Introduction and Framework 1

Statistics is a collection of methods which help us to describe, summarize, interpret, and analyse data. Drawing conclusions from data is vital in research, administration, and business. Researchers are interested in understanding whether a medical intervention helps in reducing the burden of a disease, how personality relates to decision-making, whether a new fertilizer increases the yield of crops, how a political system affects trade policy, who is going to vote for a political party in the next election, what are the long-term changes in the population of a fish species, and many more questions. Governments and organizations may be interested in the life expectancy of a population, the risk factors for infant mortality, geographical differences in energy usage, migration patterns, or reasons for unemployment. In business, identifying people who may be interested in a certain product, optimizing prices, and evaluating the satisfaction of customers are possible areas of interest.

No matter what the question of interest is, it is important to collect data in a way which allows its analysis. The representation of collected data in a **data set** or **data matrix** allows the application of a variety of statistical methods. In the first part of the book, we are going to introduce methods which help us in *describing* data, and the second and third parts of the book focus on inferential statistics, which means *drawing conclusions* from data. In this chapter, we are going to introduce the framework of statistics which is needed to properly collect, administer, evaluate, and analyse data.

1.1 Population, Sample, and Observations

Let us first introduce some terminology and related notations used in this book. The **units** on which we measure data—such as persons, cars, animals, or plants—are called **observations**. These units/observations are represented by the Greek

C. Heumann et al., *Introduction to Statistics and Data Analysis*,
DOI 10.1007/978-3-319-46162-5_1

symbol ω. The collection of all units is called **population** and is represented by Ω. When we refer to $\omega \in \Omega$, we mean a single unit out of all units, e.g. one person out of all persons of interest. If we consider a selection of observations $\omega_1, \omega_2, \ldots, \omega_n$, then these observations are called **sample**. A sample is always a subset of the population, $\{\omega_1, \omega_2, \ldots, \omega_n\} \subseteq \Omega$.

Example 1.1.1

- If we are interested in the social conditions under which Indian people live, then we would define all inhabitants of India as Ω and each of its inhabitants as ω. If we want to collect data from a few inhabitants, then those would represent a sample from the total population.
- Investigating the economic power of Africa's platinum industry would require to treat each platinum-related company as ω, whereas all platinum-related companies would be collected in Ω. A few companies $\omega_1, \omega_2, \ldots, \omega_n$ comprise a sample of all companies.
- We may be interested in collecting information about those participating in a statistics course. All participants in the course constitute the population Ω, and each participant refers to a unit or observation ω.

Remark 1.1.1 Sometimes, the concept of a population is not applicable or difficult to imagine. As an example, imagine that we measure the temperature in New Delhi every hour. A sample would then be the time series of temperatures in a specific time window, for example from January to March 2016. A population in the sense of observational units does not exist here. But now assume that we measure temperatures in several different cities; then, all the cities form the population, and a sample is any subset of the cities.

1.2 Variables

If we have specified the population of interest for a specific research question, we can think of what is of interest about our observations. A particular feature of these observations can be collected in a statistical **variable** X. Any information we are interested in may be captured in such a variable. For example, if our observations refer to human beings, X may describe marital status, gender, age, or anything else which may relate to a person. Of course, we can be interested in many different features, each of them collected in a different variable $X_i, i = 1, 2, \ldots, p$. Each observation ω takes a particular value for X. If X refers to gender, each observation, i.e. each person, has a particular value x which refers to either "male" or "female".

The formal definition of a variable is

$$\begin{aligned} X : \Omega &\to S \\ \omega &\mapsto x \end{aligned} \tag{1.1}$$

This definition states that a variable X takes a value x for each observation $\omega \in \Omega$, whereby the number of possible values is contained in the set S.

Example 1.2.1

- If X refers to gender, possible x-values are contained in $S = \{\text{male, female}\}$. Each observation ω is either male or female, and this information is summarized in X.
- Let X be the country of origin for a car. Possible values to be taken by an observation ω (i.e. a car) are $S = \{\text{Italy, South Korea, Germany, France, India, China, Japan, USA}, \ldots\}$.
- A variable X which refers to age may take any value between 1 and 125. Each person ω is assigned a value x which represents the age of this person.

1.2.1 Qualitative and Quantitative Variables

Qualitative variables are the variables which take values x that cannot be ordered in a logical or natural way. For example,

- the colour of the eye,
- the name of a political party, and
- the type of transport used to travel to work

are all qualitative variables. Neither is there any reason to list blue eyes before brown eyes (or vice versa) nor does it make sense to list buses before trains (or vice versa).

Quantitative variables represent measurable quantities. The values which these variables can take can be ordered in a logical and natural way. Examples of quantitative variables are

- size of shoes,
- price for houses,
- number of semesters studied, and
- weight of a person.

Remark 1.2.1 It is common to assign numbers to qualitative variables for practical purposes in data analyses (see Sect. 1.4 for more detail). For instance, if we consider the variable "gender", then each observation can take either the "value" male or female. We may decide to assign 1 to female and 0 to male and use these numbers instead of the original categories. However, this is arbitrary, and we could have also chosen "1" for male and "0" for female, or "2" for male and "10" for female. There is no logical and natural order on how to arrange male and female, and thus, the variable gender remains a qualitative variable, even after using numbers for coding the values that X can take.

1.2.2 Discrete and Continuous Variables

Discrete variables are variables which can only take a finite number of values. All qualitative variables are discrete, such as the colour of the eye or the region of a country. But also quantitative variables can be discrete: the size of shoes or the number of semesters studied would be discrete because the number of values these variables can take is limited.

Variables which can take an infinite number of values are called **continuous variables**. Examples are the time it takes to travel to university, the length of an antelope, and the distance between two planets. Sometimes, it is said that continuous variables are variables which are "measured rather than counted". This is a rather informal definition which helps to understand the difference between discrete and continuous variables. The crucial point is that continuous variables can, in theory, take an infinite number of values; for instance, the height of a person may be recorded as 172 cm. However, the actual height on the measuring tape might be 172.3 cm which was rounded off to 172 cm. If one had a better measuring instrument, we may have obtained 172.342 cm. But the real height of this person is a number with indefinitely many decimal places such as 172.342975328... cm. No matter what we eventually report or obtain, a variable which can take an infinite amount of values is defined to be a continuous variable.

1.2.3 Scales

The thoughts and considerations from above indicate that different variables contain different amounts of information. A useful classification of these considerations is given by the concept of the **scale** of a variable. This concept will help us in the remainder of this book to identify which methods are the appropriate ones to use in a particular setting.

Nominal scale. The values of a *nominal variable* cannot be ordered. Examples are the gender of a person (male–female) or the status of an application (pending–not pending).

Ordinal scale. The values of an *ordinal variable* can be ordered. However, the differences between these values cannot be interpreted in a meaningful way. For example, the possible values of education level (none–primary education–secondary education–university degree) can be ordered meaningfully, but the differences between these values cannot be interpreted. Likewise, the satisfaction with a product (unsatisfied–satisfied–very satisfied) is an ordinal variable because the values this variable can take can be ordered, but the differences between "unsatisfied–satisfied" and "satisfied–very satisfied" cannot be compared in a numerical way.

Continuous scale. The values of a *continuous variable* can be ordered. Furthermore, the differences between these values can be interpreted in a meaningful way. For instance, the height of a person refers to a continuous variable because the values can be ordered (170 cm, 171 cm, 172 cm, ...), and differences between these values can be compared (the difference between 170 and 171 cm is the same as the difference between 171 and 172 cm). Sometimes, the continuous scale is divided further into subscales. While in the remainder of the book we typically do not need these classifications, it is still useful to reflect on them:

Interval scale. Only differences between values, but not ratios, can be interpreted. An example for this scale would be temperature (measured in °C): the difference between $-2\,°C$ and $4\,°C$ is $6\,°C$, but the ratio of $4/-2=-2$ does not mean that $-4\,°C$ is twice as cold as $2\,°C$.

Ratio scale. Both differences and ratios can be interpreted. An example is speed: 60 km/h is 40 km/h more than 20 km/h. Moreover, 60 km/h is three times faster than 20 km/h because the ratio between them is 3.

Absolute scale. The absolute scale is the same as the ratio scale, with the exception that the values are measured in "natural" units. An example is "number of semesters studied" where no artificial unit such as km/h or °C is needed: the values are simply $1, 2, 3, \ldots$.

1.2.4 Grouped Data

Sometimes, data may be available only in a summarized form: instead of the original value, one may only know the category or group the value belongs to. For example,

- it is often convenient in a survey to ask for the income (per year) by means of groups: [€0–€20,000), [€20,000–€30,000), . . ., > €100,000;
- if there are many political parties in an election, those with a low number of voters are often summarized in a new category "Other Parties";
- instead of capturing the number of claims made by an insurance company customer, the variable "claimed" may denote whether or not the customer claimed at all (yes–no).

If data is available in grouped form, we call the respective variable capturing this information a **grouped variable**. Sometimes, these variables are also known as **categorical variables**. This is, however, not a complete definition because categorical variables refer to any type of variable which takes a finite, possibly small, number of values. Thus, any discrete and/or nominal and/or ordinal and/or qualitative variable may be regarded as a categorical variable. Any grouped or categorical variable which can only take two values is called a **binary variable**.

To gain a better understanding on how the definitions from the above sections relate to each other see Fig. 1.1. Qualitative data is always discrete, but quantitative data can be both discrete (e.g. size of shoes or a grouped variable) and continuous (e.g. temperature). Nominal variables are always qualitative and discrete (e.g. colour of the eye), whereas continuous variables are always quantitative (e.g. temperature). Categorical variables can be both qualitative (e.g. colour of the eye) and quantitative (satisfaction level on a scale from 1 to 5). Categorical variables are never continuous.

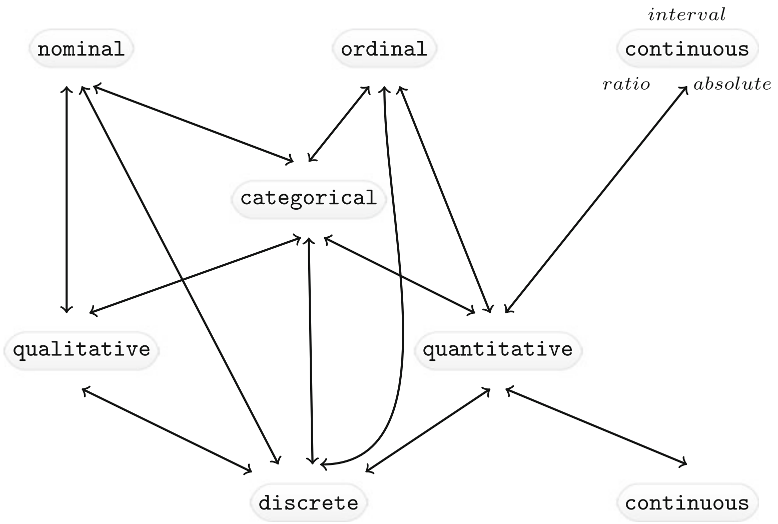


Fig. 1.1 Summary of variable classifications

1.3 Data Collection

When collecting data, we may ask ourselves how to facilitate this in detail and how much data needs to be collected. The latter question will be partly answered in Sect. 9.5; but in general, we can think of collecting data either on all subjects of interest, such as in a national census, or on a representative sample of the population. Most commonly, we gather data on a sample (described in the Part I of this book) and then draw conclusions about the population of interest (discussed in the Part III of this book). A sample might either be chosen by us or obtained through third parties (hospitals, government agencies), or created during an experiment. This depends on the context as described below.

Survey. A survey typically (but not always) collects data by asking questions (in person or by phone) or providing questionnaires to study participants (as a printout or online). For example, an opinion poll before a national election provides evidence about the future government: potential voters are asked by phone which party they are going to vote for in the next election; on the day of the election, this information can be updated by asking the same question to a sample of voters who have just delivered their vote at the polling station (so-called exit poll). A behavioural research survey may ask members of a community about their knowledge and attitudes towards drug use. For this purpose, the study coordinators can send people with a questionnaire to this community and interview members of randomly selected households.

Ideally, a survey is conducted in a way which makes the chosen sample representative of the population of interest. If a marketing company interviews people in a pedestrian zone to find their views about a new chocolate bar, then these people may not be representative of those who will potentially be interested in this product. Similarly, if students are asked to fill in an online survey to evaluate a lecture, it may turn out that those who participate are on average less satisfied than those who do not. Survey sampling is a complex topic on its own. The interested reader may consult Groves et al. (2009) or Kauermann and Küchenhoff (2011).

Experiment. Experimental data is obtained in "controlled" settings. This can mean many things, but essentially it is data which is generated by the researcher with full control over one or many variables of interest. For instance, suppose there are two competing toothpastes, both of which promise to reduce pain for people with sensitive teeth. If the researcher decided to randomly assign toothpaste A to half of the study participants, and toothpaste B to the other half, then this is an experiment because it is only the researcher who decides which toothpaste is to be used by any of the participants. It is not decided by the participant. The data of the variable toothpaste is controlled by the experimenter. Consider another example where the production process of a product can potentially be reduced by combining two processes. The management could decide to implement the new process in three production facilities, but leave it as it is in the other facilities. The production process for the different units (facilities) is therefore under control of the management. However, if each facility could decide for themselves if they wanted a change or not, it would not be an experiment because factors not directly controlled by the management, such as the leadership style of the facility manager, would determine which process is chosen.

Observational Data. Observational data is data which is collected routinely, without a researcher designing a survey or conducting an experiment. Suppose a blood sample is drawn from each patient with a particular acute infection when they arrive at a hospital. This data may be stored in the hospital's folders and later accessed by a researcher who is interested in studying this infection. Or suppose a government institution monitors where people live and move to. This data can later be used to explore migration patterns.

Primary and Secondary Data. Primary data is data we collect ourselves, i.e. via a survey or experiment. Secondary data, in contrast, is collected by someone else. For example, data from a national census, publicly available databases, previous research studies, government reports, historical data, and data from the internet, among others, are secondary data.

1.4 Creating a Data Set

There is a unique way in which data is prepared and collected to utilize statistical analyses. The data is stored in a data matrix (=data set) with p columns and n rows (see Fig. 1.2). Each row corresponds to an observation/unit ω and each column to a variable X. This means that, for example, the entry in the fourth row and second column (x_{42}) describes the value of the fourth observation on the second variable. The examples below will illustrate the concept of a data set in more detail.

$$\begin{array}{c}
\begin{array}{ccccc} \omega & \text{Variable}\,1 & \text{Variable}\,2 & \cdots & \text{Variable}\,p \end{array}\\
\left(\begin{array}{ccccc}
1 & x_{11} & x_{12} & \cdots & x_{1p}\\
2 & x_{21} & x_{22} & \cdots & x_{2p}\\
\vdots & \vdots & \vdots & & \vdots\\
n & x_{n1} & x_{n2} & \cdots & x_{np}
\end{array}\right)
\end{array}$$

Fig. 1.2 Data set or data matrix

$$\begin{array}{c}
\begin{array}{ccccc} \omega & \text{Music} & \text{Mathematics} & \text{Biology} & \text{Geography} \end{array}\\
\left(\begin{array}{ccccc}
\text{Student}\,A & 65 & 70 & 85 & 45\\
\text{Student}\,B & 77 & 82 & 80 & 60\\
\text{Student}\,C & 78 & 73 & 93 & 68\\
\text{Student}\,D & 88 & 71 & 63 & 58\\
\text{Student}\,E & 75 & 83 & 63 & 57
\end{array}\right)
\end{array}$$

Fig. 1.3 Data set of marks of five students

Example 1.4.1 Suppose five students take examinations in music, mathematics, biology, and geography. Their marks, measured on a scale between 0 and 100 (where 100 is the best mark), can be written down as illustrated in Fig. 1.3. Note that each row refers to a student and each column to a variable. We consider a larger data set in the next example.

