable diameter should be included into the model, a conclusion that already was drawn based on the t_1-value of 17.82. The added variable plot for length, however, is more difficult to interpret. Although it shows a slight upward trend, there is no clear linear relationship. This too reflects the above found t_2-value 2.61.

As we saw in Example 8.3 above, the partial correlation coefficients for diameter and length, which are the correlation coefficients of the two corresponding added variable plots, are 0.96 and 0.44, respectively.
***

8.2.2 The residuals

Not only inspection of scatter plots of Y against the different X_i, scatter plots of the X_i against each other, and added variable plots, but also graphical investigation of the residuals may yield information about the quality of the model. Indeed, the values of the residuals can be seen as estimates of the realizations of the measurement errors $e_1, \ldots, e_n$. Because the measurement errors are assumed to be independent and normally distributed with expectation 0 and variance σ^2, any systematic pattern in the residuals, for example a relationship with another variable, is suspect. Moreover, as we have seen in the examples, a normal QQ-plot of the residuals gives information about the reasonability of the normality assumption. Also informative are scatter plots of

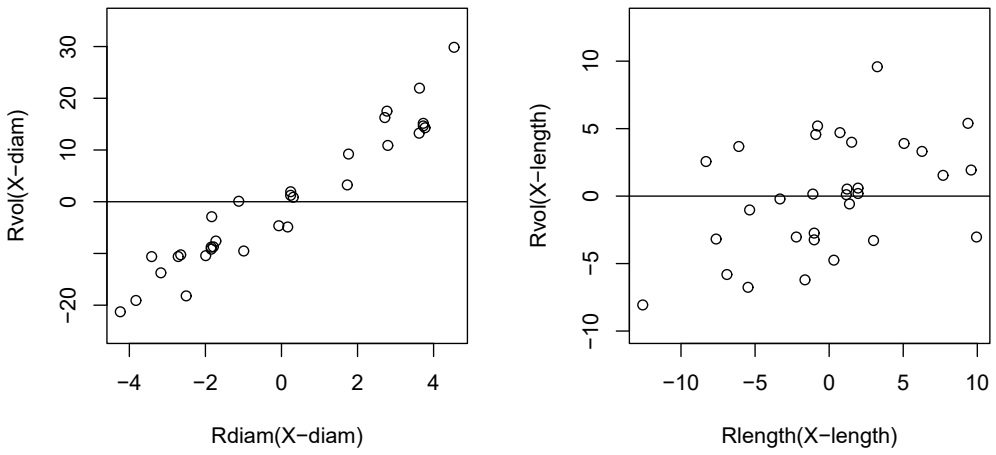


Figure 8.4: The cherry tree data: added variable plots for diameter (left) and length (right).

- the residuals against each of the explanatory variables in the model. A curvilinear graph could be an indication that a higher term of the explanatory variable needs to be included as an additional variable in the model. Moreover, each systematic relationship in the spread of the points in the plot may indicate a non-equal variance of the measurement errors.

- the residuals against explanatory variables that are not included in the model. Variables for which a clear linear relationship is seen in the plots, should be included in the model.

- the residuals against the predicted responses. If the spread of the residuals seems to increase or decrease for increasing values of the predicted response, this can be an indication of a non-equal variance of the measurement errors. A transformation could help here, or one could resort to the so-called *weighted* linear regression in which the variance of the measurement errors e_i is allowed to be different for different i. If the graph shows a nonlinear relationship, then a linear regression model is perhaps not the best model. Transformation of the response or of the explanatory variable(s), or the use of a nonlinear model could be more appropriate.

For the assessment of the quality of a model the problem is of course, that a discrepancy between model and data is not necessarily caused by inappropriate model assumptions, but can also be the result of incorrect data values. Based on a statistical analysis alone, it cannot be decided which of the two is the case. This is why it is important to always keep the experimental situation in mind. Moreover, it depends on the type of discrepancy whether or not it is easily detected. In fact, in the fitting procedure the model is fitted to all data, and hence also to deviating data. In case of deviating observations the residuals

therefore do not need to be large. In some cases a deviating observation may even lead to a very small residual for that observation.

8.2.3 Outliers

One of the assumptions in a regression analysis is that the chosen model is appropriate for all observation points. However, in practice it is often the case that one or more observation points have a response value that does not seem to match with the model that is best suited for the majority of the points. Such observation points, with extremely large or extremely small values of the response compared to what is expected under the model, are called *outliers*, a term that we already met in Chapter 5. The following example illustrates that it is always useful to start with inspecting a couple of plots for possible outliers.

Example 8.7 Around 1850 the Scottish physicist J.D. Forbes collected data about the boiling point of water (in degrees Fahrenheit) under different values of pressure (in inches of Mercury) by performing experiments at different altitudes in the Alps. The goal of these experiments was to investigate the relationship between boiling point and pressure, so that the more easily measured boiling point could serve as a predictor for pressure and ultimately for altitude. The explanatory variable was x = boiling point, and on the pressure data a transformation was performed so that the response variable became $y = 100\times \log(\text{pressure})$, with values:

x	194.5	194.3	197.9	198.4	199.4	199.9	200.9	201.1	201.4
y	131.79	131.79	135.02	135.55	136.46	136.83	137.82	138.00	138.06
x	201.3	203.6	204.6	209.5	208.6	210.7	211.9	212.2	
y	138.04	140.04	142.44	145.47	144.34	146.30	147.54	147.80	

Figure 8.5 gives a couple of relevant plots for these data. The upper left plot is a scatter plot of y- against x-values. There obviously is a strong linear relationship between the two variables, although one of the points, observation 12, does not lie on the line with the other points. If the residuals are plotted against the corresponding x values (top-right), then we see that most of the residuals are small, except for the value of the 12-th point. Because the linear relationship between x and y is very strong, a plot of the residuals against the predicted responses, $\hat{y}$, gives the same picture, see the lower left plot. The lower right plot shows a normal QQ-plot of the residuals. Also here the 12-th observation is not in line with the others.
***

For a simple linear regression problem, like in the example above, to detect outliers, it is usually sufficient to inspect a scatter plot of the response Y against $X = X_1$, the single

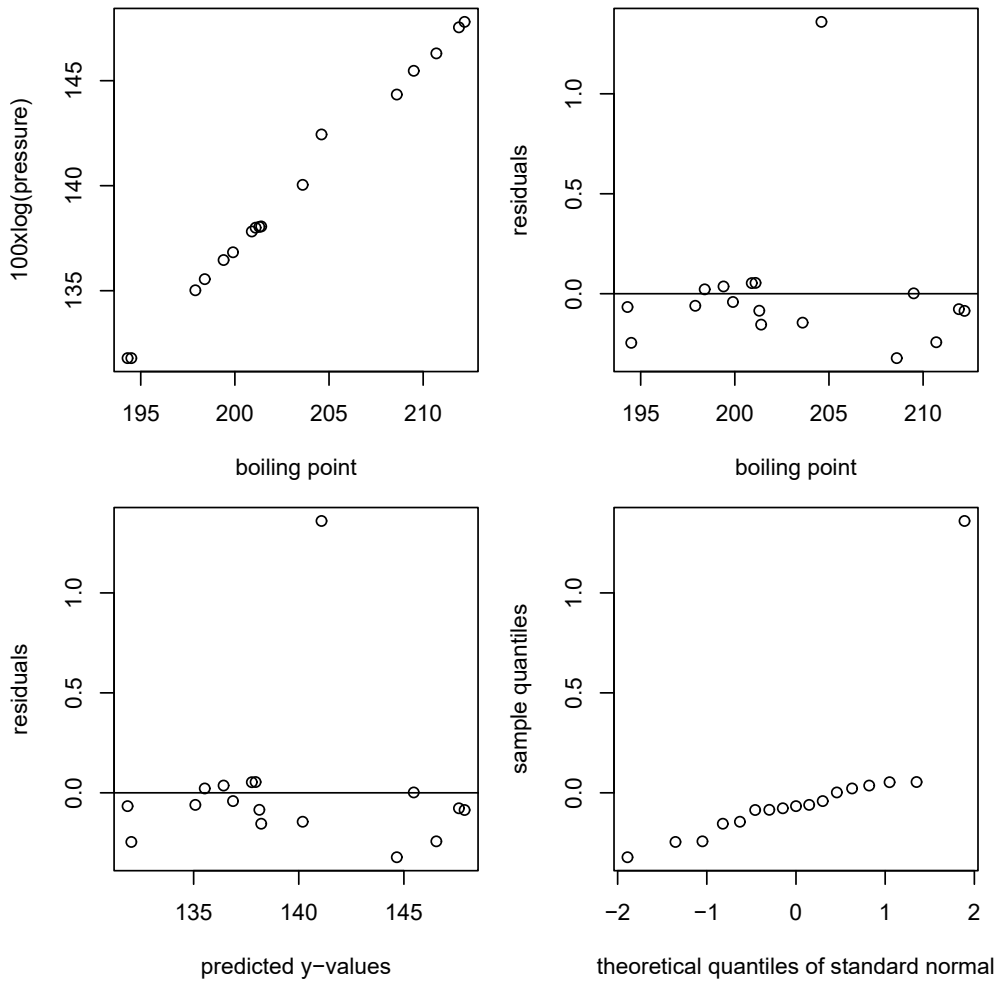


Figure 8.5: Forbes' data: scatter plots of y against x, residuals against x, residuals against $\hat{y}$ and a normal QQ-plot of the residuals.

explanatory variable. In a multiple linear regression setting scatter plots of the response against the explanatory variables as well as added variable plots may give information about this, but the problem is obviously more complex.

