Appendix A
Statistics

Since classical statistics provides many data analysis methods and supports and justifies a lot of others, we provide in this appendix a brief review of some basics of statistics. We discuss descriptive statistics, inferential statistics, and needed fundamentals from probability theory. Since we strove to make this appendix as self-contained as possible, some overlap with the chapters of this book is unavoidable. However, material is not simply repeated here but presented from a slightly different point of view, emphasizing different aspects and using different examples.

In [14] (classical) **statistics** is characterized as follows:

> Statistics is the art to acquire and collect data, to depict them, and to analyze and interpret them in order to gain new knowledge.

This characterization already indicates that statistics is very important for data analysis. Indeed: there is a vast collection of statistical procedures with which the tasks described in Sect. 1.3 (see page 11) can be tackled or which are needed to support or justify other methods. Some of these methods are discussed in this appendix. However, we do not claim to have provided a complete overview. For a more detailed review of (classical) statistics and its procedures, an interested reader is referred to standard textbooks and references like [3, 4, 10].

The statistical concepts and procedures we are going to discuss can roughly be divided into two categories corresponding to the two main areas of statistics:

- **descriptive statistics** (Sect. A.2) and
- **inferential statistics** (Sect. A.4).

In descriptive statistics it is tried to make data more comprehensible and interpretable by representing them in tables, charts, and diagrams, and to summarize them by computing certain characteristic measures. In inferential statistics, on the other hand, it is tried to draw inferences about the data generating process, like estimating the parameters of the process or selecting a model that fits it. The basis of many procedures of inferential statistics is **probability theory** (see Sect. A.3); its goal is usually to prepare for and to support decision making.

M.R. Berthold et al., *Guide to Intelligent Data Analysis,*
Texts in Computer Science 42,
DOI 10.1007/978-1-84882-260-3, © Springer-Verlag London Limited 2010

A.1 Terms and Notation

Before we can turn to statistical procedures, we have to introduce some terms and notions, together with some basic notation, with which we can refer to data.

- **object**, **case**
  Data describe objects, cases, people etc. For example, medical data usually describes patients, stockkeeping data usually describes components, devices or generally products, etc. The objects or cases are sometimes called the *statistical units*.
- **(random) sample**
  The set of objects or cases that are described by a data set is called a *sample*, its size (number of elements) is called the *sample size*. If the objects or cases are the outcomes of a random experiment (for example, drawing the numbers in a lottery), we call the sample a *random sample*.
- **attribute**, **feature**
  The objects or cases of the sample are characterized by attributes or features that refer to different properties of these objects or cases. Patients, for example, may be described by the attributes sex, age, weight, blood group, etc., component parts may have features like their physical dimensions or electrical parameters.
- **attribute value**
  The attributes, by which the objects/cases are described, can take different *values*. For example, the sex of a patient can be *male* or *female*, its age can be a positive integer number, etc. The set of values an attribute can take is called its *domain*.

Depending on the kind of the attribute values, one distinguishes different **scale types** (also called *attribute types*). This distinction is important, because certain *characteristic measures* (which we study in Sect. A.2.3) can be computed only for certain scale types. Furthermore, certain statistical procedures presuppose attributes of specific scale types. Table A.1 shows the most important scale types **nominal**, **ordinal**, and **metric** (or **numerical**), together with the core operations that are possible on them and a few examples of attributes having the scale type.

Table A.1 The most important scale types

Scale type	Possible operations	Examples
nominal	test for equality	sex
(categorical, qualitative)		blood type
ordinal	test for equality	school grade
(rank scale, comparative)	greater/less than	wind strength
metric (numerical)	test for equality	length
(interval scale, quantitative)	greater/less than	weight
	difference	time
	maybe ratio	temperature

Within nominal scales, one sometimes distinguishes according to the number of possible values. Attributes with only two values are called *dichotomous*, *alternatives* or *binary*, while attributes with more than two values are called *polytomous*. Within metric scales, one distinguishes whether only differences (temperature, calender time) are meaningful or whether it makes sense to compute ratios (length, weight, duration). One calls the former case *interval scale* and the latter *ratio scale*. In the following, however, we will not make much use of these additional distinctions.

From the above explanations of notions and expressions it already follows that a data set is the joint statement of attribute values for the objects or cases of a sample. The number of attributes that is used to describe the sample is called its *dimension*. One-dimensional data sets will be denoted by lowercase letters from the end of the alphabet, that is, for example x, y, z. These letters denote the attribute that is used to describe the objects or cases. The elements of the data set (the sample values are denoted by the same lowercase letter, with an index that states their position in the data set. For instance, we write $x = (x_1, x_2, \ldots, x_n)$ for a sample of size n. (A data set is written as a vector and not as a set, since several objects or cases may have the same sample value.) Multidimensional data sets are written as vectors of lowercase letters from the end of the alphabet. The elements of such data sets are vectors themselves. For example, a two-dimensional data set is written as $(x, y) = ((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n))$, where x and y are the two attributes by which the sample is described.

A.2 Descriptive Statistics

The task of descriptive statistics is to describe states and processes on the basis of observed data. The main tools to tackle this task are tabular and graphical representations and the computation of characteristic measures.

A.2.1 Tabular Representations

Tables are used to display observed data in a clearly arranged form, and also to collect and display characteristic measures. The simplest tabular representation of a (one-dimensional) data set is the **frequency table**, which is the basis for many graphical representations. A frequency table records for every attribute value its (absolute and/or relative) frequency in a sample, where the **absolute frequency** f_k is simply the occurrence frequency of an attribute value a_k in the sample, and the **relative frequency** r_k is defined as $r_k = \frac{f_k}{n}$ with the sample size n. In addition, columns for the cumulated (absolute and/or relative) frequencies (also simply referred to as *frequency sums*) may be present. As an example, we consider the data set

$$x = (3, 4, 3, 2, 5, 3, 1, 2, 4, 3, 3, 4, 4, 1, 5, 2, 2, 3, 5, 3, 2, 4, 3, 2, 3),$$

Table A.2 A simple frequency table showing the absolute frequencies f_k, the relative frequencies r_k, and the cumulated absolute and relative frequencies $\sum_{i=1}^{k} h_i$ and $\sum_{i=1}^{k} r_i$, respectively

a_k	f_k	r_k	$\sum_{i=1}^{k} f_i$	$\sum_{i=1}^{k} r_i$
1	2	$\frac{2}{25} = 0.08$	2	$\frac{2}{25} = 0.08$
2	6	$\frac{6}{25} = 0.24$	8	$\frac{8}{25} = 0.32$
3	9	$\frac{9}{25} = 0.36$	17	$\frac{17}{25} = 0.68$
4	5	$\frac{5}{25} = 0.20$	22	$\frac{22}{25} = 0.88$
5	3	$\frac{3}{25} = 0.12$	25	$\frac{25}{25} = 1.00$

Table A.3 A contingency table for two attributes A and B

	a_1	a_2	a_3	a_4	$\sum$
b_1	8	3	5	2	18
b_2	2	6	1	3	12
b_3	4	1	2	7	14
$\sum$	14	10	8	12	44

which may be, for instance, the grades of a written exam at school.[1] A frequency table for this data set is shown in Table A.2. Obviously, this table provides a much better view of the data than the raw data set as it is shown above, which only lists the sample values (an not even in a sorted fashion).

A two- or generally multidimensional frequency table, into which the (relative and/or absolute) frequency of every attribute value *combinations* is entered, is also called a **contingency table**. An example of a contingency table for two attribute A and B (with absolute frequencies), which also records the row and column sums, that is, the frequencies of the values of the individual attributes, is shown in Table A.3.

A.2.2 Graphical Representations

Graphical representations serve the purpose to make tabular data more easily comprehensible. The main tool to achieve this is to use geometric quantities—like lengths, areas, and angles—to represent numbers, since such geometric properties are more quickly interpretable for humans than abstract numbers. The most important types of graphical representations are:

[1] In most of Europe it is more common to use numbers for grades, with 1 being the best and 6 being the worst possible, while in the United States it is more common to use letters, with A being the best and F being the worst possible. However, there is an obvious mapping between the two scales. We chose numbers here to emphasize that nominal scales may use numbers and thus may look deceptively metric.

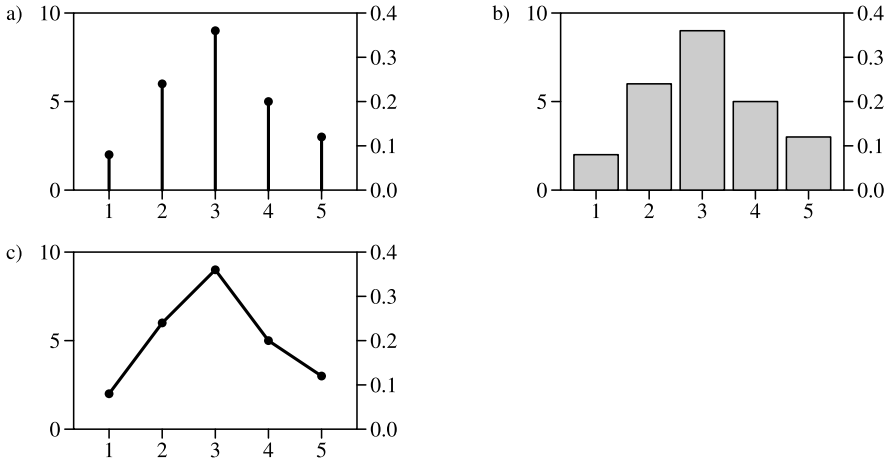
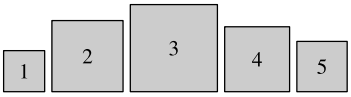


Fig. A.1 Pole (**a**) and bar chart (**b**) and frequency polygons (**c**) for the data shown in Table A.2

Fig. A.2 Area chart for the data shown in Table A.2



- **pole/stick/bar chart**
  Numbers, which may be, for instance, the frequencies of different attribute values in a sample, are represented by the lengths of poles, sticks, or bars. In this way a good impression especially of ratios can be achieved (see Figs. A.1a and b, in which the frequencies of Table A.2 are displayed).
- **area and volume charts**
  Area and volume charts are closely related to pole and bar charts: the difference is merely that they use areas and volumes instead of lengths to represent numbers and their ratios (see Fig. A.2, which again shows the frequencies of Table A.2). However, area and volume charts are usually less comprehensive (maybe except if the represented quantities are actually areas and volumes), since human beings usually have trouble comparing areas and volumes and often misjudge their numerical ratios. This can already be seen in Fig. A.2: only very few people correctly estimate that the area of the square for the value 3 (frequency 9) is three times as large as that of the square for the value 5 (frequency 3).
- **frequency polygons** and **line chart**
  A *frequency polygon* results if the ends of the poles of pole diagram are connected by lines, so that a polygonal course results. This can be advantageous if the attribute values have an inherent order and one wants to show the development of the frequency along this order (see Fig. A.1c). In particular, it can be used if numbers are to be represented that depend on time. This particular case is usually referred to as a *line chart*, even though the name is not exclusively reserved for this case.

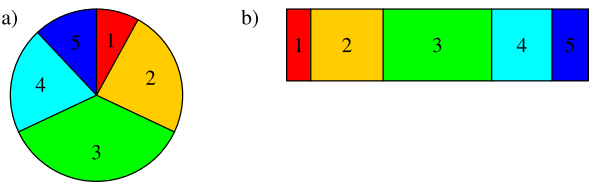


Fig. A.3 A pie chart (**a**) and a stripe chart (**b**) for the data shown in Table A.2

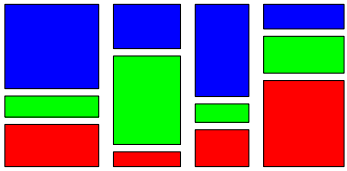


Fig. A.4 A mosaic chart for the contingency table of Table A.3

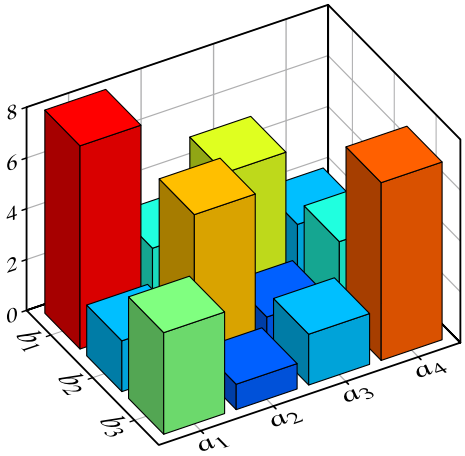


Fig. A.5 A bar chart for the contingency table of Table A.3

- **pie and stripe chart**
  Pie and stripe charts are particularly well suited if proportions or fractions of a total, for instance, relative frequencies, are to be displayed. In a pie chart proportions are represented by angles, and in a stripe chart by lengths (see Fig. A.3).

- **mosaic chart**
  Contingency tables (that is, two- or generally multidimensional frequency tables) can nicely be represented as mosaic charts. For the first attribute, the horizontal direction is divided like in a stripe diagram. Each section is then divided according to the second attribute along the vertical direction—again like in a stripe diagram (see Fig. A.4). Mosaic charts can have advantages over two-dimensional bar charts, because bars at the front can hide bars at the back, making it difficult to see their height, as shown in Fig. A.5. In principle, arbitrarily many attributes can be displayed by subdividing the resulting mosaic pieces alternatingly along the horizontal and vertical axis. However, even if one uses the widths of the gaps

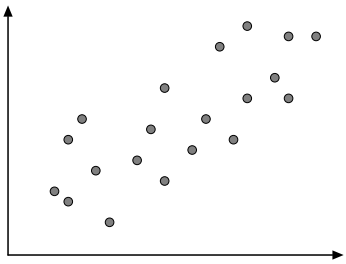


Fig. A.6 A simple scatter plot

and colors in order to help a viewer to identify attribute values, mosaic charts can easily become confusing if it is tried to use to many attributes.

- **histogram**
  In principle, a histogram looks like a bar chart, with the only difference that the domain of the underlying attribute is metric (numerical). As a consequence, it is usually impossible to simply enumerate the frequencies of the individual attribute values (because there are usually too many different values), but one has to form counting intervals, which are usually called *bins* or *buckets*. The width (or, if the domain is fixed, equivalently the number) of these bins has to be chosen by a user. All bins should have the same width, since histograms with varying bin widths are usually more difficult to read—for the same reasons why area charts are more difficult to interpret than bar charts (see above). In addition, a histogram may only provide a good impression of the data if an appropriate bin width has been chosen and onto which values the borders of the bins fall (see Sect. 4.3.1).
- **scatter plot**
  A scatter plot displays a two-dimensional data set of metric attributes by interpreting the sample values as coordinates of a point in a metric space (see Fig. A.6). A scatter plot is very well suited if one wants to see whether the two represented quantities depend on each other or vary independently (see also Sects. A.2.4 and 8.3).

Examples how graphical representations can be misleading—a property that is sometimes (all too often actually) exploited to convey a deceptively favorable or unfavorable impression, in particular in the press and in advertisements—can be found in the highly recommended books [6, 8].

A.2.3 Characteristic Measures for One-Dimensional Data

The goal of computing characteristic measures is to summarize the data set, that is, to capture characteristic and relevant properties in as few quantities as possible. There are basically three types of characteristic measures:

- **location measures**
  As their name clearly indicates, location measures specify the location of the (majority of) the data in the domain of an attribute by a single number or attribute values. Thus location measures summarize the data heavily.

- **dispersion measures**
  Given the value of a location measure, dispersion measures specify how much the data scatter around this value (how much they deviate from it) and thus characterize how well the location measure captures the location of the data.
- **shape measures**
  Given the values of a location and a dispersion measure, shape measures characterize the distribution of the data by comparing its shape to a reference shape. The most common reference shape is the normal distribution (see Sect. A.3.5.7).

In the following we study these characteristic measures, which will turn out to be very useful in inferential statistics (see Sect. A.4), in more detail.

A.2.3.1 Location Measures

As already mentioned, a location measure characterizes the location of the data in the domain of the underlying attribute (recall that we are currently concerned only with one-dimensional data sets) by a single attribute value. This value should be as representative for the data as possible. We may require, for example, that the sum of the deviations of the individual sample values from the value of the location measure should be as small as possible. The most important location measures are the *mode*, the *median* (also called the *central value*), and its generalization, the so-called *quantiles* and the *mean*, which is the most common location measure.

Mode An attribute value is called the (empirical) *mode* x^* of a data set if it is the value that occurs most frequently in the sample. As a consequence, it need not be uniquely determined, since there may be several values that have the frequency. Modes can be determined for any scale type, because the only operation needed to compute them is a test for equality. Therefore the mode is the most general location measure. However, for metric data, it is most of the time less well suited than other measures, because the usually large number of possible values of metric attributes obviously poses some problems. However, in many cases one can amend this situation by choosing the middle of the highest bar in an appropriate histogram as the mode of the distribution of a numerical attribute.

Median (Central Value) The (empirical) *median* or *central value* $\tilde{x}$ can be introduced as a value that minimizes the sum of the absolute deviations. That is, a median $\tilde{x}$ is any value that satisfies

$$\sum_{i=1}^{n} |x_i - \tilde{x}| = \min.$$

In order to find a value for $\tilde{x}$, we take the derivative of the left-hand side and equate the result to zero (since the derivative must vanish at the minimum). In this way we obtain

$$\sum_{i=1}^{n} \operatorname{sgn}(x_i - \tilde{x}) = 0,$$

where sgn is the sign function (which is -1 if its argument is negative, $+1$ if its argument is positive, and 0 if its argument is 0).[2] Therefore a median is a value that lies "in the middle of the data." That is, in the data set there are as many values greater than $\tilde{x}$ as smaller than $\tilde{x}$ (this justifies the expression *central value* as an alternative to *median*).

With the above characterization, the median is not always uniquely determined. For example, if all sample values are distinct, there is only a unique middle element if the sample size is odd. If it is even, there may be several values that satisfy the above defining equations. As an example, consider the data set $(1, 2, 3, 4)$. Any value in the interval $[2, 3]$ minimizes the sum of absolute deviations. In order to obtain a unique value, one usually defines the median as the arithmetic mean of the two sample values in the middle of the (sorted) data set in such a case. In the above example, this would result in $\tilde{x} = \frac{2+3}{2} = \frac{5}{2}$. Note that the median is always uniquely determined for even sample size if the two middle values are equal.

Formally the median is defined as follows: let $x = (x_{(1)}, \ldots, x_{(n)})$ be a sorted data set, that is, we have $\forall i, j : (j > i) \rightarrow (x_{(j)} \geq x_{(i)})$. Then

$$\tilde{x} = \begin{cases} x_{(\frac{n+1}{2})} & \text{if } n \text{ is odd,} \\ \frac{1}{2}\left(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}\right) & \text{if } n \text{ is even,} \end{cases}$$

is called the *(empirical) median* of the data set x.

The median can be computed for ordinal and metric attributes, because all it requires is a test for greater or less than, namely for sorting the sample values. For ordinal values, computing the arithmetic mean of the two middle values for even sample size is replaced by simply choosing one of them, thus eliminating the need for the computation. Note, however, that the above characterization of the median as the minimizer of the absolute deviations can, of course, not be used for ordinal attributes as they do not allow for computing differences. We used it here nevertheless in order to show the analogy to the mean, which is considered below.

Quantiles We have seen in the preceding section that the median is an attribute value such that half of the sample values are less than it, and the other half is greater. This idea can easily be generalized by finding an attribute value such that a certain fraction p, $0 < p < 1$, of the sample values is less than this attribute value (and a fraction of $1 - p$ of the sample values are greater). These values are called (empirical) p-quantiles. The median in particular is the (empirical) $\frac{1}{2}$-quantile of a data set.

Other important quantiles are the first, second, and third quartiles, for which $p = \frac{1}{4}$, $\frac{2}{4}$, and $\frac{3}{4}$, respectively, of the data set are smaller (therefore the median is also identical to the second quartile), and the deciles (k tenths of the data set are smaller) and the percentiles (k hundredths of the data set are smaller).

Note that for metric attributes, it may be necessary, depending on the sample size and the exact sample values, to introduce adaptations that are analogous to the computation of the arithmetic mean of the middle values for the median.

[2] Note that, with the standard definition of the sign function, this equation cannot always be satisfied. In this case one confines oneself with the closest possible approximation to zero.

Mean While the median minimizes the *absolute deviations* of the sample values, the (empirical) mean $\bar{x}$ can be defined as the value that minimizes the sum of the squares of the deviations of the sample values. That is, the mean is the attribute value that satisfies

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \min.$$

In order to find a value for $\bar{x}$, we take the derivative of the left-hand side and equate it to zero (since the derivative must vanish at a minimum). In this way we obtain

$$\sum_{i=1}^{n}(x_i - \bar{x}) = \sum_{i=1}^{n} x_i - n\bar{x} = 0,$$

and thus

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Therefore the mean of a sample is the arithmetic mean of the sample values.

Even though the mean is the most commonly used location measure for metric attributes (note that it cannot be applied for ordinal attributes as it requires summation and thus an interval scale), the median should be preferred for

- few measurement values,
- asymmetric (skewed) data distributions, and
- likely presence of outliers,

since the median is more robust in these cases and conveys a better impression of the data. In order to make the mean more robust against outliers, it is often computed by eliminating the largest and the smallest sample values (a typical procedure for averaging the ratings of the judges in sports events) or even multiple extreme values, like all values before the 1st and beyond the 99th percentile.

A.2.3.2 Dispersion Measures

As already mentioned in the general overview of characteristic measures, dispersion measures specify how broadly the sample values scatter around a location parameter. Hence they characterize how well the data are captured by the location parameter. The reason for introducing dispersion measures is that a location measure alone does not tell us anything about the size of the deviations and thus one may be deceived about the true situation. This possibility is captured well in the old statistics joke:

> A man with his head in the freezer and feet in the oven is *on the average* quite comfortable.

Range The range of a data set is simply the difference between that largest and the smallest sample value:

$$R = x_{\max} - x_{\min} = \max_{i=1}^{n} x_i - \min_{i=1}^{n} x_i .$$

The range is a very intuitive dispersion measure. However, it is very sensitive against outliers, which tend to corrupt one or even both of the values it is computed from.

Interquantile Range The difference between the (empirical) $(1 - p)$- and the (empirical) p-quantiles of a data set is called the p-interquantile range, $0 < p < \frac{1}{2}$. Commonly used interquantile ranges are the interquartile range ($p = \frac{1}{4}$, difference between the third and first quartiles), the interdecile range ($p = \frac{1}{10}$, difference between the 9th and the 1st deciles), and the interpercentile range ($p = \frac{1}{100}$, difference between the 99th and the 1st percentiles). For small p, the p-interquantile range transfers the idea to make the mean more robust by eliminating extreme values to the range.