Example 1.4.2 Consider the data set described in Appendix A.4. A pizza delivery service captures information related to each delivery, for example the delivery time, the temperature of the pizza, the name of the driver, the date of the delivery, the name of the branch, and many more. To capture the data of all deliveries during one month, we create a data matrix. Each row refers to a particular delivery, therefore representing the observations of the data. Each column refers to a variable. In Fig. 1.4, the variables X_1 (delivery time in minutes), X_2 (temperature in °C), and X_{12} (name of branch) are listed.

$$\begin{array}{c}
\begin{array}{ccccc} \text{Delivery} & \text{Delivery Time} & \text{Temperature} & \cdots & \text{Branch} \end{array}\\
\left(\begin{array}{ccccc}
1 & 35.1 & 68.3 & \cdots & \text{East}\,(1)\\
2 & 25.2 & 71.0 & \cdots & \text{East}\,(1)\\
\vdots & \vdots & \vdots & & \vdots\\
1266 & 35.7 & 60.8 & \cdots & \text{West}\,(2)
\end{array}\right)
\end{array}$$

Fig. 1.4 Pizza data set

Table 1.1 Coding list for branch

Variable	Values	Code
Branch	East	1
	West	2
	Centre	3
	Missing	4

The first row tells us about the features of the first pizza delivery: the delivery time was 35.1 min, the pizza arrived with a temperature of 68.3 °C, and the pizza was delivered from the branch in the East of the city. In total, there were $n = 1266$ deliveries. For nominal variables, such as branch, we may decide to produce a coding list, as illustrated in Table 1.1: instead of referring to the branches as "East", "West", and "Centre", we may simply call them 1, 2, and 3. As we will see in Chap. 11, this has benefits for some analysis methods, though this is not needed in general.

If some values are missing, for example because they were never captured or even lost, then this requires special attention. In Table 1.1, we assign missing values the number "4" and therefore treat them as a separate category. If we work with statistical software (see below), we may need other coding such as `NA` in the statistical software *R* or in Stata. More detail can be found in Appendix A.

Another consideration when collecting data is that of **transformations**: we may have captured the velocity of cars in kilometres per hour, but may need to present the data in miles per hour; we have captured the temperature in degrees Celsius, whereas we need to communicate results in degrees Fahrenheit, or we have created a satisfaction score which we want to range from -5 to $+5$, while the score currently runs from 0 to 20. This is not a problem at all. We can simply create a new variable which reflects the required transformation. However, valid transformations depend on the scale of a variable. Variables on an interval scale can use transformations of the following kind:

$$g(x) = a + bx, \quad b > 0. \tag{1.2}$$

For ratio scales, only the following transformations are valid:

$$g(x) = bx, \quad b > 0. \tag{1.3}$$

In the above equation, a is set to 0 because ratios only stay the same if we respect a variable's natural point of origin.

Example 1.4.3 The temperature in °F relates to the temperature in °C as follows:

$$\begin{array}{rcccc} \text{Temperature in }^\circ\text{F} = & 32 & + & 1.8 & \text{Temperature in }^\circ\text{C} \\ g(x) = & a & + & b & x \end{array}$$

This means that 25 °C relates to $(32 + 1.8 \cdot 25)\,^\circ\text{F} = 77\,^\circ\text{F}$. If X_1 is a variable representing temperature by °C, we can simply create a new variable X_2 which is temperature in °F. Since temperature is measured on an interval scale, this transformation is valid.

Changing currencies is also possible. If we would like to represent the price of a product not in South African Rand but in €, we simply apply the transformation

$$\text{Price in South African Rand} = b \cdot \text{ Price in } €$$

whereby b is the currency exchange rate.

1.4.1 Statistical Software

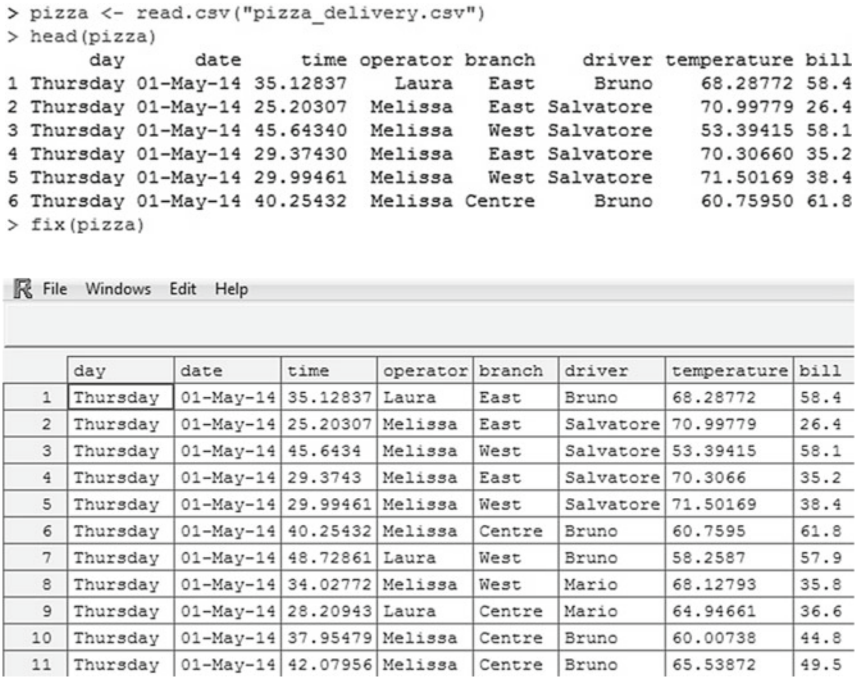
There are number of statistical software packages which allow data collection, management, and–most importantly–analysis. In this book, we focus on the statistical software *R* which is freely available at http://cran.r-project.org/. A gentle introduction to *R* is provided in Appendix A. A data matrix can be created manually using commands such as `matrix()`, `data.frame()`, and others. Any data can be edited using `edit()`. However, typically analysts have already typed their data into databases or spreadsheets, for example in Excel, Access, or MySQL. In most of these applications, it is possible to save the data as an ASCII file (*.dat*), as a tab-delimited file (*.txt*), or as a comma-separated values file (*.csv*). All of these formats allow easy switching between different software and database applications. Such data can easily be read into *R* by means of the following commands:

```
setwd('C:/directory')
read.table('pizza_delivery.dat')
read.table('pizza_delivery.txt')
read.csv('pizza_delivery.csv')
```

where `setwd` specifies the working directory. Alternatively, loading the library `foreign` allows the import of data from many different statistical software packages, notably Stata, SAS, Minitab, SPSS, among others. A detailed description of data import and export can be found in the respective *R* manual available at http://cran.r-project.org/doc/manuals/r-release/R-data.pdf. Once the data is read into *R*, it can be viewed with

```
fix()     # option 1
View()    # option 2
```

We can also can get an overview of the data directly in the *R*-console by displaying only the top lines of the data with `head()`. Both approaches are visualized in Fig. 1.5 for the pizza data introduced in Example 1.4.2.



```
> pizza <- read.csv("pizza_delivery.csv")
> head(pizza)
       day      date     time operator branch    driver temperature bill
1 Thursday 01-May-14 35.12837    Laura   East     Bruno    68.28772 58.4
2 Thursday 01-May-14 25.20307  Melissa   East Salvatore    70.99779 26.4
3 Thursday 01-May-14 45.64340  Melissa   West Salvatore    53.39415 58.1
4 Thursday 01-May-14 29.37430  Melissa   East Salvatore    70.30660 35.2
5 Thursday 01-May-14 29.99461  Melissa   West Salvatore    71.50169 38.4
6 Thursday 01-May-14 40.25432  Melissa Centre     Bruno    60.75950 61.8
> fix(pizza)
```

	day	date	time	operator	branch	driver	temperature	bill
1	Thursday	01-May-14	35.12837	Laura	East	Bruno	68.28772	58.4
2	Thursday	01-May-14	25.20307	Melissa	East	Salvatore	70.99779	26.4
3	Thursday	01-May-14	45.6434	Melissa	West	Salvatore	53.39415	58.1
4	Thursday	01-May-14	29.3743	Melissa	East	Salvatore	70.3066	35.2
5	Thursday	01-May-14	29.99461	Melissa	West	Salvatore	71.50169	38.4
6	Thursday	01-May-14	40.25432	Melissa	Centre	Bruno	60.7595	61.8
7	Thursday	01-May-14	48.72861	Laura	West	Bruno	58.2587	57.9
8	Thursday	01-May-14	34.02772	Melissa	West	Mario	68.12793	35.8
9	Thursday	01-May-14	28.20943	Laura	Centre	Mario	64.94661	36.6
10	Thursday	01-May-14	37.95479	Melissa	Centre	Bruno	60.00738	44.8
11	Thursday	01-May-14	42.07956	Melissa	Centre	Bruno	65.53872	49.5

Fig. 1.5 Viewing data in *R*

1.5 Key Points and Further Issues

Note:

✓ The scale of variables is not only a formalism but an essential framework for choosing the correct analysis methods. This is particularly relevant for association analysis (Chap. 4), statistical tests (Chap. 10), and linear regression (Chap. 11).

✓ Even if variables are measured on a nominal scale (i.e. if they are categorical/qualitative), we may choose to assign a number to each category of this variable. This eases the implementation of some analysis methods introduced later in this book.

✓ Data is usually stored in a data matrix where the rows represent the observations and the columns are variables. It can be analysed with statistical software. We use *R* (R Core Team 2016) in this book. A gentle introduction is provided in Appendix A and throughout the book. A more comprehensive introduction can be found in other books, for example in Albert and Rizzo (2012), Crawley (2013), or Ligges (2008). Even advanced books, e.g. Adler (2012) or Everitt and Hothorn (2011), can offer insights to beginners.

1.6 Exercises

Exercise 1.1 Describe both the population and the observations for the following research questions:

(a) Evaluation of the satisfaction of employees from an airline.
(b) Description of the marks of students from an assignment.
(c) Comparison of two drugs which deal with high blood pressure.

Exercise 1.2 A national park conducts a study on the behaviour of their leopards. A few of the park's leopards are registered and receive a GPS device which allows measuring the position of the leopard. Use this example to describe the following concepts: population, sample, observation, value, and variable.

Exercise 1.3 Which of the following variables are qualitative, and which are quantitative? Specify which of the quantitative variables are discrete and which are continuous:

> Time to travel to work, shoe size, preferred political party, price for a canteen meal, eye colour, gender, wavelength of light, customer satisfaction on a scale from 1 to 10, delivery time for a parcel, blood type, number of goals in a hockey match, height of a child, subject line of an email.

Exercise 1.4 Identify the scale of the following variables:

(a) Political party voted for in an election
(b) The difficulty of different levels in a computer game
(c) Production time of a car
(d) Age of turtles
(e) Calender year
(f) Price of a chocolate bar
(g) Identification number of a student
(h) Final ranking at a beauty contest
(i) Intelligence quotient.

Exercise 1.5 Make yourself familiar with the pizza data set from Appendix A.4.

(a) First, browse through the introduction to *R* in Appendix A. Then, read in the data.
(b) View the data both in the *R* data editor and in the *R* console.
(c) Create a new data matrix which consists of the first 5 rows and first 5 variables of the data. Print this data set on the *R* console. Now, save this data set in your preferred format.
(d) Add a new variable "NewTemperature" to the data set which converts the temperature from °C to °F.

(e) Attach the data and list the values from the variable "NewTemperature".
(f) Use "?" to make yourself familiar with the following commands: `str, dim, colnames, names, nrow, ncol, head, and tail`. Apply these commands to the data to get more information about it.

Exercise 1.6 Consider the research questions of describing parents' attitudes towards immunization, what proportion of them wants immunization against chicken pox for their last-born child, and whether this proportion differs by gender and age.

(a) Which data collection method is the most suitable one to answer the above questions: survey or experiment?
(b) How would you capture the attitudes towards immunization in a single variable?
(c) Which variables are needed to answer all the above questions? Describe the scale of each of them.
(d) Reflect on what an appropriate data set would look like. Now, given this data set, try to write down the above research questions as precisely as possible.

→ Solutions to all exercises in this chapter can be found on p. 321

2 Frequency Measures and Graphical Representation of Data

In Chap. 1, we highlighted that different variables contain different levels of information. When summarizing or visualizing one or more variable(s), it is this information which determines the appropriate statistical methods to use.

Suppose we are interested in studying the employment opportunities and starting salaries of university graduates with a master's degree. Let the variable X denote the starting salaries measured in €/year. Now suppose 100 graduate students provide their initial salaries. Let us write down the salary of the first student as x_1, the salary of the second student as x_2, and so on. We therefore have 100 observations $x_1, x_2, \ldots, x_{100}$. How can we summarize those 100 values best to extract meaningful information from them? The answer to this question depends upon several aspects like the nature of the recorded data, e.g. how many observations have been obtained (either small in number or large in number) or how the data was recorded (either exact values were obtained or the values were obtained in intervals). For example, the starting salaries may be obtained as exact values, say 51,500 €/year, 32,350 €/year, etc. Alternatively, these values could have been summarized in categories such as low income (<30,000 €/year), medium income (30,000–50,000 €/year), high income (50,000–70,000 €/year), and very high income (>70,000 €/year). Another approach is to ask whether the students were employed or not after graduating and record the data in terms of "yes" or "no". It is evident that the latter classification is less detailed than the grouped income data which is less detailed than the exact data. Depending on which conceptualization of "starting salary" we use, we need to choose the approach to summarize the data, that is the 100 values relating to the 100 graduated students.

2.1 Absolute and Relative Frequencies

Discrete Data. Let us first consider a simple example to illustrate our notation.

C. Heumann et al., *Introduction to Statistics and Data Analysis*,
DOI 10.1007/978-3-319-46162-5_2

Example 2.1.1 Suppose there are ten people in a supermarket queue. Each of them is either coded as “F” (if the person is female) or “M” (if the person is male). The collected data may look like

M, F, M, F, M, M, M, F, M, M.

There are now two categories in the data: male (M) and female (F). We use a_1 to refer to the male category and a_2 to refer to the female category. Since there are seven male and three female students, we have 7 values in category a_1, denoted as $n_1 = 7$, and 3 values in category a_2, denoted as $n_2 = 3$. The number of observations in a particular category is called the **absolute frequency**. It follows that $n_1 = 7$ and $n_2 = 3$ are the absolute frequencies of a_1 and a_2, respectively. Note that $n_1 + n_2 = n = 10$, which is the same as the total number of collected observations. We can also calculate the **relative frequencies** of a_1 and a_2 as $f_1 = f(a_1) = \frac{n_1}{n} = \frac{7}{10} = 0.7 = 70\,\%$ and $f_2 = f(a_2) = \frac{n_2}{n} = \frac{3}{10} = 0.3 = 30\,\%$, respectively. This gives us information about the proportions of male and female customers in the queue.

We now extend these concepts to a general framework for the summary of **data on discrete variables**. Suppose there are k categories denoted as $a_1, a_2, \ldots, a_k$ with $n_j (j = 1, 2, \ldots, k)$ observations in category a_j. The **absolute frequency** n_j is defined as the number of units in the jth category a_j. The sum of absolute frequencies equals the total number of units in the data: $\sum_{j=1}^{k} n_j = n$. The **relative frequencies** of the jth class are defined as

$$f_j = f(a_j) = \frac{n_j}{n}, \quad j = 1, 2, \ldots, k. \tag{2.1}$$

The relative frequencies always lie between 0 and 1 and $\sum_{j=1}^{k} f_j = 1$.

Grouped Continuous Data. Data on continuous variables usually has a large number (k) of different values. Sometimes k may even be the same as n and in such a case the relative frequencies become $f_j = \frac{1}{n}$ for all j. However, it is possible to define intervals in which the observed values are contained.

Example 2.1.2 Consider the following $n = 20$ results of the written part of a driving licence examination (a maximum of 100 points could be achieved):

28, 35, 42, 90, 70, 56, 75, 66, 30, 89, 75, 64, 81, 69, 55, 83, 72, 68, 73, 16.

We can summarize the results in class intervals such as 0–20, 21–40, 41–60, 61–80, and 81–100, and the data can be presented as follows:

Class intervals	0–20	21–40	41–60	61–80	81–100
Absolute frequencies	$n_1 = 1$	$n_2 = 3$	$n_3 = 3$	$n_4 = 9$	$n_5 = 4$
Relative frequencies	$f_1 = \frac{1}{20}$	$f_2 = \frac{3}{20}$	$f_3 = \frac{3}{20}$	$f_4 = \frac{9}{20}$	$f_5 = \frac{5}{20}$

We have $\sum_{j=1}^{5} n_j = 20 = n$ and $\sum_{j=1}^{5} f_j = 1$.