Once one or more possible outliers are visually detected, one could apply a more formal procedure to decide whether or not these points should be considered extreme. One such procedure makes use of a t-test like the one discussed above, combined with a so-called *mean shift outlier model*. Suppose that the k-th point is suspected to be an outlier. Then one could think that the multivariate regression model is correct for all points, except for this one. To adjust for this, the original model is adapted for the k-th point in such a way that the expected mean for the k-th response under the original model, namely $x_k^T\beta$, is changed with an unknown amount δ so that the resulting model could fit the k-th point

too. This resulting model is the mean shift outlier model:

$$Y_i = \begin{cases} x_i^T\beta + e_i, & i \neq k \\ x_i^T\beta + \delta + e_i, & i = k, \end{cases}$$

or, in matrix notation,

$$Y = X\beta + u\delta + e = X_*\beta_* + e, \tag{8.14}$$

where u is a fictional "explanatory" variable with $u_i = 0$ for $i \neq k$ and $u_k = 1$; the new design matrix X_* is the $n \times (p+2)$-matrix (X, u), and the new parameter vector $\beta_* = (\beta, \delta)^T$.

If model (8.14) holds, the k-th observation point is an outlier and the corresponding response has expectation $x_k^T\beta + \delta$. If the value of δ is large enough, then indeed it is plausible that the k-th point is an outlier. In other words, if δ is significantly different from zero, the k-th point is an outlier. But then we may just use the t-test (8.13) to test the significance of δ in the model (8.14). Generally, the sign of δ is known from the plots, and a one-sided test should be used in this situation. Let us assume for simplicity that $\delta > 0$; for $\delta < 0$ the procedure is analogous. We test the following hypotheses:

H_0: $\delta = 0$, β arbitrary,

H_1: $\delta > 0$, β arbitrary.

For this testing problem a test with significance level α is: reject H_0 if

$$t_{p+1} = \frac{\hat{\delta}}{\sqrt{\widehat{\mathrm{Cov}}(\hat{\beta}_*)_{p+1,p+1}}} \geq t_{(n-p-2);1-\alpha},$$

where $\hat{\delta}$ in the numerator of the test statistic is the least squares estimate of the shift δ in the model (8.14), and $\widehat{\mathrm{Cov}}(\hat{\beta}_*)_{p+1,p+1}$ in the denominator is the $(p+2)$-th diagonal element of $\widehat{\mathrm{Cov}}(\hat{\beta}_*) = \hat{\sigma}_*^2(X_*^TX_*)^{-1}$ with

$$\hat{\sigma}_*^2 = \frac{RSS_*}{(n-p-2)} = \frac{(Y - X_*\hat{\beta})^T(Y - X_*\hat{\beta})}{(n-p-2)}.$$

The model (8.14) and the corresponding test procedure can be extended in the obvious way for more than one outlier. Evidently, once it is decided that a certain point is an outlier, it should be investigated what could be the cause for this. If it is likely that measurement error is the cause, then the analysis could be repeated without the point. If measurement error is not a likely cause, or if the cause is unknown, it is advisable to try other type of models. In general, it should always be mentioned in the report of the analysis and its results which points, if any, were omitted from the analysis.

8.2.4 Leverage points and the hat matrix

For the Forbes' data in Example 8.7 it is easy to see from the plots that the 12-th point stands out from the others, because the corresponding value of the explanatory variable lies in the middle of the range of the x-values. It will become more difficult to determine whether a point is an outlier when its explanatory variable has an extreme value. Such points *potentially* have a large influence on the regression results.

Example 8.8 Huber's data set consists of the following fictive data:

x :	-4	-3	-2	-1	0	10
y :	2.48	0.73	-0.04	-1.44	-1.32	0.00

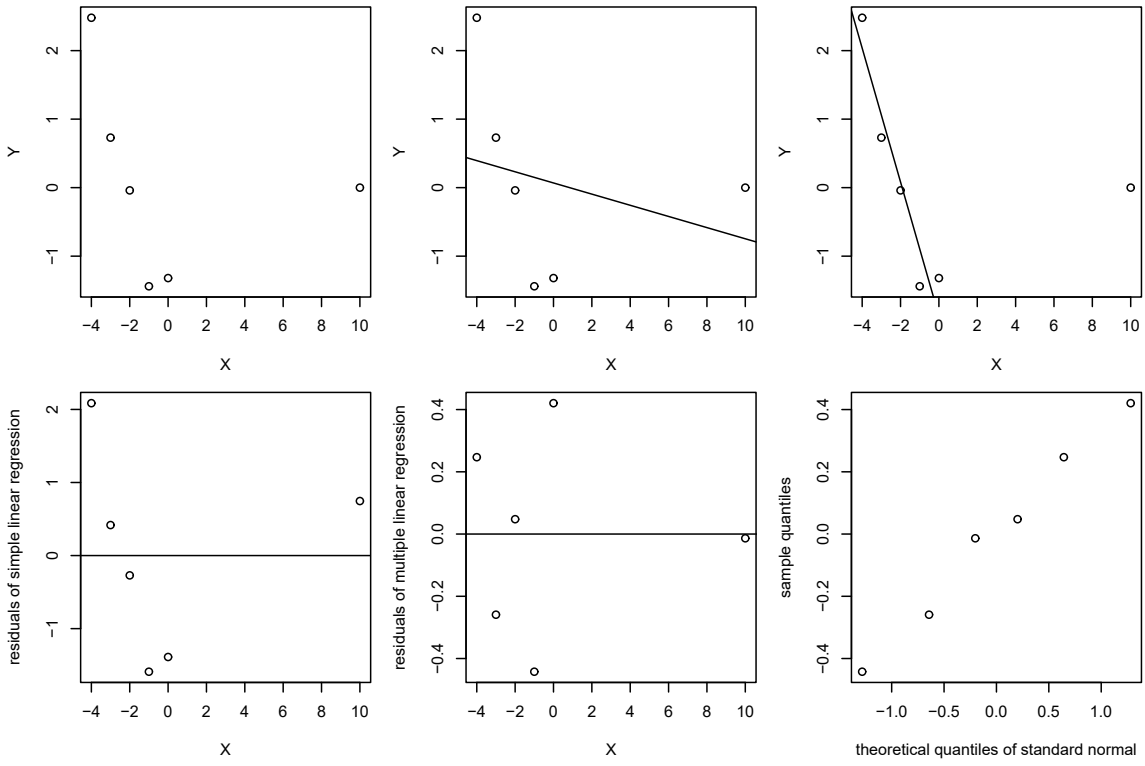


Figure 8.6: Several plots for Huber's data.

Figure 8.6 shows several plots for these data. The scatter plot of y against x (top-left) suggests that if the 6-th observation is not taken into consideration, a straight

line is a reasonable model for the other 5 points. If a straight line is fitted through all points, then the resulting line (top-middle) is rather different from the line obtained by performing linear regression without the 6-th point (top-right). In a plot of the residuals of the regression on all points against the x-values (bottom-left) we see that there is one large residual, but this is not the one corresponding to observation 6 but the one of observation 1.

If there is a good reason not to ignore observation 6, then the scatter plot of the data suggests that we may investigate a model in which a quadratic term is included. If x^2 is added as a second explanatory variable, the t-test for testing whether the corresponding coefficient is significantly different from zero yields a p-value smaller than 0.01, and therefore the coefficient is significantly different from zero. For this multiple linear regression model with the two explanatory variables x and x^2 the residuals are smaller than the residuals from the simple linear regression and a plot of these residuals against x looks good (bottom-middle), although the residual of the 6-th observation is still a bit small compared to the other residuals. The latter may be caused by the fact that the quadratic nature of the relationship between x and y is mainly determined by this one point.