Mean Absolute Deviation The mean absolute deviation is the arithmetic mean of the absolute deviations of the sample values from the (empirical) median or mean:

$$d_{\tilde{x}} = \frac{1}{n} \sum_{i=1}^{n} |x_i - \tilde{x}|$$

is the mean absolute deviation from the median, and

$$d_{\bar{x}} = \frac{1}{n} \sum_{i=1}^{n} |x_i - \bar{x}|$$

is the mean absolute deviation from the mean. It is always $d_{\tilde{x}} \leq d_{\bar{x}}$, because the median minimizes the sum and thus also the mean of the absolute deviations.

Variance and Standard Deviation In analogy to the absolute deviation, one may also compute the mean squared deviation. (Recall that the mean minimizes the sum of the squared deviations.) However, instead of

$$m^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2,$$

it is more common to employ

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

as a dispersion measure, which is called the *(empirical) variance* of the sample. The reason for the value $n - 1$ in the denominator is provided by inferential statistics, in which the characteristic measures of descriptive statistics are related to certain parameters of probability distributions and density functions (see Sect. A.4).

A detailed explanation will be provided in Sect. A.4.2, which deals with parameter estimation (unbiasedness of an estimator for the variance of a normal distribution).

The positive square root of the variance, that is,

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

is called the *(empirical) standard deviation* of the sample.

Not that the (empirical) variance can often be computed more conveniently with formula that is obtained with the following transformation:

$$\begin{aligned} s^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i^2 - 2x_i\bar{x} + \bar{x}^2\right) \\ &= \frac{1}{n-1}\left(\sum_{i=1}^{n}x_i^2 - 2\bar{x}\sum_{i=1}^{n}x_i + \sum_{i=1}^{n}\bar{x}^2\right) = \frac{1}{n-1}\left(\sum_{i=1}^{n}x_i^2 - 2n\bar{x}^2 + n\bar{x}^2\right) \\ &= \frac{1}{n-1}\left(\sum_{i=1}^{n}x_i^2 - n\bar{x}^2\right) = \frac{1}{n-1}\left(\sum_{i=1}^{n}x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n}x_i\right)^2\right). \end{aligned}$$

The advantage of this formula is that it allows us to compute both the mean and the variance of a sample with one pass through the data, by computing the sum of sample values and the sum of their squares. A computation via the original formula, on the other hand, needs two passes: in the first pass the mean is computed, and in the second pass the variance is computed from the sum of the squared deviations.

A.2.3.3 Shape Measures

If one plots a histogram of observed metric data, one often obtains a bell shape. In practice, this bell shape usually differs more or less from the reference of an ideal Gaussian bell curve (normal distribution, see Sect. A.3.5.7). For example, the empirical distribution, as shown by the histogram, is asymmetric or differently curved. With shape measures one tries to capture these deviations.

Skewness The skewness or simply skew α_3 states whether, and if yes, by how much a distribution differs from a symmetric distribution.[3] The skewness is computed as

$$\alpha_3 = \frac{1}{n \cdot s^3}\sum_{i=1}^{n}(x_i - \bar{x})^3 = \frac{1}{n}\sum_{i=1}^{n}z_i^3 \quad \text{with} \quad z_i = \frac{x_i - \bar{x}}{s},$$

that is, z is the z-score normalized variable. For a symmetric distribution, $\alpha_3 = 0$. If the skew is positive, the distribution is steeper on the left, and if it is negative, the distribution is steeper on the right (see Fig. A.7).

[3] The index 3 indicates that the skew is the 3rd moment of the sample around the mean—see the defining formula, which employs a third power.

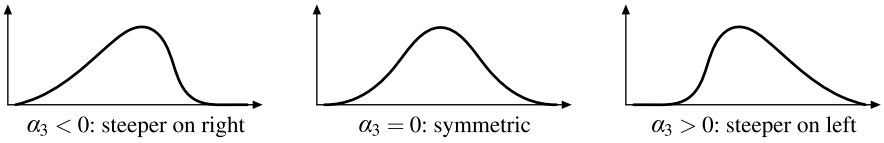


Fig. A.7 Illustration of the shape measure *skewness*

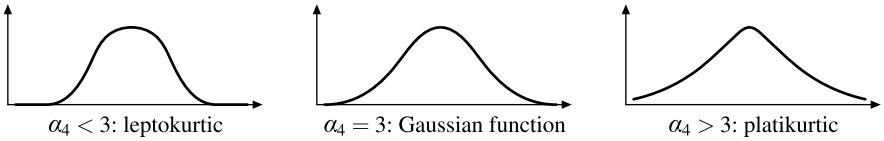


Fig. A.8 Illustration of the shape measure *kurtosis*

Kurtosis The kurtosis α_4 describes how strongly a bell-shaped distribution is curved and thus how steep the peak is[4] (compared to the ideal Gaussian bell curve). The kurtosis is computed as

$$\alpha_4 = \frac{1}{n \cdot s^4} \sum_{i=1}^{n} (x_i - \bar{x})^4 = \frac{1}{n} \sum_{i=1}^{n} z_i^4 \quad \text{with} \quad z_i = \frac{x_i - \bar{x}}{s}.$$

An ideal Gaussian function as a kurtosis of 3. If the kurtosis is smaller than 3, the distribution is more peaked (leptokurtic) than a Gaussian bell curve; if it is greater than 3, the distribution is less peaked (platikurtic) than a Gaussian bell curve (see Fig. A.8). Sometimes the kurtosis is defined as $\alpha_4 - 3$, so that the Gaussian functions has a kurtosis of 0 and the sign indicates whether the distribution under consideration is more (negative, leptokurtic) or less peaked (positive, platikurtic) than a Gaussian function.

A.2.3.4 Box Plots

Some characteristic measures, namely the median, the mean, the range, and the interquartile range are often displayed jointly in a so-called box plot (see Fig. A.9): the outer lines show the range and the box in the middle, which gives this diagram form its name, indicates the interquartile range. Into the box the median is drawn as a solid, and the mean as a dashed line (alternatively mean and median can be drawn in different colors). The range may be replaced by the interpercentile range. In this case the extreme values outside this range are depicted as individual dots. Sometimes the box that represents the interquartile range is drawn constricted at the location of the mean in order to emphasize the location of the mean. Obviously this simple diagram provides a good compact impression of the rough shape of the data distribution.

[4]The index 4 indicates that the kurtosis is the 4th moment around the mean—see the defining formula, which employs a fourth power.

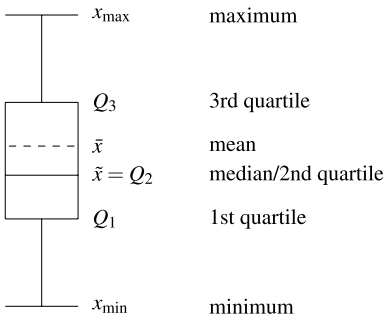


Fig. A.9 A box plot is a simple diagram that captures the most important characteristic measures

A.2.4 Characteristic Measures for Multidimensional Data

Several of the characteristic measures that we introduced in the preceding section for one-dimensional data sets can easily be transferred to multidimensional data by simply executing the computations with vectors instead of scalars (simple numbers). Here we consider as examples the transfer of the mean and the variance, which will lead us to the covariance (matrix). By normalizing the covariance, we obtain the important measure of the correlation coefficient.

Mean For multidimensional data, the mean turns into the vector mean of the data points. For example, for two-dimensional data, we have

$$\overline{(x, y)} = \frac{1}{n} \sum_{i=1}^{n} (x_i, y_i) = (\bar{x}, \bar{y}).$$

It should be noted that one obtains the same result if one forms the vector that consists of the means of the individual attributes. Hence, for computing the mean, the attributes can be treated independently.

Covariance and Correlation It is equally simple to transfer dispersion measure *variance* to multidimensional data by simply executing the computations with vectors instead of scalars. The only problem consists in squaring the differences between the sample data points and the mean vector, since these differences are now vectors. In order to compute this square, the so-called outer product or matrix product of the difference vector with itself is computed. This outer product is defined as $\mathbf{v}\mathbf{v}^\top$ (where $\top$ denotes a transposed vector) and yields a square matrix. These matrices (one for each sample data point) are summed and (like the standard scalar variance) divided by $n - 1$, where n is the sample size. The result is a square, symmetric,[5] and positive definite[6] matrix, the so-called covariance matrix. For two-dimensional data, the covariance matrix is defined as

[5] A square matrix $\mathbf{M} = (m_{ij})_{1 \le i \le m, 1 \le j \le m}$ is called symmetric if $\forall i, j : m_{ij} = m_{ji}$.

[6] A matrix $\mathbf{M}$ is called *positive definite* if for all vectors $\mathbf{v} \neq \mathbf{0}$, $\mathbf{v}^\top \mathbf{M} \mathbf{v} > 0$.

$$\Sigma = \frac{1}{n-1}\sum_{i=1}^{n}\left(\begin{pmatrix}x_i\\y_i\end{pmatrix}-\begin{pmatrix}\bar{x}\\\bar{y}\end{pmatrix}\right)\left(\begin{pmatrix}x_i\\y_i\end{pmatrix}-\begin{pmatrix}\bar{x}\\\bar{y}\end{pmatrix}\right)^{\top} = \begin{pmatrix}s_x^2 & s_{xy}\\ s_{xy} & s_y^2\end{pmatrix},$$

where

$$s_x^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right) \qquad \text{(variance of } x\text{)},$$

$$s_y^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right) \qquad \text{(variance of } y\text{)},$$

$$s_{xy} = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}\right) \qquad \text{(covariance of } x \text{ and } y\text{)}.$$

In addition to the variances of the individual attributes, a covariance matrix contains an additional quantity, the so-called *covariance*. It yields information about the strength of the (linear) dependence of the two attributes. However, since its value also depends on the variance of the individual attributes, it is normalized by dividing it by the standard deviations of the individual attributes, which yields the so-called correlation coefficient (more precisely, Pearson's product moment correlation coefficient; see Sect. 4.4 for alternatives, especially for ordinal attributes),

$$r = \frac{s_{xy}}{s_x s_y}.$$

It should be noted that the correlation coefficient is identical to the covariance of the two attributes if their values are first normalized to mean 0 and standard deviation 1. The correlation coefficient has a value between -1 and $+1$ and characterizes the strength of the *linear dependence* of two metric quantities: if all data points lie exactly on an ascending straight line, its value is $+1$. If they lie exactly on a descending straight line, its value is -1. In order to convey an intuition of intermediate values, Fig. A.10 shows some examples.

Note that it does not mean that two measures are (stochastically) independent if their correlation coefficient vanishes. For example, if the data points lie symmetrically on a parabola, the correlation coefficient is $r = 0$. Nevertheless there is, of course, a clear and exact functional dependence of the two measures. If the correlation coefficient is zero, this only means that this dependence is not linear.

Since the covariance and correlation describe the linear dependence of two measures, it is not surprising that they can be used to fit a straight line to the data, that is, to determine a so-called regression line. This line is defined as

$$(y - \bar{y}) = \frac{s_{xy}}{s_x^2}(x - \bar{x}) \quad \text{or} \quad y = \frac{s_{xy}}{s_x^2}(x - \bar{x}) + \bar{y}.$$

The regression line can be seen as kind of mean function, which assigns a mean of the y-values to each of the x-values (conditional mean). This interpretation is supported by the fact that the regression line minimizes the sum of the squares of the deviations of the data points (in y-direction), just like the mean. More details about regression and the method of least squares, together with generalizations to larger function classes, can be found in Sect. 8.3.

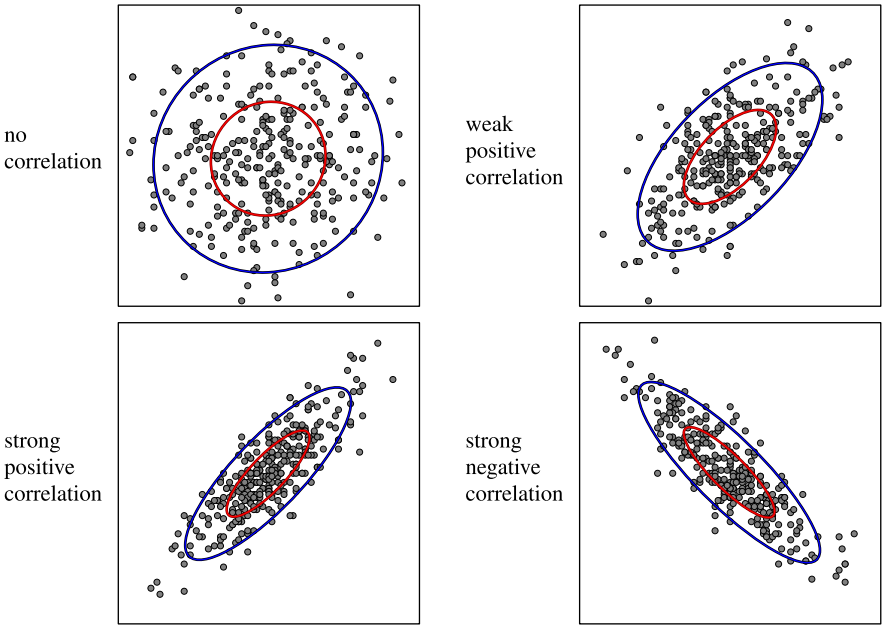


Fig. A.10 Illustration of the meaning of the correlations coefficient

A.2.5 Principal Component Analysis

Correlations between the attributes of a data set can be used to reduce its dimension: if an attribute is (strongly) correlated with another, then this attribute is essentially a linear function of the other (plus some noise). In such a case it is often sufficient to consider only one of the two attributes, since the other can be reconstructed (approximately) via the regression line. However, this approach has the disadvantage that it is not trivial to decide which of several correlated attributes should be kept and which can be discarded.

A better approach to reduce the dimension of a data set is the so-called principal component analysis (PCA; see also Sect. 4.3.2.1). The basic idea of this procedure is not to select a subset of the features of the data set, but to construct a small number of new features as linear combinations of the original ones. These new quantities are supposed to capture the greater part of the information in the data set, where the information content is measured by the (properly normalized) variance: the larger the variance, the more important the (constructed) feature.

In order to find the linear combinations that define the new features, the data is first normalized to mean 0 and standard deviation 1 in all original attributes, so that the scale of the attributes (that is, for example, the units in which they were measured) does not influence the result. In the next step one tries to find a new basis for the data space, that is, perpendicular directions. This is done in such a way that the first direction is the one in which the (normalized) data exhibits the largest

variance. The second direction is the one which is perpendicular to the first and in which the data exhibits the largest variance among all directions perpendicular to the first, and so on. Finally, the data is transformed to the new basis of the data space, and some of the constructed features are discarded, namely those, for which the transformed data exhibits the lowest variances. How many features are discarded is decided based on the sum of the variances of the kept features relative to the total sum of the variance of all features.

Formally, the perpendicular directions referred to above can be found with a mathematical method that is known as principal axes transformation. This transformation is applied to the correlation matrix, that is, the covariance matrix of the data set that has been normalized to mean 0 and standard deviation 1 in all features. That is, one finds a rotation of the coordinate system such that the correlation matrix becomes a diagonal matrix. The elements of this diagonal matrix are the variances of the data set w.r.t. the new basis of the data space, while all covariances vanish. This is also a fundamental goal of principal component analysis: one wants to obtain features that are linearly independent. As is well known from linear algebra (see, for example, [5, 11]), a principal axes transformation consists basically in computing the eigenvalues and eigenvectors of a matrix. The eigenvalues show up on the diagonal of the transformed correlation matrix, the eigenvectors (which can be obtained in parallel with appropriate methods) indicate the desired directions in the data space. The directions w.r.t. which the data is now described are selected based on the eigenvalues, and finally the data is projected to the subspace chosen in this way.

The following physical analog may make the idea of principal axes transformation clearer: how a solid body reacts to a rotation around a given axis, can be described by the so-called tensor of inertia [9]. Formally, this tensor is a symmetric 3×3 matrix,

$$\Theta = \begin{pmatrix} \Theta_{xx} & \Theta_{xy} & \Theta_{xz} \\ \Theta_{xy} & \Theta_{yy} & \Theta_{yz} \\ \Theta_{xz} & \Theta_{yz} & \Theta_{zz} \end{pmatrix}.$$

The diagonal elements of this matrix are the *moments of inertia* of the body w.r.t. the axes that pass through its center of gravity[7] and are parallel to the axes of the coordinate system that we use to describe the rotation. The remaining (off-diagonal) elements of the matrix are called deviation moments and describe the forces that act perpendicular to the axes during the rotation.[8] However, for any solid body, regardless of its shape, there are three axes w.r.t. which the deviation moments vanish, the

[7]The inertia behavior of axes that do not pass through the center of gravity can easily be described with Steiner's law. However, this goes beyond the scope of this discussion, see standard textbooks on theoretical mechanics like [9] for details.

[8]These forces result from the fact that generally the vector of angular momentum is not parallel to the vector of angular velocity. However, this again leads beyond the scope of this discussion.

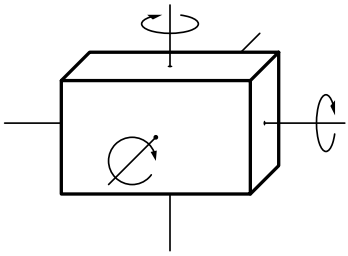


Fig. A.11 The principal axes of inertia of a box

so-called *principal axes of inertia*.[9] As an example, Fig. A.11 shows the principal axes of inertia of a box. The principal axes of inertia are always perpendicular to each other. In the coordinate system that is spanned by the principal axes of inertia, the tensor of inertia is a diagonal matrix.

Formally, the principal axes of inertia are found by carrying out a principal axes transformation of the tensor of inertia (given w.r.t. an arbitrary coordinate system): its eigenvectors are the directions of the principal axes of inertia.

In the real world, the deviation moments cause shear forces, which lead to vibrations and jogs in the bearings of the axis. Since such vibrations and jogs naturally lead to quick abrasion of the bearings, it is tried to minimize the deviation moments. As a consequence, a car mechanic who balances a wheel can be seen as carrying out a principal axes transformation (though not in mathematical form), because he/she tries to equate the rotation axis with a principal axis of inertia. However, he/she does not do so by changing the direction of the rotation axis, as this is fixed in the wheel. Rather, he/she changes the mass distribution by adding, removing, and shifting small weights so that the deviation moments vanish.

Based on this analog, we may say that a statistician looks, in the first step of a principal component analysis, for axes around which a mass distribution with unit weights at the locations of the data points can be rotated without vibrations or jogs in the bearings. Afterwards, he selects a subset of the axes by removing those axes around which the rotation needs the most energy, that is, those axes for which the moments of inertia are largest (in the direction of these axes, the variance is smallest, and perpendicularly to them, the variance is largest).

Formally, the axes are selected via the percentage of explained variance. It can be shown that the sum of the eigenvalues of a correlation matrix equals the dimension m of the data set, that is, it is equal to the number of features (see, for example, [5, 11]). In this case it is plausible to define that the proportion of the total variance that is captured by the jth principal axis as

$$p_j = \frac{\lambda_j}{m} \cdot 100\%,$$

where λ_j is the eigenvalue corresponding to the jth principal axis.

[9] Note that a body may possess more than three axes w.r.t. which the deviation moments vanish. For a sphere with homogeneous mass distribution, for example, any axis that passes through the center is such an axis. However, any body, regardless of its shape, has at least three such axes.

Let $p_{(1)}, \ldots, p_{(m)}$ a sequence of these percentages that is sorted descendingly. For this sequence, one determines the smallest value k such that

$$\sum_{j=1}^{k} p_{(j)} \geq \alpha \cdot 100\%$$

with a proportion α that has to be chosen by a user (for example, $\alpha = 0.9$). The corresponding k principal axes are chosen as the new features, and the data points are projected to them. Alternatively, one may specify to how many features one desires to reduce the data set to and then chooses the axes following a descending proportion $\frac{\lambda_j}{m}$. In this case the above sum provides information about how much information contained in the original data is lost.

A.2.5.1 Example of a Principal Component Analysis

As a very simple example of a principal component analysis, we consider the reduction of a data set of two correlated quantities to a one-dimensional data set. Let the following data be given:

x	5	15	21	29	31	43	49	51	61	65
y	33	35	24	21	27	16	18	10	4	12

Even a quick look at this table already reveals that the two features are strongly correlated, an even clearer impression is provided by Fig. A.12. As a consequence, it can be expected that this data set can be reduced to a one-dimensional data set without much loss of information.

In the first step we compute the correlation matrix. To this end we normalize the data to mean 0 and standard deviation 1 and compute the covariance matrix of the normalized data. Thus we obtain the normalized data set

x'	−1.6	−1.1	−0.8	−0.4	−0.3	0.3	0.6	0.7	1.2	1.4
y'	1.3	1.5	0.4	0.1	0.7	−0.4	−0.2	−1.0	−1.6	−0.8

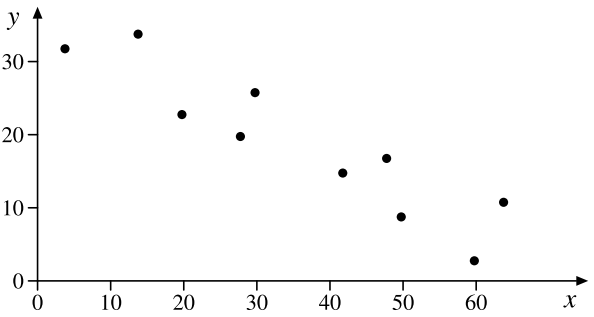


Fig. A.12 Data of an example for the principal component analysis as a scatter plot

The covariance of these normalized features is the correlation coefficient

$$r = s_{x'y'} = \frac{1}{9}\sum_{i=1}^{10} x_i\, y_i = \frac{-8.28}{9} = -\frac{23}{25} = -0.92.$$

Therefore the correlation matrix is

$$\Sigma = \frac{1}{9}\begin{pmatrix} 9 & -8.28 \\ -8.28 & 9 \end{pmatrix} = \begin{pmatrix} 1 & -\frac{23}{25} \\ -\frac{23}{25} & 1 \end{pmatrix}.$$

(Note that the diagonal elements of this matrix are the correlation coefficients of the features with themselves. Hence they are necessarily 1.)

For this correlation matrix, we have to carry out a principal axes transformation. In order to do so, we compute the eigenvalues and eigenvectors of this matrix, that is, those values λ_i and vectors $\mathbf{v}_i$, $i = 1, 2$, for which

$$\Sigma\mathbf{v}_i = \lambda_i\mathbf{v}_i \quad \text{or} \quad (\Sigma - \lambda_i\mathbf{I})\mathbf{v}_i = \mathbf{0},$$

where $\mathbf{I}$ is the unit matrix. In order to compute the eigenvalues λ_i, we rely here on the simple method of finding the roots of the characteristic polynomial[10]

$$c(\lambda) = |\Sigma - \lambda\mathbf{E}| = (1-\lambda)^2 - \frac{529}{625}.$$

The roots of this polynomial are the eigenvalues

$$\lambda_{1/2} = 1 \pm \sqrt{\frac{529}{625}} = 1 \pm \frac{23}{25}; \quad \text{hence} \quad \lambda_1 = \frac{48}{25} \text{ and } \lambda_2 = \frac{2}{25}.$$

The corresponding eigenvectors are

$$\mathbf{v}_1 = \left(\frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{2}\right) \quad \text{and} \quad \mathbf{v}_2 = \left(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right)$$

that can be found by simply inserting the eigenvalues into

$$(\Sigma - \lambda_i\mathbf{E})\mathbf{v}_i = \mathbf{0}$$

and solving the resulting underdetermined linear equation system.[11] Therefore the principal axes transformation is given by the orthogonal matrix that has the eigenvectors as its columns, that is,

$$\mathbf{T} = \begin{pmatrix} \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \\ -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{pmatrix}.$$

[10] Note, however, that for larger matrices, this method is numerically unstable and should be replaced by some other approach; see, for example, [13].