Table 2.1 Frequency distribution for discrete data

Class intervals (a_j)	a_1	a_2	...	a_k
Absolute frequencies (n_j)	n_1	n_2	...	n_k
Relative frequencies (f_j)	f_1	f_2	...	f_k

Now, suppose the n observations can be classified into k class intervals $a_1, a_2, \dots, a_k$, where $a_j \, (j = 1, 2, \dots, k)$ contains n_j observations with $\sum_{j=1}^{k} n_j = n$. The relative frequency of the jth class is $f_j = n_j/n$ and $\sum_{j=1}^{k} f_j = 1$. Table 2.1 displays the **frequency distribution** of a discrete variable X.

Example 2.1.3 Consider the pizza delivery service data (Example 1.4.2, Appendix A.4). We are interested in the pizza deliveries by branch and generate the respective frequency table, showing the distribution of the data, using the `table` command in *R* (after reading in and attaching the data) as

```
table(branch)                 # absolute frequencies
table(branch)/length(branch)  # relative frequencies
```

a_j	Centre	East	West
n_j	421	410	435
f_j	$\frac{421}{1266} \approx 0.333$	$\frac{410}{1266} \approx 0.323$	$\frac{435}{1266} \approx 0.344$

We have $n = \sum_j n_j = 1266$ deliveries and $\sum_j f_j = 1$. We can see from this table that each branch has a similar absolute number of pizza deliveries and each branch contributes to about one-third of the total number of deliveries.

2.2 Empirical Cumulative Distribution Function

Another approach to summarize and visualize the (frequency) distribution of variables is the **empirical cumulative distribution function**, often abbreviated as "ECDF". As the name itself suggests, it gives us an idea about the cumulative relative frequencies up to a certain point. For example, say we want to know how many people scored up to 60 points in Example 2.1.2. Then, this can be calculated by adding the number of people in the class intervals 0–20, 21–40, and 41–60, which corresponds to $n_1 + n_2 + n_3 = 1 + 3 + 3 = 7$ and is the **cumulative frequency**. If we want to know the relative frequency of people obtaining up to 60 points, we have to add the relative frequencies of the people in the class intervals 0–20, 21–40, and 41–60 as $f_1 + f_2 + f_3 = \frac{1}{20} + \frac{3}{20} + \frac{3}{20} = \frac{7}{20}$.

Before discussing the empirical cumulative distribution function in a more general framework, let us first understand the concept of ordered values. Suppose the values of height of four people are observed as $x_1 = 180$ cm, $x_2 = 160$ cm, $x_3 = 175$ cm, and $x_4 = 170$ cm. We arrange these values in an order, say ascending order, i.e. first the smallest value (denoted as $x_{(1)}$) and lastly the largest value (denoted as $x_{(4)}$). We obtain

$$x_{(1)} = x_2 = 160 \text{ cm}, \quad x_{(2)} = x_4 = 170 \text{ cm},$$
$$x_{(3)} = x_3 = 175 \text{ cm}, \quad x_{(4)} = x_1 = 180 \text{ cm}.$$

The values $x_{(1)}, x_{(2)}, x_{(3)}$, and $x_{(4)}$ are called **ordered values** for which $x_{(1)} < x_{(2)} < x_{(3)} < x_{(4)}$ holds. Note that x_1 is not necessarily the smallest value but $x_{(1)}$ is necessarily the smallest value. In general, if we have n observations $x_1, x_2, \ldots, x_n$, then the ordered data is $x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$.

Consider n observations $x_1, x_2, \ldots, x_n$ of a variable X, which are arranged in ascending order as $x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$ (and are thus on an at least ordinal scale). The **empirical cumulative distribution function** $F(x)$ is defined as the cumulative relative frequencies of all values a_j, which are smaller than, or equal to, x:

$$F(x) = \sum_{a_j \le x} f(a_j). \tag{2.2}$$

This definition implies that $F(x)$ is a monotonically non-decreasing function, $0 \le F(x) \le 1$, $\lim_{x\to-\infty} F(x) = 0$ (the lower limit of F is 0), $\lim_{x\to+\infty} F(x) = 1$ (the upper limit of F is 1), and $F(x)$ is right continuous.

2.2.1 ECDF for Ordinal Variables

The empirical cumulative distribution function of ordinal variables is a **step function**.

Example 2.2.1 Consider a customer satisfaction survey from a car service company. The 200 customers who had a car service done within the last 30 days were asked to respond regarding their overall level of satisfaction with the quality of the car service on a scale from 1 to 5 based on the following options: 1 = not satisfied at all, 2 = unsatisfied, 3 = satisfied, 4 = very satisfied, and 5 = perfectly satisfied. Based on the frequency of each option, we can calculate the relative frequencies and then plot the empirical cumulative distribution function, either manually (takes longer) or by using *R* (quick):

Satisfaction level (a_j)	$j=1$	$j=2$	$j=3$	$j=4$	$j=5$
n_j	4	16	90	70	20
f_j	4/200	16/200	90/200	70/200	20/200
F_j	4/200	20/200	110/200	180/200	200/200

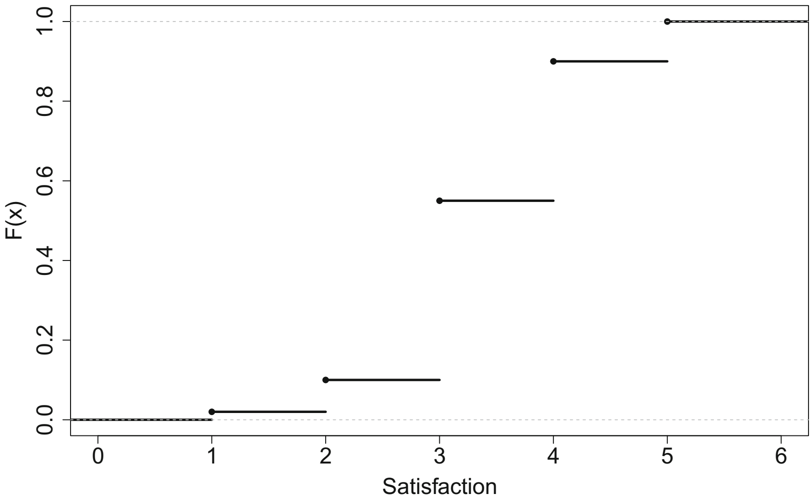


Fig. 2.1 ECDF for the satisfaction survey

The F_j's are calculated as follows:

$$
\begin{aligned}
F_1 &= f_1, & F_3 &= f_1 + f_2 + f_3, \\
F_2 &= f_1 + f_2, & F_4 &= f_1 + f_2 + f_3 + f_4.
\end{aligned}
$$

The ECDF for this data can be obtained by summarizing the data in a vector and using the `plot.ecdf()` function in *R*, see Fig. 2.1:

```
sv <- c(rep(1,4),rep(2,16),rep(3,90),rep(4,70),rep(5,20))
plot.ecdf(sv)
```

The ECDF can be used to obtain the relative frequencies for values contained in certain intervals as

$$H(c \le x \le d) = \text{relative frequency of values } x \text{ with } c \le x \le d.$$

It further follows that:

$$
\begin{aligned}
H(x \le a_j) &= F(a_j) && (2.3) \\
H(x < a_j) &= H(x \le a_j) - f(a_j) = F(a_j) - f(a_j) && (2.4) \\
H(x > a_j) &= 1 - H(x \le a_j) = 1 - F(a_j) && (2.5) \\
H(x \ge a_j) &= 1 - H(X < a_j) = 1 - F(a_j) + f(a_j) && (2.6) \\
H(a_{j_1} \le x \le a_{j_2}) &= F(a_{j_2}) - F(a_{j_1}) + f(a_{j_1}) && (2.7) \\
H(a_{j_1} < x \le a_{j_2}) &= F(a_{j_2}) - F(a_{j_1}) && (2.8) \\
H(a_{j_1} < x < a_{j_2}) &= F(a_{j_2}) - F(a_{j_1}) - f(a_{j_2}) && (2.9) \\
H(a_{j_1} \le x < a_{j_2}) &= F(a_{j_2}) - F(a_{j_1}) - f(a_{j_2}) + f(a_{j_1}) && (2.10)
\end{aligned}
$$

Example 2.2.2 Suppose, in Example 2.2.1, we want to know how many customers are not satisfied with their car service. Then, using the data relating to the responses "1" *and* "2", we observe from the ECDF that $(16+4)/200\,\% = 10\,\%$ of the customers were not satisfied with the car service. This relates to using rule (2.3): $H(X \leq 2) = F(2) = 0.1$ or 10 %. Similarly, the proportion of customers who are more than satisfied can be obtained using (2.5) as $H(X > 3) = 1 - H(x \leq 3) = 1 - 110/200 = 0.45\,\%$ or 45 %.

2.2.2 ECDF for Continuous Variables

In general, we can apply formulae (2.2)–(2.10) to continuous data as well. However, before demonstrating their use, let us consider a somewhat different setting. Let us assume that a continuous variable of interest is only available in the form of grouped data. We may assume that the observations within each group, i.e. each category or each interval, are distributed uniformly over the entire interval. The ECDF then consists of straight lines connecting the lower and upper values of the ECDF in each of the intervals. To understand this concept in more detail, we introduce the following notation:

k	number of groups (or intervals),
e_{j-1}	lower limit of jth interval,
e_j	upper limit of jth interval,
$d_j = e_j - e_{j-1}$	width of the jth interval,
n_j	number of observations in the jth interval.

Under the assumption that all values in a particular interval are distributed uniformly within this interval, the empirical cumulative distribution function relates to a **polygonal chain** connecting the points $(0,0)$, $\big(e_1, F(e_1)\big)$, $\big(e_2, F(e_2)\big)$, $\ldots$, $(e_k, 1)$. The ECDF can then be defined as

$$F(x) = \begin{cases} 0, & x < e_0 \\ F(e_{j-1}) + \dfrac{f_j}{d_j}(x - e_{j-1}), & x \in [e_{j-1}, e_j) \\ 1, & x \geq e_k \end{cases} \tag{2.11}$$

with $F(e_0) = 0$. The idea behind (2.11) is presented in Fig. 2.2. For any interval $[e_{j-1}, e_j)$, the respective lower and upper limits of the ECDF are $F(e_j)$ and $F(e_{j-1})$. If we assume values to be distributed uniformly over this interval, we can connect $F(e_j)$ and $F(e_{j-1})$ with a straight line. To obtain $F(x)$ with $x > e_{j-1}$ and $x < e_j$, we simply add the height of the ECDF between $F(e_{j-1})$ and $F(x)$ to $F(e_{j-1})$.

Example 2.2.3 Consider Example 2.1.3 of the pizza delivery service. Suppose we are interested in determining the distribution of the pizza delivery times. Using the function `plot.ecdf()` in *R*, we obtain the ECDF of the continuous data, see Fig. 2.3a. Note that the structure of the curve is a step function but now almost looks like a continuous curve. The reason for this is that when the number of observations is large, then the lengths of class intervals become small. When these small lengths are

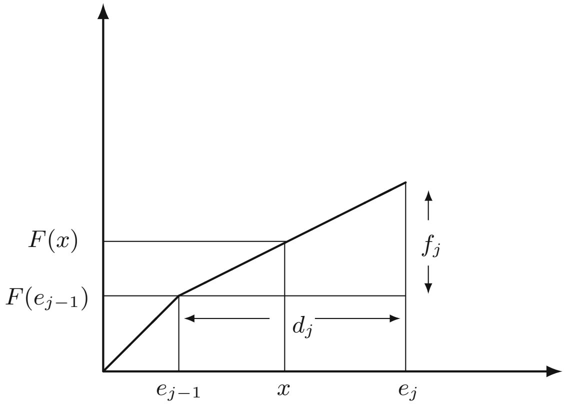


Fig. 2.2 Illustration of the ECDF for continuous data available in groups/intervals*

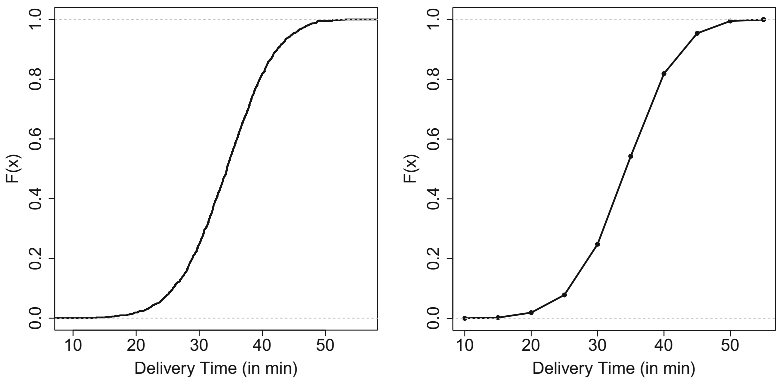
joined together, they appear like a continuous curve. As the number of observations increases, the smoothness of the curve increases too. If the number of observations is not large, e.g. suppose the data is reported as a summary from the drivers, i.e. whether the delivery took <15 min, between 15 and 20 min, between 20 and 25 min, and so on, then we can construct the ECDF by creating a table summarizing the data features as in Table 2.2.

Figure 2.3b shows the ECDF based on the grouped data evaluated in Table 2.2. It is interesting to see that the graphs emerging from the use of the grouped data and ungrouped data are similar in this specific example.

Suppose we are interested in calculating how many deliveries were completed within the desired time limit of 30 min, with a tolerance of maximum 10 % deviation, i.e. a deviation of 3 min. We can evaluate the ECDF at $x = 33$ min.

Table 2.2 The values needed to calculate the ECDF for the grouped pizza delivery time data in Example 2.2.3

Delivery time	j	e_{j-1}	e_j	n_j	f_j	$F(e_j)$
[0; 10]	1	0	10	0	0.0000	0.0000
(10; 15]	2	10	15	3	0.0024	0.0024
(15; 20]	3	15	20	21	0.0166	0.0190
(20; 25]	4	20	25	75	0.0592	0.0782
(25; 30]	5	25	30	215	0.1698	0.2480
(30; 35]	6	30	35	373	0.2946	0.5426
(35; 40]	7	35	40	350	0.2765	0.8191
(40; 45]	8	40	45	171	0.1351	0.9542
(45; 50]	9	45	50	52	0.0411	0.9953
(50; 55]	10	50	55	6	0.0047	1.0000



(a) Continuous data (b) Grouped continuous data

Fig. 2.3 Empirical cumulative distribution function for pizza delivery time

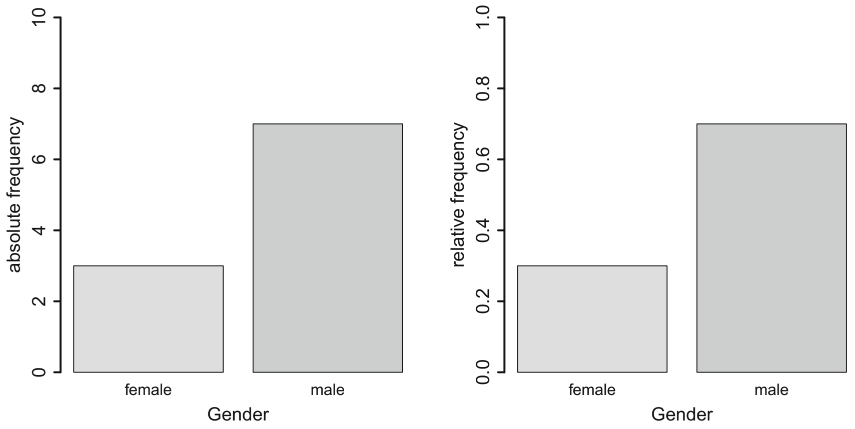
Based on (2.11), we calculate $H(X \leq 33) = F(33) = F(30) + f(6)/5(33 - 30) = 0.2480 + 0.2946/5 \cdot 3 = 0.42476$. Thus, we conclude, based on the grouped data, that only about 42 % of the deliveries were completed in the desired time frame.

2.3 Graphical Representation of a Variable

Frequency tables and empirical cumulative distribution functions are useful in providing a numerical summary of a variable. Graphs are an alternative way to summarize a variable's information. In many situations, they have the advantage of conveying the information hidden in the data more compactly. Similarly, someone's mood can be more easily understood when looking at a smiley ☺ than by reading an essay about one's mood in a long paragraph.