Finally, a normal QQ-plot of the residuals of the quadratic regression shows a straight line (bottom-right). We may conclude that for the complete data set the multiple linear regression model with x and x^2 as explanatory variables is a reasonable model.
***

The example above shows that for detecting outliers in the explanatory variables checking for large residuals is not sufficient, because observation 6, which is deviating in some way, has a small residual for both the simple and the quadratic model. A small residual may be a good sign if the model fits the data well and the point with the small residual is just one of the points where the model fits very well. However, it may also be a bad sign if the small residual is the result of the regression "being forced" into a certain direction by extreme values of the corresponding point, like in the middle plot of the top row in Figure 8.6. To see how it can be investigated whether there is a potential problem due to one or more extreme values of the explanatory variables for the general multiple linear regression situation, let us once again consider the multiple regression model (8.2), and the predicted response

$$\hat{Y} = X\hat{\beta} = X(X^T X)^{-1} X^T Y = HY, \tag{8.15}$$

where

$$H = X(X^T X)^{-1} X^T. \tag{8.16}$$

Because the response vector Y is transformed by the matrix H into the vector of predicted responses $\hat{Y}$, which is usually called 'Y-hat', the matrix H is called the *hat matrix*. The

hat matrix, and in particular its diagonal elements $h_{ii} = x_i^T(X^TX)^{-1}x_i$, $i = 1, \ldots, n$, can tell us something about the potential influence an extreme value can have on the regression. To see this, we note that H is an orthogonal projection which projects on the column space of X. This means that h_{ii} is a measure of the remoteness of the i-th observation from the other $(n-1)$ observations in the space of the columns of X. The sum of the h_{ii} equals the rank of X, which is $p+1$, and $\frac{1}{n} \leq h_{ii} \leq 1$[1], and

$$h_{ii} = \sum_i h_{ij}^2 = h_{ii}^2 + \sum_{j \neq i} h_{ij}^2.$$

If h_{ii} is far away from the mean of all diagonal elements, $(p+1)/n$, then the i-th observation is in some way deviating from the other data points, and caution is in place. In particular, if h_{ii} equals the maximum possible value 1, then all h_{ij}, $j \neq i$, will be 0, and the prediction $\hat{Y}_i$, which from (8.15) can be seen to satisfy

$$\hat{Y}_i = h_{ii}Y_i + \sum_{j \neq i} h_{ij}Y_j,$$

will be equal to Y_i. If h_{ii} is large, but not exactly equal to the maximum possible value 1, then with large probability $\hat{Y}_i$ will be very close to Y_i, but whether or not this will actually happen depends on where the other points lie with respect to the i-th point. Observation points with large h_{ii} are called *potential* or *leverage points*; h_{ii} is called the *potential* or *leverage* of the i-th observation point. For a particular h_{ii} to be considered large, sometimes the condition $h_{ii} > 2(p+1)/n$, i.e., larger than twice the mean of all diagonal elements of H, is used.

As explained, an observation point with large h_{ii} is not necessarily influential but has the potential to be so. In particular, two groups of n observation points with the same values for the explanatory variables, and hence the same H, but with different realizations of the response variables may have different influences on the regression. For instance, for Huber's data set, imagine what would have been the regression line and the size of the residuals if the 6-th response would have been -8.00 instead of 0.00. The notion of 'potential' influence can also be understood by considering the variances of the residuals. For the vector of residual R it holds that

$$R = Y - \hat{Y} = (I_{n \times n} - H)Y. \tag{8.17}$$

If $e_1, \ldots, e_n$ are independent and normally distributed, then the same holds for $Y_1, \ldots, Y_n$ and it follows from (8.17) that R is normally distributed with expectation vector $\mathrm{E}(R) = 0$ and covariance matrix $\mathrm{Cov}(R) = \sigma^2(I_{n \times n} - H)$. Hence, the residuals may be correlated

[1]The term $1/n$ on the left stems from the fact that we consider models with an intercept; for models without an intercept this term would vanish.

and have different variances:

$$\mathrm{Var}(R_i) = \sigma^2(1 - h_{ii}). \tag{8.18}$$

But (8.18) means that for observation points with a large value, close to 1, of h_{ii}, the variance $\mathrm{Var}(R_i)$ of the corresponding residual R_i is small. Because $\mathrm{E}\, R_i = 0$, the realized value of the residual for such points will almost always be small, so that $\hat{Y}_i$ will be very close to Y_i no matter what value Y_i has.

Example 8.9 Let us illustrate the use of the hat matrix for detection of leverage points by considering Huber's data set again. In the case of simple linear regression the diagonal elements of H satisfy

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2},$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. If $x_i = \bar{x}$, h_{ii} attains the minimum value $\frac{1}{n}$. The larger the distance of x_i to $\bar{x}$ is, the larger is h_{ii}. The 6×6 hat matrix for simple linear regression for Huber's data is:

$$H = \begin{pmatrix} 0.290 & 0.259 & 0.228 & 0.197 & 0.167 & -0.141 \\ 0.259 & 0.236 & 0.213 & 0.190 & 0.167 & -0.064 \\ 0.228 & 0.213 & 0.197 & 0.182 & 0.167 & 0.013 \\ 0.197 & 0.190 & 0.182 & 0.174 & 0.167 & 0.090 \\ 0.167 & 0.167 & 0.167 & 0.167 & 0.167 & 0.167 \\ -0.141 & -0.064 & 0.013 & 0.090 & 0.167 & 0.936 \end{pmatrix}.$$

We see that h_{66} is close to 1, which agrees with the impression from the plots that the 6th point is far away from the other points. Because $x_5 = \bar{x}$, h_{55} equals the minimum value $\frac{1}{n} = \frac{1}{6}$. We also see that the further away x_i is from $\bar{x} = x_5$, the larger is h_{ii}. For the predicted responses we have, for example,

$$\begin{aligned} \hat{y}_1 &= 0.290y_1 + 0.259y_2 + 0.228y_3 + 0.197y_4 + 0.167y_5 - 0.141y_6 \\ \hat{y}_6 &= -0.141y_1 - 0.064y_2 + 0.013y_3 + 0.090y_4 + 0.167y_5 + 0.936y_6, \end{aligned}$$

so that $\hat{y}_1$ depends on all y-values, but $\hat{y}_6$ mainly on y_6.
***

Instead of checking the values of the h_{ii} separately, it may be handy to make an *index plot* for the leverages: the leverage of the i-th point is plotted against i. This may be especially useful for large data sets. Figure 8.7 shows the index plot for the leverages of Huber's data.

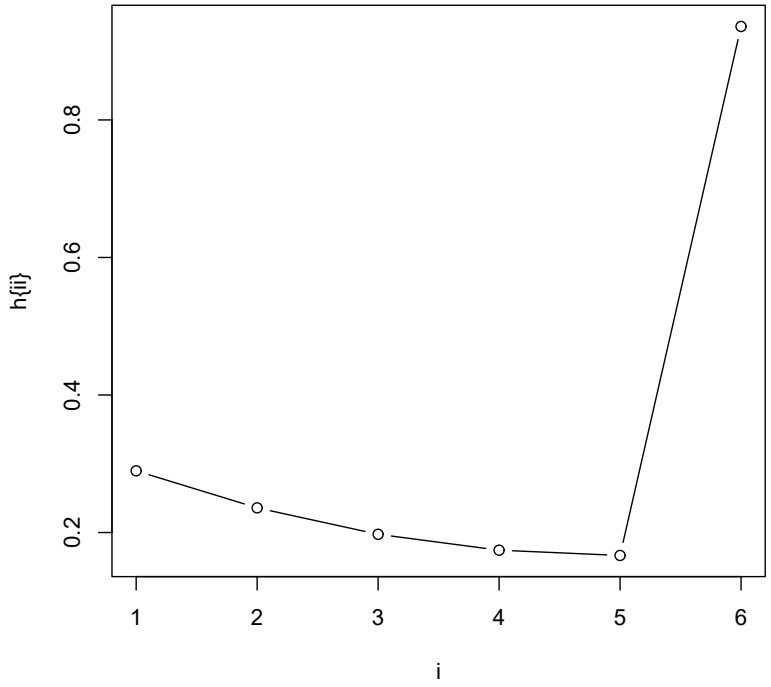


Figure 8.7: Index plot for the leverages of Huber's data for simple linear regression.