[11] Note that for two variables, due to the special form of the characteristic polynomial, the eigenvectors are always exactly these two vectors (for the normalized data), independent of original data.

This matrix satisfies

$$\mathbf{T}^\top \Sigma \mathbf{T} = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}.$$

However, in order to obtain simple numbers for the transformed data, we multiply the data points with $\sqrt{2}\mathbf{T}^\top$ instead of $\mathbf{T}^\top$, that is, we compute

$$\begin{pmatrix} x'' \\ y'' \end{pmatrix} = \sqrt{2} \cdot \mathbf{T}^\top \cdot \begin{pmatrix} x' \\ y' \end{pmatrix}.$$

With the help of this transformation, the data points are projected to the principal axes. Intuitively, the multiplication with the matrix is equivalent to dropping perpendiculars from each data point to the axes that are given by the eigenvectors and using the distances of the feet of the perpendiculars from the origin as new coordinates. The transformed data set is

x''	−2.9	−2.6	−1.2	−0.5	−1.0	0.7	0.8	1.7	2.8	2.2
y''	−0.3	0.4	−0.4	−0.3	0.4	−0.1	0.4	−0.3	−0.4	0.6

This data set describes the data points in the space that is spanned by the principal axes. Since the y'' values vary only fairly little compared to the x'' values (without the factor $\sqrt{2}$ that we added to the transformation, the eigenvalues $\lambda_1 = \frac{23}{25}$ and $\lambda_2 = \frac{2}{25}$ would be the variances in these dimensions, with the factor they are twice as large), we can confine ourselves to considering the x'' values. Thus we have reduced the data set to one dimension, without losing much information.

Note that the final data set shown above can be obtained directly from the original data set by the transformation

$$\begin{pmatrix} x'' \\ y'' \end{pmatrix} = \sqrt{2} \cdot \mathbf{T}^\top \cdot \begin{pmatrix} s_x^{-1} & 0 \\ 0 & s_y^{-1} \end{pmatrix} \cdot \left(\begin{pmatrix} x \\ y \end{pmatrix} - \begin{pmatrix} \bar{x} \\ \bar{y} \end{pmatrix} \right),$$

which combines the normalization and the projection of the data to the principal axes. This clearly demonstrates that the new features are merely linear combinations of the original features—as stated at the beginning.

A.3 Probability Theory

Inferential statistics, as it is considered in Sect. A.4, is firmly based on probability theory. It uses notions and methods from probability theory in order to obtain conclusions and judgments about the data generating process. Therefore this section recalls some essential notions and theorems of probability theory.

A.3.1 Probability

Probability theory is concerned with **random events**. That is, it is known what specific events can occur in principle, but it is uncertain which of the possible events

will actually occur in any given instance. Examples are the throw of a coin or the cast of die. In probability theory a numerical quantity, called *probability*, is assigned to random events, which is intended to capture the chance or the tendency of the event to actually occur, at least in relation to the other possible outcomes.

A.3.1.1 Intuitive Notions of Probability

In order to achieve a first intuition, we start by considering the classical definition of probability. This definition is actually a procedure to compute probabilities and was developed out of analyses of the possible outcomes of gambles:

The probability of an event A is the ratio of the number N_A of favorable events (where an event is called favorable for A if A occurs with it) to the total number N of mutually exclusive, equally possible events:

$$P(A) = \frac{N_A}{N}.$$

That the events are equally possible is usually derived from symmetry considerations. The mutually exclusive, equally possible events are called **elementary events**. Together they form the **sample space**. Other random events can be represented as combinations of elementary events, generally as sets of elementary events. An essential set of tools for determining probabilities on the basis of the above definition are the methods of **combinatorics**, which can be used to count the number of all possible and the number of favorable events (see Sect. A.3.2.1).

As an example, we consider a symmetric die consisting of homogeneous material. The sample space consists of the elementary events "The number x was cast," $x \in \{1, 2, 3, 4, 5, 6\}$. The event "An even number was cast" consists of the elementary events "The number x was cast," $x \in \{2, 4, 6\}$, and thus occurs in three out of six equally possible cases. Therefore its probability is $\frac{3}{6} = \frac{1}{2}$.

The classical notion of probability can be generalized by dropping the requirement that the elementary events are equally possible. Instead, a sample space is defined, and each elementary event is associated with an **elementary probability**. Obviously, the classical notion then results as a special case of equal elementary events, demonstrating the term "equally possible" is actually a trick to avoid having to speak of "equally probable," which would result in a circular definition.

Another intuitive concept related to the probability of an event is the relative frequency of the event. Suppose that some experiment with a random outcome is executed n times under equal conditions. If A is a random event, then A either occurs or does not occur in each execution of the experiment. The number $f_n(A)$ of executions of the experiment in which A occurs is called the **absolute frequency**, and the ratio $r_n(A) = \frac{f_n(A)}{n}$ is called the **relative frequency** of A (cf. page 305).

In [12] von Mises tried to define the probability $P(A)$ of an event as the limit that is approached by the relative frequencies as the sample size n goes to infinity, that is,

$$P(A) = \lim_{n \to \infty} r_n(A).$$

However, this does not work. Even though it is necessary for the above definition to be valid, it is not possible to ensure that sequences of experiments occur in which

$$\forall n \geq n_0(\varepsilon): \quad P(A) - \varepsilon \leq r_n(A) \leq P(A) + \varepsilon$$

does not hold, regardless of how large n_0 is chosen. (Even though it is highly unlikely, it is not impossible that repeated throws of a die result only in, say, ones.) Nevertheless the intuition of a probability as a relative frequency is helpful, especially if it is interpreted as our estimate of the relative frequency of the event (in future executions of the same experiment).

As an example, consider the sex of a newborn child. Usually the number of girls roughly equals the number of boys. Hence the relative frequency of a girl being born is equal to the relative frequency of a boy being born. Therefore we say that the probability that a girl is born (and also the probability that a boy is born) equals $\frac{1}{2}$. Note that this probability cannot be derived from considerations of symmetry (as we can derive the probabilities of heads and tails when throwing a coin).[12]

A.3.1.2 The Formal Definition of Probability

The classical notion of probability and the interpretation as a relative frequency are deeply rooted in our intuition. However, modern mathematics is based on the axiomatic method, because one has realized that relying too much on intuition can introduce hidden assumptions, which can reduce the generality of obtained results or even invalidate them. The axiomatic method abstracts from the meaning of the objects that one considers. It takes the objects as given and having initially no other properties than their identity (that is, we can somehow distinguish one object from another). Then it studies merely the structure of the relations between these objects as they follow from the axioms (basic statements) that are laid down.

As a consequence, modern probability theory is also based on an axiomatic approach, which relies on **Kolmogorov's axioms** [7]. In this approach an event is simply a set of elementary events, which are distinguishable, that is, which have an identity. A probability is a number that is assigned to such events, so that the resulting system of numbers satisfies certain conditions, which are specified in the axioms. However, before we take a look at these axioms, we define the more basic notions of event algebra and σ-algebra, which are referred to in the axioms.

Definition A.1 Let Ω be a base set of elementary events (the sample space). Any subset $E \subseteq \Omega$ is called an **event**. A system $\mathcal{S} \subseteq 2^{\Omega}$ of events is called an **event algebra** iff

1. $\mathcal{S}$ contains the **certain event** Ω and the **impossible event**.
2. If an event A belongs to $\mathcal{S}$, then the event $\overline{A} = \Omega - A$ also belongs to S.

[12] It should be noted, though, that explanations from evolutionary biology for the usually equal probabilities of the two sexes indeed make references to symmetry.

3. If events A and B belong to $\mathcal{S}$, then the events $A \cap B$ and $A \cup B$ also belong to $\mathcal{S}$.

In addition the following condition may be satisfied:

3′. If for all $i \in \mathbb{N}$, the event A_i belongs to $\mathcal{S}$,
then the events $\bigcup_{i=1}^{\infty} A_i$ and $\bigcap_{i=1}^{\infty} A_i$ also belong to $\mathcal{S}$.

In this case $\mathcal{S}$ is called a σ**-algebra**.

The notion of a *probability* is now defined by the following axioms:

Definition A.2 (Kolmogorov's axioms) Let $\mathcal{S}$ be an event algebra that is defined on a finite sample space Ω.

1. The **probability** $P(A)$ of an event $A \in \mathcal{S}$ is a uniquely determined, nonnegative real number, which can be at most one, that is, $0 \leq P(A) \leq 1$.
2. The certain event Ω has probability 1, that is, $P(\Omega) = 1$.
3. **Additivity**: If A and B are two mutually exclusive events (that is, $A \cap B = \emptyset$), then $P(A \cup B) = P(A) + P(B)$.

If the sample space Ω has infinitely many elements, $\mathcal{S}$ must be a σ-algebra, and Axiom 3 must be replaced by:

3′. **Extended additivity**: If $A_1, A_2, \ldots$ are countably infinitely many, pairwise mutually exclusive events, then

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} A_i .$$

Hence the probability $P(A)$ can be seen as a function that is defined on an event algebra or on a σ-algebra and that has certain properties. From the above definition several immediate consequences follow:

1. For every event A, we have $P(\overline{A}) = 1 - P(A)$.
2. The impossible event has probability 0, that is, $P(\emptyset) = 0$.
3. From $A \subseteq B$ it follows that $P(A) \leq P(B)$.
4. For arbitrary events A and B, we have
$P(A - B) = P(A \cap \overline{B}) = P(A) - P(A \cap B)$.
5. For arbitrary events A and B, we have
$P(A \cup B) = P(A) + P(B) - P(A \cap B)$.
6. Since $P(A \cap B)$ cannot be negative, it follows that
$P(A \cup B) \leq P(A) + P(B)$.
7. By simple induction we infer from the additivity axiom:
if $A_1, \ldots, A_m$ are (finitely many) pairwise mutually exclusive events, we have
$P(A_1 \cap \cdots \cap A_m) = \sum_{i=1}^{m} P(A_i)$.

Kolmogorov's system of axioms is consistent, because there are actual systems that satisfy these axioms. This system of axioms allows us to construct probability theory as part of measure theory and to interpret a probability as a nonnegative, normalized, additive set function and thus as a measure (in the sense of measure theory).

Definition A.3 Let Ω be a set of elementary events, $\mathcal{S}$ a σ-algebra defined on a sample space Ω, and P a probability that is defined on $\mathcal{S}$. Then the triple $(\Omega, \mathcal{S}, P)$ is called a **probability space**.

A.3.2 Basic Methods and Theorems

After we have defined probability, we now turn to basic methods and theorems of probability theory, which we explain with simple examples. Among these are the computation of probabilities with combinatorial and geometrical approaches, the notions of conditional probability and (conditionally) independent events, the product law, the theorem of total probability, and finally the very important Bayes' rule.

A.3.2.1 Combinatorial Methods

As already mentioned, on the basis of the classical definition of probability, combinatorics provides many tools to compute probabilities. We confine ourselves to a simple example, namely the famous birthday problem:

Let m people be randomly chosen. What is the probability of the event A_m that at least two of these people have their birthday on the same day (day and month in a year, but not the same year)? In order not to over-complicate things, we neglect leap years. In addition, we assume that every day of a year is equally possible as a birthday.[13] It is immediately clear that for $m \geq 366$, at least two people must have their birthday on the same day. Hence we have

$$\forall m \geq 366: \quad P(A_m) = 1.$$

For $m \leq 365$, we consider the complementary event $\overline{A_m}$, which occurs if all m people have their birthday on different days. This switch to the complementary event is a very common technique when computing with probabilities, which often leads to considerable simplifications. If we number the people, the first person may have its birthday on 365, the second on 364 (since the birthday of the first is excluded), for the mth, $365 - m + 1$ days (since the birthdays of the $m - 1$ preceding people are excluded. This yields the number of favorable cases for $\overline{A_m}$. In all, there are 365^m possible cases (because any person could have its birthday on any day of the year). Therefore we have

$$P(\overline{A_m}) = \frac{365 \cdot 364 \cdots (365 - m + 1)}{365^m},$$

and consequently

$$P(A_m) = 1 - \frac{365 \cdot 364 \cdots (365 - m + 1)}{365^m}.$$

[13] This is not quite correct, since surveys show that the frequency of births is not quite uniformly distributed over the year, but a reasonable approximation.

The surprising property of this formula is that for m as low as 23, we already have $P(A_{23}) \approx 0.507$. Hence for 23 or more people, it is more likely that two people have their birthday on the same day than that all have their birthdays on different days.

A.3.2.2 Geometric Probabilities

Geometric probabilities are a generalization of the classical definition of probability. One no longer counts the favorable and all possible cases and then forms their ratio. Rather, the counts are replaced by geometric quantities, like lengths or areas.

As an example, we consider the game franc-carreau as it was studied in [2]: In this game a coin is thrown onto an area that is divided into rectangles of equal shape and size. Let the coin have the radius r, and the rectangles the side lengths a and b, with $2r \leq a$ and $2r \leq b$, so that the coin fits completely into a rectangle. We desire to find the probability that the coin lies on at least one of the two sides of a rectangle. If we inscribe into each rectangle a smaller one with side lengths $a - 2r$ and $b - 2r$ in such a way that the centers of the rectangles coincide and the sides are parallel, it is clear that the coin does *not* lie on any side of a rectangle if and only if its center lies inside the inner rectangle. Since the area of the inner rectangle is $(a - 2r)(b - 2r)$ and the area of the outer ab, the desired probability is

$$P(A) = 1 - \frac{(a - 2r)(b - 2r)}{ab}.$$

A.3.2.3 Conditional Probability and Independent Events

In many cases one has to determine the probability of an event A when it is already known that some other event B has occurred. Such probabilities are called conditional probabilities and denoted $P(A \mid B)$. In a strict sense the "unconditional" probabilities we considered up to now are also conditional, because they always refer to specific frame conditions and circumstances. For example, we assumed that the die we throw is symmetric and made of homogeneous material. Only under these, and possibly other, silently adopted frame conditions (like no electromagnetic influence on the die, etc.) we stated that the probability of each number is $\frac{1}{6}$.

Definition A.4 Let A and B be two arbitrary events with $P(B) > 0$. Then

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

is called the **conditional probability** of A given B.

A simple example: two dice are cast. What is the probability that one of the dice displays a five if it is known that the sum of the pips is eight. If two dice are cast, we have 36 elementary events, five of which satisfy that the sum of the pips is eight ($4 + 4$, $5 + 3$, and $6 + 2$, where the last two have to be counted twice due to the two

possible distributions of the numbers to the two dice). That is, we have $P(B) = \frac{5}{36}$. The event "The sum of the pips is eight, and one of the dice shows a five" can be obtained from two elementary events: either the first die shows a five, and the second a three, or vice versa. Therefore $P(A \cap B) = \frac{2}{36}$, and thus the desired conditional probability is $P(A \mid B) = \frac{2}{5}$.

Theorem A.1 (product law) *For arbitrary events A and B, we have*

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

This theorem follows immediately from the definition of conditional probability together with the obvious relation that $P(A \cap B) = 0$ if $P(B) = 0$.[14] By simple induction over the number of events, we obtain the generalization for m events:

$$P\left(\bigcap_{i=1}^{m} A_i\right) = \prod_{i=1}^{m} P\left(A_i \,\middle|\, \bigcap_{k=1}^{i-1} A_k\right).$$

A conditional probability has all properties of a normal probability, that is, it satisfies Kolmogorv's Axioms. Therefore we have:

Theorem A.2 *For an arbitrary, but fixed, event B with $P(B) > 0$, the function P_B that is defined as $P_B(A) = P(A \mid B)$ is a probability which satisfies $P_B(\overline{B}) = 0$.*

With the help of the notion of a conditional probability, we can now define the notion of (stochastic) independence of events. This notion can be motivated as follows: if, for example, smoking had no influence on the development of lung cancer, then the proportion people with lung cancer among smokers should be (roughly) equal to the proportion of people with lung cancer among nonsmokers.

Definition A.5 Let B be an event with $0 < P(B) < 1$. Then an event A is called **(stochastically) independent** of B iff

$$P(A \mid B) = P(A \mid \overline{B}).$$

The following two relations, which are usually easier to handle, are equivalent:

Theorem A.3 *An event A is (stochastically) independent of an event B with $0 < P(B) < 1$ iff*

$$P(A \mid B) = P(A)$$

or equivalently iff

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

[14] Formally this argument is not quite valid, though, since for $P(B) = 0$, the conditional probability $P(A \mid B)$ is undefined (see Definition A.4). However, since the equation holds for any value that may be fixed for $P(A \mid B)$ in case that $P(B) = 0$, we allow ourselves to be slightly sloppy here.

Note that the relation of (stochastic) independence is symmetric, that is, if A is (stochastically) independent of B, then B is also (stochastically) independent of A (provided that $0 < P(A) < 1$). In addition, the notion of (stochastic) independence can easily be extended to more than two events:

Definition A.6 m events $A_1, \ldots, A_m$ are called completely (stochastically) independent if for any selection $A_{i_1}, \ldots, A_{i_t}$ of t events with $\forall r, s; 1 \leq r, s \leq t : i_r \neq i_s$, we have

$$P\left(\bigcap_{k=1}^{t} A_{i_k}\right) = \prod_{k=1}^{t} P(A_{i_k}).$$

Note that for the complete (stochastic) independence of more than two events, their pairwise independence is necessary but not sufficient.

Let us consider a simple example: A white and a red die are cast. Let A be the event "The number of pips shown by the white die is even," B the event "The number of pips shown by the red die is odd," and C the event "The sum of the pips is even." It is easy to check that A and B are pairwise (stochastically) independent, as well as B and C, and also A and C. However, due to $P(A \cap B \cap C) = 0$ (since the sum of an even number and an odd number must be odd), they are not completely independent.

Another generalization of (stochastic) independence can be achieved by introducing another condition for all involved probabilities:

Definition A.7 (conditionally (stochastically) independent) Two events A and B are called conditionally (stochastically) independent given a third event C with $0 < P(C) < 1$ iff

$$P(A \cap B \mid C) = P(A \mid C) \cdot P(B \mid C).$$

Note that two events A and B may be conditionally independent but not unconditionally independent and vice versa. To see this, consider again the example of the red and white die discussed above. A and B are independent but not conditionally independent given C, because if C holds, only one of A and B can be true, even though either of them can be true (provided that the other is false). Hence the joint probability $P(A \cap B \mid C) = 0$, while $P(A \mid C) \cdot P(B \mid C) > 0$. Examples for the reverse case (conditional independence, but unconditional dependence) are also easy to find.

A.3.2.4 Total Probability and Bayes' Rule

Often we face situations in which the probabilities of disjoint events A_i are known, which together cover the whole sample space. In addition, we know the conditional probabilities of an event B given these A_i. Desired is the (unconditional) probability of the event B. As an example, consider a plant that has a certain number of

machines to produce the same product. Suppose that the capacities of the individual machines and their probabilities (rates) of producing faulty products are known. The total probability (rate) of faulty products is to be computed. This rate can easily be found by using the law of total probability. However, before we turn to it, we formally define the notion of an event partition.

Definition A.8 m events $A_1, \ldots, A_m$ form an **event partition** iff all pairs A_i, A_k, $i \neq k$, are mutually exclusive (that is, $A_i \cap A_k = \emptyset$ for $i \neq k$) and if $A_1 \cup \cdots \cup A_m = \Omega$, that is, all events together cover the whole sample space.

Theorem A.4 (law of total probability) *Let $A_1, \ldots, A_m$ be an event partition with $\forall i; 1 \leq i \leq m : P(A_i) > 0$ (and, as follows from the additivity axiom, $\sum_{i=1}^{m} P(A_i) = 1$). Then the probability of an arbitrary event B is*

$$P(B) = \sum_{i=1}^{m} P(B \mid A_i) P(A_i).$$

The law of total probability can be derived by applying the product rule (see Theorem A.1) to the relation

$$P(B) = P(B \cap \Omega) = P\left(B \cap \bigcup_{i=1}^{m} A_i\right) = P\left(\bigcup_{i=1}^{m} (B \cap A_i)\right) = \sum_{i=1}^{m} P(B \cap A_i),$$

the last step of which follows from the additivity axiom.

With the help of this theorem we can easily derive the important Bayes' rule. To do so, it is merely necessary to realize that the product rule can be applied in two ways to the simultaneous occurrence of two events A and B:

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

Dividing the right-hand side by $P(B)$ (which, of course, must be positive to be able to do so), yields the simple form of Bayes' rule. By applying the law of total probability to the denominator, we obtain the extended form.

Theorem A.5 (Bayes' rule) *Let $A_1, \ldots, A_m$ be an event partition with $\forall i; 1 \leq i \leq m : P(A_i) > 0$, and B an arbitrary event with $P(B) > 0$. Then*

$$P(A_i \mid B) = \frac{P(B \mid A_i) P(A_i)}{P(B)} = \frac{P(B \mid A_i) P(A_i)}{\sum_{k=1}^{m} P(B \mid A_k) P(A_k)}.$$

This rule[15] is also called the formula for the probability of hypotheses, since it can be used to compute the probability of hypotheses (for example, the probability that a patient suffers from a certain disease) if the probabilities are known with

[15] Note that Thomas Bayes (1702–1761) did not derive this formula, despite the fact that it bears his name. In the form given here it was stated only later by Pierre-Simon de Laplace (1749–1827). This supports a basic law of the history of science: a law or an effect that bears the name of a person was found by somebody else.

which the hypotheses lead to the considered events A_i (for example, medical symptoms).

As an example, we consider five urns with the following contents:

- two urns with the contents A_1 with two white and three black balls each,
- two urns with the contents A_2 with one white and four black balls each,
- one urn with the contents A_3 with four white and one black ball.

Suppose that an urn is chosen at random and a ball is drawn from it, also at random. Let this ball be white: this is the event B. What is the (posterior) probability that the ball stems from an urn with the contents A_3?