2.3.1 Bar Chart

A simple tool to visualize the relative or absolute frequencies of observed values of a variable is a **bar chart**. A bar chart can be used for nominal and ordinal variables, as long as the number of categories is not very large. It consists of one bar for each category. The height of each bar is determined by either the absolute frequency or the relative frequency of the respective category and is shown on the y-axis. If the variable is measured on an ordinal level, then it is recommended to arrange the bars on the x-axis according to their ranks or values. If the number of categories is large, then the number of bars will be large too and the bar chart, in turn, may not remain informative.



(a) With absolute frequencies

(b) With relative frequencies

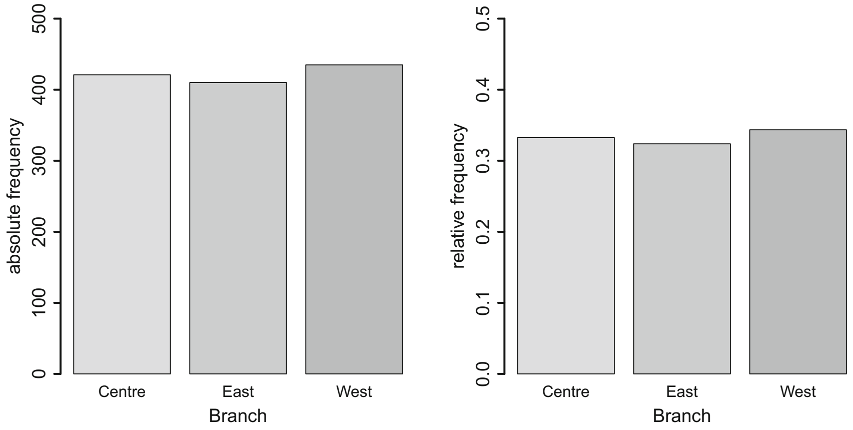
Fig. 2.4 Bar charts

Example 2.3.1 Consider Example 2.1.1 in which ten people, queueing in a supermarket, were classified as being either male (M) or female (F). The absolute frequencies for males and females are $n_1 = 7$ and $n_2 = 3$, respectively. Since there are two categories, M and F, two bars are needed to construct the chart—one for the male category and another for the female category. The heights of the bars are determined as either $n_1 = 7$ and $n_2 = 3$ or $f_1 = 0.7$ and $f_2 = 0.3$. These graphs are shown in Fig. 2.4.

Example 2.3.2 Consider the data in Example 2.1.3, where the pizza delivery times for each branch are recorded over a period of 1 month. The frequency table forms the basis for the bar chart, either using the absolute or relative frequencies on the y-axis. Figure 2.5 shows the bar charts for the number and proportion of pizza deliveries per branch. The graphs can be produced in *R* by applying the `barplot` command to a frequency table:

```
barplot(table(branch))
barplot(table(branch)/length(branch))
```

Remark 2.3.1 Instead of vertical bars, horizontal bars can be drawn using the optional argument `horiz=TRUE` in the `barplot` command.



(a) With absolute frequencies (b) With relative frequencies

Fig. 2.5 Bar charts for the pizza deliveries per branch

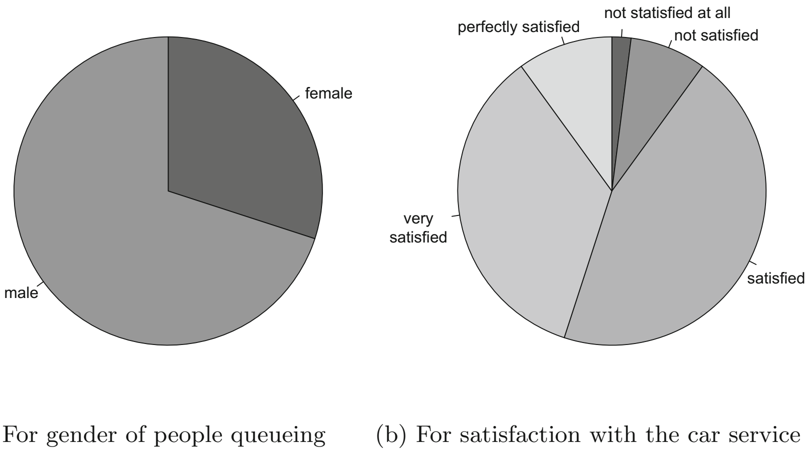
2.3.2 Pie Chart

Pie charts are another option to visualize the absolute and relative frequencies of nominal and ordinal variables. A pie chart is a circle partitioned into segments, where each of the segments represents a category. The size of each segment depends upon the relative frequency and is determined by the angle $f_j \cdot 360°$.

Example 2.3.3 To illustrate the construction of a pie chart, let us consider again Example 2.1.1 in which ten people in a supermarket queue were classified as being either male (M) or female (F): M, F, M, F, M, M, M, F, M, M. The pie chart for this data will have two segments: one for males and another one for females. The relative frequencies are $f_1 = 7/10$ and $f_2 = 3/10$, respectively. The size of the segment for the first category (M) is $f_1 \cdot 360° = (7/10) \cdot 360° = 252°$, and the size of the segment for the second category (F) is $f_2 \cdot 360° = (3/10) \cdot 360° = 108°$. The pie chart is shown in Fig. 2.6a.

Example 2.3.4 Consider again Example 2.2.1, where 200 customers were asked about their level of satisfaction (5 categories) with their car service. The pie chart for this example consists of five segments representing the categories 1, 2, 3, 4, and 5. The size of the jth segment is $f_j \cdot 360°$, $j = 1, 2, 3, 4, 5$. For example, for category 1, there are 4 out of 200 customers, who are not satisfied at all. The angle of the segment "not satisfied at all" therefore is $f_1 \cdot 360° = 4/200 \cdot 360° = 7.2°$. Similarly, we can calculate the angle of the other segments and obtain a pie chart as shown in Fig. 2.6b using the pie command in *R*

```
pie(table(sv))
```



(a) For gender of people queueing (b) For satisfaction with the car service

Fig. 2.6 Pie charts

Remark 2.3.2 Note that the area of a segment is *not* proportional to the absolute frequency of the respective category. Instead, the area of the segment is proportional to the angle $f_j \cdot 360°$ (and depends also on the radius of the whole circle). It has been argued that this may cause improper interpretations as the human eye may catch the segment's area more easily than the angle of a segment. Pie charts should therefore be used with caution.

2.3.3 Histogram

If a variable consists of a large number of different values, the number of categories used to construct bar charts will consequently be large too. A bar chart may thus not give a clear summary when applied to a continuous variable. Instead, a **histogram** is the appropriate choice to represent the distribution of values of continuous variables. It is based on the idea to categorize the data into different groups and plot the bars for each category with height $h_j = f_j/d_j$, where $d_j = e_j - e_{j-1}$ denotes the width of the jth class interval or category. An important consideration for this concept is that the area of the bars (=height × width) is proportional to the relative frequency. This means that the widths of the bars need not necessarily to be the same because different widths can be adjusted with different heights of the bars.

Example 2.3.5 Consider Example 2.1.2, where $n = 20$ people were divided into five class intervals 0–20, 21–40, 41–60, 61–80, and 81–100 based on their performance in a written driving licence examination. The frequency table is given as

Class intervals	0–20	21–40	41–60	61–80	81–100
Absolute freq	$n_1 = 1$	$n_2 = 3$	$n_3 = 3$	$n_4 = 9$	$n_5 = 4$
Relative freq	$f_1 = \frac{1}{20}$	$f_2 = \frac{3}{20}$	$f_3 = \frac{3}{20}$	$f_4 = \frac{9}{20}$	$f_5 = \frac{5}{20}$
Height f_j/d_j	$h_1 = \frac{1}{400}$	$h_2 = \frac{3}{400}$	$h_3 = \frac{3}{400}$	$h_4 = \frac{9}{400}$	$h_5 = \frac{4}{400}$

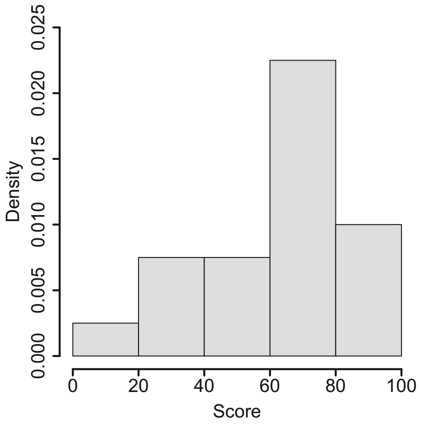


Fig. 2.7 Histogram for the scores of the people

The histogram for this grouped data set has five categories and therefore it has five bars. Since the widths of class intervals are the same, the heights of the bars are proportional to the relative frequency of the respective category. The resulting histogram is displayed in Fig. 2.7.

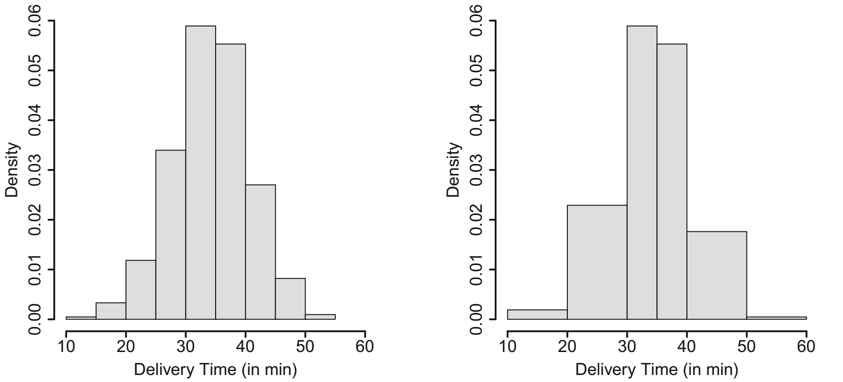
Example 2.3.6 Recall Example 2.2.3 and the variable "pizza delivery time". Table 2.3 shows the summary of the grouped data and the values needed to calculate the histogram. Figure 2.8a shows the histogram with equal widths of delivery time intervals. We see a symmetric distribution of the pizza delivery times, but many delivery times exceeding the target time of 30 min. If the histogram is required to have different widths for different bars, i.e. different delivery time intervals for different categories, then it can also be constructed as shown in Fig. 2.8b. This representation is different from Fig. 2.8a. The following commands in *R* are used to construct the histograms for absolute and relative frequencies, respectively:

```
hist(time)            # show abs. frequencies
hist(time, freq=F)    # show rel. frequencies
```

Remark 2.3.3 The *R* command `truehist()` from the library `MASS` presents an alternative to the `hist()` command. The default specifications are somewhat different, and many users prefer it to the command `hist`.

Table 2.3 Values needed to calculate the histogram for the grouped pizza delivery time data

Delivery time	j	e_{j-1}	e_j	d_j	f_j	h_j
[0; 10]	1	0	10	10	0.0000	0.00000
(10; 15]	2	10	15	5	0.0024	0.00047
(15; 20]	3	15	20	5	0.0166	0.00332
(20; 25]	4	20	25	5	0.0592	0.01185
(25; 30]	5	25	30	5	0.1698	0.03397
(30; 35]	6	30	35	5	0.2946	0.05893
(35; 40]	7	35	40	5	0.2765	0.05529
(40; 45]	8	40	45	5	0.1351	0.02701
(45; 50]	9	45	50	5	0.0411	0.00821
(50; 55]	10	50	55	5	0.0047	0.00094

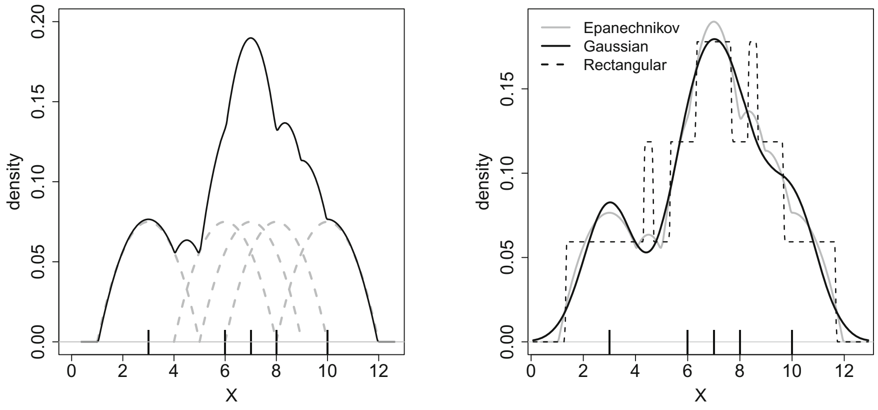


(a) With same width for each category (b) With different widths per category

Fig. 2.8 Histogram for pizza delivery time

2.4 Kernel Density Plots

A disadvantage of histograms is that continuous data is categorized artificially. The choice of the class intervals is crucial for the final look of the graph. A more elegant way to deal with this problem is to smooth the histogram in the sense that each observation may contribute to different classes with different weights, and the distribution is represented by a continuous function rather than a step function. A **kernel density plot** can be produced by using the following function:



(a) Epanechnikow kernel density estimate

(b) Different kernel choices

Fig. 2.9 Construction of kernel density plots

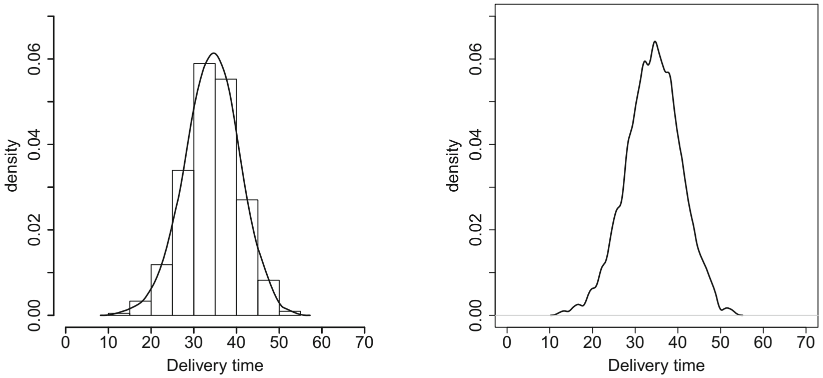
$$\hat{f}_n(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right), \quad h > 0, \tag{2.12}$$

where n is the sample size, h is the bandwidth, and K is a kernel function, for example

$$K(x) = \begin{cases} \frac{1}{2} & \text{if } -1 \leq x \leq 1 \\ 0 & \text{elsewhere} \end{cases} \qquad \text{(rectangular kernel)}$$

$$K(x) = \begin{cases} \frac{3}{4}(1 - x^2) & \text{if } |x| < 1 \\ 0 & \text{elsewhere.} \end{cases} \qquad \text{(Epanechnikov kernel)}$$

To better understand this concept, consider Fig. 2.9a. The tick marks on the x-axis represent five observations: 3, 6, 7, 8, and 10. On each observation x_i as well as its surrounding values, we apply a kernel function, which is the Epanechnikov kernel in the figure. This means that we have five functions (grey, dashed lines), which refer to the five observations. These functions are largest at the observation itself and become gradually smaller as the distance from the observation increases. Summing up the functions, as described in Eq. (2.12), yields the solid black line, which is the kernel density plot of the five observations. It is a smooth curve, which represents the data distribution. The degree of smoothness can be controlled by the bandwidth h, which is chosen as 2 in Fig. 2.9a.



(a) Default bandwidth

(b) Shorter bandwidth

Fig. 2.10 Kernel density plot for delivery time

The choice of the kernel may affect the overall look of the plot. Above, we have given the functions for the rectangular and Epanechnikov kernels. However, another common function for kernel density plots is the normal distribution function, which is introduced in Sect. 8.2.2, see Fig. 2.9b for a comparison of different kernels. The kernel which is based on the normal distribution is called the "Gaussian kernel" and is the default in *R*, where a kernel density plot can be produced combining the `plot` and `density` commands:

```
example <- c(3,6,7,8,10)
plot(density(example, kernel='gaussian'))
```

Please note that kernel functions are not defined arbitrarily and need to satisfy certain conditions, such as those required for probability density functions as explained in Chap. 7, Theorem 7.2.1.