8.2.5 Influence points

One way to investigate the effect of a leverage point, or more generally of any observation point, on the results of a regression analysis is to fit models with and without this point. If in Example 8.8 one of the observations 1-5 is omitted, this barely has an effect on the results, whereas deletion of observation 6 has the effect that the quadratic term drops out. An observation point with such a large effect on the fitting results is called an *influence point*. We already remarked that a point with high leverage is not necessarily an influence point. If in Example 8.8 the response value of point 6 would have been such that the point would have been on the line with the other observations, then deletion of this point would have had little effect. In that case, the point still would have been a leverage point, but not an influence point. Obviously, not only leverage points, but also outliers—observation points with an extreme value of the response variable—can be influence points.

Instead of studying the effect of one particular observation point, one could make use of a quantitative measure for the influence that one point has on the regression, and compare this measure for the different points. This is especially useful when it is not completely clear beforehand which observation points should be investigated. After comparing the measures, one could decide which points should be inspected more closely. In the sequel an index (i) means "with the i-th observation point omitted". For example, $\hat{\beta}_{(i)}$ is the

estimator of β computed without the i-th point:

$$\hat{\beta}_{(i)} = (X_{(i)}^T X_{(i)})^{-1} X_{(i)}^T Y_{(i)}.$$

One way to quantify the influence of the i-th point is by comparing $\hat{\beta}$ and $\hat{\beta}_{(i)}$. The *Cook's distance* exactly does this. To introduce this measure, we first note that under model (8.2) and the independence and normality assumptions of the measurement errors, the set of vectors b defined by

$$\{b : \frac{(b-\hat{\beta})^T (X^T X)(b-\hat{\beta})}{(p+1)\hat{\sigma}^2} < F_{(p+1),(n-p-1);1-\alpha}\} \tag{8.19}$$

is a $(1-\alpha)100\%$ confidence region for the parameter vector β. This set is centered around $\hat{\beta}$ and has the form of an ellipsoid. When, for fixed n and p, α is chosen smaller, the set becomes larger and the probability that the region contains the real value of β also becomes larger. Hence, these regions form some kind of metric that can tell us whether a particular value of b lies far away from $\hat{\beta}$. Now we are interested whether $\hat{\beta}_{(i)}$ lies far away from $\hat{\beta}$. It follows that if we could find a value of α such that $\hat{\beta}_{(i)}$ would lie outside the $(1-\alpha)100\%$ confidence region for β around $\hat{\beta}$ if the difference between $\hat{\beta}_{(i)}$ and $\hat{\beta}$ is too large, then we would have found a means to quantify the influence of the i-th point. Experience has shown that a value of α of around 0.5, which yields values of $F_{(p+1),(n-p-1);1-0.5}$ in (8.19) around 1 for most n and p, satisfies this requirement.

The definition of the Cook's distance makes use of this argumentation: the Cook's distance D_i for the i-th point is defined as

$$D_i = \frac{(\hat{\beta}_{(i)} - \hat{\beta})^T (X^T X)(\hat{\beta}_{(i)} - \hat{\beta})}{(p+1)\hat{\sigma}^2} = \frac{(\hat{Y}_{(i)} - \hat{Y})^T (\hat{Y}_{(i)} - \hat{Y})}{(p+1)\hat{\sigma}^2}, \tag{8.20}$$

where $\hat{Y} = X\hat{\beta}$ and $\hat{Y}_{(i)} = X\hat{\beta}_{(i)}$. Observation points with a large Cook's distance have substantial influence on $\hat{\beta}$ and, as can be seen from the second equality in the definition (8.20), also on the predicted values. Deletion of these points can rigorously change the conclusions of the statistical analysis. In view of the argumentation following (8.19), D_i is most often considered large if it is close to or much larger than 1.

In general, the advice is to list the D_i values for all points. Then to perform regression with and without the point(s) with the largest Cook's distance(s), even if these are smaller than 1, and to check whether the results differ substantially or not.

Example 8.10 Below the Cook's distances for a couple of data sets are listed with the relatively large values in bold.

Cherry tree data, regression on diameter and length:
0.098, 0.148, 0.167, 0.000, 0.004, 0.008, 0.001, 0.001, 0.003, 0.000, 0.009, 0.000, 0.000, 0.001, 0.021, 0.027, 0.019, 0.178, 0.041, 0.108, 0.000, 0.020, 0.001, 0.038, 0.009, 0.047, 0.031, 0.070, 0.018, 0.026, **0.605**.

Cherry tree data, regression on diameter:
0.110, 0.049, 0.022, 0.000, 0.004, 0.006, 0.015, 0.001, 0.016, 0.000, 0.021, 0.000, 0.000, 0.001, 0.025, 0.037, 0.028, 0.009, 0.045, 0.064, 0.000, 0.011, 0.000, 0.061, 0.018, 0.065, 0.050, 0.076, 0.028, 0.040, **0.888**.

Forbes' data:
0.064, 0.005, 0.001, 0.000, 0.000, 0.001, 0.001, 0.001, 0.006, 0.002, 0.005, **0.469**, 0.000, 0.054, 0.051, 0.007, 0.010.

Huber's data:
0.520, 0.015, 0.005, 0.135, 0.096, **26.399**.

We see that in only one case, in Huber's data set, the Cook's distance is much larger than 1. Still, the last observation point of the cherry tree data, which has the largest influence on $\hat{\beta}$ and the predicted responses in both multiple and simple regression, should be investigated. This is because the difference in Cook's distance with the other points is quite remarkable.

For Forbes' data the 12-th observation point, which we found to be an outlier, is not an influence point according to the above criterion. It does have a Cook's distance which is larger than that of the other points, though, and therefore has relatively more influence than the other points.

Finally, the leverage point in Huber's data, clearly is an influence point.

*****

8.3 Collinearity

In this section we discuss the important concept of *collinearity*. If two explanatory variables have an exact linear relationship with each other, then it is intuitively clear that if one of them is already in the model, it is not necessary any more to include the other. This is because the variation in the response variable that can be explained by the first variable is exactly the same as the variation that can be explained by the second variable. Hence, if the first one is already in the model, no variation that can be explained by the second is left over. Moreover, if both variables are included, the design matrix X is no longer of maximal rank, $X^T X$ is no longer invertible, and the corresponding $\hat{\beta}$ cannot be determined. If there is an *almost* exact linear relationship between two variables, the above is almost true: in this case there may be numerical problems with computation of the inverse $(X^T X)^{-1}$ and determining the estimate $\hat{\beta}$ may be difficult in the sense

that one of the estimators $\hat{\beta}_j$ has a large variance, so that the corresponding estimate may be unreliable. The same can happen if more than two explanatory variables have a strong linear relationship. If two or more explanatory variables have a very strong linear relationship it is said that they are (approximately) *collinear*. Formally, a set of variables $X_1, \ldots, X_m$ is called (exactly) collinear if there exist constants $c_0, c_1, \ldots, c_m$, not all equal to 0, such that

$$c_0 + c_1 X_1 + \cdots + c_m X_m = 0. \tag{8.21}$$

A strong linear relationship between two explanatory variables is the simplest form of a collinearity. A collinearity may also exist between more than two variables, and within a group of more than four variables more than one collinearity may be present.

Obviously, besides a potential cause for problems with numerically solving the equation $X^T Y = X^T X\beta$, collinearity in X is also a statistical problem. Therefore, it is crucial to determine whether there exist collinearities between groups of explanatory variables, not only to prevent numerical problems, but, more importantly, to prevent bad estimates. In the context of a statistical data analysis one generally only needs to be concerned about a collinearity if it causes inefficient estimation of one or more of the coefficients β_j. Indeed, in case of collinearity in X the variance of one or more $\hat{\beta}_j$ can be unpleasantly large, so that the estimate $\hat{\beta}_j$ is useless. If there is a collinearity between a group of explanatory variables, at least one of them is approximately a linear combination of the others, so that this variable does not give more information than is already contained in the others together. If in this case the whole group is included in the model, at least one of the components of the least squares estimator $\hat{\beta}$ cannot be trusted.

A remedy for collinearity in a design matrix is generally not found by mere numerical arguments. For example, from a numerical point of view the problem could be solved by deleting one or more columns of X. However, it is the question whether the resulting model is still useful. From a statistical point of view, the interpretation of the model should always be taken into account. An important question in this respect is what could be the cause of the collinearity. For instance, there is a big difference between collinearities caused by measurement error and collinearities caused by a true linear relationship. Therefore it is not only important to detect collinearities, but also to investigate their cause and to study the effect of possible remedies. In the following we discuss a couple of measures that give information about the amount of collinearity within a set of explanatory variables.

8.3.1 Variance inflation factors

As suggested before, it is helpful to start with making scatter plots of each pair of explanatory variables or to compute their correlation coefficients. In this way, only the *pairwise* linear relationships can be detected.