According to our presuppositions, we have:

$$P(A_1) = \frac{2}{5}, \qquad P(A_2) = \frac{2}{5}, \qquad P(A_3) = \frac{1}{5},$$
$$P(B \mid A_1) = \frac{2}{5}, \quad P(B \mid A_2) = \frac{1}{5}, \quad P(B \mid A_3) = \frac{4}{5}.$$

We start by applying the law of total probability in order to find the probability $P(B)$:

$$\begin{aligned} P(B) &= P(B \mid A_1)P(A_1) + P(B \mid A_2)P(A_2) + P(B \mid A_2)P(A_3) \\ &= \frac{2}{5} \cdot \frac{2}{5} + \frac{1}{5} \cdot \frac{2}{5} + \frac{4}{5} \cdot \frac{1}{5} = \frac{10}{25}. \end{aligned}$$

Using Bayes' rule, we then obtain

$$P(A_3 \mid B) = \frac{P(B \mid A_3)P(A_3)}{P(B)} = \frac{\frac{4}{5} \cdot \frac{1}{5}}{\frac{10}{25}} = \frac{2}{5}.$$

Likewise, we can obtain

$$P(A_1 \mid B) = \frac{2}{5} \quad \text{and} \quad P(A_2 \mid B) = \frac{1}{5}.$$

A.3.2.5 Bernoulli's Law of Large Numbers

In Sect. A.3.1 we already considered the relation between the probability $P(A)$ and the relative frequency $r_n(A) = \frac{h_n(A)}{n}$ of an event A, where $h_n(A)$ is the absolute frequency of this event in n trials. We saw that it is not possible to define the probability of A as the limit of the relative frequency as $n \to \infty$. However, a slightly weaker statement holds, namely the famous law of large numbers.

Definition A.9 A random experiment in which the event A can occur is repeated n times. Let A_i be the event that A occurs in the ith trial. Then the sequence of experiments of length n is called a **Bernoulli experiment**[16] for the event A iff the following conditions are satisfied:

[16]The notion "Bernoulli experiment" was introduced in recognition of the Swiss mathematician Jakob Bernoulli (1654–1705).

1. $\forall 1 \leq i \leq n : P(A_i) = p$.
2. The events $A_1, \ldots, A_n$ are fully independent.

Theorem A.6 (Bernoulli's law of large numbers) *Let $h_n(A)$ be the number of occurrences of an event A in n independent trials of a Bernoulli experiment, where in each of the trials the probability $P(A)$ of the occurrence of A equals an arbitrary, but fixed, value p, $0 \leq p \leq 1$. Then for any $\varepsilon > 0$,*

$$\lim_{n \to \infty} P\left(\left| \frac{h_n(A)}{n} - p \right| < \varepsilon \right) = \frac{1}{\sqrt{2\pi}} \int e^{-\frac{z^2}{2}} \, \mathrm{d}z = 1.$$

This property of the relative frequency $r_n(A) = \frac{h_n(A)}{n}$ can be interpreted as follows: even though $p = P(A)$ is not the limit of the relative frequency for infinite sample size, as von Mises [12] tried to define, but it can be seen as very probable (practically certain) that in a Bernoulli experiment of sufficiently large size n, the relative frequency $r_n(A)$ differs only very little from a fixed value, the probability p. This is often referred to by saying that the relative frequency $r_n(A)$ **converges in probability** to the probability $P(A) = p$. With the above law we have the fundamental relationship between the relative frequency and the probability of an event A.

A.3.3 Random Variables

In many situations we are not interested in the complete set of elementary events and their individual probabilities, but in the probabilities of events that result from a partition of the sample space. That is, these events are mutually exclusive, but together cover the whole sample space (an *event partition*, see above). The probabilities of such events are commonly described by so-called *random variables*, which can be seen as transformations from one sample space into another [10].

Definition A.10 A function X that is defined on a sample space Ω and has the domain $\mathrm{dom}(X)$ is called **random variable** if the preimage of any subset of its domain possesses a probability. Here the preimage of a subset $U \subseteq \mathrm{dom}(X)$ is defined as

$$X^{-1}(U) = \{\omega \in \Omega \mid X(\omega) \in U\}.$$

A.3.3.1 Real-Valued Random Variables

The simplest random variable is obviously one that has the elementary events as its possible values. However, in principle, any set can be the domain of a random variable. However, most often the domain is the set of real numbers.

Definition A.11 A function X that maps a sample space Ω to the real numbers is called a **real-valued random variable** if it possesses the following properties: for

any $x \in \mathbb{R}$ and any interval $(a, b]$, $a < b$, (where $a = -\infty$ is possible), the events $A_x = \{\omega \in \Omega \mid X(\omega) = x\}$ and $A_{(a,b]} = \{\omega \in \Omega \mid a < X(\omega) \leq b\}$ possess probabilities.

Sometimes the required properties are stated with an interval $[a, b)$ that is open on the right. This does not lead to any significant differences.

Definition A.12 Let X be a real-valued random variable. The real-valued function

$$F(x) = P(X \leq x)$$

is called the **distribution function** of X.

A.3.3.2 Discrete Random Variables

Definition A.13 A random variable X with finite or only countable infinite domain $\mathrm{dom}(X)$ is called a **discrete random variable**. The total of all pairs $(x, P(X = x))$, $x \in \mathrm{dom}(X)$, is called the **(probability) distribution** of a discrete random variable X.

If the probabilities $P(X = x)$ can be stated as a function, the distribution of a discrete random variable is often stated as a function $v_X(x) = P(X = x)$. Possible parameters (in order to select a function from a parameterized family of functions) are separated by a semicolon from the function argument. For example, the binomial distribution

$$b_X(x; p, n) = \binom{n}{x} p^x (1 - p)^{n-x}$$

has the probability p of the occurrence of the considered event in a single trial and the size n of the sample as parameters. (The binomial distribution is considered in more detail in Sect. A.3.5.1; later sections consider other important distributions.)

The values of the distribution function of a discrete, real-valued random variable can be computed as follows from the values of the probability distribution:

$$F(x) = P(X \leq x) = \sum_{x' \leq x} P\big(X = x'\big).$$

Every discrete real-valued random variable X has a step function F as its distribution function, which has jumps of height $P(X = x')$ only at those values x' that are in the domain $\mathrm{dom}(X)$. From $x < y$ it follows that $F(x) \leq F(y)$, that is, F is monotone nondecreasing. The function values $F(x)$ become arbitrarily small if only x is chosen small enough, while the values $F(x)$ get arbitrarily close to 1 for growing x. Therefore we have

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

Vice versa, from any step function F, which satisfies the above conditions, the distribution $(x, P(X = x))$, $x \in \mathrm{dom}(X)$, of a real-valued random variable can be derived.

With the help of a distribution function F it is very simple to compute the probability that X assumes a value from a given interval:

Theorem A.7 *Let F be the distribution function of a discrete real-valued random variable X. Then the following relations hold*:

$$P(a < X \leq b) = F(b) - F(a),$$
$$P(a \leq X \leq b) = F(b) - F_L(a),$$
$$P(a < X) = 1 - F(a),$$

where $F_L(a)$ is the left limit of $F(x)$ at the location a. This limit equals $F(a)$, if a is not the location of a discontinuity, and otherwise equal to the value of the step directly to the left of a, or formally, $F_L(a) = \sup_{x<a} F(x)$.

A.3.3.3 Continuous Random Variables

In contrast to discrete random variables, continuous (real-valued) random variables are defined as random variables with a super-countably infinite domain. Obviously, this requires to replace the sum in the distribution function by an integral.

Definition A.14 A real-valued random variable X is called **continuous** if there exists a nonnegative, integrable function f such that for its distribution function $F(x) = P(X \leq x)$, the integral representation

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(u)\,\mathrm{d}u$$

holds. The function f is called the **(probability) density function** (or simply density) of the random variable X.

Since $P(-\infty < X < \infty) = P(\{\omega \in \Omega \mid -\infty < X(\omega) < \infty\}) = P(\Omega)$, the density function f satisfies

$$\int_{-\infty}^{\infty} f(u)\,\mathrm{d}u = 1.$$

For continuous random variables, similar relations hold as for discrete real-valued random variables.

Theorem A.8 *If X is a continuous random variable with density function f, then for arbitrary numbers $a, b, c \in \mathbb{R}$ with $a < b$, we have*

$$P(a < X \leq b) = F(b) - F(a) = \int_{a}^{b} f(u)\,\mathrm{d}u,$$
$$P(X > c) = 1 - F(c) = \int_{c}^{\infty} f(u)\,\mathrm{d}u.$$

A.3.3.4 Random Vectors

Up to now we have only considered single random variables. In the following we expand our study to several random variables and their interaction, that is, their joint distribution and their dependence or independence. In order to do so, we define the notion of a random vector or, equivalently, of a multidimensional random variable.

Definition A.15 Let $X_1, \ldots, X_m$ be m random variables that are defined on the same probability space $(\Omega, \mathcal{S}, P)$, that is, on the same sample space Ω with the same σ-algebra $\mathcal{S}$ and probability P. In this case the vector $(X_1, \ldots, X_m)$ is called a **random vector** or an m**-dimensional random variable**.

In order to keep things simple, we consider here only two-dimensional random variables. However, all definitions and theorems can easily be transferred to multi-dimensional random variables (random vectors with finite length m). In addition we confine ourselves, as in the preceding sections, to real-valued random variables.

Definition A.16 Let X and Y be two real-valued random variables. The function F which is defined for all pairs $(x, y) \in \mathbb{R}^2$ as

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

is called the **distribution function** of the two-dimensional random variable (X, Y).

The one-dimensional distribution functions

$$F_1(x) = P(X \leq x) \quad \text{and} \quad F_2(y) = P(Y \leq y)$$

are called **marginal distribution functions**.

For discrete random variables, the notion of their joint distribution is defined in an analogous way.

Definition A.17 Let X and Y be two discrete random variables. Then the total of pairs $\forall x \in \mathrm{dom}(X) : \forall y \in \mathrm{dom}(Y) : ((x, y), P(X = x, Y = y))$ is called the **joint distribution** of X and Y. The one-dimensional distributions $\forall x \in \mathrm{dom}(X) : (x, \sum_y P(X = x, Y = y))$ and $\forall y \in \mathrm{dom}(Y) : (y, \sum_x P(X = x, Y = y))$ are **marginal distributions**.

Continuous random variables are treated in a similar way, by simply replacing the joint distribution with the joint density function.

Definition A.18 The two-dimensional random variable (X, Y) is called continuous if there exists a nonnegative function $f(x, y)$ such that for every $(x, y) \in \mathbb{R}^2$, we have

$$F(x, y) = P(X \leq x, Y \leq y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(u, v)\, \mathrm{d}u\, \mathrm{d}v.$$

The function $f(x, y)$ is called the **joint density function** or simply **joint density** of the random variables X and Y. The one-dimensional density functions

$$f_1(x) = \int_{-\infty}^{+\infty} f(x, y)\,\mathrm{d}y \quad \text{and} \quad f_2(y) = \int_{-\infty}^{+\infty} f(x, y)\,\mathrm{d}x$$

are called **marginal density functions** or simply **marginal densities**.

By extending the notion of the independence of events one can define the notion of the independence of random variables.

Definition A.19 Two continuous real-valued random variables X and Y with two-dimensional distribution function $F(x, y)$ and marginal distribution functions $F_1(x)$ and $F_2(y)$ are called **(stochastically) independent** if for all pairs of values $(x, y) \in \mathbb{R}^2$, we have

$$F(x, y) = P(X \leq x, Y \leq y) = P(X \leq x) \cdot P(Y \leq y) = F_1(x) \cdot F_2(y).$$

Definition A.20 Two discrete random variables X and Y with joint distribution $\forall x \in \mathrm{dom}(X) : \forall y \in \mathrm{dom}(Y) : ((x, y), P(X = x, Y = y))$ and marginal distributions $\forall x \in \mathrm{dom}(X) : (x, P(X = x))$ and $\forall y \in \mathrm{dom}(Y) : (y, P(Y = y))$ are called **(stochastically) independent**, if

$$\forall x \in \mathrm{dom}(X) : \forall y \in \mathrm{dom}(Y): \quad P(X = x, Y = y) = P(X = x) \cdot P(Y = y).$$

As for events, the notion of (stochastic) independence can easily be generalized to conditional (stochastic) independence. In this case the distributions and distribution functions are replaced by conditional distributions and conditional distribution functions. For example, for discrete random variables, we obtain:

Definition A.21 Let X, Y, and Z be three discrete random variables. Let the X and Y have the conditional joint distribution $\forall x \in \mathrm{dom}(X) : \forall y \in \mathrm{dom}(Y) : \forall z \in \mathrm{dom}(Z) : ((x, y, z), P(X = x, Y = y \mid Z = z))$ given Z and the conditional marginal distributions $\forall x \in \mathrm{dom}(X){:}\forall z \in \mathrm{dom}(Z) : ((x, z), P(X = x \mid Z))$ and $\forall y \in \mathrm{dom}(Y) : \forall z \in \mathrm{dom}(Z) : ((y, z), P(Y = y \mid Z))$. X and Y are **conditionally (stochastically) independent** if

$$\forall x \in \mathrm{dom}(X) : \forall y in\, \mathrm{dom}(Y) : \forall z \in \mathrm{dom}(Z):$$
$$P(X = x, Y = y \mid Z = z) = P(X = x \mid Z = z) \cdot P(Y = y \mid Z = z).$$

For continuous random variables, the definition is generalized in an analogous way but is formally a bit tricky, and that is why we omit it here. In order to distinguish normal (stochastic) independence from conditional (stochastic) independence, the former is often also called **marginal independence**.

Since the notion of conditional independence is a very important concept, we try to convey a better intuition of it with a simple example. Consider the scatter

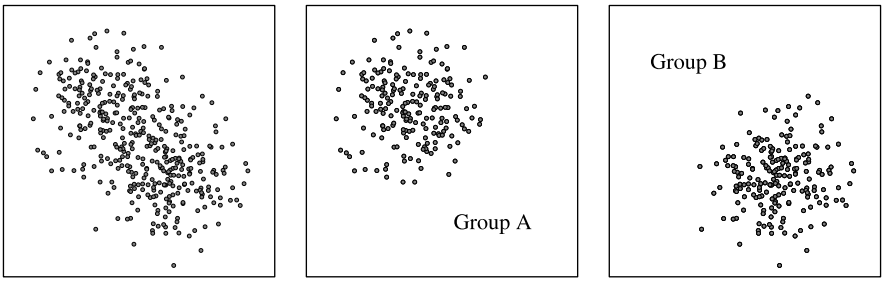


Fig. A.13 Illustration of marginal dependence and conditional independence: the two measures describing the points are marginally dependent (*left*), but if a third, binary variable is fixed and thus the data points are split into two groups, the dependence vanishes (*middle* and *right*)

plot of a sample from two continuous random variables that is shown in Fig. A.13 on the left. Obviously, the two quantities are not independent, because there is a clear tendency for the Y variable (vertical axis) to take lower values if X (horizontal axis) has higher values. Hence X and Y are not marginally independent. To make the example more vivid, we can interpret the horizontal axis as the average number of cigarettes a person smoked per day and the vertical axis as the age of death of that person. Of course, this is a fictitious example, with artificially generated data points, but medical surveys usually show such a dependence.[17] From such an observed dependence it is usually concluded that smoking is a health hazard.

However, this need not be conclusive. There may be a third variable that couples the two, which, if we fix its value, renders the two quantities independent. A passionate smoker may claim, for example, that this third variable is whether a person is exposed to severe stress at work. Such stress is certainly a health hazard and it causes, as our passionate smoker may argue, both: a shorter life span (due to the strain on the person's health by the stress it is exposed to) and a higher cigarette consumption (due to the fact that smoking has a calming effect and thus can help to cope with stress). If this argument were correct,[18] the dependence should vanish if we consider people that are exposed to stress at work and those who are not separately. That is, if the argument were correct, we should see the separate data as depicted in Fig. A.13 in the middle (people that are not exposed to stress at work and thus smoke less and live longer) and on the right (people that are exposed to stress at work and thus smoke more and live less long). In both cases the dependence between the two quantities has vanished, and thus they are conditionally independent given the third variable (stress at work or not).

[17] However, we do not claim that the actual dependence looks like our data.

[18] We do not believe it is. The claim that smoking harms your health is much better supported than just by an observation of a correlation like the one depicted in Fig. A.13, even though such correlations are part of the argument.

A.3.4 Characteristic Measures of Random Variables

In analogy to Sect. A.2.3, where we defined characteristic measures for data sets, random variables can be described by analogous measures. While for data sets, these measures are derived from the sample values, measures for random variables are derived from their distributions and distribution functions. The analogy is actually very close: the unit weight of each sample data point is replaced by the probability mass that is assigned to the different values in the domain of a random variable.

A.3.4.1 Expected Value

If the notion of a random variable is applied to gambling, where the value of the random variable could be, for example, the gains or losses connected to different outcomes, the idea suggests itself to consider a kind of average or expected win (or loss) if a sufficiently large number of gambles is played. This idea leads to the notion of an expected value.

Definition A.22 Let X be a discrete real-valued random variable with distribution $(x, P(X = x))$, $x \in \mathrm{dom}(X)$. If $\sum_x |x| P(X = x)$ is finite, the limit (which must exist in this case)

$$\mu = E(X) = \sum_{i=1}^{\infty} x_i \, P(X = x_i)$$

is called the **expected value** of X.

As an example, consider the expected value of the winnings in the classical game of roulette. We do not bet on a so-called *simple chance* (*Rouge* vs. *Noir*, *Pair* vs. *Impair*, *Manque* vs. *Passe*), in order to avoid the difficulties that result from the special rules applying to them, but bet on a column (one of the sets of numbers 1–12, 13–24, and 14–36). In case of a win we receive three times our wager. Since in roulette 37 numbers can occur (0–36),[19] all of which are equally likely if we assume perfect conditions, winning with a bet on a column has the probability $\frac{12}{37}$, and losing has the probability $\frac{25}{37}$. Let us assume that the wager consists of m chips. In case of a win we have twice the wager as a net gain (the paid out win has to be reduced by the initially waged m chips), that is, $2m$ chips, whereas in case of a failure we lose m chips. As a consequence, the expected value is

$$E(X) = 2m \cdot \frac{12}{37} - m \cdot \frac{25}{37} = -\frac{1}{37} m \approx -0.027m.$$

On average we thus lose 2.7% of our wager in every gamble.

In order to define the expected value of continuous random variables, we only have to replace the sum by an integral, and the probabilities of the distribution by the density function.

[19] In certain types of American roulette even 38, as these have 0 and 00 (double zero).

Definition A.23 If X be a continuous random variable with density function f and if the integral

$$\int_{-\infty}^{\infty} |x| f(x)\,\mathrm{d}x = \lim_{\substack{a\to -\infty \\ b\to \infty}} \int_{a}^{b} |x| f(x)\,\mathrm{d}x$$

exists, then the integral (which must exist in this case)

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

is called the expected value of X.

A.3.4.2 Properties of the Expected Value

In this section some properties of the expected value are collected, which can often be exploited when one has to compute the expected value.

Theorem A.9 *Let X be a discrete random variable that takes no other values than some constant c. Then its expected value is equal to c: $\mu = E(X) = c$.*

Theorem A.10 (linearity of the expected value) *Let X be a (discrete or continuous) real-valued random variable with expected value $E(X)$. Then the expected value of the random variable $Y = aX + b$, $a, b \in \mathbb{R}$, is*

$$E(Y) = E(aX + b) = aE(X) + b.$$

The statement can easily be checked by simply inserting the expression $aX + b$ into the definition of the expected value, once for a discrete and once for a continuous random variable.

As an example, we consider the distribution of a random variable X, which describes the sum of the pips of two dice. This sum is clearly symmetric w.r.t. the value 7, that is, $\forall k \in \{0, 1, \ldots, 5\} : P(X = 7 + k) = P(X = 7 - k)$. It follows that the expected values of the random variables $Y_1 = X - 7$ and $Y_2 = -(X - 7) = 7 - X$ must be identical, since they have, due to the symmetry, the same distribution. Therefore,

$$E(Y_1) = E(X - 7) = E(7 - X) = E(Y_2).$$

Applying Theorem A.10 yields $E(X) - 7 = 7 - E(X)$, and therefore $E(X) = 7$. From this example we can conclude generally that the point of symmetry of a distribution, provided that it has one, must be its expected value. This also holds for continuous random variables.

Next, we turn to the expected value of functions of two random variables, namely their sum and their product.

Theorem A.11 (expected value of a sum of random variables) *The expected value of a sum $Z = X + Y$ of two arbitrary real-valued random variables X and Y, whose*

expected values $E(X)$ and $E(Y)$ both exist, is equal to the sum of their expected values,

$$E(Z) = E(X + Y) = E(X) + E(Y).$$

Theorem A.12 (expected value of a product of random variables) *The expected value of a product $Z = X \cdot Y$ of two independent real-valued random variables X and Y, whose expected values $E(X)$ and $E(Y)$ exist, is equal to the product of their expected values*,

$$E(Z) = E(X \cdot Y) = E(X) \cdot E(Y).$$

Again the validity of these theorems can easily be checked by inserting the sum/product into the definition of the expected value, in the case of a product of random variables by also exploiting the definition of independence. It should be clear that both theorems can easily be generalized to sums and products of finitely many (independent) random variables. Never forget about the presupposition of independence in the second theorem, since it does not hold for dependent random variables.

A.3.4.3 Variance and Standard Deviation

The expected value alone does not sufficiently characterize a random variable. We must consider also what deviation from the expected value can occur on average (see Sect. A.2.3.2). This dispersion is described by variance and standard deviation.

Definition A.24 If μ is the expected value of a discrete real-valued random variable X, then the value (provided that it exists)

$$\sigma^2 = D^2(X) = E\big([X - \mu]^2\big) = \sum_{i=1}^{\infty} (x_i - \mu)^2 P(X = x_i)$$

is called the **variance**, and the positive square root $\sigma = D(X) = +\sqrt{\sigma^2}$ is called the **standard deviation** of X

Let us again consider roulette as an example. If we bet m chips on a column, the variance is

$$D^2(X) = \left(2m + \frac{1}{37}m\right)^2 \cdot \frac{12}{37} + \left(-m + \frac{1}{37}m\right)^2 \cdot \frac{25}{37} = \frac{99900}{50653}m^2 \approx 1.97m^2.$$

Hence the standard deviation $D(X)$ is about $1.40m$. In comparison, the variance of a bet on a *plain chance*, that is, on a single number, has the same expected value, but the variance

$$D^2(X) = \left(35m + \frac{1}{37}m\right)^2 \cdot \frac{1}{37} + \left(-m + \frac{1}{37}m\right)^2 \cdot \frac{36}{37} = \frac{1726272}{50653}m^2 \approx 34.1m^2,$$

and thus a standard deviation $D(X)$ of about $5.84m$. Despite the same expected value, the average deviation from the expected value is about 4 times as large for a bet on a plain chance than for a bet on a column.

In order to define the variance of a continuous random variable, we only have to replace the sum by an integral—just as we did for the expected value.

Definition A.25 If μ is the expected value of a continuous random variable X, then the value (provided that it exists)

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

is called the **variance** of X, and $\sigma = D(X) = +\sqrt{\sigma^2}$ is called the **standard deviation** of X.