Example 2.4.1 Let us consider the pizza data which we introduced earlier and in Appendix A.4. We can summarize the delivery time by using a kernel density plot using the *R* command `plot(density(time))` and compare it with a histogram, see Fig. 2.10a. We see that the delivery times are symmetric around 35 min. If we shorten the bandwidth to a half of the default bandwidth (option `adjust=0.5`), the kernel density plot becomes more wiggly, which is illustrated in Fig. 2.10b.

2.5 Key Points and Further Issues

Note:

- ✓ Bar charts and histograms are not the same graphical tools. Bar charts visualize the categories of nominal or ordinal variables whereas histograms visualize the distribution of continuous variables. A bar chart does not require to have ordered values on the x-axis, but a histogram always requires the values on the x-axis to be on a continuous scale and to be ordered. The interpretation of a histogram is simplified if the class intervals are equally sized, since then the heights of the rectangles of the histogram are proportional to the absolute or relative frequencies.
- ✓ The ECDF can be used only for ordinal and continuous variables, see Sect. 7.2 for the theoretical background of the cumulative distribution function.
- ✓ A pie chart summarizes observations from a discrete (nominal, ordinal or grouped continuous) variable. It is only useful if the number of different values (categories) is small. It is to be kept in mind that the area of each segment is not proportional to the absolute frequency of the respective category. The angle of the segment is proportional to the relative frequency of the respective category.
- ✓ Other possibilities to visualize the distribution of variables are, for example, box plots (Sect. 3.3) and stratified plots (Sects. 4.1.3, 4.3.1, and 4.4).

2.6 Exercises

Exercise 2.1 Consider the results of the national elections in South Africa in 2014 and 2009:

Party		Results 2014 (%)	Results 2009 (%)
ANC	(African National Congress)	62.15	65.90
DA	(Democratic Alliance)	22.23	16.66
EFF	(Economic Freedom Fighters)	6.35	–
IFP	(Inkatha Freedom Party)	2.40	4.55
COPE	(Congress of the People)	0.67	7.42
Others		6.20	5.47

(a) Summarize the results of the 2014 elections in a bar chart. Do it manually and by using R.

(b) How would you compare the results of the 2009 and 2014 elections? Offer a simple solution that can be represented in a single plot. Construct this plot in R.

Exercise 2.2 Consider a variable X describing the time until the first goal was scored in the matches of the 2006 football World Cup competition. Only matches with at least one goal are considered, and goals during the xth minute of extra time are denoted as $90 + x$:

6	24	90+1	8	4	25	3	83	89	34	25	24	18	6
23	10	28	4	63	6	60	5	40	2	22	26	23	26
44	49	34	2	33	9	16	55	23	13	23	4	8	26
70	4	6	60	23	90+5	28	49	6	57	33	56	7	

(a) What is the scale of X?
(b) Write down the frequency table of X based on the following categories: $[0, 15)$, $[15, 30)$, $[30, 45)$, $[45, 60)$, $[60, 75)$, $[75, 90)$, $[90, 96)$.
(c) Draw the histogram for X with intervals relating to the groups from the frequency table.
(d) Now use *R* to reproduce the histogram. Compare the histogram to a kernel density plot of your choice.
(e) Calculate the empirical cumulative distribution function for the grouped data.
(f) Use *R* to plot the ECDF (via a step function) for

(i) the original data and
(ii) the grouped data.

(g) Consider the grouped data. Now assume that the values within each interval are distributed uniformly. Determine the proportion of first goals which occurred

(i) in the first half, i.e. during the first 45 min,
(ii) in the last 10 min or during the extra time,
(iii) between the 20th and 65th min, i.e. what is $H(20 \leq X \leq 65)$?

(h) Determine the time point at which in 80 % of the matches the first goal was scored at or before this time point.

Exercise 2.3 Suppose we have the following information to construct a histogram for a continuous variable with 2000 observations:

j	e_{j-1}	e_j	d_j	h_j
1	0	1	1	0.125
2	1	4	3	0.125
3	4	7	3	0.125
4	7	8	1	0.125

(a) Determine the relative frequencies for each interval.
(b) Determine the absolute frequencies.

Exercise 2.4 A university survey was conducted on 500 first-year students to obtain knowledge about the size of their accommodation (in square metres).

j	Size of accommodation (m^2) $e_{j-1} \leq x \leq e_j$	$F(x)$
1	8–14	0.25
2	14–22	0.40
3	22–34	0.75
4	34–50	0.97
5	50–82	1.00

(a) Determine the absolute frequencies for each category.
(b) What proportion of people live in a flat of at least 34 m^2?

Exercise 2.5 Consider a survey in which 100 people were asked to rate on a scale from 1 to 10 how much they agree with the statement that "there is too much football on television". The results are summarized below:

Score	0	1	2	3	4	5	6	7	8	9	10
Responses	0	1	3	8	8	27	30	11	6	4	2

(a) Calculate and draw the ECDF of the scores.
(b) Determine $F(3)$ and $F(9)$.
(c) Consider the situation, where the data is summarized in the two categories "disagree" (score ≤ 5) and "agree" (score > 5). What would the ECDF look like under the approach outlined in (2.11)? Determine $F(3)$ and $F(9)$ for the summarized data.
(d) Explain the differences between (b) and (c).

Exercise 2.6 It is possible to produce professional graphics in *R*. However, it is advantageous to go beyond the default options. To demonstrate this, consider Example 2.1.3 about the pizza delivery data, which is described in Appendix A.4.

(a) Set the working directory in *R* (`setwd()`), read in the data (`read.csv()`), and attach the data. Draw a histogram of the variable "temperature". Type `?hist`, and view the options. Adjust the histogram so that you are satisfied with (i) axes labelling, (ii) axes range, and (iii) colour. Now use the `lines()` command to add a dashed vertical line at 65 °C (which is the minimum temperature the pizza should have at the time of delivery).
(b) Consider a different approach, which constructs plots by means of multiple layers using `ggplot2`. You need an Internet connection to install the package using the command `install.packages('ggplot2')`. Browse through the help

pages on http://docs.ggplot2.org/current/. Look specifically at the examples for `ggplot, qplot, scale_histogram, and scale_y_continuous`. Try to understand the roles of "aesthetics" and "geoms". Now, after loading the library via `library(ggplot2)`, create a ggplot object for the pizza data, which declares "temperature" to be the x-variable. Now add a layer with `geom_histogram` to create a histogram with interval width of 2.5 and dark grey bars which are 50 % transparent. Change the y-axis labelling by adding the relevant layer using `scale_y_continuous`. Plot the graph.

(c) Now create a normal bar chart for the variable "driver" in R. Type `?barplot` and `?par` to see the options one can pass on to `barchart()` to adjust the graph. Make the graph look good.

(d) Now create the same bar chart with `ggplot2`. Use `qplot` instead of `ggplot` to create the plot. Use an option which makes each bar to consist of segments relating to the day of delivery, so that one can see the number of deliveries by driver to highlight during which days the drivers delivered most often. Browse through "themes" and "scales" on the help page, and add layers that make the background black and white and the bars on a grey scale.

→ Solutions to all exercises in this chapter can be found on p. 325

**Source* Toutenburg, H., Heumann, C., *Deskriptive Statistik*, 7th edition, 2009, Springer, Heidelberg

3 Measures of Central Tendency and Dispersion

A data set may contain many variables and observations. However, we are not always interested in each of the measured values but rather in a summary which interprets the data. Statistical functions fulfil the purpose of summarizing the data in a meaningful yet concise way.

Example 3.0.1 Suppose someone from Munich (Germany) plans a holiday in Bangkok (Thailand) during the month of December and would like to get information about the weather when preparing for the trip. Suppose last year's maximum temperatures during the day (in degrees Celsius) for December 1–31 are as follows:

$$22, 24, 21, 22, 25, 26, 25, 24, 23, 25, 25, 26, 27, 25, 26,$$
$$25, 26, 27, 27, 28, 29, 29, 29, 28, 30, 29, 30, 31, 30, 28, 29.$$

How do we draw conclusions from this data? Looking at the individual values gives us a feeling about the temperatures one can experience in Bangkok, but it does not provide us with a clear summary. It is evident that the average of these 31 values as "Sum of all values/Total number of observations" $(22 + 24 + \cdots + 28 + 29)/31 = 26.48$ is meaningful in the sense that we know what temperature to expect "on average". To choose the right clothing for the holidays, we may also be interested in knowing the temperature range to understand the variability in temperature, which is between 21 and 31 °C. Summarizing 31 individual values with only three numbers (26.48, 21, and 31) will provide sufficient information to plan the holidays.

In this chapter, we focus on the most important statistical concepts to summarize data: these are measures of central tendency and variability. The applications of each measure depend on the scale of the variable of interest, see Appendix D.1 for a detailed summary.

C. Heumann et al., *Introduction to Statistics and Data Analysis*,
DOI 10.1007/978-3-319-46162-5_3

3.1 Measures of Central Tendency

A natural human tendency is to make comparisons with the “average”. For example, a student scoring 40 % in an examination will be happy with the result if the average score of the class is 25 %. If the average class score is 90 %, then the student may not feel happy even if he got 70 % right. Some other examples of the use of “average” values in common life are mean body height, mean temperature in July in some town, the most often selected study subject, the most popular TV show in 2015, and average income. Various statistical concepts refer to the “average” of the data, but the right choice depends upon the nature and scale of the data as well as the objective of the study. We call statistical functions which describe the average or centre of the data **location parameters** or **measures of central tendency**.

3.1.1 Arithmetic Mean

The **arithmetic mean** is one of the most intuitive measures of central tendency. Suppose a variable of size n consists of the values $x_1, x_2, \ldots, x_n$. The arithmetic mean of this data is defined as

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i . \tag{3.1}$$

In informal language, we often speak of “the average” or just “the mean” when using the formula (3.1).

To calculate the arithmetic mean for grouped data, we need the following frequency table:

Class intervals a_j	$a_1 = e_0 - e_1$	$a_2 = e_1 - e_2$	$\ldots$	$a_k = e_{k-1} - e_k$
Absolute freq. n_j	n_1	n_2	$\ldots$	n_k
Relative freq. f_j	f_1	f_2	$\ldots$	f_k

Note that $a_1, a_2, \ldots, a_k$ are the k class intervals and each interval a_j $(j = 1, 2, \ldots, k)$ contains n_j observations with $\sum_{j=1}^{k} n_j = n$. The relative frequency of the jth class is $f_j = n_j/n$ and $\sum_{j=1}^{k} f_j = 1$. The mid-value of the jth class interval is defined as $m_j = (e_{j-1} + e_j)/2$, which is the mean of the lower and upper limits of the interval. The **weighted arithmetic mean** for grouped data is defined as

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{k} n_j m_j = \sum_{j=1}^{k} f_j m_j . \tag{3.2}$$

Example 3.1.1 Consider again Example 3.0.1 where we looked at the temperature in Bangkok during December. The measurements were

$$22, 24, 21, 22, 25, 26, 25, 24, 23, 25, 25, 26, 27, 25, 26,$$
$$25, 26, 27, 27, 28, 29, 29, 29, 28, 30, 29, 30, 31, 30, 28, 29\,.$$

The arithmetic mean is therefore

$$\bar{x} = \frac{22 + 24 + 21 + \cdots + 28 + 29}{31} = 26.48\,^{\circ}\mathrm{C}.$$

In *R*, the arithmetic mean can be calculated using the `mean` command:

```
weather <- c(22,24,21,,30,28,29)
mean(weather)
[1] 26.48387
```

Let us assume the data in Example 3.0.1 is summarized in categories as follows:

Class intervals	< 20	$(20 - 25]$	$(25, 30]$	$(30, 35]$	> 35
Absolute frequencies	$n_1 = 0$	$n_2 = 12$	$n_3 = 18$	$n_4 = 1$	$n_5 = 0$
Relative frequencies	$f_1 = 0$	$f_2 = \frac{12}{31}$	$f_3 = \frac{18}{31}$	$f_4 = \frac{1}{31}$	$f_5 = 0$

We can calculate the (weighted) arithmetic mean as

$$\bar{x} = \sum_{j=1}^{k} f_j m_j = 0 + \frac{12}{31} \cdot 22.5 + \frac{18}{31} \cdot 27.5 + \frac{1}{31} 32.5 + 0 \approx 25.7.$$

In *R*, we use the `weighted.mean` function to obtain the result. The function requires to specify the (hypothesized) means for each group, for example the middle values of the class intervals, as well as the weights.

```
weighted.mean(c(22.5,27.5,32.5),c(12/31,18/31,1/31))
```

Interestingly, the results of the mean and the weighted mean differ. This is because we use the middle of each class as an approximation of the mean within the class. The implication is that we assume that the values are uniformly distributed within each interval. This assumption is obviously not met. If we had knowledge about the mean in each class, like in this example, we would obtain the correct result as follows:

$$\bar{x} = \sum_{j=1}^{k} f_j \bar{x}_j = 0 + \frac{12}{31} \cdot 23.83333 + \frac{18}{31} \cdot 28 + \frac{1}{31} 32.5 + 0 = 26.48387.$$

However, the weighted mean is meant to estimate the arithmetic mean in those situations where only grouped data is available. It is therefore typically used to obtain an approximation of the true mean.

Properties of the Arithmetic Mean.

(i) The sum of the deviations of each variable around the arithmetic mean is zero:

$$\sum_{i=1}^{n}(x_i - \bar{x}) = \sum_{i=1}^{n} x_i - n\bar{x} = n\bar{x} - n\bar{x} = 0. \tag{3.3}$$

(ii) If the data is linearly transformed as $y_i = a + bx_i$, where a and b are known constants, it holds that

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{n}(a + bx_i) = \frac{1}{n}\sum_{i=1}^{n} a + \frac{b}{n}\sum_{i=1}^{n} x_i = a + b\bar{x}. \tag{3.4}$$

Example 3.1.2 Recall Examples 3.0.1 and 3.1.1 where we considered the temperatures in December in Bangkok. We measured them in degrees Celsius, but someone from the USA might prefer to know them in degrees Fahrenheit. With a linear transformation, we can create a new temperature variable as

$$\text{Temperature in } ^\circ\text{F} = 32 + 1.8\,\text{Temperature in } ^\circ\text{C}.$$

Using $\bar{y} = a + b\bar{x}$, we get $\bar{y} = 32 + 1.8 \cdot 26.48 \approx 79.7\,^\circ$F.

3.1.2 Median and Quantiles

The median is the value which divides the observations into two equal parts such that at least 50 % of the values are greater than or equal to the median and at least 50 % of the values are less than or equal to the median. The median is denoted by $\tilde{x}_{0.5}$; then, in terms of the empirical cumulative distribution function, the condition $F(\tilde{x}_{0.5}) = 0.5$ is satisfied. Consider the n observations $x_1, x_2, \ldots, x_n$ which can be ordered as $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$. The calculation of the median depends on whether the number of observations n is odd or even. When n is odd, then $\tilde{x}_{0.5}$ is the middle ordered value. When n is even, then $\tilde{x}_{0.5}$ is the arithmetic mean of the two middle ordered values:

$$\tilde{x}_{0.5} = \begin{cases} x_{((n+1)/2)} & \text{if } n \text{ is odd} \\ \frac{1}{2}(x_{(n/2)} + x_{(n/2+1)}) & \text{if } n \text{ is even.} \end{cases} \tag{3.5}$$

Example 3.1.3 Consider again Examples 3.0.1–3.1.2 where we evaluated the temperature in Bangkok in December. The ordered values $x_{(i)}$, $i = 1, 2, \ldots, 31$, are as follows:

°C	21	22	22	23	24	24	25	25	25	25	25	25	26	26	26	26
(i)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
°C	27	27	27	28	28	28	29	29	29	29	29	30	30	30	31	
(i)	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	

We have $n = 31$, and therefore $\tilde{x}_{0.5} = x_{((n+1)/2)} = x_{((31+1)/2)} = x_{(16)} = 26$. Therefore, at least 50 % of the 31 observations are greater than or equal to 26 and at least 50 % are less than or equal to 26. If one value was missing, let us say the last observation, then the median would be calculated as $\frac{1}{2}(x_{(30/2)} + x_{(30/2+1)}) = \frac{1}{2}(26 + 26) = 26$. In R, we would have obtained the results using the `median` command:

```
median(weather)
```

If we deal with grouped data, we can calculate the median under the assumption that the values within each class are equally distributed. Let $K_1, K_2, \dots, K_k$ be k classes with observations of size $n_1, n_2, \dots, n_k$, respectively. First, we need to determine which class is the median class, i.e. the class that includes the median. We define the median class as the class K_m for which

$$\sum_{j=1}^{m-1} f_j < 0.5 \quad \text{and} \quad \sum_{j=1}^{m} f_j \geq 0.5 \tag{3.6}$$

hold. Then, we can determine the median as

$$\tilde{x}_{0.5} = e_{m-1} + \frac{0.5 - \sum_{j=1}^{m-1} f_j}{f_m} d_m \tag{3.7}$$

where e_{m-1} denotes the lower limit of the interval K_m and d_m is the width of the interval K_m.