A frequently used method to find out which of the $\hat{\beta}_j$ should not be trusted, is to compute the *variance inflation factors*. The j-th variance inflation factor VIF_j is defined

by

$$VIF_j = \frac{1}{1-\mathcal{R}_j^2}, \qquad j = 1, \ldots, p, \tag{8.22}$$

where $\mathcal{R}_j^2$ is the determination coefficient of the regression of the explanatory variable X_j on the other explanatory variables. To see why the VIFs are useful for detecting unreliable $\hat{\beta}_j$, remember that the determination coefficient $\mathcal{R}_j^2$ is equal to the square of the multiple correlation coefficient between X_j and the other variables, and therefore the vector $(\mathcal{R}_1^2, \ldots, \mathcal{R}_p^2)^T$ of all such determination coefficients gives information about possible multiple collinearities between the explanatory variables. If $\mathcal{R}_j^2 = 0$ and thus $VIF_j = 1$, X_j is orthogonal to the others, and if $\mathcal{R}_j^2 = 1$ and hence $VIF_j = \infty$, there is a perfect linear correlation of X_j and (some of) the other variables. The definition and name of the VIF_j become clear from the following. It can be proved that for $j = 1, \ldots, p$, the variance of $\hat{\beta}_j$ equals

$$\mathrm{Var}(\hat{\beta}_j) = \sigma^2 \frac{1}{(1-\mathcal{R}_j^2)} \frac{1}{\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2}. \tag{8.23}$$

This expression shows that if, due to a collinearity, VIF_j is large, the variance of $\hat{\beta}_j$ is inflated with precisely this factor VIF_j. Note that (8.23) does not hold for $j = 0$, because $\mathcal{R}_0^2$ is not defined.

In conclusion, the VIF_j's represent the increase of the variance of the estimators due to a linear relationship between the explanatory variables, i.e., due to collinearity. As such they indicate which estimates should be distrusted: if VIF_j is close to 1, the estimate $\hat{\beta}_j$ can be trusted; if VIF_j is large, $\hat{\beta}_j$ is unreliable. However, based on the values of the VIF_j's it is difficult to distinguish the different collinearities, so that it is not clear which X_j are involved in which collinearities. The VIF_j's therefore do not give a solution for the problem.

Example 8.11 The two variance inflation factors for the regression of the cherry tree volumes on diameter and length are both equal to 1.369. Therefore, based on the variance inflation factors there is no reason to distrust the estimates of β_1 and β_2.
***

8.3.2 The condition number

Since in the case of exact collinearity it holds that the determinant of X^TX is equal to 0, it is tempting to use the value of $\mathrm{Det}(X^TX)$ for detecting collinearity between columns of X. However, the size of $\mathrm{Det}(X^TX)$ generally does not reflect to what extent the

structure of X leads to incorrect regression results. It is better to consider the eigenvalues of X^TX instead. Because for each exact collinearity precisely one eigenvalue is equal to 0, an eigenvalue close to 0 tells us something about potential problems with inverting the matrix, and the number of small eigenvalues indicates the number of collinearities. The first measure for collinearity that is derived from this is the *condition number*. Let $\nu_1, \ldots, \nu_{p+1}$ be the eigenvalues of X^TX, then the condition number $\kappa(X)$ of X is defined as

$$\kappa(X) = \sqrt{\frac{\max_j \nu_j}{\min_j \nu_j}}. \tag{8.24}$$

It holds that $\kappa(X) \geq 1$; the lower bound 1 is attained when the columns of X are orthogonal. Large values of $\kappa(X)$ indicate collinearity in X and potential sensitivity of a solution (vector) b of the linear system $Xb = c$ to small changes in c and X. If $\kappa(X)$ is large, X is called *ill-conditioned*. As such it is also used outside the context of regression analysis.

Example 8.12 Consider the matrices

$$A = \begin{pmatrix} 1 & a \\ a & 1 \end{pmatrix} \text{ and } B = \begin{pmatrix} b & 0 \\ 0 & b \end{pmatrix}.$$

For $a \neq 1$, we have $\kappa(A) = (1+a)(1-a)^{-1}$ and for $b \neq 0$, $\kappa(B) = bb^{-1} = 1$. If a tends to 1, then $\kappa(A)$ tends to infinity. Hence, for values of a close to 1 the matrix A is ill-conditioned. On the other hand, if b tends to 0, $\kappa(B)$ keeps being equal to 1, provided $b \neq 0$. This means that even for values of b close to 0, the matrix B is well conditioned.

If we would have limited ourselves to only consider $\text{Det}(A^TA)$ and $\text{Det}(B^TB)$, we would have found: $\text{Det}(A^TA) = (1-a^2)^2$ and $\text{Det}(B^TB) = b^4$, from which it follows that $\text{Det}(A^TA)$ and $\text{Det}(B^TB)$ are close to 0 for values of a close to 1, and for values of b close to 0, respectively. A value of the determinant close to 0 therefore does not necessarily correspond to the matrix being ill-conditioned.

***

A matrix X which has one or more columns of predominantly very small or predominantly very large elements relative to the other columns may cause numerical problems. It is often advised in this case to scale the columns of X such that they have the same (Euclidean) length, for example length 1. Such scaling usually decreases the condition number of the matrix and diminishes the risk of numerical problems. However, in a regression analysis the columns of the design matrix X generally have a specific meaning and one has to make sure that the scaled version of a column still has the same meaning in the practical context of the regression problem. Moreover, scaling of the design matrix does not solve the problem of too large variances of the $\hat{\beta}_j$, because scaling does not have

any effect on the collinearities. Another approach that is often taken to avoid numerical problems, namely centering of the design matrix–subtracting of each column its mean–and consecutively performing regression without an intercept, has the same drawbacks.

The condition number can be used to get a first impression about whether or not further investigation for collinearity is needed. A frequently used, but not very well founded, criterion for further investigation is $\kappa(X) > 30$ (for columns of roughly equal Euclidean length). The condition number does not give any indication about where the collinearities are. In that sense it is a weaker measure than the variance inflation factors that show which columns of X are approximately orthogonal to the others, and which $\hat{\beta}_j$ should be distrusted.

Example 8.13 The condition number for the cherry tree data for regression on diameter and length is equal to 959.4; scaling of the columns of the design matrix yields 32.2. We will see below that the large condition number is not caused by collinearity of diameter and length, which we saw in Figure 8.3 to have some linear relationship, but not a very strong one. Instead it is due to the fact that the length values have relatively little variation and, as a result, the corresponding column in the design matrix is approximately collinear to the columns of ones which represents the intercept.
*****

Instead of from the eigenvalues of X^TX, $\kappa(X)$ can also be computed from the so-called *singular value decomposition* of X. The singular value decomposition is a decomposition of the form

$$X = UDV^T \tag{8.25}$$

with the $n \times (p+1)$- matrix U and $(p+1) \times (p+1)$ matrix V satisfying $U^TU = V^TV = I_{p+1}$ and the $(p+1) \times (p+1)$ matrix D being a diagonal matrix with nonnegative diagonal elements $\mu_1, \ldots, \mu_{p+1}$. The $\mu_1, \ldots, \mu_{p+1}$ are called the *singular values* of X. The singular value decomposition is not unique, because a permutation of the diagonal elements of D, with the columns of U and V simultaneously permuted in the same way, also yields a singular value decomposition of X. It is easy to see that the squares of the singular values of X are the eigenvalues of X^TX, so that

$$\kappa(X) = \sqrt{\frac{\max_j \nu_j}{\min_j \nu_j}} = \frac{\max_j \mu_j}{\min_j \mu_j}. \tag{8.26}$$

8.3.3 Condition indices and variance decomposition

The condition number makes use of only a small part of the information contained in the singular value decomposition of X, namely of the values of the largest and smallest

singular values relative to each other. We now consider the values of all singular values relative to the largest one and see what this can tell us.

As mentioned earlier, the number of small eigenvalues of X^TX, or equivalently the number of singular values of X, relates to the number of collinearities. The question is: how small is 'small' in this respect? To answer this question we compare, like in the definition of the condition number, the values of all singular values to the largest one: for $k = 1, \ldots, p+1$ we define the k-th *condition index* $\kappa_k(X)$ of X by

$$\kappa_k(X) = \frac{\max_j \mu_j}{\mu_k}. \tag{8.27}$$

Evidently, the condition indices satisfy $1 \leq \kappa_k(X) \leq \kappa(X),\ k = 1, \ldots, p+1$. A large value of $\kappa_k(X)$ corresponds to a collinearity. Also here $\kappa_k(X) > 30$ is a frequently used criterion for scaled X.