A.3.4.4 Properties of the Variance

In this section we collect a few useful properties of the variance.

Theorem A.13 *Let X be a discrete random variable which takes no other values than a constant c. Then its variance is* 0: $\sigma^2 = D^2(X) = 0$.

Theorem A.14 *Let X be a (discrete or continuous) real-valued random variable with variance $D^2(X)$. Then the variance of the random variable $Y = aX + b, a, b \in \mathbb{R}$, is*

$$D^2(Y) = D^2(aX+b) = a^2 D^2(X),$$

and therefore, for the standard deviation, we have

$$D(Y) = D(aX+b) = |a| D(X).$$

The validity of this theorem (like the validity of the next theorem) can easily be checked by inserting the given expressions into the definition of the variance, once for discrete and once for continuous random variables.

Theorem A.15 *The variance σ^2 of a (discrete or continuous) real-valued random variable satisfies*

$$\sigma^2 = E\left(X^2\right) - \mu^2.$$

Theorem A.16 (variance of a sum of random variables, covariance) *If X and Y are two (discrete or continuous) real-valued random variables, whose variances $D^2(X)$ and $D^2(Y)$ exist, then*

$$D^2(X+Y) = D^2(X) + D^2(Y) + 2[E(X \cdot Y) - E(X) \cdot E(Y)].$$

The expression $E(X \cdot Y) - E(X) \cdot E(Y) = E[(X - E(X))(Y - E(Y))]$ *is called the* **covariance** *of* X *and* Y. *From the* (*stochastic*) *independence of* X *and* Y *it follows that*

$$D^2(Z) = D^2(X + Y) = D^2(X) + D^2(Y),$$

that is, *the covariance of independent random variables vanishes*.

Again the validity of this theorem can easily be checked by inserting the sum into the definition of the variance. By simple induction it can easily be generalized to finitely many random variables.

A.3.4.5 Quantiles

Quantiles are defined in direct analogy to the quantiles of a data set, with the fraction of the data set replaced by the fraction of the probability mass. For continuous random variables, quantiles are often also called **percentage points**.

Definition A.26 Let X be a real-valued random variable. Then any value x_α, $0 < \alpha < 1$, with

$$P(X \leq x_\alpha) \geq \alpha \quad \text{and} \quad P(X \geq x_\alpha) \geq 1 - \alpha$$

is called an α-**quantile** of X (or of its distribution).

Note that for discrete random variables, several values may satisfy both inequalities, because their distribution function is piecewise constant. It should also be noted that the pair of inequalities is equivalent to the double inequality

$$\alpha - P(X = x) \leq F_X(x) \leq \alpha,$$

where $F_X(x)$ is the distribution function of a random variable X. For a continuous random variable X, it is usually more convenient to define that the α-quantile is the value x that satisfies $F_X(x) = \alpha$. In this case a quantile can be computed from the inverse of the distribution function F_X (provided that it exists and can be specified in closed form).

A.3.5 Some Special Distributions

In this section we study some special distributions, which are often needed in applications (see Sect. A.4 about inferential statistics).

A.3.5.1 The Binomial Distribution

Let X be a random variable that describes the number of trials of a Bernoulli experiment of size n in which an event A occurs with probability $p = P(A)$ in each trial.

Then X has the distribution $\forall x \in \mathbb{N} : (x; P(X = x))$ with

$$P(X = x) = b_X(x; p, n) = \binom{n}{x} p^x (1-p)^{n-x}$$

and is said to be **binomially distributed** with parameters p and n. This formula is also known as **Bernoulli's formula**. The expression $\binom{n}{x} = \frac{n!}{x!(n-x)!}$ (pronounced "n choose x") is called a **binomial coefficient**.

The distribution satisfies the recursive relation

$$\forall x \in \mathbb{N}_0 : b_X(k+1; p, n) = \frac{(n-x)p}{(x+1)(1-p)} b_X(x; p, n)$$
$$\text{with} \quad b_X(0; p, n) = (1-p)^n.$$

For the expected value and variance, we have

$$\mu = E(X) = np; \qquad \sigma^2 = D^2(X) = np(1-p).$$

A.3.5.2 The Polynomial Distribution

Bernoulli experiments can easily be generalized to more than two mutually exclusive events. In this way one obtains the polynomial distribution, which is a multidimensional distribution: a random experiment is executed independently n times. Let $A_1, \ldots, A_k$ be mutually exclusive events, of which in each trial exactly one must occur, that is, let $A_1, \ldots, A_k$ be an event partition. In every trial each event A_i occurs with constant probability $p_i = P(A_i)$, $1 \leq i \leq k$. Then the probability that in n trials the event A_i, $i = 1, \ldots, k$, occurs x_i times, $\sum_{i=1}^{k} x_i = n$, is equal to

$$P(X_1 = x_1, \ldots, X_k = x_k) = \binom{n}{x_1 \ldots x_k} p_1^{x_1} \cdots p_k^{x_k} = \frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k}.$$

The total of all probabilities of all vectors $(x_1, \ldots, x_k)$ with $\sum_{i=1}^{k} x_i = n$ is called the (k-dimensional) **polynomial distribution** with parameters $p_1, \ldots, p_k$ and n. The binomial distribution is obviously a special case of for $k = 2$. The expression $\binom{n}{x_1 \ldots x_k} = \frac{n!}{x_1! \cdots x_k!}$ is called a **polynomial coefficient**, in analogy to the binomial coefficient $\binom{n}{x} = \frac{n!}{x!(n-x)!}$.

A.3.5.3 The Geometric Distribution

Let X be a random variable that describes the number of trials in a Bernoulli experiment that are needed until an event A, which occurs with $p = P(A) > 0$ in each trial, occurs for the first time. Then X has the distribution $\forall x \in \mathbb{N} : (x; P(X = x))$ with

$$P(X = x) = g_X(x; p) = p(1-p)^{x-1}$$

and is said to be **geometrically distributed** with parameter p. In order to compute the probabilities the recursive relation

$$\forall x \in \mathbb{N} : P(X = x + 1) = (1 - p)P(X = x) \quad \text{with} \quad P(X = 1) = p$$

can be useful. The expected value and variance are

$$\mu = E(X) = \frac{1}{p}; \qquad \sigma^2 = D^2(X) = \frac{1 - p}{p^2}.$$

A.3.5.4 The Hypergeometric Distribution

From an urn which contains M black and $N - M$ white, and thus in total N balls, n balls are drawn without replacement. Let X be the random variable that describes the number of black balls that have been drawn. Then X has the distribution $\forall x; \max(0, n - (N - M)) \leq x \leq \min(n, M) : (x; P(X = x))$ with

$$P(X = x) = h_X(x; n, M, N) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$$

and is said to be **hypergeometrically distributed** with parameters n, M and N. This distribution satisfies the recursive relation

$$\begin{aligned} &\forall x; \max(0, n - (N - M)) \leq x \leq \min(n, M) : \\ &\quad h_X(x + 1; n, M, N) = \frac{(M - x)(n - x)}{(x + 1)(N - M - n + x + 1)} h_X(x; n, M, N) \\ &\quad \text{with} \quad h_X(1; n, M, N) = \frac{M}{N}. \end{aligned}$$

With $p = \frac{M}{N}$ and $q = 1 - p$, the expected value and variance are

$$\mu = E(X) = np; \qquad \sigma^2 = D^2(X) = npq\frac{N - n}{N - 1}.$$

A.3.5.5 The Poisson Distribution

A random variable X with the distribution $\forall x \in \mathbb{N} : (x; P(X = x))$ where

$$P(X = x) = \Lambda_X(x; \lambda) = \frac{\lambda^x}{x!} e^{-\lambda}$$

is said to be **Poisson distributed**[20] with parameter λ. This distribution satisfies the recursive relation

$$\forall x \in \mathbb{N}_0 : \Lambda_X(x + 1; \lambda) = \frac{\lambda}{x + 1} \Lambda_X(x; \lambda) \quad \text{with} \quad \Lambda_X(0; \lambda) = e^{-\lambda}.$$

[20] This distribution bears its name in recognition of the French mathematician Siméon-Denis Poisson (1781–1840).

The expected value and variance are

$$\mu = E(X) = \lambda; \qquad \sigma^2 = D^2(X) = \lambda.$$

A Poisson distribution describes the occurrence frequency of a certain type of events in a fixed duration of time, for example, the number of deadly traffic accidents.

For rare events A and a large number n of trials, the binomial distribution can be approximated by a Poisson distribution, because the following relation holds: if in a binomial distribution, n goes to infinity so that $np = \lambda$ stays constant, then

$$\forall x \in \mathbb{N}: \lim_{\substack{n\to\infty \\ np=\lambda}} b_X(x; p, n) = \frac{\lambda^x}{x!} e^{-\lambda},$$

and thus for large n and small p, we obtain the approximation

$$\forall x \in \mathbb{N}: b_X(x; p, n) \approx \frac{np^x}{x!} e^{-np}.$$

For Poisson distributed random variables, the following reproduction law holds:

Theorem A.17 *Let X and Y be two (stochastically) independent Poisson-distributed random variables with parameters λ_X and λ_Y, respectively. Then the sum $Z = X + Y$ is also Poisson distributed with parameter $\lambda_Z = \lambda_X + \lambda_Y$.*

A.3.5.6 The Uniform Distribution

A random variable X with the density function

$$f_X(x; a, b) = \begin{cases} \frac{1}{b-a} & \text{for } x \in [a, b], \\ 0 & \text{otherwise,} \end{cases}$$

with $a, b \in \mathbb{R}$, $a < b$, is said to be **uniformly distributed** in $[a, b]$. Its distribution function F_X is

$$F_X(x; a, b) = \begin{cases} 0 & \text{for } x \le a, \\ \frac{x-a}{b-a} & \text{for } a \le x \le b, \\ 1 & \text{for } x \ge b. \end{cases}$$

The expected value and variance are

$$\mu = E(X) = \frac{a+b}{2}; \qquad \sigma^2 = D^2(X) = \frac{(b-a)^2}{12}.$$

A.3.5.7 The Normal Distribution

A random variable X with the density function

$$N_X\left(x; \mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{(x-\mu)^2}{2\sigma^2}}$$

is said to be **normally distributed** with parameters μ and σ^2.

The expected value and variance are

$$E(X) = \mu; \qquad D^2(X) = \sigma^2.$$

The normal distribution with expected value $\mu = 0$ and variance $\sigma^2 = 1$ is called standard normal distribution.

The density function $f_X(x; \mu, \sigma^2)$ has its maximum at μ and inflection points at $x = \mu \pm \sigma$. The distribution function of X does not possess a closed-form representation. As a consequence, it is usually tabulated, most commonly for the standard normal distribution, from which the values of arbitrary normal distributions can be easily obtained by simple linear transformations.

However, in practice one often faces the reversed problem, namely to find the argument of the distribution function of the standard normal distribution for which is has a given value (or, in other words, one desires to find a quantile of the normal distribution). In order to solve this problem one may just as well use tabulated values. However, the inverse function can be approximated fairly well by the ratio of two polynomials, which is usually employed in computer programs (e.g., [15]).

The normal distribution is certainly the most important continuous distribution, since many random processes, especially measurements of physical quantities, can be described well by this distribution. The theoretical justification for this observed fact is the important central limit theorem:

Theorem A.18 (central limit theorem) *Let $X_1, \ldots, X_m$ be m (stochastically) independent real-valued random variables. In addition, let them satisfy the so-called Lindeberg condition, that is, if $F_i(x)$ are the distribution functions of the random variables X_i, $i = 1, \ldots, m$, μ_i their expected values, and σ_i^2 their variances, then for every $\varepsilon > 0$, it is*

$$\lim_{m\to\infty} \frac{1}{V_m^2} \sum_{i=1}^{m} \int_{|x_i-\mu_i|>\varepsilon V_m^2} (x_i - \mu_i)^2 \,\mathrm{d}F_i(x) = 0$$

with $V_m^2 = \sum_{i=1}^m \sigma_i^2$. Then the standardized sums

$$S_m = \frac{\sum_{i=1}^m (X_i - \mu_i)}{\sqrt{\sum_{i=1}^m \sigma_i^2}}$$

(*that is, standardized to expected value* 0 *and variance* 1) *satisfy*

$$\forall x \in \mathbb{R}: \quad \lim_{m\to\infty} P(S_m \le x) = \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} \,\mathrm{d}t,$$

where $\Phi(x)$ is the distribution function of the standard normal distribution.

Intuitively this theorem says that the sum of a large number of almost arbitrarily distributed random variables (the Lindeberg condition is a very weak restriction) is approximately normally distributed. Since physical measurements are usually affected by a large number of random influences from several independent sources,

which all add up to form the total measurement error, the result is often approximately normally distributed. The central limit theorem thus explains why normally distributed quantities are so common in practice.

Like the Poisson distribution, the normal distribution can be used as an approximation of the binomial distribution, even if the probabilities p are not small.

Theorem A.19 (limit theorem of de Moivre–Laplace)[21] *If the probability of the occurrence of an event A in n independent trials is constant and equal to p, $0 < p < 1$, then the probability $P(X = x)$ that, in these trials, the event A occurs exactly x times satisfies as $n \to \infty$ the relation*

$$\sqrt{np(1-p)}\,P(X=x)\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}y^2} \to 1, \quad \text{where} \quad y = \frac{x-np}{\sqrt{np(1-p)}}.$$

The convergence is uniform for all x for which y lies in an arbitrary finite interval (a, b).

This theorem allows us to approximate the probabilities of the binomial distribution for large n by

$$\begin{aligned}
&\forall x;\, 0 \le x \le n\text{:}\\
&\quad P(X=x) = \binom{n}{x} p^x (1-p)^{n-x}\\
&\qquad\qquad \approx \frac{1}{\sqrt{2\pi np(1-p)}} \exp\left(-\frac{(x-np)^2}{2np(1-p)}\right) \quad \text{or}\\
&\quad P(X=x) \approx \Phi\left(\frac{x-np+\frac{1}{2}}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{x-np-\frac{1}{2}}{\sqrt{np(1-p)}}\right) \quad \text{and}\\
&\forall x_1, x_2;\, 0 \le x_1 \le x_2 \le n\text{:}\\
&\quad P(x_1 \le X \le x_2) \approx \Phi\left(\frac{x_2-np+\frac{1}{2}}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{x_1-np-\frac{1}{2}}{\sqrt{np(1-p)}}\right),
\end{aligned}$$

where Φ is the distribution function of the standard normal distribution. The approximation is reasonably good for $np(1-p) > 9$.

A.3.5.8 The χ^2 Distribution

If one forms the sum of m independent, standard normally distributed random variables (expected value 0 and variance 1), one obtains a random variable X with the density function

$$f_X(x;m) = \begin{cases} 0 & \text{for } x < 0,\\ \frac{1}{2^{\frac{m}{2}} \cdot \Gamma(\frac{m}{2})} \cdot x^{\frac{m}{2}-1} \cdot e^{-\frac{x}{2}} & \text{for } x \ge 0, \end{cases}$$

[21] This theorem bears its name in recognition of the French mathematicians Abraham de Moivre (1667–1754) and Pierre-Simon de Laplace (1749–1827).

where Γ is the so-called **Gamma function** (generalization of the factorial)

$$\Gamma(x) = \int_0^\infty t^{x-1} e^t \, \mathrm{d}t$$

for $x > 0$. This random variable is said to be χ^2-distributed with m degrees of freedom. The expected value and variance are

$$E(X) = m; \qquad D^2(X) = 2m.$$

The χ^2 distribution plays an important role in the statistical theory of hypothesis testing (see Sect. A.4.3), for example, for independence tests.

A.3.5.9 The Exponential Distribution

A random variable X with the density function

$$f_X(x; \alpha) = \begin{cases} 0 & \text{for } x \leq 0, \\ \alpha e^{-\alpha x} & \text{for } x > 0, \end{cases}$$

with $\alpha > 0$ is said to be **exponentially distributed** with parameter α. Its distribution function F_X is

$$F_X(x; \alpha) = \begin{cases} 0 & \text{for } x \leq 0, \\ 1 - e^{-\alpha x} & \text{for } x > 0. \end{cases}$$

The expected value and variance are

$$\mu = E(X) = \frac{1}{\alpha}; \qquad \sigma^2 = D^2(X) = \frac{1}{\alpha^2}.$$

The exponential distribution is commonly used to model the durations between the arrivals of people or jobs that enter a queue to wait for service or processing.

A.4 Inferential Statistics

With inferential statistics one tries to answer the question whether observed phenomena are typical or regular or whether they may also be caused by random influences. In addition, one strives to find probability distributions that are good models of the data-generating process and tries to estimate their parameters. Therefore inferential statistics is concerned with the following important tasks:

- **parameter estimation**
  Given a model of the data-generating process, especially an assumption about the family of distributions of the underlying random variables, the parameters of this model are estimated from the data.
- **hypothesis testing**
  One or more hypotheses about the data-generating process are checked on the basis of the given data. Special types of hypothesis tests are:

  - parameter test: a test whether a parameter of a distribution can have a certain value or whether the parameter of the distributions underlying two different data sets are equal.
  - goodness-of-fit test: a test whether a certain distribution assumption fits the data or whether observed deviations from the expected characteristics of the data (given the distribution assumption) can be explained by random influences.
  - dependence test: a test whether two features are dependent, or whether observed deviations from an independent distribution can be explained by random influences.
- **model selection**
  From several models, which could be used to explain the data, select the best fitting one, paying attention to the complexity of the models.

Note that parameter estimation can be seen as a special case of model selection, in which the class of models from which one can select is highly restricted, so that they differ only in the value(s) of a (set of) parameter(s). The goodness-of-fit test resides on the border between hypothesis testing and model selection, because it serves the purpose to check whether a model of a certain class is appropriate to explain the data. We do not discuss more complex model selection in this appendix.

A.4.1 Random Samples

In Sect. A.2 we considered so-called *samples* (vectors of observed or measured feature values) and represented them in a clearly arranged form in tables and charts. In inferential statistics we also consider samples: we employ the mathematical tools of probability calculus in order to gain (new) knowledge or to check hypotheses on the basis of data samples. In order for this to be possible, the sample values must be obtained as the outcomes of random experiments, so that probability theory is applicable. Such specific samples are called **random samples**.

The random variable which yields the sample value x_i when carrying out the corresponding random experiment is denoted X_i. The value x_i is called a **realization** of the random variable X_i, $i = 1, \ldots, n$. Therefore a random sample $x = (x_1, \ldots, x_n)$ can be seen as a *realization of the random vector* $X = (X_1, \ldots, X_n)$. A random sample is called **independent** if the random variables $X_1, \ldots, X_n$ are (stochastically) independent, that is, if

$$\forall c_1, \ldots, c_n \in \mathbb{R}: \quad P\left(\bigwedge_{i=1}^{n} X_i \leq c_i\right) = \prod_{i=1}^{n} P(X_i \leq c_i).$$

An independent random sample is called **simple** if all random variables $X_1, \ldots, X_n$ have the same distribution function. Likewise we will call the corresponding random vector *independent* and *simple*, respectively.

A.4.2 Parameter Estimation

As already pointed out, parameter estimation rests on an assumption of a model for the data-generating process, in particular an assumption about the family of distribution functions of the underlying random variables. Given this assumption, it estimates the parameters of this model from the given data.

Given:
- A data set and
- a family of equally shaped, parameterized distribution functions $f_X(x; \theta_1, \ldots, \theta_k)$.

Such families may be, for example:
The family of binomial distributions $b_X(x; p, n)$ with parameters p, $0 \leq p \leq 1$, and $n \in \mathbb{N}$, where n, however, is already implicitly given by the sample size.
The family of Poisson distributions $\Lambda_X(x; \lambda, n)$ with parameters $\lambda > 0$ and $n \in \mathbb{N}$, where n is again given by the sample size.
The family of normal distributions $N_X(x; \mu, \sigma^2)$ with parameters μ (expected value) and σ^2 (variance).

Assumption: The process that has generated the given data can appropriately be described by an element of the considered family of distribution functions: distribution assumption.

Desired: There is an element of the considered family of distribution functions (determined by the values of the parameters) that is the best model of the given data (w.r.t. certain quality criteria).

Estimators for the parameters are **statistics**, that is, functions of the sample values of a given data set. Hence they are functions of (realizations of) random variables and therefore (realizations of) random variables themselves. As a consequence, the whole set of tools that probability theory provides for examining the properties of random variables can be applied to estimators.

One distinguishes mainly two types of parameter estimation:

- **point estimators**
  determine the best individual value of a parameter w.r.t. the given data and certain quality criteria;
- **interval estimators**
  yield a so-called confidence interval, in which the true value of the parameter lies with high certainty, with the degree of certainty to be chosen by a user.

A.4.2.1 Point Estimation

It is immediately clear that not every statistic, that is, not every function computed from the sample values, is a usable point estimator for an examined parameter θ. Rather, a statistic should have certain properties, in order to be a reasonable estimator. Desirable properties are:

- **consistency**
  If the amount of available data grows, the estimated value should become closer and closer to the actual value of the estimated parameter, at least with higher and higher probability. This can be formalized by requiring that for growing sample size, the estimation function converges in probability to the true value of the parameter. For example, if T is an estimator for the parameter θ, it should be

  $$\forall \varepsilon > 0: \quad \lim_{n \to \infty} P(|T - \theta| < \varepsilon) = 1,$$

  where n is the sample size. This condition should be satisfied by every point estimator; otherwise we have no reason to assume that the estimated value is in any way related to the true value.
- **unbiasedness**
  An estimator should not tend to generally under- or over-estimate the parameter, but should, on average, yield the right value. Formally, this means that the expected value of the estimator should coincide with the true value of the parameter. For example, if T is an estimator for the parameter θ, it should be

  $$E(T) = \theta,$$

  independently of the sample size.
- **efficiency**
  The estimation should be as precise as possible, that is, the deviation from the true value of the parameter should be as small as possible. Formally, one requires that the variance of the estimator should be as small as possible, since the variance is a natural measure for the precision of the estimation. For example, let T and U be unbiased estimators for a parameter θ. Then T is called more efficient than U iff

  $$D^2(T) < D^2(U).$$

  However, it is rarely possible to show that an estimator achieves the highest possible efficiency for a given estimation problem.
- **sufficiency**
  An estimation function should exploit the information that is contained in the data in an optimal way. This can be made more precise by requiring that different samples, which yield the same estimated value, should be equally likely (given the estimated value of the parameter). The reason is that if they are not equally likely, it must be possible to derive additional information about the parameter value from the data. Formally, this means that an estimator T for a parameter θ is called *sufficient* if for all random samples $x = (x_1, \ldots, x_n)$ with $T(x) = t$, the expression

  $$\frac{f_{X_1}(x_1; \theta) \cdots f_{X_n}(x_n; \theta)}{f_T(t; \theta)}$$

  does not depend on θ [10].