Example 3.1.4 Recall Example 3.1.1 where we looked at the grouped temperature data:

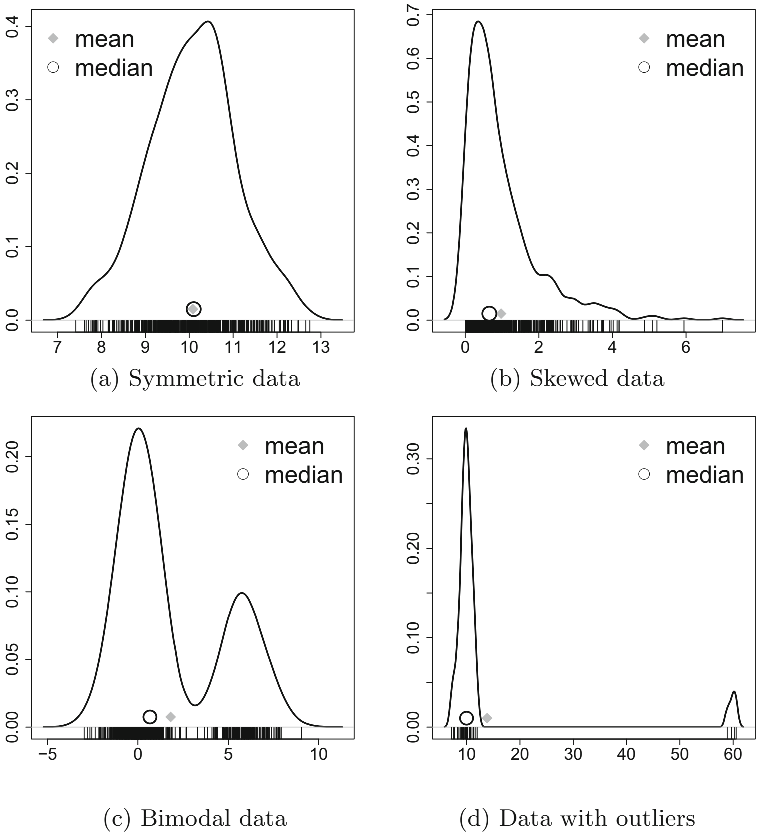
Class intervals	<20	(20–25]	(25, 30]	(30, 35]	>35
n_j	$n_1 = 0$	$n_2 = 12$	$n_3 = 18$	$n_4 = 1$	$n_5 = 0$
f_j	$f_1 = 0$	$f_2 = \frac{12}{31}$	$f_3 = \frac{18}{31}$	$f_4 = \frac{1}{31}$	$f_5 = 0$
$\sum_j f_j$	0	$\frac{12}{31}$	$\frac{30}{31}$	1	1

For the third class ($m = 3$), we have

$$\sum_{j=1}^{m-1} f_j = \frac{12}{31} < 0.5 \quad \text{and} \quad \sum_{j=1}^{m} f_j = \frac{30}{31} \geq 0.5.$$

We can therefore calculate the median as

$$\tilde{x}_{0.5} = e_{m-1} + \frac{0.5 - \sum_{j=1}^{m-1} f_j}{f_m} d_m = 25 + \frac{0.5 - \frac{12}{31}}{\frac{18}{31}} \cdot 5 \approx 25.97.$$



(a) Symmetric data

(b) Skewed data

(c) Bimodal data

(d) Data with outliers

Fig. 3.1 Arithmetic mean and median for different data

Comparing the Mean with the Median. In the above examples, the mean and the median turn out to be quite similar to each other. This is because we looked at data which is symmetrically distributed around its centre, i.e. on average, we can expect 26 °C with deviations that are similar above and below the average temperature. A similar example is given in Fig. 3.1a: we see that the raw data is summarized by using ticks at the bottom of the graph and by using a kernel density estimator. The mean and the median are similar here because the distribution of the observations is symmetric around the centre. If we have skewed data (Fig. 3.1b), then the mean and the median may differ. If the data has more than one centre, such as in Fig. 3.1c, neither the median nor the mean has meaningful interpretations. If we have outliers (Fig. 3.1d), then it is wise to use the median because the mean is sensitive to outliers. These examples show that depending on the situation of interest either the mean, the median, both or neither of them can be useful.

Quantiles. Quantiles are a generalization of the idea of the median. The median is the value which splits the data into two equal parts. Similarly, a quantile partitions the data into other proportions. For example, a 25 %-quantile splits the data into two parts such that at least 25 % of the values are less than or equal to the quantile and at least 75 % of the values are greater than or equal to the quantile. In general, let α be a number between zero and one. The $(\alpha \times 100)\%$-quantile, denoted as $\tilde{x}_\alpha$, is defined as the value which divides the data in proportions of $(\alpha \times 100)\%$ and $(1-\alpha) \times 100\,\%$ such that at least $\alpha \times 100\,\%$ of the values are less than or equal to the quantile and at least $(1-\alpha) \times 100\,\%$ of the values are greater than or equal to the quantile. In terms of the empirical cumulative distribution function, we can write $F(\tilde{x}_\alpha) = \alpha$. It follows immediately that for n observations, at least $n\alpha$ values are less than or equal to $\tilde{x}_\alpha$ and at least $n(1-\alpha)$ observations are greater than or equal to $\tilde{x}_\alpha$. The median is the 50 %-quantile $\tilde{x}_{0.5}$. If α takes the values $0.1, 0.2, \dots, 0.9$, the quantiles are called **deciles**. If $\alpha \cdot 100$ is an integer number (e.g. $\alpha \times 100 = 95$), the quantiles are called **percentiles**, i.e. the data is divided into 100 equal parts. If α takes the values 0.2, 0.4, 0.6, and 0.8, the quantiles are known as **quintiles** and they divide the data into five equal parts. If α takes the values 0.25, 0.5, and 0.75, the quantiles are called **quartiles**.

Consider n ordered observations $x_{(1)} \le x_{(2)} \le \dots \le x_{(n)}$. The $\alpha \cdot 100\,\%$-quantile $\tilde{x}_\alpha$ is calculated as

$$\tilde{x}_\alpha = \begin{cases} x_{(k)} & \text{if } n\alpha \text{ is not an integer number,} \\ & \text{choose } k \text{ as the smallest integer } > n\alpha, \\ \frac{1}{2}(x_{(n\alpha)} + x_{(n\alpha+1)}) & \text{if } n\alpha \text{ is an integer.} \end{cases} \qquad (3.8)$$

Example 3.1.5 Recall Examples 3.0.1–3.1.4 where we evaluated the temperature in Bangkok in December. The ordered values $x_{(i)}$, $i = 1, 2, \dots, 31$ are as follows:

°C	21	22	22	23	24	24	25	25	25	25	25	25	26	26	26	26
(i)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
°C	27	27	27	28	28	28	29	29	29	29	29	30	30	30	31	
(i)	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	

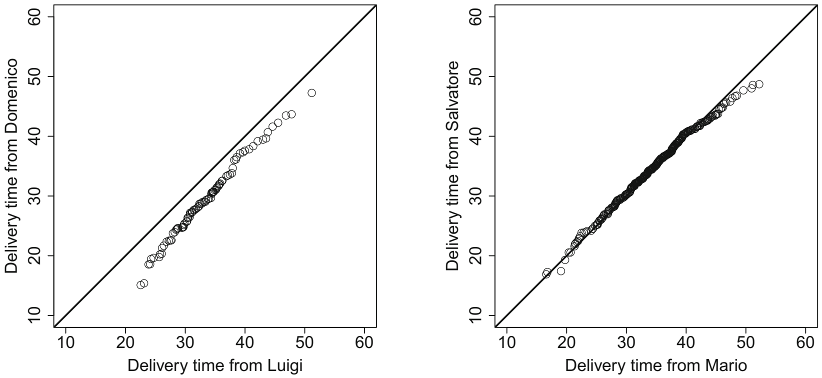
To determine the quartiles, i.e. the 25, 50, and 75 % quantiles, we calculate $n\alpha$ as $31 \cdot 0.25 = 7.75$, $31 \cdot 0.5 = 15.5$, and $31 \cdot 0.75 = 23.25$. Using (3.8), it follows that

$$\tilde{x}_{0.25} = x_{(8)} = 25, \qquad \tilde{x}_{0.5} = x_{(16)} = 26,$$
$$\tilde{x}_{0.75} = x_{(24)} = 29.$$

In *R*, we obtain the same results using the `quantile` function. The `probs` argument is used to specify α. By default, the quartiles are reported.

```
quantile(weather)
quantile(weather, probs=c(0,0.25,0.5,0.75,1))
```

However, please note that *R* offers nine different ways to obtain quantiles, each of which can be chosen by the `type` argument. See Hyndman and Fan (1996) for more details.



(a) QQ-plot for Luigi and Domenico (b) QQ-plot for Mario and Salvatore

Fig. 3.2 QQ-plots for the pizza delivery time for different drivers

3.1.3 Quantile–Quantile Plots (QQ-Plots)

If we plot the quantiles of two variables against each other, we obtain a Quantile–Quantile plot (QQ-plot). This provides a simple summary of whether the distributions of the two variables are similar with respect to their location or not.

Example 3.1.6 Consider again the pizza data which is described in Appendix A.4. We may be interested in the delivery time for different drivers to see if their performance is the same. Figure 3.2a shows a QQ-plot for the delivery time of driver Luigi and the delivery time of driver Domenico. Each point refers to the $\alpha\%$ quantile of both drivers. If the point lies on the bisection line, then they are identical and we conclude that the quantiles of the both drivers are the same. If the point is below the line, then the quantile is higher for Luigi, and if the point is above the line, then the quantile is lower for Luigi. So if all the points lie exactly on the line, we can conclude that the distributions of both the drivers are the same. We see that all the reported quantiles lie below the line, which implies that all the quantiles of Luigi have higher values than those of Domenico. This means that not only on an average, but also in general, the delivery times are higher for Luigi. If we look at two other drivers, as displayed in Fig. 3.2b, the points lie very much on the bisection line. We can therefore conclude that the delivery times of these two drivers do not differ much.

In *R*, we can generate QQ-plots by using the `qqplot` command:

```
qqplot()
```

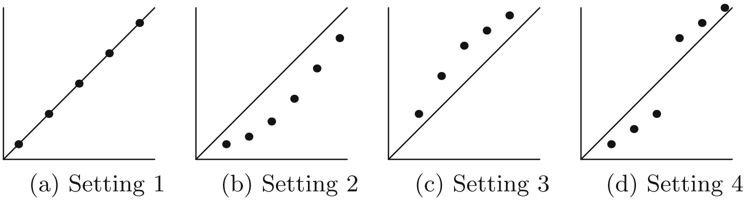


Fig. 3.3 Different patterns for a QQ-plot

As a summary, let us consider four important patterns:

(a) If all the pairs of quantiles lie (nearly) on a straight line at an angle of 45 % from the x-axis, then the two samples have similar distributions (Fig. 3.3a).
(b) If the y-quantiles are lower than the x-quantiles, then the y-values have a tendency to be lower than the x-values (Fig. 3.3b).
(c) If the x-quantiles are lower than the y-quantiles, then the x-values have a tendency to be lower than the y-values (Fig. 3.3c).
(d) If the QQ-plot is like Fig. 3.3d, it indicates that there is a break point up to which the y-quantiles are lower than the x-quantiles and after that point, the y-quantiles are higher than the x-quantiles.

3.1.4 Mode

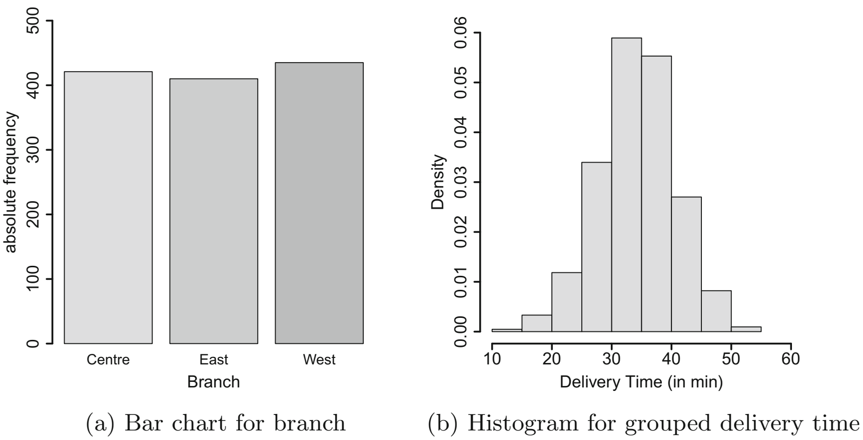
Consider a situation in which an ice cream shop owner wants to know which flavour of ice cream is the most popular among his customers. Similarly, a footwear shop owner may like to find out what design and size of shoes are in highest demand. To answer this type of questions, one can use the mode which is another measure of central tendency.

The mode $\bar{x}_M$ of n observations $x_1, x_2, \ldots, x_n$ is the value which occurs the most compared with all other values, i.e. the value which has maximum absolute frequency. It may happen that two or more values occur with the same frequency in which case the mode is not uniquely defined. A formal definition of the mode is

$$\bar{x}_M = a_j \Leftrightarrow n_j = \max\{n_1, n_2, \ldots, n_k\}. \tag{3.9}$$

The mode is typically applied to any type of variable for which the number of different values is not too large. If continuous data is summarized in groups, then the mode can be used as well.

Example 3.1.7 Recall the pizza data set described in Appendix A.4. The pizza delivery service has three branches, in the East, West, and Centre, respectively. Suppose we want to know which branch delivers the most pizzas. We find that most of the deliveries have been made in the West, see Fig. 3.4a; therefore the mode is $\bar{x}_M$ = West. Similarly, suppose we also want to find the mode for the categorized pizza delivery time: if we group the delivery time in intervals of 5 min, then we see that the most frequent delivery time is the interval "30−35" min, see Fig. 3.4b. The mode is therefore $\bar{x}_M = [30, 35)$.



(a) Bar chart for branch (b) Histogram for grouped delivery time

Fig. 3.4 Results from the pizza data set

3.1.5 Geometric Mean

Consider n observations $x_1, x_2, \ldots, x_n$ which are all positive and collected on a quantitative variable. The geometric mean $\bar{x}_G$ of this data is defined as

$$\bar{x}_G = \sqrt[n]{\prod_{i=1}^{n} x_i} = \left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}} . \tag{3.10}$$

The geometric mean plays an important role in fields where we are interested in products of observations, such as when we look at percentage changes in quantities. We illustrate its interpretation and use by looking at the average growth of a quantity in the sense that we allow a starting value, such as a certain amount of money or a particular population, to change over time. Suppose we have a starting value at some baseline time point 0 (zero), which may be denoted as B_0. At time t, this value may have changed and we therefore denote it as B_t, $t = 1, 2, \ldots, T$. The ratio of B_t and B_{t-1},

$$x_t = \frac{B_t}{B_{t-1}},$$

is called the tth growth factor. The growth rate r_t is defined as

$$r_t = ((x_t - 1) \cdot 100)\,\%$$

and gives us an idea about the growth or decline of our value at time t. We can summarize these concepts in the following table:

Time	Inventory	Growth factor	Growth rate
t	B_t	x_t	r_t
0	B_0	–	–
1	B_1	$x_1 = B_1/B_0$	$((x_1 - 1) \cdot 100)\,\%$
2	B_2	$x_2 = B_2/B_1$	$((x_2 - 1) \cdot 100)\,\%$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
T	B_T	$x_T = B_T/B_{T-1}$	$((x_T - 1) \cdot 100)\,\%$

We can calculate B_t $(t = 1, 2, \ldots, T)$ by using the growth factors:

$$B_t = B_0 \cdot x_1 \cdot x_2 \cdot \ldots \cdot x_t.$$

The average growth factor from B_0 to B_T is the geometric mean or geometric average of the growth factors:

$$\begin{aligned} \bar{x}_G &= \sqrt[T]{x_1 \cdot x_2 \cdot \ldots \cdot x_T} \\ &= \sqrt[T]{\frac{B_0 \cdot x_1 \cdot x_2 \cdot \ldots \cdot x_T}{B_0}} \\ &= \sqrt[T]{\frac{B_T}{B_0}}. \end{aligned} \tag{3.11}$$

Therefore, B_t at time t can be calculated as $B_t = B_0 \cdot \bar{x}_G^t$.