Not only the number of collinearities, but also the structure of a collinearity can be inferred from the singular value decomposition $X = UDV^T$. To see this, we note that if the condition index $\kappa_k(X)$ is large, and correspondingly μ_k small, then some linear algebra yields

$$\sum_{j=0}^{p} v_{jk} X_j = XV_k = U_k \mu_k \approx 0, \tag{8.28}$$

where v_{jk} is the element (j,k) of V, V_k the k-th column of matrix V, U_k the k-th column of U. It follows from (8.21) and (8.28) that the columns X_j for which the corresponding j-th component v_{jk} of V_k is substantially different from 0 take part in the collinearity corresponding to μ_k.

The singular value decomposition (8.25) not only tells us which variables take part in a collinearity, but it can also help us to determine which collinearities we should worry about. As was explained at the beginning of this section, in principle we only need to be concerned about a collinearity if it causes unreliable estimates of one or more of the β_j. Whether or not β_j can be estimated accurately can be investigated by inspecting the variance of its estimator $\hat{\beta}_j$, like we did in the subsection about variance inflation factors. The variance inflation factors, however, could only tell us which variances are 'inflated', but not which variables constitute the collinearity that possibly caused the inflation and therefore could not help us to find a solution for the problem of inaccurate estimation.

To see how the singular value decomposition can help indicating which collinearities should be worried about, we first remind that if a set of explanatory variables is almost collinear, each of the corresponding β_j can be derived from the others, which makes it impossible to accurately estimate the individual $\hat{\beta}_j$ that are related to this collinearity. Reversely, if the variances of a set of $\hat{\beta}_j$s to a large extent are determined by the same small singular value, the variables corresponding to this set of $\hat{\beta}_j$s form the collinearity that corresponds to this small singular value. Hence, it would be helpful to find out which

variances are affected by which singular value. Now, the singular value decomposition (8.25) gives rise to the following useful decomposition of the variances of the $\hat{\beta}_j$ that exactly shows this. We have

$$\mathrm{Var}(\hat{\beta}) = \sigma^2 (X^T X)^{-1} = \sigma^2 V D^{-2} V^T,$$

so that for $j = 0, 1, \ldots, p$,

$$\mathrm{Var}(\hat{\beta}_j) = \sigma^2 \sum_{k=1}^{p+1} \frac{v_{jk}^2}{\mu_k^2}. \tag{8.29}$$

Note that each of the components of the sum in (8.29) is associated with exactly one of the singular values μ_k. Because μ_k occurs in the denominator, the contribution of the k-th component will be large when μ_k is small, or equivalently, when $\kappa_k(X)$ is large. Of course, the magnitude of this contribution also depends on the size of v_{jk} in the numerator. Therefore, if the variance of $\hat{\beta}_j$ is large, then it can be deduced from (8.29) which components have contributed to this large variance: the relative contribution of μ_k to $\mathrm{Var}(\hat{\beta}_j)$ is

$$\pi_{kj} = \frac{v_{jk}^2/\mu_k^2}{\sum_{k=1}^{p+1} v_{jk}^2/\mu_k^2}, \qquad k = 1, \ldots, p+1;\ j = 0, 1, \ldots, p. \tag{8.30}$$

The numbers $\pi_{1j}, \ldots, \pi_{p+1,j}$, are called the *variance decomposition proportions* of $\hat{\beta}_j$. If π_{kj} is large, and μ_k small, we may conclude that the magnitude of $\mathrm{Var}(\hat{\beta}_j)$ is to a large extent caused by the collinearity that belongs to μ_k, or equivalently to the collinearity that belongs to the large condition index $\kappa_k(X)$. Hence, if we have detected all $\mathrm{Var}(\hat{\beta}_j)$ whose sizes to a large extent are determined by the collinearity that belongs to the large condition index $\kappa_k(X)$, we know that the variables that correspond to these β_j form that particular collinearity.

Whether or not a variance decomposition proportion can be considered sufficiently large depends on the context. A proportion $\pi_{kj} = 1$ means that a (possibly large) variance is completely caused by one collinearity. This can easily happen if one of the condition indices is larger than the others. A large variance may also be caused by two collinearities of nearly equal strength. In that case proportions between 0.3 and 1 are not unusual. We stress that a large variance decomposition proportion $\pi_{kj} = 1$ is only problematic if $\mathrm{Var}(\hat{\beta}_j)$ is also large.

From the above it follows that to investigate the nature of a collinearity based on the singular value decomposition, two approaches can be used. In both cases it is first determined which singular values μ_k cause a large variance, $\mathrm{Var}(\hat{\beta}_l)$ say. In most situations there will be only one, and $\pi_{lj} \approx 1$. With the first approach, one next considers for each such μ_k the column V_k of V. According to (8.28) the coordinates of this column indicate the linear combination of the columns of X which approximately equals zero, and thus form the collinearity belonging to μ_k.

The second approach is generally easier. Instead of V_k, it considers for each μ_k that causes a large variance, the variance decompositions of the other $\hat{\beta}_j$. If besides $\mathrm{Var}(\hat{\beta}_l)$ also the variances of $\hat{\beta}_{j_1}, \ldots, \hat{\beta}_{j_m}$ are largely determined by μ_k, then this is an indication that the columns X_l and $X_{j_1}, \ldots, X_{j_m}$ are collinear. An operational description of the second approach is: write down the matrix of π_{kj}. If μ_k is small (or κ_k large), then search in the k-th row for the $\pi_{kj_1}, \ldots, \pi_{kj_m}$ that are close to 1 (or at least differ substantially from 0). It is important to note that at least two columns are needed to form a collinearity: besides π_{kl} at least one other π_{kj} should be relatively large for a collinearity to exist.

Finally, we remark that the existence of collinearities only depends on the structure of the design matrix, but that the variance of an estimator being large or not also depends on the structure of the response vector. This means that for step one of the two approaches all data need to be considered, whereas to perform the second step it suffices to investigate the design matrix. Once it is determined that a collinearity leads to serious problems, the only solution is to change the design matrix X.

Example 8.14 Consider the so-called Bauer matrix

$$B = \begin{pmatrix} -74 & 80 & 18 & -56 & -112 \\ 14 & -69 & 21 & 52 & 104 \\ 66 & -72 & -5 & 764 & 1528 \\ -12 & 66 & -30 & 4096 & 8192 \\ 3 & 8 & -7 & -13276 & -26552 \\ 4 & -12 & 4 & 8421 & 16842 \end{pmatrix}.$$

The 5th column of this matrix equals twice the 4th column. Moreover, the last two columns are orthogonal to the first three, which themselves are not orthogonal to each other. Although this matrix does not contain a column of ones and thus differs from the general design matrix that we use, the (second step of the) two approaches explained above can well be illustrated with it.

The singular values of B are

$$\begin{aligned} \mu_1 &= 170.7 \\ \mu_2 &= 60.5 \\ \mu_3 &= 7.6 \\ \mu_4 &= 36368.4 \\ \mu_5 &= 0.0 \end{aligned}.$$

Because the last two columns are exactly collinear, one of the singular values, μ_5, is exactly equal to 0. Also the third, μ_3, is relatively small compared to the others, indicating the existence of one more collinearity. The singular values are permuted so that the corresponding matrix V in the singular value decomposition of B reflects

the orthogonality structure of B by blocks of zeros:

$$V = \begin{pmatrix} 0.540 & 0.625 & -0.556 & 0.000 & 0.000 \\ -0.836 & 0.383 & -0.393 & 0.000 & 0.000 \\ 0.033 & -0.680 & -0.733 & 0.000 & 0.000 \\ 0.000 & 0.000 & 0.000 & -0.447 & -0.894 \\ 0.000 & 0.000 & 0.000 & -0.894 & 0.447 \end{pmatrix}.$$

(Note that the reported values of the μ_k and V are rounded for clarity. Of course, these numerically determined values will never turn out to be exactly equal to 0.) For the small value μ_5 we conclude from the fact that only the values of the 4th and 5th element in the 5th column of V are substantially different from 0, that the 4th and the 5th column of the "design matrix" B are collinear. Furthermore, for the relatively small value μ_3 we see from the values of the elements in the 3rd column of V that the first three columns of the "design matrix" B are collinear. This illustrates the second step of the first approach.