Note that the estimators used in the definition of efficiency must be unbiased, since otherwise arbitrary constants (variance $D^2 = 0$) would be efficient estimators. Con-

sistency, on the other hand, can often be neglected as an additional condition, since an unbiased estimator T for a parameter θ which also satisfies

$$\lim_{n\to\infty} D^2(T) = 0$$

is consistent (not surprisingly).

A.4.2.2 Point Estimation Examples

Given: A family of uniform distributions on the interval $[0, \theta]$, that is,

$$f_X(x;\theta) = \begin{cases} \frac{1}{\theta} & \text{if } 0 \le x \le \theta, \\ 0 & \text{otherwise.} \end{cases}$$

Desired: An estimator for the unknown parameter θ.

(a) Estimate the parameter θ as the maximum of the sample values, that is, choose $T = \max\{X_1, \ldots, X_n\}$ as the estimation function.
In order to check the properties of this estimator, we first determine its probability density,[22] from which we can then derive other properties:

$$\begin{aligned} f_T(t;\theta) &= \frac{\mathrm{d}}{\mathrm{d}t} F_T(t;\theta) = \frac{\mathrm{d}}{\mathrm{d}t} P(T \le t) \\ &= \frac{\mathrm{d}}{\mathrm{d}t} P\big(\max\{X_1, \ldots, X_n\} \le t\big) \\ &= \frac{\mathrm{d}}{\mathrm{d}t} P\left(\bigwedge_{i=1}^{n} X_i \le t\right) = \frac{\mathrm{d}}{\mathrm{d}t} \prod_{i=1}^{n} P(X_i \le t) \\ &= \frac{\mathrm{d}}{\mathrm{d}t} (F_X(t;\theta))^n = n \cdot (F_X(t;\theta))^{n-1} f_X(t,\theta), \end{aligned}$$

where

$$F_X(x;\theta) = \int_{-\infty}^{x} f_X(x;\theta)\,\mathrm{d}x = \begin{cases} 0 & \text{if } x \le 0, \\ \frac{x}{\theta} & \text{if } 0 \le x \le \theta, \\ 1 & \text{if } x \ge \theta. \end{cases}$$

It follows that

$$f_T(t;\theta) = \frac{n \cdot t^{n-1}}{\theta^n} \quad \text{for } 0 \le t \le \theta.$$

T is a consistent estimator for θ:

$$\begin{aligned} \lim_{n\to\infty} P(|T-\theta| < \varepsilon) &= \lim_{n\to\infty} P(T > \theta - \varepsilon) \\ &= \lim_{n\to\infty} \int_{\theta-\varepsilon}^{\theta} \frac{n \cdot t^{n-1}}{\theta^n}\,\mathrm{d}t = \lim_{n\to\infty} \left[\frac{t^n}{\theta^n}\right]_{\theta-\varepsilon}^{\theta} \end{aligned}$$

[22] Recall that estimators are functions of random variables and thus random variables themselves. As a consequence, they have a probability density.

$$= \lim_{n\to\infty}\left(\frac{\theta^n}{\theta^n} - \frac{(\theta-\varepsilon)^n}{\theta^n}\right)$$
$$= \lim_{n\to\infty}\left(1 - \left(\frac{\theta-\varepsilon}{\theta}\right)^n\right) = 1.$$

However, T is not unbiased. This is already intuitively clear, since T can only underestimate the value of θ, but never overestimate it. Therefore the estimated value will "almost always" be too small. Formally, we have:

$$E(T) = \int_{-\infty}^{\infty} t \cdot f_T(t;\theta)\,\mathrm{d}t = \int_0^{\theta} t \cdot \frac{n \cdot t^{n-1}}{\theta^n}\,\mathrm{d}t$$
$$= \left[\frac{n \cdot t^{n+1}}{(n+1)\theta^n}\right]_0^{\theta} = \frac{n}{n+1}\theta < \theta \quad \text{for } n < \infty.$$

(b) Choose $U = \frac{n+1}{n}T = \frac{n+1}{n}\max\{X_1, \ldots, X_n\}$ as the estimation function.
The statistic U is a consistent and unbiased estimator for the parameter θ. We omit a formal proof, which can be obtained along the same lines as in a). However, we will later need the probability density of this estimator, which is

$$f_U(u;\theta) = \frac{n^{n+1}}{(n+1)^n}\frac{u^{n-1}}{\theta^n}.$$

Given: A family of normal distributions $N(x;\mu,\sigma)$, i.e.,

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

Desired: Estimators for the unknown parameters μ and σ^2.

(a) The (empirical) median and the (empirical) mean are both consistent and unbiased estimators for the parameter μ. The median is less efficient than the mean. Since it only exploits order information from the data, it is naturally also not sufficient. Even though the median is preferable for small sample size due to the fact that it is less sensitive to outliers, its variance is larger than that of the mean (provided that the sample size is sufficiently large).

(b) The function $V^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ is a consistent, but not unbiased, estimator for the parameter σ^2, since this function tends to underestimating the variance. The (empirical) variance $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$, however, is a consistent and unbiased estimator for the parameter σ^2. (Due to the square root, however, $\sqrt{S^2}$ is *not* an unbiased estimator for the standard deviation.)

Given: A family of polynomial distributions, that is,

$$f_{X_1,\ldots,X_k}(x_1,\ldots,x_k;\theta_1,\ldots,\theta_k,n) = \frac{n!}{\prod_{i=1}^{k} x_i!}\prod_{i=1}^{k}\theta_i^{x_i},$$

where θ_i is the probability that values a_i occurs, and the random variable X_i describes how often the value a_i occurs in the sample.

Desired: Estimators for the unknown parameters $\theta_1, \ldots, \theta_k$.

The relative frequencies $R_i = \frac{X_i}{n}$ of the feature values in the sample are consistent, unbiased, most efficient, and sufficient estimators for the unknown parameters θ_i, $i = 1, \ldots, k$. This is the reason why relative frequencies are used in basically all cases to estimate the probabilities of nominal values.

A.4.2.3 Maximum Likelihood Estimation

Up to now we have simply stated estimation functions for parameters. This is possible because for many standard problems, consistent, unbiased, and efficient estimators are known, so that they can be looked up in standard textbooks. Nevertheless, we consider briefly how one can find estimation functions in principle.

Besides the method of moments, which we omit here, maximum likelihood estimation, as it was developed by R.A. Fisher,[23] is one of the most popular methods for finding estimation functions. The underlying principle is very simple: choose the value of the parameter to estimate (or the set of values of the parameters to estimate if there are several) that renders the given random sample most likely. This is achieved as follows: if the parameter(s) of the true underlying distribution were known, we could easily compute the probability of a random experiment generating the observed random sample. However, this probability can also be written with unknown parameters (though not necessarily be numerically computed). The result is a function that describes the likelihood of a random sample given the unknown parameters. This function is called a **likelihood function**. By taking partial derivatives of this function w.r.t. the parameters to estimate and setting them equal to zero (since the derivative must vanish at a maximum), estimation functions are derived.

A.4.2.4 Maximum Likelihood Estimation Example

Given: A family of normal distributions $N_X(x; \mu, \sigma^2)$, that is,

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

Desired: Estimators for the unknown parameters μ and σ^2.

The likelihood function of a simple random sample $x = (x_1, \ldots, x_n)$, which is the realization of a vector $X = (X_1, \ldots, X_n)$ of normally distributed random variables with parameters μ and σ^2, is

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

[23] However, R.A. Fisher did not invent this method as is often believed. Earlier on C.F. Gauß and D. Bernoulli already made use of it, but Fisher was the first to study it systematically and to establish it in statistics [10].

It describes the probability of the sample (the data set) depending on the parameters μ and σ^2. By exploiting the known rules for computing with exponential functions, this expression can be transformed into

$$L\big(x_1,\ldots,x_n;\mu,\sigma^2\big) = \frac{1}{(\sqrt{2\pi\sigma^2})^n}\exp\bigg(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\bigg).$$

In order to determine the maximum of this function w.r.t. parameters μ and σ^2, it is convenient to take the natural logarithm (in order to eliminate the exponential function). Since the logarithm is a monotone function, this does not change the location of the maximum. We obtain the **log-likelihood function**

$$\ln L\big(x_1,\ldots,x_n;\mu,\sigma^2\big) = -n\ln\Big(\sqrt{2\pi\sigma^2}\Big) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2.$$

In order to find the maximum, we set the partial derivatives w.r.t. μ and σ^2 equal to 0. The partial derivative w.r.t. μ is

$$\frac{\partial}{\partial\mu}\ln L\big(x_1,\ldots,x_n;\mu,\sigma^2\big) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu) \stackrel{!}{=} 0,$$

from which

$$\sum_{i=1}^{n}(x_i-\mu) = \left(\sum_{i=1}^{n}x_i\right) - n\mu \stackrel{!}{=} 0,$$

and thus

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n}x_i$$

follows as an estimate for the parameter μ. The partial derivative of the log-likelihood function w.r.t. σ^2 yields

$$\frac{\partial}{\partial\sigma^2}\ln L\big(x_1,\ldots,x_n;\mu,\sigma^2\big) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2 \stackrel{!}{=} 0.$$

By inserting the estimated value $\hat{\mu}$ for the parameter μ, we obtain the estimator

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\big(x_i-\hat{\mu}\big)^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - \frac{1}{n^2}\left(\sum_{i=1}^{n}x_i\right)^2$$

for the parameter σ^2. Note that the result is not unbiased. (Recall that, as we mentioned above, the empirical variance with a factor of $\frac{1}{n}$ instead of $\frac{1}{n-1}$ is not unbiased.) This shows that there is no estimator for the variance of a normal distribution that has all desirable properties. Among those that are unbiased, the data is not maximally likely, and the one that makes the data maximally likely is not unbiased.

A.4.2.5 Maximum A Posteriori Estimation

An alternative for maximum likelihood estimation is maximum a posteriori estimation, which rest on Bayes' rule. This estimation method assumes a prior distribution on the domain of the parameter(s) and computes, with the help of this prior distribution and the given data, a posterior distribution via Bayes' rule. The parameter value with the greatest posterior probability (density) is chosen as the estimated value. That is, one chooses the value for θ that maximizes

$$f(\theta \mid D) = \frac{f(D \mid \theta) f(\theta)}{f(D)} = \frac{f(D \mid \theta) f(\theta)}{\int_{-\infty}^{\infty} f(D \mid \theta) f(\theta) \mathrm{d}\theta},$$

where D are the given data, and $f(\theta)$ is the assumed prior distribution. Compare this to maximum likelihood estimation, which chooses the value of the parameter θ that maximizes $f(D \mid \theta)$, that is, the likelihood of the data.

A.4.2.6 Maximum A Posteriori Estimation Example

In order to illustrate why it can be useful to assume a prior distribution on the possible values of the parameter θ, we consider three situations:

- A drunkard claims to be able to predict the side onto which a tossed coin will land (head or tails). On ten trials he always states the correct side beforehand.
- A tea lover claims that she is able to taste whether the tea or the milk was poured into the cup first. On ten trials she always identifies the correct order.
- An expert of classical music claims to be able to recognize from a single sheet of music whether the composer was Mozart or somebody else. On ten trials he is indeed correct every time.

Let θ be the (unknown) parameter that states the probability that a correct prediction is made. The data is formally identical in all three cases: 10 correct, 0 wrong predictions. Nevertheless we are reluctant to treat these three cases equally, as maximum likelihood estimation does. We hardly believe that the drunkard can actually predict the side a tossed coin will land on but assume that he was simply "lucky." The tea lover we also view sceptically, even though our skepticism is less pronounced as in the case of the drunkard. Maybe there are certain chemical processes that depend on the order in which tea and milk are poured into the cup and which change the taste slightly and thus are noticeable to a passionate tea drinker. We just see this possibility as unlikely. On the other hand, we are easily willing to believe the music expert. Clearly, there are differences in the style of different composers that may allow a knowledgeable music expert to see even from a single sheet of music whether it was composed by Mozart or not.

The three attitudes with which we see the three situations can be expressed by prior distribution on the domain of the parameter θ. In the case of the drunkard we ascribe a nonvanishing probability density only to the value 0.5.[24] In the case of the

[24] Formally: Dirac pulse at $\theta = 0.5$.

tea lover we may choose a prior distribution, which ascribes values close to 0.5 a high probability density, which quickly declines towards 1. In the case of the music expert, however, we ascribe a significant probability densities also to values closer to 1. In effect, this means that in the case of the drunkard we always estimate θ as 0.5, regardless of the data. In the case of the tea lover only fairly clear evidence in favor of her claim will make us accept higher values for θ. In the case of the music expert, however, few positive examples suffice to obtain a fairly high value for θ.

Obviously, the prior distribution contains background knowledge about the data-generating process and expresses which parameter values we expect and how easily we are willing to accept them. However, how to choose the prior distribution is a tricky and critical problem, since it has to be chosen subjectively. Depending on their experience, different people will choose different distributions.

A.4.2.7 Interval Estimation

A parameter value that is estimated with a point estimator from a data set usually deviates from the true value of the parameter. Therefore it is useful if one can make statements about these unknown deviations and their expected magnitude. The most straightforward approach is certainly to provide a point-estimated value t and the standard deviation $D(T)$ of the estimator, that is,

$$t \pm D(T) = t \pm \sqrt{D^2(T)}.$$

However, a better possibility consists in determining intervals—so-called confidence intervals—that contain the true value of the parameter with high probability.

The boundaries of these confidence intervals are computed by certain rules from the sample values. Hence they are also statistics, and thus, like point estimators, (realizations of) random variables. Therefore they can be treated analogously. Formally, they are defined as follows:

Let $X = (X_1, \ldots, X_n)$ be a simple random vector the random variables of which have the distribution function $F_{X_i}(x_i; \theta)$ with (unknown) parameter θ. Furthermore, let $A = g_A(X_1, \ldots, X_n)$ and $B = g_B(X_1, \ldots, X_n)$ be two estimators defined on X such that

$$P(A < \theta < B) = 1 - \alpha, \qquad P(\theta \leq A) = \frac{\alpha}{2}, \qquad P(\theta \geq B) = \frac{\alpha}{2}.$$

Then the random interval $[A, B]$ (or a realization $[a, b]$ of this random interval) is called a $(1 - \alpha) \cdot 100\%$ **confidence interval** for the (unknown) parameter θ. The value $1 - \alpha$ is called **confidence level**.

Note the term "confidence" refers to the *method* and *not* to the *result* of the procedure (that is, to a realization of the random interval). *Before* data has been collected, a $(1 - \alpha) \cdot 100\%$ confidence interval contains the true parameter value with probability $1 - \alpha$. However, *after* the data has been collected and the interval boundaries have been computed, the interval boundaries are not random variables anymore. Therefore the interval either contains the true value of the parameter θ

or it does not (probability 1 or 0—even though it is not known which of the two possibilities is obtained).

The above definition of a confidence interval is not specific enough to derive a computation procedure from it. Indeed, the estimators A and B are not uniquely determined: the sets of realizations of the random vectors $X_1, \ldots, X_n$ for which $A \geq \theta$ and $B \leq \theta$ hold merely have to be disjoint and must possess the probability $\frac{\alpha}{2}$. In order to derive a procedure to obtain the boundaries A and B of a confidence interval, the estimators are restricted as follows: they are not defined as general functions of the random vector but rather as functions of a chosen point estimators T for the parameter θ. That is,

$$A = h_A(T) \quad \text{and} \quad B = h_B(T).$$

In this way confidence intervals can be determined generally, namely by replacing an investigation of $A < \theta < B$ with the corresponding event w.r.t. the estimator T, that is, $A^* < T < B^*$. Of course, this is only possible if we can derive the functions $h_A(T)$ and $h_B(T)$ from the inverse functions $A^* = h_A^{-1}(\theta)$ and $B^* = h_B^{-1}(\theta)$ that we have to consider w.r.t. T.

$$\begin{aligned}
\text{Idea:} \quad & P\left(A^* < T < B^*\right) = 1 - \alpha \\
\Rightarrow \quad & P\left(h_A^{-1}(\theta) < T < h_B^{-1}(\theta)\right) = 1 - \alpha \\
\Rightarrow \quad & P(h_A(T) < \theta < h_B(T)) = 1 - \alpha \\
\Rightarrow \quad & P(A < \theta < B) = 1 - \alpha.
\end{aligned}$$

Unfortunately, this is not always possible (in a sufficiently simple way).

A.4.2.8 Interval Estimation Examples

Given: A family of uniform distributions on the interval $[0, \theta]$, that is,

$$f_X(x; \theta) = \begin{cases} \frac{1}{\theta} & \text{if } 0 \leq x \leq \theta, \\ 0 & \text{otherwise.} \end{cases}$$

Desired: A confidence interval for the unknown parameter θ.

A confidence interval can be computed in this case by starting from the unbiased point estimator $U = \frac{n+1}{n} \max\{X_1, \ldots, X_n\}$:

$$P(U \leq B^*) = \int_0^{B^*} f_U(u; \theta)\, \mathrm{d}u = \frac{\alpha}{2} \quad \text{and}$$

$$P(U \geq A^*) = \int_{A^*}^{\frac{n+1}{n}\theta} f_U(u; \theta)\, \mathrm{d}u = \frac{\alpha}{2}.$$

As we know from the section on point estimation,

$$f_U(u; \theta) = \frac{n^{n+1}}{(n+1)^n} \frac{u^{n-1}}{\theta^n}.$$

Thus we obtain

$$B^* = \sqrt[n]{\frac{\alpha}{2}}\frac{n+1}{n}\theta \quad \text{and} \quad A^* = \sqrt[n]{1-\frac{\alpha}{2}}\frac{n+1}{n}\theta,$$

that is,

$$P\left(\sqrt[n]{\frac{\alpha}{2}}\frac{n+1}{n}\theta < U < \sqrt[n]{1-\frac{\alpha}{2}}\frac{n+1}{n}\theta\right) = 1-\alpha,$$

from which we can derive easily

$$P\left(\frac{U}{\sqrt[n]{1-\frac{\alpha}{2}}\frac{n+1}{n}} < \theta < \frac{U}{\sqrt[n]{\frac{\alpha}{2}}\frac{n+1}{n}}\right) = 1-\alpha.$$

This expression allows us to read the values A and B directly.

Given: A family of binomial distributions, that is,

$$b_X(x;\theta,n) = \binom{n}{x}\theta^x(1-\theta)^{n-x}.$$

Desired: A confidence interval for the unknown parameter θ.

For an exact computation of a confidence interval for the parameter θ, we start in analogy to the above example with

$$P(X \geq A^*) = \sum_{i=A^*}^{n}\binom{n}{i}\theta^i(1-\theta)^{n-i} \leq \frac{\alpha}{2},$$

$$P(X \leq B^*) = \sum_{i=0}^{B^*}\binom{n}{i}\theta^i(1-\theta)^{n-i} \leq \frac{\alpha}{2}.$$

However, these expressions are often difficult to evaluate. Hence one often chooses an alternative approach, namely the approximate computation of a confidence interval based on the central limit theorem (see Theorem A.19 on page 348). This theorem allows us to approximate the binomial distribution by a standard normal distribution:

$$b_X(x;\theta,n) \approx N\left(\frac{x-n\theta}{\sqrt{n\theta(1-\theta)}};0,1\right) = \frac{1}{\sqrt{2\pi n\theta(1-\theta)}}\exp\left(-\frac{(x-n\theta)^2}{2n\theta(1-\theta)}\right).$$

This approximation is fairly good already for $n\theta(1-\theta) > 9$.

Similar to the approach above, we now start from

$$P\left(\frac{X-n\theta}{\sqrt{n\theta(1-\theta)}} \leq B^*\right) = \frac{\alpha}{2} \quad \text{and}$$

$$P\left(\frac{X-n\theta}{\sqrt{n\theta(1-\theta)}} \geq A^*\right) = \frac{\alpha}{2}.$$

Note that this expression does not contain the estimator (here $T = \frac{X}{n}$) for the unknown parameter θ itself but a function of this estimator (because of the approximation used).

Due to the symmetry of the normal distribution, the computations become fairly simple. For example, due to this symmetry, we know that $B^* = -A^*$. Hence we can write

$$\begin{aligned} & P\left(-A^* < \frac{X - n\theta}{\sqrt{n\theta(1-\theta)}} < A^*\right) \\ &= \int_{-A^*}^{A^*} \frac{1}{\sqrt{2\pi n\theta(1-\theta)}} \exp\left(-\frac{(x-n\theta)^2}{2n\theta(1-\theta)}\right) \mathrm{d}x \\ &= \Phi\left(A^*\right) - \Phi\left(-A^*\right) = 2\Phi\left(A^*\right) - 1 = 1 - \alpha, \end{aligned}$$

where Φ is the distribution function of the standard normal distribution. This function cannot be computed analytically but is available in tabulated form, so that one can easily find the value x that corresponds to a given value $\Phi(x)$. Thus we only have to derive an expression $P(A < \theta < B)$ from the above expression. This is done as follows:

$$\begin{aligned} & -A^* < \frac{X - n\theta}{\sqrt{n\theta(1-\theta)}} < A^* \\ \Rightarrow \quad & |X - n\theta| < A^* \sqrt{n\theta(1-\theta)} \\ \Rightarrow \quad & (X - n\theta)^2 < \left(A^*\right)^2 n\theta(1-\theta) \\ \Rightarrow \quad & \theta^2\left(n\left(A^*\right)^2 + n^2\right) - \theta\left(2nX + \left(A^*\right)^2 n\right) + X^2 < 0. \end{aligned}$$

From the resulting quadratic equation we easily obtain the values of A and B as

$$A/B = \frac{1}{n + (A^*)^2}\left(X + \frac{(A^*)^2}{2} \mp A^* \sqrt{\frac{X(n-X)}{n} + \frac{(A^*)^2}{4}}\right),$$

where $\Phi(A^*) = 1 - \frac{\alpha}{2}$.

A.4.3 Hypothesis Testing

A hypothesis test is a statistical procedure where a decision is made between two contrary hypotheses about the data generating process. The hypotheses may refer to the value of a parameter *(parameter test)*, to a distribution assumption *(goodness-of-fit test)*, or to the dependence or independence of two quantities *(dependence test)*. One of the two hypotheses is preferred, that is, in case of doubt the decision is made in its favor. The preferred hypothesis is called the **null hypothesis** H_0, and the other is called the **alternative hypothesis** H_a. Only if sufficiently strong evidence is available against the null hypothesis, then the alternative hypothesis is accepted (and thus the null hypothesis is rejected). One also says that the null hypothesis receives the benefit of the doubt.[25]

[25] Alternatively, one may say that a court trial is held against the null hypothesis, where the data (sample) act as evidence. In case of doubt the defendant is acquitted (the null hypothesis is accepted). Only if the evidence is sufficiently incriminating, the defendant is convicted (the null hypothesis is rejected).

The test decision is made on the basis of a **test statistic**, that is, a function of the sample values of the given data set. The null hypothesis is rejected if the value of the test statistic lies in the so-called **critical region** C. The development of a statistical test consists in choosing, for a given distribution assumption and a parameter, an appropriate test statistic and then to determine, for a user-specified *significance level* (see the next section), the corresponding critical region C (see the following sections).