Example 3.1.8 Suppose someone wants to deposit money, say €1000, in a bank. The bank advisor proposes a 5-year savings plan with the following plan for interest rates: 1 % in the first year, 1.5 % in the second year, 2.5 % in the third year, and 3 % in the last 2 years. Now he would like to calculate the average growth factor and average growth rate for the invested money. The concept of the geometric mean can be used as follows:

Year	Euro	Growth factor	Growth rate (%)
0	1000	–	–
1	1010	1.01	1.0
2	1025.15	1.015	1.5
3	1050.78	1.025	2.5
4	1082.30	1.03	3.0
5	1114.77	1.03	3.0

The geometric mean is calculated as

$$\bar{x}_G = (1.01 \cdot 1.015 \cdot 1.025 \cdot 1.03 \cdot 1.03)^{\frac{1}{5}} = 1.021968$$

which means that he will have on average about 2.2 % growth per year. The savings after 5 years can be calculated as

$$€\ 1000 \ \cdot \ 1.021968^5 = €\ 1114.77.$$

It is easy to compare two different saving plans with different growth strategies using the geometric mean.

3.1.6 Harmonic Mean

The harmonic mean is typically used whenever different x_i contribute to the mean with a different weight w_i, i.e. when we implicitly assume that the weight of each x_i is not one. It can be calculated as

$$\bar{x}_H = \frac{w_1 + w_2 + \cdots + w_k}{\frac{w_1}{x_1} + \frac{w_2}{x_2} + \cdots + \frac{w_k}{x_k}} = \frac{\sum_{i=1}^k w_i}{\sum_{i=1}^k \frac{w_i}{x_i}}. \tag{3.12}$$

For example, when calculating the average speed, each weight relates to the relative distance travelled, n_i/n, with speed x_i. Using $w_i = n_i/n$ and $\sum_i w_i = \sum_i n_i/n = 1$, the harmonic mean can be written as

$$\bar{x}_H = \frac{1}{\sum_{i=1}^k \frac{w_i}{x_i}}. \tag{3.13}$$

Example 3.1.9 Suppose an investor bought shares worth €1000 for two consecutive months. The price for a share was €50 in the first month and €200 in the second month. What is the average purchase price? The number of shares purchased in the first month is $1000/50 = 20$. The number of shares purchased in the second month is $1000/200 = 5$. The total number of shares purchased is thus $20 + 5 = 25$, and the total investment is €2000. It is evident that the average purchase price is $2000/25 =$ €80. This is in fact the harmonic mean calculated as

$$\bar{x}_H = \frac{1}{\frac{0.5}{50} + \frac{0.5}{200}} = 80$$

because the weight of each purchase is $n_i/n = 1000/2000 = 0.5$. If the investment was €1200 in the first month and €800 in the second month, then we could use the harmonic mean with weights $1200/2000 = 0.6$ and $800/2000 = 0.4$, respectively, to obtain the results.

3.2 Measures of Dispersion

Measures of central tendency, as introduced earlier, give us an idea about the location where most of the data is concentrated. However, two different data sets may have the same value for the measure of central tendency, say the same arithmetic means, but they may have different concentrations around the mean. In this case, the location measures may not be adequate enough to describe the distribution of the data. The concentration or dispersion of observations around any particular value is another property which characterizes the data and its distribution. We now introduce statistical methods which describe the **variability** or **dispersion** of data.

Example 3.2.1 Suppose three students Christine, Andreas, and Sandro arrive at different times in the class to attend their lectures. Let us look at their arrival time in the class after or before the starting time of lecture, i.e. let us look how early or late they were (in minutes).

Week	1	2	3	4	5	6	7	8	9	10
Christine	0	0	0	0	0	0	0	0	0	0
Andreas	−10	+10	−10	+10	−10	+10	−10	+10	−10	+10
Sandro	3	5	6	2	4	6	8	4	5	7

We see that Christine always arrives on time (time difference of zero). Andreas arrives sometimes 10 min early and sometimes 10 min late. However, the arithmetic mean of both students is the same—on average, they both arrive on time! This interpretation is obviously not meaningful. The difference between both students is the variability in arrival times that cannot be measured with the mean or median. For this reason, we need to introduce measures of dispersion (variability). With the knowledge of both location and dispersion, we can give a much more nuanced comparison between the different arrival times. For example, consider the third student Sandro. He is always late; sometimes more, sometimes less. However, while on average he comes late, his behaviour is more predictable than that of Andreas. Both location and dispersion are needed to give a fair comparison.

Example 3.2.2 Consider another example in which a supplier for the car industry needs to deliver 10 car doors with an exact width of 1.00 m. He supplies 5 doors with a width of 1.05 m and the remaining 5 doors with a width of 0.95 m. The arithmetic mean of all the 10 doors is 1.00 m. Based on the arithmetic mean, one may conclude that all the doors are good but the fact is that none of the doors are usable as they will not fit into the car. This knowledge can be summarized by a measure of dispersion.

The above examples highlight that the distribution of a variable needs to be characterized by a measure of dispersion in addition to a measure of location (central tendency). Now we introduce various measures of dispersion.

3.2.1 Range and Interquartile Range

Consider a variable X with n observations $x_1, x_2, \ldots, x_n$. Order these n observations as $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$. The range is a measure of dispersion defined as the difference between the maximum and minimum value of the data as

$$R = x_{(n)} - x_{(1)}. \tag{3.14}$$

The **interquartile range** is defined as the difference between the 75th and 25th quartiles as

$$d_Q = \tilde{x}_{0.75} - \tilde{x}_{0.25}. \tag{3.15}$$

It covers the centre of the distribution and contains 50 % of the observations.

Remark 3.2.1 Note that the interquartile range is defined as the interval $[\tilde{x}_{0.25}; \tilde{x}_{0.75}]$ in some literature. However, in line with most of the statistical literature, we define the interquartile range to be a measure of dispersion, i.e. the difference between $\tilde{x}_{0.75}$ and $\tilde{x}_{0.25}$.

Example 3.2.3 Recall Examples 3.0.1–3.1.5 where we looked at the temperature in Bangkok during December. The ordered values $x_{(i)}$, $i = 1, \ldots, 31$, are as follows:

°C	21	22	22	23	24	24	25	25	25	25	25	25	26	26	26	26
(i)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
°C	27	27	27	28	28	28	29	29	29	29	29	30	30	30	31	
(i)	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	

We obtained the quantiles in Example 3.1.5 as $\tilde{x}_{0.25} = 25$ and $\tilde{x}_{0.75} = 29$. The interquartile range is therefore $d_Q = 29 - 25 = 4$, which means that 50 % of the data is centred between 25 and 29 °C. The range is $R = 31 - 21 = 10$ °C, meaning that the temperature is varying at most by 10 °C. In *R*, there are several ways to obtain quartiles, minimum and maximum values, e.g. by using `min`, `max`, `quantiles`, `range`, among others. All numbers can be easily obtained by the `summary` command which we recommend using.

```
summary(weather)
```

3.2.2 Absolute Deviation, Variance, and Standard Deviation

Another measure of dispersion is the variance. The variance is one of the most important measures in statistics and is needed throughout this book. We use the idea of "absolute deviation" to give some more background and motivation for understanding the variance as a measure of dispersion, followed by some examples.

Consider the deviations of n observations around a certain value "A" and combine them together, for instance, via the arithmetic mean of all the deviations:

$$D = \frac{1}{n}\sum_{i=1}^{n}(x_i - A). \tag{3.16}$$

This measure has the drawback that the deviations $(x_i - A)$, $i = 1, 2, \ldots, n$, can be either positive or negative and, consequently, their sum can potentially be very small or even zero. Using D as a measure of variability is therefore not a good idea since D may be small even for a large variability in the data.

Using absolute values of the deviations solves this problem, and we introduce the following measure of dispersion:

$$D(A) = \frac{1}{n}\sum_{i=1}^{n} |x_i - A|. \tag{3.17}$$

It can be shown that the absolute deviation attains its minimum when A corresponds to the median of the data:

$$D(\tilde{x}_{0.5}) = \frac{1}{n}\sum_{i=1}^{n} |x_i - \tilde{x}_{0.5}|. \tag{3.18}$$

We call $D(\tilde{x}_{0.5})$ the **absolute median deviation**. When $A = \bar{x}$, we speak of the **absolute mean deviation** given by

$$D(\bar{x}) = \frac{1}{n}\sum_{i=1}^{n} |x_i - \bar{x}|. \tag{3.19}$$

Another solution to avoid the positive and negative signs of deviation in (3.16) is to consider the squares of deviations $x_i - A$, rather than using the absolute value. This provides another measure of dispersion as

$$s^2(A) = \frac{1}{n}\sum_{i=1}^{n} (x_i - A)^2 \tag{3.20}$$

which is known as the **mean squared error** (MSE) with respect to A. The MSE is another important measure in statistics, see Chap. 9, Eq. (9.4), for details. It can be shown that $s^2(A)$ attains its minimum value when $A = \bar{x}$. This is the (sample) **variance**

$$\tilde{s}^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2. \tag{3.21}$$

After expanding $\tilde{s}^2$, we can write (3.21) as

$$\tilde{s}^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2. \tag{3.22}$$

The positive square root of the variance is called the (sample) **standard deviation**, defined as

$$\tilde{s} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2}. \tag{3.23}$$

The standard deviation has the same unit of measurement as the data whereas the unit of the variance is the square of the units of the observations. For example, if X is weight, measured in kg, then $\bar{x}$ and $\tilde{s}$ are also measured in kg, while $\tilde{s}^2$ is measured in kg^2 (which may be more difficult to interpret). The variance is a measure which we use in other chapters to obtain measures of association between variables and to draw conclusions from a sample about a population of interest; however, the standard deviation is typically preferred for a descriptive summary of the dispersion of data.

The standard deviation measures how much the observations vary or how they are dispersed around the arithmetic mean. A low value of the standard deviation indicates that the values are highly concentrated around the mean. A high value of the standard deviation indicates lower concentration of the observations around the mean, and some of the observed values may even be far away from the mean. If there are extreme values or outliers in the data, then the arithmetic mean is more sensitive to outliers than the median. In such a case, the absolute median deviation (3.18) may be preferred over the standard deviation.

Example 3.2.4 Consider again Example 3.2.1 where we evaluated the arrival times of Christine, Andreas, and Sandro in their lecture. Using the arithmetic mean, we concluded that both Andreas and Christine arrive on time, whereas Sandro is always late; however, we saw that the variation of arrival times differs substantially among the three students. To describe and quantify this variability formally, we calculate the variance and absolute median deviation:

$$\begin{aligned}
\tilde{s}_C^2 &= \frac{1}{10}\sum_{i=1}^{10}(x_i-\bar{x})^2 = \frac{1}{10}((0-0)^2+\cdots+(0-0)^2) = 0\\
\tilde{s}_A^2 &= \frac{1}{10}\sum_{i=1}^{10}(x_i-\bar{x})^2 = \frac{1}{10}((-10-0)^2+\cdots+(10-0)^2) \approx 111.1\\
\tilde{s}_S^2 &= \frac{1}{10}\sum_{i=1}^{10}(x_i-\bar{x})^2 = \frac{1}{10}((3-5)^2+\cdots+(7-5)^2) \approx 3.3
\end{aligned}$$

$$\begin{aligned}
D(\tilde{x}_{0.5,C}) &= \frac{1}{10}\sum_{i=1}^{n}|x_i-\tilde{x}_{0.5}| = |0-0|+\cdots+|0-0| = 0\\
D(\tilde{x}_{0.5,A}) &= \frac{1}{10}\sum_{i=1}^{n}|x_i-\tilde{x}_{0.5}| = |-10-0|+\cdots+|10-0| = 10\\
D(\tilde{x}_{0.5,S}) &= \frac{1}{10}\sum_{i=1}^{n}|x_i-\tilde{x}_{0.5}| = |3-5|+\cdots+|7-5| = 1.4.
\end{aligned}$$

We observe that the variation/dispersion/variability is the lowest for Christine and highest for Andreas. Both median absolute deviation and variance allow a comparison between the two students. If we take the square root of the variance, we obtain the standard deviation. For example, $\tilde{s}_S = \sqrt{3.3} \approx 1.8$, which means that the average difference of the observations from the arithmetic mean is 1.8.

In *R*, we can use the `var` command to calculate the variance. However, note that *R* uses $1/(n-1)$ instead of $1/n$ in calculating the variance. The idea behind the multiplication by $1/(n-1)$ in place of $1/n$ is discussed in Chap. 9, see also Theorem 9.2.1.

Variance for Grouped Data. The variance for grouped data can be calculated using

$$s_b^2 = \frac{1}{n}\sum_{j=1}^{k} n_j(a_j - \bar{x})^2 = \frac{1}{n}\left(\sum_{j=1}^{k} n_j a_j^2 - n\bar{x}^2\right) = \frac{1}{n}\sum_{j=1}^{k} n_j a_j^2 - \bar{x}^2, \tag{3.24}$$

where a_j is the middle value of the jth interval. However, when the data is artificially grouped and the knowledge about the original ungrouped data is available, we can also use the arithmetic mean of the jth class:

$$s_b^2 = \frac{1}{n}\sum_{j=1}^{k} n_j(\bar{x}_j - \bar{x})^2. \tag{3.25}$$

The two expressions (3.24) and (3.25) represent the **variance between the different classes**, i.e. they describe the variability of the class specific means $\bar{x}_j$, weighted by the size of each class n_j, around the overall mean $\bar{x}$. It is evident that the variance *within* each class is not taken into account in these formulae. The variability of measurements in each class, i.e. the variability of $\forall x_i \in K_j$, is another important component to determine the overall variance in the data. It is therefore not surprising that using only the between variance $\tilde{s}_b^2$ will underestimate the total variance and therefore

$$s_b^2 \leq s^2. \tag{3.26}$$

If the data within each class is known, we can use the Theorem of Variance Decomposition (see p. 136 for the theoretical background) to determine the variance. This allows us to represent the total variance as the sum of the **variance between the different classes** and the **variance within the different classes** as

$$\tilde{s}^2 = \underbrace{\frac{1}{n}\sum_{j=1}^{k} n_j(\bar{x}_j - \bar{x})^2}_{\text{between}} + \underbrace{\frac{1}{n}\sum_{j=1}^{k} n_j\tilde{s}_j^2}_{\text{within}}. \tag{3.27}$$

In (3.27), $\tilde{s}_j^2$ is the variance of the jth class:

$$\tilde{s}_j^2 = \frac{1}{n_j}\sum_{x_i \in K_j} (x_i - \bar{x}_j)^2. \tag{3.28}$$

The proof of (3.27) is given in Appendix C.1, p. 423.

Example 3.2.5 Recall the weather data used in Examples 3.0.1–3.2.3 and the grouped data specified as follows:

Class intervals	<20	(20–25]	(25, 30]	(30, 35]	>35
n_j	$n_1 = 0$	$n_2 = 12$	$n_3 = 18$	$n_4 = 1$	$n_5 = 0$
$\bar{x}_j$	–	23.83	28	31	–
$\tilde{s}_j^2$	–	1.972	2	0	–

We know that $\bar{x} = 26.48$ and $n = 31$. The first step is to calculate the mean and variances in each class using (3.28). We then obtain $\bar{x}_j$ and s_j^2 as listed above. The within and between variances are as follows:

$$\frac{1}{n}\sum_{j=1}^{k} n_j \tilde{s}_j^2 = \frac{1}{31}(12 \cdot 1.972 + 18 \cdot 2 + 1 \cdot 0) \approx 1.925$$

$$\frac{1}{n}\sum_{j=1}^{k} n_j(\bar{x}_j - \bar{x})^2 = \frac{1}{31}(12 \cdot [23.83 - 26.48]^2 + 18 \cdot [28 - 26.48]^2$$
$$+1 \cdot [31 - 26.48]^2) \approx 4.71.$$

The total variance is therefore $\tilde{s}^2 \approx 6.64$. Estimating the variance using all 31 observations would yield the same results. However, it becomes clear that without knowledge about the variance within each class, we cannot reliably estimate $\tilde{s}^2$. In the above example, the variance between the classes is 3 times lower than the total variance which is a serious underestimation.