The condition indices and variance decomposition proportions for B are

$$\begin{pmatrix} \textit{cond.ind.} & vdp(\hat{\beta}_1) & vdp(\hat{\beta}_2) & vdp(\hat{\beta}_3) & vdp(\hat{\beta}_4) & vdp(\hat{\beta}_5) \\ 2.131e+02 & 0.002 & 0.000 & 0.000 & 0.000 & 0.000 \\ 6.008e+02 & 0.020 & 0.015 & 0.013 & 0.000 & 0.000 \\ 4.784e+03 & 0.979 & 0.977 & 0.987 & 0.000 & 0.000 \\ 1.000e+00 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 \\ \infty & 0.000 & 0.000 & 0.000 & 1.000 & 1.000 \end{pmatrix}.$$

As expected, the condition index corresponding to μ_5 is very large. This index corresponds to two variance proportions equal to 1: one of the variance of $\hat{\beta}_4$ and one of the variance of $\hat{\beta}_5$. Hence, the corresponding columns of the "design matrix" are collinear. These are the 4th and the 5th column of B, which agrees with our earlier observation. We also see a large condition index in the 3rd row, the one corresponding to the singular value μ_3. This condition index relates to three variance proportions close to 1; the columns 1, 2 and 3 of B are also strongly collinear. The same conclusions would have been drawn if another singular value decomposition would have been chosen. This would only have resulted in a permutation of the rows of the matrix with condition indices and variance proportions. This illustrates the second step of the second approach.

If the second approach is being performed on a scaled version of B, we obtain

$$\begin{pmatrix} \textit{cond.ind.} & vdp(\hat{\beta}_1) & vdp(\hat{\beta}_2) & vdp(\hat{\beta}_3) & vdp(\hat{\beta}_4) & vdp(\hat{\beta}_5) \\ 1.326e+00 & 0.001 & 0.001 & 0.047 & 0.000 & 0.000 \\ 1.600e+01 & 0.994 & 0.994 & 0.953 & 0.000 & 0.000 \\ 1.039e+00 & 0.005 & 0.005 & 0.000 & 0.000 & 0.000 \\ 1.000e+00 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 \\ \infty & 0.000 & 0.000 & 0.000 & 1.000 & 1.000 \end{pmatrix}.$$

We see that in this case the largest condition index is still infinitely large. This should not surprise, because an exact collinearity of the last two columns does not disappear by scaling of the columns of a matrix. The other condition indices have become smaller, as expected, but still the same conclusions as for the unscaled matrix may be drawn. This illustrates that scaling may reduce numerical problems, but does not make collinearities disappear.
***

Example 8.15 For the cherry tree data we find:

$$\begin{pmatrix} \textit{cond.ind.} & \textit{vdp}(\hat{\beta}_0) & \textit{vdp}(\hat{\beta}_1) & \textit{vdp}(\hat{\beta}_2) \\ 1.000 & 0.000 & 0.000 & 0.005 \\ 29.252 & 0.000 & 0.964 & 0.122 \\ 959.377 & 1.000 & 0.036 & 0.874 \end{pmatrix}.$$

The third row of this matrix contains a large condition index and two large variance proportions. The latter belong to $\hat{\beta}_0$ and $\hat{\beta}_2$. Apparently, the column of ones which represents the intercept, and length are collinear. The large variance of the estimator of the intercept for regression of volume on diameter and length that we found in Example 8.1 is most likely caused by this collinearity.

Scaling of the design matrix makes its condition slightly better:

$$\begin{pmatrix} \textit{cond.ind.} & \textit{vdp}(\hat{\beta}_1) & \textit{vdp}(\hat{\beta}_2) & \textit{vdp}(\hat{\beta}_3) \\ 1.000 & 0.001 & 0.004 & 0.001 \\ 9.964 & 0.055 & 0.827 & 0.015 \\ 32.178 & 0.945 & 0.169 & 0.984 \end{pmatrix}.$$

Scaling therefore may be better from a numerical point of view, but it does not make the collinearity between the columns of ones and length disappear.

Due to collinearity the estimator of the intercept for regression of volume on diameter and length has a larger variance than for regression on diameter only (Example 8.1). Addition of a variable that has an approximate linear relationship with the column of ones is addition of a variable with approximately all equal elements — the variable 'length' in this case. This results in a larger uncertainty about the value of the intercept. Because addition of length barely improves the regression fit (see Example 8.3) this is another reason not to include this variable in the regression model.

On the other hand, it should be noted that the intercept in this problem does not have a physical meaning. If the model would be valid for the whole range of values of diameter and length, then the intercept would even be zero: for diameter and length zero, the volume is zero too. This suggests that a different model could yield a better description of reality, for example a linear model for linear transformations of the three variables. In view of the meaning of the variables, the model

$$\log(\text{volume}) = \beta_0 + \beta_1 \log(\text{diameter}) + \beta_2 \log(\text{length}) + e \tag{8.31}$$

is a plausible one. However, within the limited range of the given diameters and lengths the simpler linear regression suffices. This range is of course more or less the range that is of practical interest too. Moreover, if we are mostly interested in predicting the volume $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ and to a lesser extent in the exact values of β_0, β_1 and β_2, then we do not even need to be concerned about the collinearity of the intercept and length.
***

8.3.4 Remedies

As mentioned earlier, it is impossible to give general guidelines for solving problems with a regression analysis due to collinearity. Deletion of one or more columns or replacement of the columns involved in a collinearity by a suitable linear combination of these columns could solve the problems. However, whether the new model still makes sense depends on the context, that is, the real situation that the model aims to describe and the goal of the study. Thorough knowledge of the experimental situation and common sense are the best guidelines here.

We remark that in the case of collinearity the use of least squares estimators is not recommended. Although they are still the best *unbiased* estimators, there exist better biased ones. We will not elaborate on this.

We conclude with two examples that illustrate the complexity of the collinearity issue.

Example 8.16 An agriculturist has asked a student to perform a small experiment on four trial fields in order to test the effect of addition of two chemical compounds, X_1 and X_2, to the soil on the yield of product Y. The student chooses the following quantities (in kg) of X_1 and X_2 for the four fields:

$field$	X_1	X_2
1	0	0
2	0	1
3	1	0
4	1	1

.

One month after the experiment started it turns out that one more trial field is available. The student decides to use this fifth field to obtain additional information. Because the chemical compounds came in 50 kg bags, the student chooses to use the remainder and spread on the fifth field 48 kg of both compounds. The design

matrix for this experiment becomes

$$\begin{pmatrix} X_0 & X_1 & X_2 \\ 1 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \\ 1 & 48 & 48 \end{pmatrix}.$$

Since the correlation coefficient between X_1 and X_2 is 0.999, the last two columns of this matrix are highly collinear. This is confirmed by the values of the other collinearity measures: for the variance inflation factors of the above matrix we find $VIF_1 = VIF_2 = 1/(1 - \rho^2(X_1, X_2)) = 903$, and the variance decomposition proportions are

$$\begin{pmatrix} \textit{cond.ind.} & \textit{prop}(\hat{\beta}_0) & \textit{prop}(\hat{\beta}_1) & \textit{prop}(\hat{\beta}_2) \\ 1.0 & 0 & 0 & 0 \\ 34.3 & 1 & 0 & 0 \\ 73.2 & 0 & 1 & 1 \end{pmatrix}.$$

The collinearity in the design matrix is solely due to the high values of the compounds for the fifth field: the last two columns of the design matrix for the experiment with only the four original fields have correlation coefficient 0. Scaling of the matrix is no option here, for the only sensible scaling would be the one in which the column of ones is made of the same length as that of the other two. All other scalings would mean that regression would be performed on quantities that are not actually tested and thus would not give a realistic picture.

This example also shows why centering of the design matrix, another frequently used approach for avoiding numerical problems, is often not useful: the means of X_1 and X_2 are 10, so that centering would yield negative values of compound quantities. The resulting new columns have no practical meaning any more.

What now is the best thing that the student could have done with the fifth field? If the model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e$$

holds in all five cases, then how should, given the first four points, the additional point be chosen such that the most accurate estimates of β would have been obtained? It turns out, that this is achieved indeed by making X_1 and X_2 for the fifth field as large as possible. For example, if the point $(x_{51}, x_{52}) = (48, 48)$ is added, then one gets $\mathrm{Var}(\hat{\beta}_1) = \mathrm{Var}(\hat{\beta}_2) = 0.5\sigma^2$, whereas if $(x_{51}, x_{52}) = (0.5, 0.5)$, it holds that $\mathrm{Var}(\hat{\beta}_1) = \mathrm{Var}(\hat{\beta}_2) = \sigma^2$. The real problem in this case is not the collinearity in the design matrix, but that the linear model most likely is not suited for all five situations. It is highly probable that the model is a reasonable description of reality for small values of X_1 and X_2, but not for large ones.

***

When the collinearity measures do not indicate collinearity, this necessarily implies that there is no collinearity present in the design matrix. On the other hand, if the collinearity measures do indicate collinearity this may be caused by a different problem than collinearity itself. In particular, leverage points can mask or induce collinearity. The following example illustrates this.