A.4.3.1 Error Types and Significance Level

Since the data on which the test decision rests is the outcome of a random process, we cannot be sure that the decision made with a hypothesis test is correct. We may decide wrongly and may do so in two different ways:

- **error of the first kind**:
  The null hypothesis H_0 is rejected, even though it is correct.
- **error of the second kind**:
  The null hypothesis H_0 is accepted, even though it is wrong.

Errors of the first kind are seen as more severe, because the null hypothesis receives the benefit of the doubt and thus is not rejected as easily as the alternative hypothesis. If the null hypothesis is rejected nevertheless, despite being correct, we commit a serious error. Therefore it is tried to limit the probability of an error of the first kind to a certain maximal value. This maximal value α is called the **significance level** of the hypothesis test. It has to be chosen by a user. Typical values of the significance level are 10%, 5%, or 1%.

A.4.3.2 Parameter Test

In a **parameter test** the contrary hypotheses make statements about the values of one or more parameters. For example, the null hypothesis may be that the true value of a parameter θ is at least (or at most) θ_0:

$$H_0: \quad \theta \geq \theta_0, \qquad H_a: \quad \theta < \theta_0.$$

In such a case the test is called **one-sided**. On the other hand, in a **two-sided** test the null hypothesis consists of a statement that the true value of a parameter lies in a certain interval or equals a specific value. Other forms of parameter tests compare the parameters of the distributions that underlie two different samples. Here we only consider a one-sided test as an example.

For a one-sided test, like the one described above, one usually chooses a point estimator T for the parameter θ as a test statistic. In such a case we will reject the null hypothesis H_0 only if the value of the point estimator T has a value c, which does not exceed the **critical value**. Therefore the critical region is $C = (-\infty, c]$. Hence it is clear that the value c must lie to the left of θ_0, because we will not be

able to reasonably reject H_0 if even the value of the point estimators T exceeds θ_0. However, even a value that is only slightly smaller than θ_0 will not be sufficient to make the probability of an error of the first kind (the null hypothesis H_0 is rejected even though it is correct) sufficiently small. Therefore c must lie at some distance to the left of θ_0. Formally, the critical value c is determined as follows: We consider

$$\beta(\theta) = P_\theta(H_0 \text{ is rejected}) = P_\theta(T \in C),$$

which can be simplified to $\beta(\theta) = P(T \leq c)$ for a one-sided test. The quantity $\beta(\theta)$ is also called the **power** of the test. It describes the probability of a rejection of H_0 dependently on the value of the parameter θ. For all values θ that satisfy the null hypothesis, the value of $\beta(\theta)$ must be less than the significance level α. The reason is that it the null hypothesis is true, we want to reject it at most with probability α in order to commit an error of the first kind at most with this probability. Therefore we must have

$$\max_{\theta:\theta \text{ satisfies } H_0} \beta(\theta) \leq \alpha.$$

For the test we consider here, it is easy to see that the power $\beta(\theta)$ of the test reaches its maximum for $\theta = \theta_0$: the larger the true value of θ, the less likely it is that the test statistic (the point estimator T) yields a value of at most c. Hence we must choose the smallest value θ that satisfies the null hypothesis $H_0 : \theta \geq \theta_0$. The expression reduces to

$$\beta(\theta_0) \leq \alpha.$$

At this point all that is left to do to complete the test is to determine $\beta(\theta_0)$ from the distribution assumption and the point estimator T.

A.4.3.3 Parameter Test Example

As an example, for a parameter test, we consider a one-sided test of the expected value μ of a normal distribution $N(\mu, \sigma^2)$ with known variance σ^2 [1]. That is, we consider the hypotheses

$$H_0\colon \quad \mu \geq \mu_0, \qquad H_a\colon \quad \mu < \mu_0.$$

As a test statistic, we use the standard point estimator for the mean (expected value) of a normal distribution, namely

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i,$$

that is, the arithmetic mean of the sample values. (n is the sample size.) As one can easily check, this estimator has the probability density

$$f_{\bar{X}}(x) = N\left(x; \mu, \frac{\sigma^2}{n}\right).$$

Therefore it is

$$\alpha = \beta(\mu_0) = P_{\mu_0}(\bar{X} \le c) = P\left(\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \le \frac{c - \mu_0}{\sigma/\sqrt{n}}\right) = P\left(Z \le \frac{c - \mu_0}{\sigma/\sqrt{n}}\right)$$

with standard normally distributed random variable Z. (The third step in the above transformation served the purpose to obtain a statement about such a random variable.) Thus we have

$$\alpha = \Phi\left(\frac{c - \mu_0}{\sigma/\sqrt{n}}\right),$$

where Φ is the distribution function of the standard normal distribution, which can be found in a tabulated form in many textbooks. From such a table we obtain the value z_α for which $\Phi(z_\alpha) = \alpha$. Then the critical value is

$$c = \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}}.$$

Note that due to the small value of α, the value of z_α is negative, and therefore c, as already made plausible above, is smaller than μ_0.

In order to give a numeric example, we choose [1] $\mu_0 = 130$ and $\alpha = 0.05$. In addition, let $\sigma = 5.4$, $n = 125$, and $\bar{x} = 128$. From a table of the standard normal distribution we obtain $z_{0.05} \approx -1.645$ and arrive at

$$c_{0.05} \approx 130 - 1.645\frac{5.4}{\sqrt{25}} \approx 128.22.$$

Since $\bar{x} = 128 < 128.22 = c$, the null hypothesis H_0 is rejected. If we had chosen $\alpha = 0.01$ instead, we would have obtained (with $z_{0.01} \approx -2.326$)

$$c_{0.01} \approx 130 - 2.326\frac{5.4}{\sqrt{25}} \approx 127.49,$$

and thus H_0 would not have been rejected.

As an alternative, the significance level can be left unspecified. Instead, one provides the value α from which upward the null hypothesis H_0 is rejected. This value α is also called p**-value**. For the above example, it has the value

$$p = \Phi\left(\frac{128 - 130}{5.4/\sqrt{25}}\right) \approx 0.032.$$

That is, the null hypothesis H_0 is rejected for a significance level above 0.032 but accepted for a significance level less than 0.032. Note, however, that one must **not** choose the significance level **after** computing the p-value as this would undermine the validity of the test. The p-value is only a convenience in order to accommodate the different attitudes of users, some of which are more cautious and thus choose lower significance levels α, while other are more daring and thus choose higher significance levels. From the p-value all users can see whether they would reject or accept the null hypothesis and thus need not follow the choice of the writer.

A.4.3.4 Goodness-of-Fit Test

With a goodness of fit test, it is checked whether two distributions, two empirical distributions, or one empirical and one theoretical coincide. Often a goodness-of-fit test is used to check a distribution assumption, as it is needed for parameter estimation. As an example, we consider the χ^2 goodness-of-fit test for a polynomial distribution: let a one-dimensional data set of size n be given for k attribute values $a_1, \ldots, a_k$. In addition, let p_i^*, $1 \leq i \leq k$, be an assumption about the probabilities with which the attribute values a_i occur. We want to check whether the hypothesis fits the data set, that is, whether the actual probabilities p_i coincide with the hypothetical p_i^*, $1 \leq i \leq k$, or not. Thus we contrast the hypotheses

$$H_0 : \forall i, 1 \leq i \leq k : p_i = p_i^* \quad \text{and} \quad H_a : \exists i, 1 \leq i \leq k : p_i \neq p_i^*.$$

An appropriate test statistic can be derived from the following theorem about polynomially distributed random variables, which describe the frequency of the occurrence of the different values a_i in a sample.

Theorem A.20 *Let $(X_1, \ldots, X_k)$ be a k-dimensional polynomially distributed random variable with parameters $p_1, \ldots, p_k$ and n. Then the random variable*

$$Y = \sum_{i=1}^{k} \frac{(X_i - np_i)^2}{np_i}$$

is approximately χ^2-distributed with $k - 1$ degrees of freedom. (*In order for this approximation to be sufficiently good, it should be $\forall i, 1 \leq i \leq k : np_i \geq 5$. This can always be achieved by combining attribute values and/or random variables.*)

In the expression for calculating the random variable Y, the values of the random variables X_i are compared to their expected values np_i, the deviations are squared (among other reasons, so that positive and negative deviations do not cancel), and summed weighted, with a deviation being weighted the lower, the smaller the expected value is. Since Y is χ^2 distributed, large values are unlikely.

The degrees of freedom result from the number of free parameters of the distribution. The number n is not a free parameter, since it is fixed by the size of the sample. From the k parameters $p_1, \ldots, p_k$ only $k - 1$ can be chosen freely, since it must be $\sum_{i=1}^{k} p_i = 1$. Hence only $k - 1$ of the $k + 1$ parameters of the polynomial distribution remain that determine the degrees of freedom.

By replacing the actual probabilities p_i by the hypothetical p_i^* and replacing the random variables X_i by their realizations (absolute frequency of the occurrence of a_i in the sample), we obtain a test statistic for the goodness-of-fit test, namely

$$y = \sum_{i=1}^{k} \frac{(x_i - np_i^*)^2}{np_i^*}.$$

If the null hypothesis H_0 is correct, that is, if all hypothetical probabilities coincide with the actual ones, it is very unlikely that y takes a large value, since y

is a realization of the random variable Y, which is χ^2 distributed. Therefore the null hypothesis H_0 is rejected if the value of y exceeds a certain critical value c, which depends of the significance level. Hence the critical region is $C = [c, \infty)$. The critical value c is determined from the χ^2 distribution with $k - 1$ degrees of freedom, namely as the value for which $P(Y > c) = \alpha$ (or equivalently, for which $P(Y \leq c) = F_Y(c) = 1 - \alpha$), where F_Y is the distribution function of the χ^2-distribution with $k - 1$ degrees of freedom.

Note that in practice the number k of attribute values may have to be reduced by combining attribute values, in order to ensure that $\forall i, 1 \leq i \leq k : np_i \geq 5$ holds. Otherwise the approximation by the χ^2 distribution is not sufficiently good.

For continuous distributions, the domain of values can be divided into nonoverlapping regions that approximate the hypothetical distribution by a step function (constant function value in each interval). Then the goodness-of-fit test for a polynomial distribution is applied, with each interval yielding an outcome of the experiment. In this case one may either work with a completely determined hypothetical distribution (all parameters being given, test whether a specific distribution fits the data) or one may estimate all or some of the parameters from the data (test whether a distribution of a given type fits the data). In the latter case the degrees of freedom of the χ^2 distribution has to be reduced by 1 for each parameter that is estimated from the data. This is actually plausible, because with each parameter that is estimated from the data, the power of the test should go down. However, this is exactly the effect of reducing the degrees of freedom.

As an alternative of applying the χ^2 goodness-of-fit test to an interval partition, the **Kolmogorov–Smirnov test** may be used for continuous distributions, which does without an interval partition, but directly compares the empirical and hypothetical distribution functions.

A.4.3.5 Goodness-of-Fit Test Example

A die is suspected to be unfair, that is, that when tossed, the die shows the different numbers of pips with different probabilities. In order to test this hypothesis, the die is tossed 30 times and it is counted how frequently the different numbers turn up:

$$x_1 = 2,\ x_2 = 4,\ x_3 = 3,\ x_4 = 5,\ x_5 = 3,\ x_6 = 13.$$

That is, one pip turned up twice, two pips four times, etc. Now we contrast the hypotheses

$$H_0 : \forall i, 1 \leq i \leq 6 : p_i = \frac{1}{6} \quad \text{and} \quad H_a : \exists i, 1 \leq i \leq 6 : p_i \neq \frac{1}{6}.$$

Since $n = 30$, we have $\forall i : np_i = 30\frac{1}{6} = 5$, and thus the prerequisites of Theorem A.20 are satisfied. Hence the χ^2 distribution with 5 degrees of freedom is a good approximation of the random variable Y. We compute the test statistic

$$y = \sum_{i=1}^{6} \frac{(x_i - 30 \cdot \frac{1}{6})^2}{30 \cdot \frac{1}{6}} = \frac{1}{5} \sum_{i=1}^{6} (x_i - 5)^2 = \frac{67}{5} = 13.4.$$

For a significance level of $\alpha_1 = 0.05$ (5% probability for an error of the first kind), the critical value is $c \approx 11.07$, since a χ^2 distributed random variable Y with five degrees of freedom satisfies

$$P(Y \leq 11.07) = F_Y(11.07) = 0.95 = 1 - \alpha_1,$$

as one may easily obtain from tables of the χ^2 distribution. Since $13.4 > 11.07$, the null hypothesis that the die is fair can be rejected on a significance level of $\alpha_1 = 0.05$. However, it cannot be rejected on a significance level of $\alpha_2 = 0.01$, since

$$P(Y \leq 15.09) = F_Y(15.09) = 0.99 = 1 - \alpha_2$$

and $13.4 < 15.09$. The p-value is

$$p = 1 - F_Y(13.4) \approx 1 - 0.9801 = 0.0199.$$

That is, for a significance level of 0.0199 and above, the null hypothesis H_0 is rejected, while for a significance level below 0.0199, however, it is accepted.

A.4.3.6 (In)Dependence Test

With a dependence test, it is checked whether two quantities are dependent. In principle, any goodness-of-fit test can easily be turned into a dependence test: simply compare the empirical joint distribution of two quantities with a hypothetical independent distribution that has the same marginals. In such a case the marginal distributions are usually estimated from the data.

As an example, we consider the χ^2 **dependence test** for two nominal values, which is derived from the χ^2 goodness-of-fit test. Let X_{ij}, $1 \leq i \leq k_1$, $1 \leq j \leq k_2$, be random variables that describe the absolute frequency of the joint occurrence of the values a_i and b_j of two attributes A and B, respectively. Furthermore, let $X_{i.} = \sum_{j=1}^{k_2} X_{ij}$ and $X_{.j} = \sum_{i=1}^{k_1} X_{ij}$ be the marginal frequencies (absolute frequencies of the attribute values a_i and b_j). Then, as a test statistic, we compute

$$y = \sum_{i=1}^{k_1} \sum_{j=1}^{k_2} \frac{(x_{ij} - \frac{1}{n} x_{i.} x_{.j})^2}{\frac{1}{n} x_{i.} x_{.j}} = \sum_{i=1}^{k_1} \sum_{j=1}^{k_2} n \frac{(p_{ij} - p_{i.} p_{.j})^2}{p_{i.} p_{.j}}$$

from the realizations x_{ij}, $x_{i.}$, and $x_{.j}$ of these random variables, which are counted in a sample of size n, or from the estimated joint probabilities $p_{ij} = \frac{x_{ij}}{n}$ and marginal probabilities $p_{i.} = \frac{x_{i.}}{n}$ and $p_{.j} = \frac{x_{.j}}{n}$. The critical value c is determined with the help of the chosen significance level from a χ^2 distribution with $(k_1 - 1)(k_2 - 1)$ degrees of freedom. The degrees of freedom are justified as follows: for the $k_1 \cdot k_2$ probabilities p_{ij}, $1 \leq i \leq k_1$, $1 \leq j \leq k_2$, and for the occurrence of the different combinations of a_i and b_j, it must be $\sum_{i=1}^{k_1} \sum_{j=1}^{k_2} p_{ij} = 1$. Thus $k_1 \cdot k_2 - 1$ free parameters remain. From the data we estimate the k_1 probabilities $p_{i.}$ and the k_2 probabilities $p_{.j}$. However, they must also satisfy $\sum_{j=1}^{k_2} p_{i.} = 1$ and $\sum_{i=1}^{k_1} p_{.j} = 1$, so that the degrees of freedom are reduced by only $(k_1 - 1) + (k_2 - 1)$. In total, we have $k_1 k_2 - 1 - (k_1 - 1) - (k_2 - 1) = (k_1 - 1)(k_2 - 1)$ degrees of freedom.

Appendix B
The R Project

R is an open-source statistics and data analysis software available under the General Public License (GPL). This means especially that R can be downloaded, used, and distributed freely.

R is based on a very simple command-line language that can be used interactively, but also for writing programs in R. The sections in this book referring to R are in no way intended to give a comprehensive introduction to R and do not claim to be complete in anyway. The main purpose of these sections is to enable the reader to apply methods introduced in the "theoretical chapters" directly to their own data. For most of the methods whose usage is explained in R, one or two commands will be sufficient.

This appendix explains how to get started with R a provides quick overview on the very basics of R. More details can be found at the website for R

http://www.r-project.org

B.1 Installation and Overview

R can be downloaded from the above-mentioned website where versions suitable for standard operating systems are available. All is needed to download the corresponding installation file, unzip and start it and then follow the installation instructions. The simplest way is to stick to the proposed default settings.

Once R is installed, double click the R symbol on the desktop that should have been created during installation.

Figure B.1 shows a screenshot of R after the command

```
> plot(iris)
```

has been entered in the console window. Note that the prompt symbol > does not belong to the command. We will always display the prompt symbol in order to distinguish R commands from outputs generated after a command has been entered. Outputs will be shown without the prompt symbol.

M.R. Berthold et al., *Guide to Intelligent Data Analysis,*
Texts in Computer Science 42,
DOI 10.1007/978-1-84882-260-3, © Springer-Verlag London Limited 2010

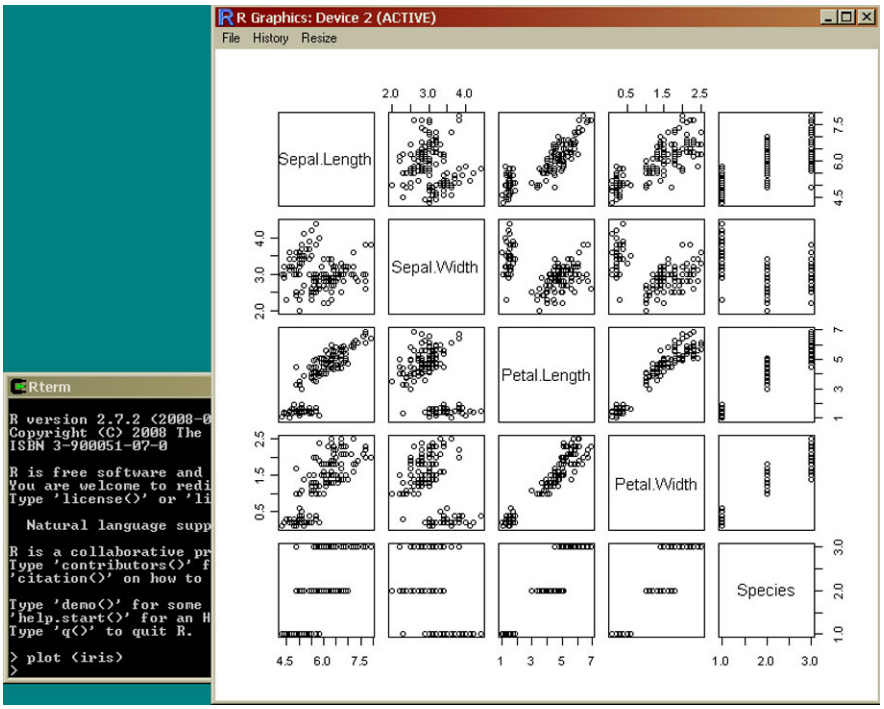


Fig. B.1 A screenshot of R

The graphics displayed on the right in Fig. B.1 is generated by the `plot` command.

B.2 Reading Files and R Objects

R is a type-free language. This means that variables need not to be declared before their use. A variable might be used to store just a single number, but it can also be a complex object with many attributes. In most cases, the objects in this book will contain data sets or analysis results. Assignments in R are denoted by the two symbols `<-`. Before we can analyze a data set with R, we have to load the data set into R. The easiest way to achieve this uses the function `read.table`:

```
> mydata <- read.table(file.choose(),header=TRUE)
```

This command will open a file chooser window to find and select the file with the data set to be analyzed. After the file has been chosen, it will be stored in the variable `mydata`. The specification `header=TRUE` when calling the function `read.table` tells R that the first line in the file should not be interpreted as data,

but rather as names for the attributes. It is therefore assumed that the structure of the file looks like

```
 x      y    z
1.3    2.8   a
3.4    1.9   b
2.7    4.2   a
...    ...  ...
```

In this case, there are three attributes named `x`, `y`, and `z` given in the first line. The records then come in the following lines. Because three attribute names have been given in the first line, each of the following lines must also have three entries separated by an arbitrary number of blanks. If the values of the attributes are not separated by blanks but by another symbol, say a comma, then one would have to write

```
> mydata <- read.table(file.choose(),header=TRUE,sep=",")
```

Now the object named `mydata` contains the data from the file. Assume that the file contains only three records and not more as indicated by the dots above.

At least for smaller data sets, one can take a look at the data by simply typing

```
> mydata
    x   y z
1 1.3 2.8 a
2 3.4 1.9 b
3 2.7 4.2 a
```

or `print(mydata)`, which gives the same result.

The `summary` function gives main properties—in the case of an object representing a data set simple statistics—of an object:

```
> summary(mydata)
       x               y         z
 Min.   :1.300   Min.   :1.900   a:2
 1st Qu.:2.000   1st Qu.:2.350   b:1
 Median :2.700   Median :2.800
 Mean   :2.467   Mean   :2.967
 3rd Qu.:3.050   3rd Qu.:3.500
 Max.   :3.400   Max.   :4.200
```

`print` and `summary` can be applied to any R object. If the corresponding R object is not a data set but a result of some analysis method, then the main properties of the analysis result will be listed.

One can access a specific attribute of an object by writing the name of the object, followed by the symbol $ and then by the name of the variable:

```
> mydata$y
[1] 2.8 1.9 4.2
```

In this simple example, we only have three records and therefore only three values for the attribute y. If one line is not enough to list all values, R will simply continue in the next line and list the index of the first data entry in each line in square brackets. So if we had 15 records, the result might look like the following one:

```
> mydata$y
 [1] 2.8 1.9 4.2 2.4 3.0 1.7
 [7] 4.1 3.3 2.6 1.8 4.3 3.1
[13] 3.7 2.1 1.8
```

We can also assess the values of an attribute—a column in our data table—by using its index in square brackets:

```
> mydata[2]
    y
1 2.8
2 1.9
3 4.2
```

A specific record, a row in our data table, can be selected in the following way:

```
> mydata[2,]
    x   y z
2 3.4 1.9 b
```

Most of the R code examples given in the "Practical ..." section at the ends of the corresponding chapters are based on the Iris data set. It is not necessary to first load this data set into R. R provides some simple data sets, and, among them, there is the Iris data set that can be accessed via an object called `iris`. The attribute names are `Sepal.Length`, `Sepal.Width`, `Petal.Length`, `Petal.Width`, and `Species`. So, if we want to know the mean value of the sepal length, we simply need to enter

```
> mean(iris$Sepal.Length)
[1] 5.843333
```

without loading any file in advance.