Example 8.17 The top-left picture in Figure 8.8 is a scatter plot of two variables that are highly collinear. If we consider the collinearity measures, we find for the variance inflation factors $VIF_1 = VIF_2 = 1.09$, and for the condition number $\kappa = 14.6$. These values do not reflect the collinearity. It is obvious that this is most likely caused by the isolated observation point. This point is a leverage point with leverage 0.998. Ignoring this point, we obtain the top-right scatter plot in Figure 8.8, and the collinearity measures become $VIF_1 = VIF_2 = 257.02$ and $\kappa = 210.6$.

The bottom-left graph in Figure 8.8 shows a scatter plot of two variables that are definitely not collinear. The isolated point is a leverage point with leverage 0.965. The variance inflation factors are $VIF_1 = VIF_2 = 1.03$, and $\kappa = 30.68$. Ignoring the leverage point, we find $VIF_1 = VIF_2 = 1.20$ and $\kappa = 15.74$; the corresponding scatter plot is the bottom-right plot in Figure 8.8. In this case the leverage point introduces collinearity to some extent: the condition number with the point included is much larger than without the point, whereas the variance inflation factors are barely influenced by the presence of the point.
*****

Example 8.17 once again illustrates the importance of looking at the data first.

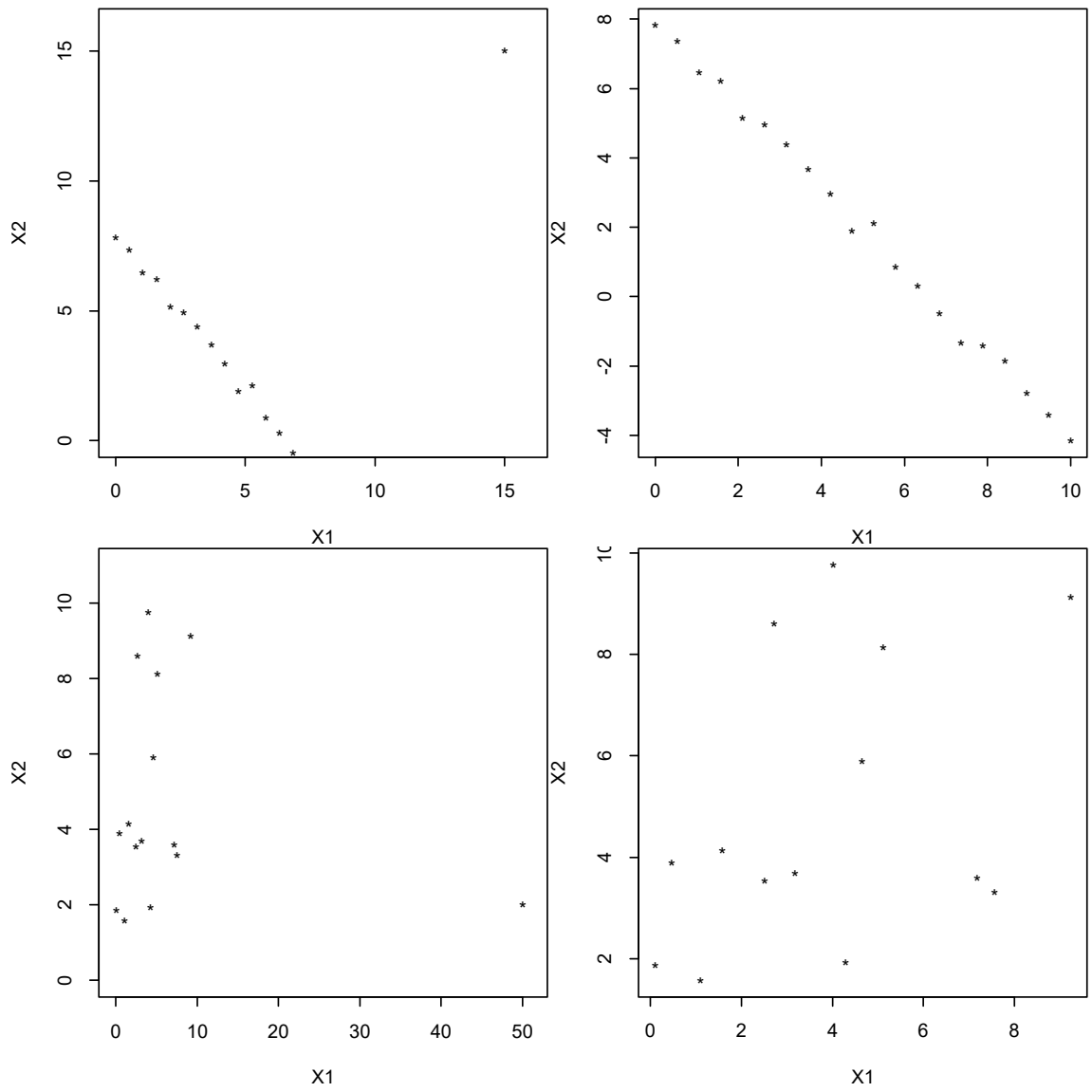


Figure 8.8: Scatter plots of data sets for which a leverage point masks (top) and induces (bottom) collinearity.

References and other useful literature

Anderson, T.W., (2003) *An Introduction to Multivariate Statistical Analysis, 3rd edn.* John Wiley & Sons, New York.

Atkinson, A.C., (1987). *Plots, Transformations and regression; an Introduction to Graphical Methods of Diagnostic Regression Analysis.* Clarendon Press, Oxford.

Bates, D.M., Watts, D.G., (2007), *Nonlinear Regression Analysis and its Applications.* John Wiley & Sons, New York.

Belsley, D.A., Kuh, E., Welsch, R.E., (2004). *Regression Diagnostics; identifying influential data and sources of collinearity.* John Wiley & Sons, New York.

Chatfield, C., Collins, A., (1981). *Introduction to Multivariate Analysis.* Chapman and Hall/CRC.

Christensen, R., (1997) *Log-Linear Models and Logistic Regression.* Springer, New York, Berlin.

Dobson, A.J., Barnett, A., (2008), *An Introduction to Generalized Linear Models*, 3rd ed. Chapman and Hall/CRC.

Duong, T., Hazelton, M.L., (2003). Plug-in bandwidth matrices for bivariate kernel density estimation. *Nonparametric Statistics* **15(1)**, 17-30.

Efron, B., Tibshirani, R., (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science* **1**, 54-75.

Efron, B., Tibshirani, R., (1993). *An Introduction to the Bootstrap.* Chapman and Hall/CRC.

Everitt, B.S., (1992). *The Analysis of Contingency Tables.* Chapman and Hall/CRC, London.

Hampel, F.R., Ronchetti, E.M., Rouseeuw, P.J., Stahel, W.A., (2005). *Robust Statistiscs: the Approach based on Influence Functions.* Wiley-Interscience.

Hoaglin, D.C., Mosteller, F., Tukey, J.W., (2000). *Understanding Robust and Exploratory Data Analysis.* Wiley-Interscience.

Hoaglin, D.C., Mosteller, F., Tukey, J.W., (2006). *Exploring Data Tables, Trends, and Shapes.* Wiley-Interscience.

Huber, P.J., (1981). *Robust Statistics.* John Wiley & Sons, New York.

Huff, D., (1993). *How to lie with statistics.* W.W. Norton & Company.

Lehmann, E.L., D'Abrera, H.J.M., (2006). *Nonparametrics: Statistical Methods based on Ranks.* Springer.

Mardia, K.V., Kent, J.T., Bibby, J.M., (1980). *Multivariate Analysis.* Academic Press, London.

McCullagh, P., Nelder, J., (1989). *Generalized Linear Models.* Chapman and Hall/CRC.

Seber, G.A.F., Lee, A.J., (2003). *Linear Regression Analysis, 2nd ed..* John Wiley & Sons, New York.

Seber, G.A.F., Wild, C.J., (2003). *Nonlinear Regression.* John Wiley & Sons, New York.

Shapiro, S.S., Wilk, M.B., (1965). *An analysis of variance test for normality. Biometrika* **52**, 591.

Shorack, G.R., Wellner, J.A., (2009). *Empirical Processes with Applications to Statistics.* Society for Industrial & Applied Mathematics.

Silverman, B.W, (1986). *Density Estimation.* Chapman and Hall/CRC.

Tukey, J.W., (1977). *Exploratory Data Analysis.* Addison-Wesley, Reading.

Weisberg, S., (2005). *Applied linear regression, 3rd ed..* John Wiley & Sons, New York.

The websites for R and RStudio:

www.r-project.org

www.rstudio.com