B.3 R Functions and Commands

Functions in R can have quite a number of parameters, but, when using a function, most of the parameters do not have to be specified, unless one wants to use a value

different from the default value of the parameter. Such a parameters can be set inside the call of the function by specifying the parameter name, followed by "=" and the value of the parameter. We have seen examples for such parameters already in the function `read.table` in the beginning of the previous section. The default value of the parameter `header` is `FALSE`, assuming that the file to be read does not contain the names of the attributes. Only if we have a file whose first line defines the names of the attributes, we have to use `header=TRUE`, as we have done using the function `read.table`. Another parameter of this function is `sep`. We do not need to specify the value of this parameter when the attribute values in a row of our file are separated by blanks. If another symbol like comma is used to separate the values, we must assign the corresponding value (symbol) to the parameter `sep`.

Not all parameters are specified in this form. As in most programming languages, the data set is just handed over to the function as normal argument as in the very first example of this appendix (`plot(iris)`).

One can browse through the history of commands or functions that have been used in an R session by the key "cursor up" and "cursor down."

B.4 Libraries/Packages

There are various libraries or packages for R for special topics or specialized methods. Some of these libraries come along with R, and other need to be downloaded. Downloading an additional package is very easy. Given that the computer is connected to the Internet, just type

```
> install.packages
```

and wait for the window asking you to choose a mirror site from which you would like to download the package. After you have clicked the mirror site, the packages will be listed in alphabetical order in a new window, and you can choose the package you need by clicking it.

Once a package—for instance, the package `cluster`—has been downloaded, it can be added to an R session by the following command:

```
> library(cluster)
```

When a package has been downloaded once, it is not necessary to download it again. However, unless the workspace (see next the section) is saved and reloaded, the packages must be added to the R session each time R is restarted.

B.5 R Workspace

The actual R session can be stored, so that it can be reloaded next time, and all the R objects, like the data that had been loaded and the analysis results that have been stored in objects, can be recovered:

```
> save(list = ls(all=TRUE), file="all.Rdata")
```

In order to load a workspace that has been stored before, the command

```
> load("all.Rdata", .GlobalEnv)
```

can be used. Of course, the file does not have to be called `all.Rdata`.

B.6 Finding Help

If detailed information about an R function or command is needed,

```
> help(...)
```

will provide the description of the R function or command that has been entered in place of the three dots.

If one does not know the command,

```
> help.search("...")
```

will help. It will list all the R functions/commands in which the specified term that should be given in place of the three dots occurs.

Even if you know the correct name of the command or function you are interested in, R will not be able to provide help if the function belongs to a package that has not been included in the corresponding R session. So

```
> help(scatterplot3d)
```

will not give any information on `scatterplot3d`, unless the package to which the function `scatterplot3d` belongs has been added to the session. (In this case, the package name is even identical with the function name.)

If you know neither the exact function name nor the corresponding package, you can simply search the R website for the topic or use a search engine and type in R the corresponding term for which you would like to find an R function.

B.7 Further Reading

There are various introductory books on R and numerous books on specialized topics like time series analysis or Bayesian statistics where R code is included. It is impossible to list all of them here. As a starting point for R and how to apply basic statistics with R, we refer to [17]. The book [19] provides details for manipulating data and connecting to databases with R. An introduction to programming with R is given in [16]. Those who are not satisfied with the simple default graphics provided by R might like to take a look at the book [18].

Appendix C
KNIME

KNIME, pronounced *[naim]*, is a modular data exploration platform that enables the user to visually create data flows (often referred to as pipelines), selectively execute some or all analysis steps, and later investigate the results through interactive views on data and models. This appendix will give a short introduction to familiarize the readers of this book with the basic usage of KNIME. Considerably more information regarding the use of KNIME is available online at

http://www.knime.org

C.1 Installation and Overview

In order to install KNIME, download one of the versions suitable for your operating system and unzip it to any directory for which you have write permissions. No other action to install KNIME is required, in particular no setup routine has to be launched. In order to start KNIME for the first time, double click the *knime.exe* file on Windows or on Linux launch *knime*.

KNIME is uninstalled from the system by simply deleting the installation directory. Per default the workspace is also in this directory. If a different location for the workspace was chosen, this directory needs to be deleted manually as well.

When KNIME is started the first time, a welcome screen opens. From here the user can

- **Open KNIME workbench**: opens the KNIME workbench to immediately start exploring KNIME, build own workflows, and explore your data.
- **Get additional nodes**: In addition to the ready-to-use basic KNIME installation, there are additional plug-ins for KNIME, e.g., an R and Weka integration, modules for image and text processing, or the integration of the Chemistry Development Kit with additional nodes for the processing of molecular structures. These features can be downloaded also later from within KNIME itself if you choose to skip this step.

M.R. Berthold et al., *Guide to Intelligent Data Analysis,*
Texts in Computer Science 42,
DOI 10.1007/978-1-84882-260-3, © Springer-Verlag London Limited 2010

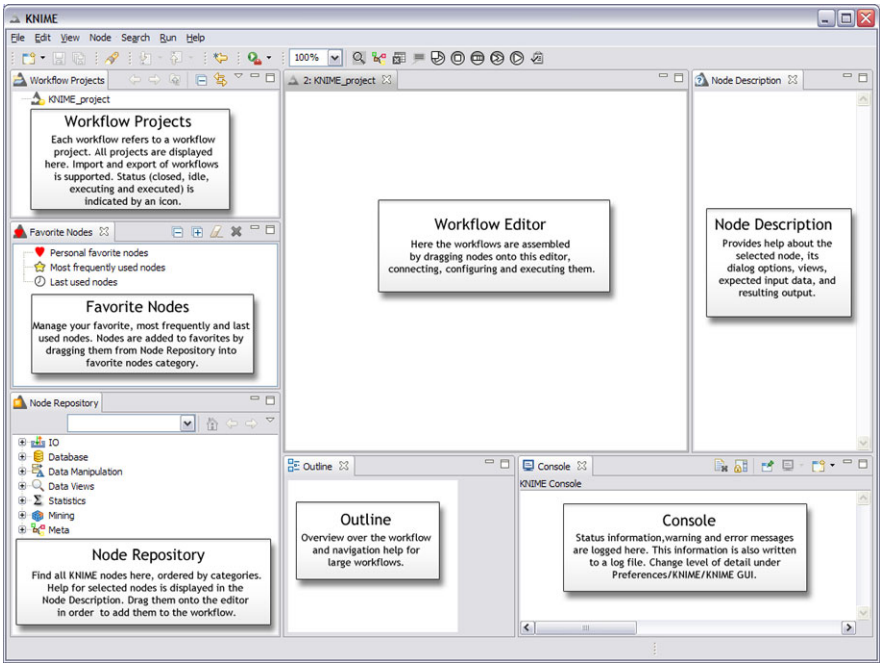


Fig. C.1 The standard outlay of the KNIME workbench

The KNIME Workbench is organized as depicted in Fig. C.1. It consists of different areas:

- Workflow Projects: All KNIME workflows are displayed in the Workflow Projects view. The status of the workflow is indicated by an icon showing whether the workflow is closed, idle, executing, or if execution is complete.
- Favorite Nodes: The Favorite Nodes view displays your favorite, most frequently used, and last used nodes. A node is added to the favorites by dragging it from the node repository into the personal favorite nodes category. Whenever a node is dragged onto the workflow editor, the last used and most frequently used categories are updated.
- Node Repository: The node repository contains all KNIME nodes ordered in categories. A category can contain another category, for example, the Read category is a subcategory of the IO category. Nodes are added from the repository to the workflow editor by dragging them to the workflow editor. Selecting a category displays all contained nodes in the node description view; selecting a node displays the help for this node. If the user knows the name of a node, you can enter parts of the name into the search box of the node repository. As you type, all nodes are filtered immediately to those that contain the entered text in their names.
- Outline: The outline view provides an overview over the whole workflow even if only a small part is visible in the workflow editor (marked in gray in the outline view). The outline view can also be used for navigation: the gray rectangle can be

moved with the mouse, which causes the editor to scroll so that the visible part matches the gray rectangle.

- Console: The console view prints out error and warning messages in order to give you a clue of what is going on under the hood. The same information is written to a log file, which is located in the workspace directory.
- Node Description: The node description displays information about the selected node (or the nodes contained in a selected category). In particular, it explains the dialog options, the available views, the expected input data, and resulting output data.
- Workflow Editor: The workflow editor is used to assemble workflows, configure and execute nodes, inspect the results, and explore your data. This section describes the interactions possible within the editor.

C.2 Building Workflows

A workflow is built by dragging nodes from the *Node Repository* onto the *Workflow Editor* and connecting them there. Nodes are the basic processing units of a workflow. Each node has a number of input and/or output ports. Data (or a model) is transferred over a connection from an out-port to the in-port(s) of other nodes.

When a node is dragged onto the workflow editor, the status light is usually red, which means that the node has to be configured in order to be able to be executed. A node is configured by right clicking it, choosing *Configure*, and adjusting the necessary settings in the node's dialog (see Fig. C.2 on the left).

When the dialog is closed by pressing the *OK* button, the node is configured, and the status light changes to yellow: the node is ready to be executed. A right-click on the node again shows an enabled *Execute* option; pressing it will execute the node,

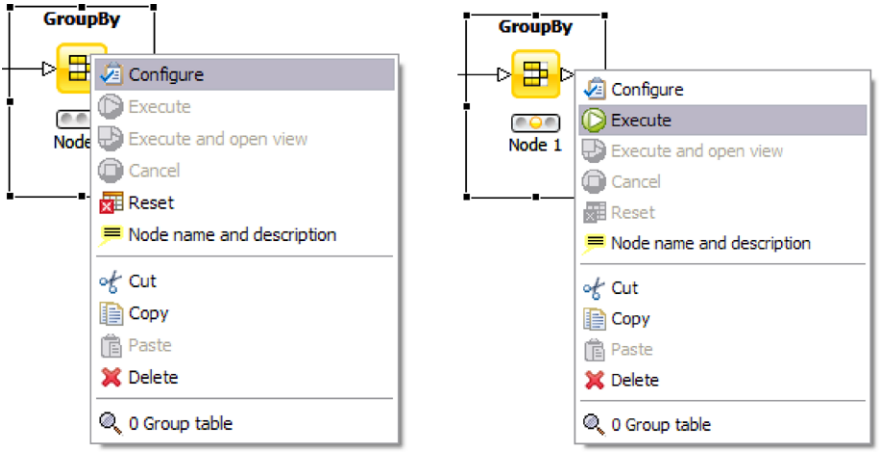


Fig. C.2 The node dialog allows one to configure (*left*) and later execute (*right*) individual nodes

Fig. C.3 Different port types: data, database, PMML, and unspecified ports

and the result of this node will be available at the out-port (see Fig. C.2 on the right). After a successful execution the status light of the node is green, indicating that the processed data is now available on the outports. The result(s) can be inspected by exploring the out-port view(s): the last entries in the context menu open them.

Ports on the left are input ports, where the data from the outport of the predecessor node is fed into the node. Ports on the right are outgoing ports. The result of the node's operation on the data is provided at the out-port to successor nodes. A tooltip gives information about the output of the node.

Nodes are typed such that only ports of the same type can be connected; Fig. C.3 shows the corresponding symbols for the following, most prominently encountered port types:

- Data Ports: The most common type is the data port (a white triangle) which transfers flat data tables from node to node.
- Database Ports: Nodes executing commands inside a database can be identified by their port color and shape (brown square):
- PMML Ports: Data Mining nodes learn a model which is passed to a model writer or predictor node via a blue squared PMML port:
- Other Ports: Whenever a node provides data which does not fit a flat data table structure, a general purpose port for structured data is used (dark cyan square). Ports that are not data, database, PMML, or ports for structured data are displayed as unknown types (gray square):

C.3 Example Flow

This section demonstrates the basic process of building a small, simple workflow: read data from an ASCII file, assign colors based on certain properties, cluster the data, and display it in a table and a scatter plot.

Start KNIME with an empty workflow and create new empty workflow by right clicking in the *Project Repository* and selecting New Project.

In the Node Repository expand the *IO* category and the *Read* subcategory as shown in Fig. C.4. Drag&drop the *File Reader* node onto the *Workflow Editor* window.

The next node will be a learning node, implementing the well-known k-means clustering algorithm. Expand the *Mining* category followed by the *Clustering* category and then drag the *K-Means* node onto the flow.

We will find the third node by using the convenient search node facility. In the search box above the *Node Repository*, enter "color" and press *Enter*. This limits the nodes shown to the ones with "color" in their name (see Fig. C.5). Pull the *Color Manger* node onto the workflow (this node will be used to define the color in the

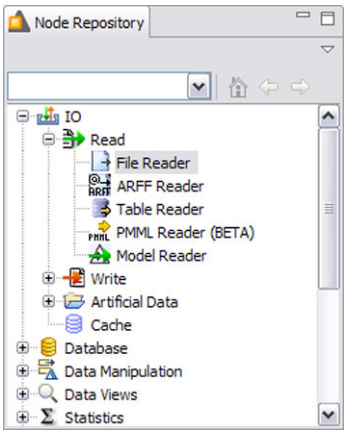


Fig. C.4 Locating the file reader in the node repository

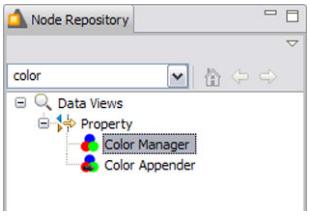


Fig. C.5 Search for nodes in the node repository

data views later). To again see all nodes in the repository, press ESC or Backspace in the search field of the *Node Repository*.

Finally, drag the *Interactive Table* and the *Scatter Plot* from the *Data Views* category to the *Workflow Editor* and position them to the right of the *Color Manager* node.

After placing all nodes, we can now connect them (one can, of course, also later drag new nodes onto the workbench). Click one output (right) port of the file reader and drag the connection to the input port of the k-means node. Then continue to connect the ports as shown in Fig. C.6. (Note that your nodes will not show a green status, as long as they are not configured and executed.)

Some of the now connected nodes may still show a red status icon, indicating that it must be configured in order to produce meaningful results. Right click the *File Reader* and select *Configure* from the menu. Navigate to the *IrisDataSet* directory located in the KNIME installation directory. Select the *data.all* file from this location. The *File Reader*'s preview table shows a sample of the data, which should match the structure of the data file correctly. Click *OK* to confirm this configuration. Once the node has been configured correctly, it switches to yellow (indicating that it is ready for execution). After that, the *K-Means* node will also turn yellow, since its default settings can be applied. To be sure that the default settings fit your needs, open the dialog and inspect the default settings.

In order to configure the *Color Manager* node, you must first execute the *K-Means* node by right clicking the node and selecting *Execute*. Note how the File

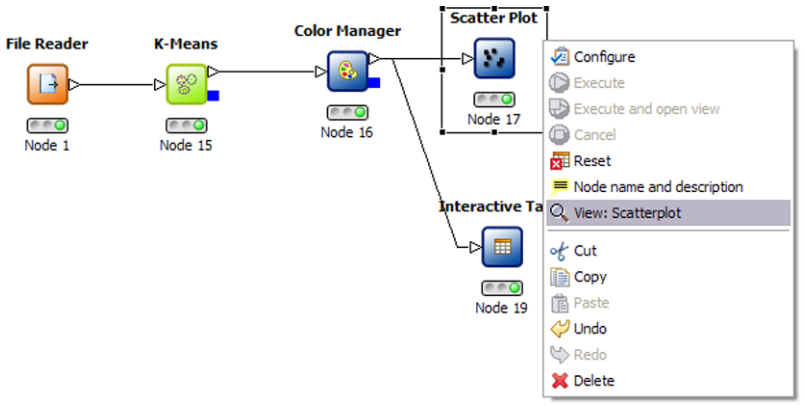


Fig. C.6 The example flow to cluster and visualize the sample data set

Reader node will automatically be executed as well. After execution all nominal values and ranges of all attributes are known at the outport of the executed node: this meta information is propagated to the successor nodes. The *Color Manager* needs this data before it can be configured. Once the *K-Means* node is executed, open the configuration dialog of the *Color Manger* node. The node will suggest to color the rows in our table based on the clustering results. Accept these default settings by clicking *OK*.

Finally, execute the *Scatter Plot*. In order to examine the data and the results, open the nodes' views. In our example, the *K-Means*, the *Interactive Table*, and the *Scatter Plot* have views. Open them from the nodes' context menus.

Select some points in the scatter plot and choose "Hilite Selected" from the "Hilite" menu. The hilited points are marked with an orange border. You will also see the hilited points in the table view. The propagation of the hilite status works for all views in all branches of the flow displaying the same data. Figure C.7 shows an example of the views with a couple of highlighted points.

C.4 R Integration

One of the nice features of KNIME is the modular, open API which allows one to easily integrate other data processing or analysis projects. From the KNIME webpage one can already download a number of such integrations of third party libraries and projects, most notable the statistical data analysis package R, and the machine learning library Weka. In addition, a number of external contributors are providing nodes integrating their own projects into KNIME.

The Weka integration is fairly straightforward to use, one simply drags the node corresponding to the desired learning algorithm onto the workbench, connects it, and opens the configuration dialog which then provides access to all appropriate parameters. If views are available, the KNIME–Weka nodes allow one to open those

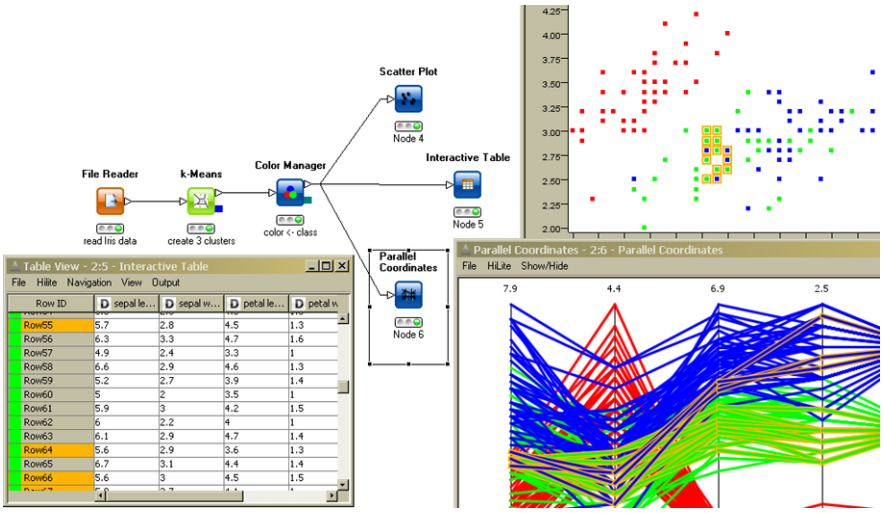


Fig. C.7 A number of open views with highlights patterns

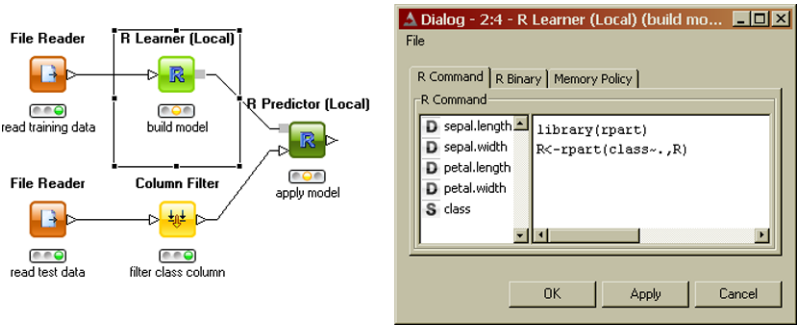


Fig. C.8 A small example workflow using an R code fragment

just like other KNIME views. Weka models can also be fed into a predictor node, similar to KNIME models and hence applied to other data.

The R integration is a bit different, however. Since R really provides more of a statistical programming language, it would require thousand of nodes to cover all the possibilities hidden within the language. KNIME therefore offer nodes which allow one to call small fragments of R code instead—which allows one to use the power of R when needed, e.g., for sophisticated statistical analyses and rely on KNIME's strengths for data loading, integration, and transformation and some of the built-in analysis routines. KNIME offers to point to a local R installation (one can actually download an integrated R installation together with the corresponding KNIME nodes), and it also allows one to access an R installation residing on a server. Dif-

ferent R nodes allow one to execute an R script on incoming data and produce again a data table, a view, or a model. The latter can then be used in the R predictor node and applied to other data. Figure C.8 shows a small example flow and the dialog for an R snippet node.

References

Appendix A

1. Berthold, M., Hand, D.: Intelligent Data Analysis. Springer, Berlin (2009)
2. Buffon, G.-L.L.: Mémoire sur le Jeu Franc-Carreau, France (1733)
3. Everitt, B.S.: The Cambridge Dictionary of Statistics, 3rd edn. Cambridge University Press, Cambridge (2006)
4. Freedman, S., Pisani, R., Purves, R.: Statistics, 4th edn. Norton, London (2007)
5. Friedberg, S.H., Insel, A.J., Spence, L.E.: Linear Algebra, 4th edn. Prentice Hall, Englewood Cliffs (2002)
6. Huff, D.: How to Lie with Statistics. Norton, New York (1954)
7. Kolmogorow, A.N.: Foundations of the Theory of Probability. Chelsea, New York (1956)
8. Krämer, W.: So lügt man mit Statistik, 7 Auflage. Campus-Verlag, Frankfurt (1997)
9. Landau, L.D., Lifshitz, E.M.: Mechanics, 3rd edn. Butterworth-Heinemann, Oxford (1976)
10. Larsen, R.J., Marx, M.L.: An Introduction to Mathematical Statistics and Its Applications, 4th edn. Prentice Hall, Englewood Cliffs (2005)
11. Lay, D.C.: Linear Algebra and Its Applications, 3rd edn. Addison Wesley, Reading (2005)
12. von Mises, R.: Wahrscheinlichkeit, Statistik und Wahrheit. Berlin (1928)
13. Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P.: Numerical Recipes in C—The Art of Scientific Computing, 2nd edn. Cambridge University Press, Cambridge (1992)
14. Sachs, L.: Angewandte Statistik—Anwendung statistischer Methoden, 11 Auflage. Springer, Berlin (2003)
15. Wichura, M.J.: Algorithm AS 241: the percentage points of the normal distribution. Appl. Stat. **37**, 477–484 (1988)

Appendix B

16. Chambers, J.: Software for Data Analysis: Programming with R. Springer, New York (2008)
17. Dalgaard, P.: Introductory Statistics with R, 2nd edn. Springer, New York (2008)
18. Murrell, P.: R Graphics. Chapman & Hall/CRC, Boca Raton (2006)
19. Spector, P.: Data Manipulation with R. Springer, New York (2008)

M.R. Berthold et al., *Guide to Intelligent Data Analysis,*
Texts in Computer Science 42,
DOI 10.1007/978-1-84882-260-3, © Springer-Verlag London Limited 2010

Index

M.R. Berthold et al., *Guide to Intelligent Data Analysis,*
Texts in Computer Science 42,
DOI 10.1007/978-1-84882-260-3, © Springer-Verlag London Limited 2010