Linear Transformations. Let us consider a linear transformation $y_i = a + bx_i$ $(b \neq 0)$ of the original data x_i, $(i = 1, 2, \ldots, n)$. We get the arithmetic mean of the transformed data as $\bar{y} = a + b\bar{x}$ and for the variance:

$$\begin{aligned} \tilde{s}_y^2 &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{b^2}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \\ &= b^2 \tilde{s}_x^2. \end{aligned} \tag{3.29}$$

Example 3.2.6 Let $x_i, i = 1, 2, \ldots, n$, denote measurements on time. These data could have been recorded and analysed in hours, but we may be interested in a summary in minutes. We can make a linear transformation $y_i = 60\,x_i$. Then, $\bar{y} = 60\bar{x}$ and $\tilde{s}_y^2 = 60^2 \tilde{s}_x^2$. If the mean and variance of the x_i's have already been obtained, then the mean and variance of the y_i's can be obtained directly using these transformations.

Standardization. A variable is called standardized if its mean is zero and its variance is 1. Standardization can be achieved by using the following transformation:

$$y_i = \frac{x_i - \bar{x}}{\tilde{s}_x} = -\frac{\bar{x}}{\tilde{s}_x} + \frac{1}{\tilde{s}_x}x_i = a + bx_i. \tag{3.30}$$

It follows that $\bar{y} = \sum_{i=1}^{n}(x_i - \bar{x})/\tilde{s}_x = 0$ and $\tilde{s}_y^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2/\tilde{s}_x^2 = 1$. There are many statistical methods which require standardization, see, for example, Sect. 10.3.1 for details in the context of statistical tests.

Example 3.2.7 Let X be a variable which measures air pollution by using the concentration of atmospheric particulate matter (in μg/m^3). Suppose we have the following 10 measurements:

$$30 \quad 25 \quad 12 \quad 45 \quad 50 \quad 52 \quad 38 \quad 39 \quad 45 \quad 33.$$

We calculate $\bar{x} = 36.9$, $\tilde{s}_x^2 = 136.09$, and $\tilde{s}_x = 11.67$. To get a standardized variable Y, we transform all the observations x_i's as

$$y_i = \frac{x_i - \bar{x}}{\tilde{s}_x} = -\frac{\bar{x}}{\tilde{s}_x} + \frac{1}{\tilde{s}_x}x_i = -\frac{36.9}{11.67} + \frac{1}{11.67}x_i = -3.16 + 0.086x_i.$$

Now $y_1 = -3.16 + 0.086 \cdot 30 = -0.58$, $y_2 = -3.16 + 0.086 \cdot 25 = -1.01$, ..., are the standardized observations. The `scale` command in *R* allows standardization, and we can obtain the standardized observations corresponding to the 10 measurements as

```
air <- c(30,25,12,45,50,52,38,39,45,33)
scale(air)
```

Please note that the `scale` command uses $1/(n-1)$ for calculating the variance, as already outlined above. Thus, the results provided by `scale` are not identical to those using (3.30).

3.2.3 Coefficient of Variation

Consider a situation where two different variables have arithmetic means $\bar{x}_1$ and $\bar{x}_2$ with standard deviations $\tilde{s}_1$ and $\tilde{s}_2$, respectively. Suppose we want to compare the variability of hotel prices in Munich (measured in euros) and London (measured in British pounds). How can we provide a fair comparison? Since the prices are measured in different units, and therefore likely have arithmetic means which differ substantially, it does not make much sense to compare the standard deviations directly. The coefficient of variation v is a measure of dispersion which uses both the standard deviation and mean and thus allows a fair comparison. It is properly defined only when all the values of a variable are measured on a ratio scale and are positive such that $\bar{x} > 0$ holds. It is defined as

$$v = \frac{s}{\bar{x}}. \tag{3.31}$$

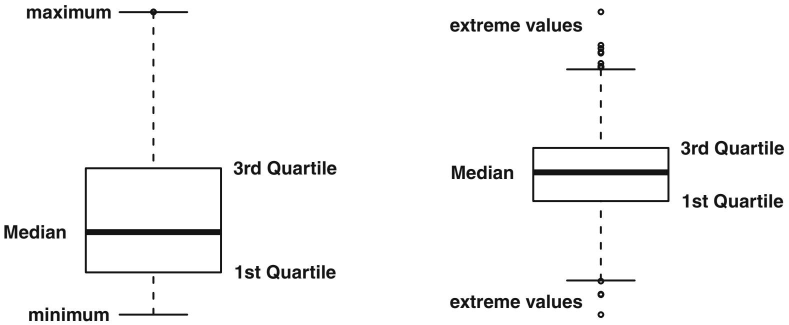
The coefficient of variation is a unit-free measure of dispersion. It is often used when the measurements of two variables are different but can be put into relation by using a linear transformation $y_i = bx_i$. It is possible to show that if all values x_i of a variable X are transformed into a variable Y with values $y_i = b \cdot x_i$, $b > 0$, then v does not change.

Example 3.2.8 If we want to compare the variability of hotel prices in two selected cities in Germany and England, we could calculate the mean prices, together with their standard deviation. Suppose a sample of prices of say 100 hotels in two selected cities in Germany and England is available and suppose we obtain the mean and standard deviations of the two cities as $x_1 = €130$, $x_2 = £230$, $s_1 = €99$, and $s_2 = £212$. Then, $v_1 = 99/130 \approx 0.72$ and $v_2 = 212/230 = 0.92$. This indicates higher variability in hotel prices in England. However, if the data distribution is skewed or bimodal, then it may be wise not to choose the arithmetic mean as a measure of central tendency and likewise the coefficient of variation.

3.3 Box Plots

So far we have described various measures of central tendency and dispersion. It can be tedious to list those measures in summary tables. A simple and powerful graph is the **box plot** which summarizes the distribution of a continuous (or sometimes an ordinal) variable by using its median, quartiles, minimum, maximum, and extreme values.

Figure 3.5a shows a typical box plot. The vertical length of the box is the interquartile range $d_Q = \tilde{x}_{0.75} - \tilde{x}_{0.25}$, which shows the region that contains 50 % of the data. The bottom end of the box refers to the first quartile, and the top end of the box refers to the third quartile. The thick line in the box is the median. It becomes immediately clear that the box indicates the symmetry of the data: if the median is in the middle of the box, the data should be symmetric, otherwise it is skewed. The *whiskers* at the end of the plot mark the minimum and maximum values of the data. Looking at the box plot as a whole tells us about the data distribution and the range and variability of observations. Sometimes, it may be advisable to understand which values are extreme in the sense that they are "far away" from the centre of the distribution. In many software packages, including *R*, values are defined to be extreme if they are greater than 1.5 box lengths away from the first or third quartile. Sometimes, they are called outliers. Outliers and extreme values are defined differently in some software packages and books.



(a) Box plot without extreme values

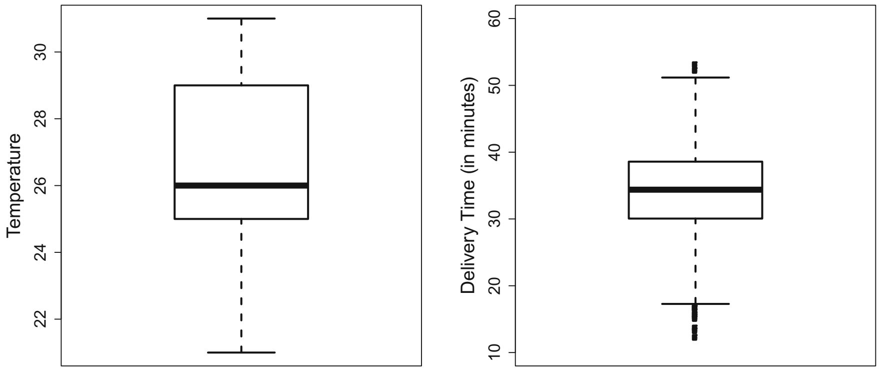
(b) Box plot with extreme values

The `boxplot` command in *R* draws a box plot. The `range` option controls whether extreme values should be plotted, and if yes, how one wants to define such values.

```
boxplot(variable, range=1.5)
```

Example 3.3.1 Recall Examples 3.0.1–3.2.5 where we looked at the temperature in Bangkok during December. We have already calculated the median (26°C) and the quartiles (25, 29°C). The minimum and maximum values are 21°C and 31°C. The box plot for this data is shown in Fig. 3.5a. One can see that the temperature distribution is slightly skewed with more variability for lower temperatures. The interquartile range is 4, and therefore, any value $>29 + 4 \times 1.5 = 35$ or $<25 - 4 \times 1.5 = 19$ would be an extreme value. However, there are no extreme values in the data.

Example 3.3.2 Consider again the pizza data described in Appendix A.4. We use *R* to plot the box plot for the delivery time via `boxplot(time)` (Fig. 3.5b). We see a symmetric distribution with a median delivery time of about 35 min. Most of the deliveries took between 30 and 40 min. The extreme values indicate that there were some exceptionally short and long delivery times.



(a) Boxplot for weather data

(b) Boxplot for pizza data

3.4 Measures of Concentration

A completely different concept used to describe a quantitative variable is the idea of concentration. For a variable X, it summarizes the proportion of each observation with respect to the sum of all observations $\sum_{i=1}^{n} x_i$. Let us look at a simple example to demonstrate its usefulness.

Table 3.1 Concentration of farmland: two different situations

Farmer (i)	x_i (Area, in hectare)
1	20
2	20
3	20
4	20
5	20
	$\sum_{i=1}^{5} x_i = 100$

Farmer (i)	x_i (Area, in hectare)
1	0
2	0
3	0
4	0
5	100
	$\sum_{i=1}^{5} x_i = 100$

Example 3.4.1 Consider a village with 5 farms. Each farmer has a farm of a certain size. How can we evaluate the land distribution? Do all farmers have a similar amount of land or do one or two farmers have a big advantage because they have considerably more space?

Table 3.1 shows two different situations: in the table on the left, we see an equal distribution of land, i.e. each farmer owns 20 hectares of farmland. This means X is *not* concentrated, rather it is equally distributed. A statistical function describing the concentration could return a value of zero in such a case. Consider another extreme where one farmer owns all the farmland and the others do not own anything, as shown on the right side of Table 3.1. This is an extreme concentration of land: one person owns everything and thus, we say the concentration is high. A statistical function describing the concentration could return a value of one in such a case.

3.4.1 Lorenz Curve

The **Lorenz curve** is a popular method to display concentrations graphically. Consider n observations $x_1, x_2, \ldots, x_n$ of a variable X. Assume that all the observations are positive. The sum of all the observations is $\sum_{i=1}^{n} x_i = n\bar{x}$ if the data is ungrouped. First, we need to order the data: $0 \leq x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$. To plot the Lorenz curve, we need

$$u_i = \frac{i}{n}, \quad i = 0, \ldots, n, \tag{3.32}$$

and

$$v_i = \frac{\sum_{j=1}^{i} x_{(j)}}{\sum_{j=1}^{n} x_{(j)}}, \quad i = 1, \ldots, n; \; v_0 := 0, \tag{3.33}$$

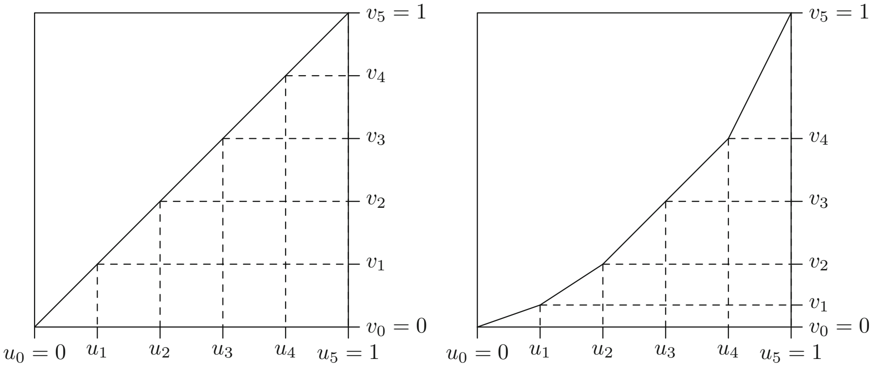


Fig. 3.5 Lorenz curves for no concentration (*left*) and some concentration (*right*)*

where $\sum_{j=1}^{i} x_{(j)}$ is the cumulative total of observations up to the ith observation. The idea is that v_i describe the contribution of all values $\leq i$ in comparison with the sum of all values. Plotting u_i against v_i for all i shows how much the sum of all x_i, for all observations $\leq i$, contributes to the total sum. In other words, the point (u_i, v_i) says that $u_i \cdot 100\,\%$ of observations contain $v_i \cdot 100\,\%$ of the sum of all x_i less than or equal to i. Obviously, if all x_i are identical, the Lorenz curve will be a straight diagonal line, also known as the identity line or **line of equality**. If the x_i are of different sizes, then the Lorenz curve falls below the line of equality. This is illustrated in the following example.

Example 3.4.2 Recall Example 3.4.1 where we looked at the distribution of farmland among 5 farmers. On the upper panel of Table 3.1, we observed an equal distribution of land among the farmers: $x_1 = 20$, $x_2 = 20$, $x_3 = 20$, $x_4 = 20$, and $x_5 = 20$. We obtain $u_1 = 1/5, u_2 = 2/5, \ldots, u_5 = 1$ and $v_1 = 20/100, v_2 = 40/100, \ldots, v_5 = 1$. This yields a Lorenz curve as displayed on the left side of Fig. 3.5: there is no concentration. We can interpret each point. For example, $(u_2, v_2) = (0.4, 0.4)$ means that 40 % of farmers own 40 % of the land.

The lower panel of Table 3.1 describes the situation with strong concentration. For this table, we obtain $u_1 = 1/5, u_2 = 2/5, \ldots, u_5 = 1$ and $v_1 = 0, v_2 = 0, \ldots, v_5 = 1$. Therefore, for example, 80 % of farmers own 0 % of the land which shows strong inequality. Most often we do not have such extreme situations. In this case, the Lorenz curve is bent towards the lower right corner of the plot, see the right side of Fig. 3.5.

We can plot the Lorenz curve in *R* using the `Lc` command in the library `ineq`. The Lorenz curve for the left table of Example 3.4.1 is plotted in *R* as follows:

```
library(ineq)
x <- c(20,20,20,20,20)
plot(Lc(x))
```

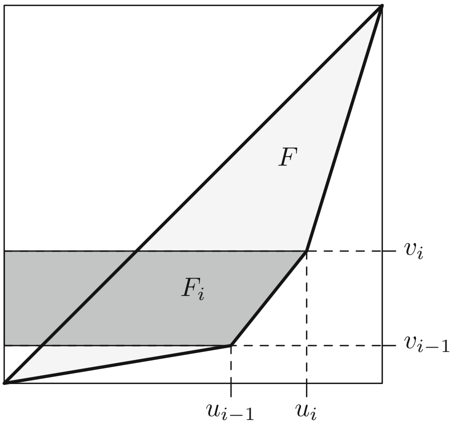


Fig. 3.6 Lorenz curve and the Gini coefficient*

We can use the same approach as above to obtain the Lorenz curve when we have grouped data. We simply describe the contributions for each class rather than for each observation and approximate the values in each class by using its mid-point. More formally we can write:

$$\tilde{u}_i = \sum_{j=1}^{i} f_j, \quad i = 1, 2, \ldots, k;\ \tilde{u}_0 := 0 \tag{3.34}$$

and

$$\tilde{v}_i = \frac{\sum_{j=1}^{i} f_j a_j}{\sum_{j=1}^{k} f_j a_j} = \frac{\sum_{j=1}^{i} n_j a_j}{n\bar{x}}, \quad i = 1, 2, \ldots, k;\ \tilde{v}_0 := 0. \tag{3.35}$$

3.4.2 Gini Coefficient

We have seen in Sect. 3.4.1 that the Lorenz curve corresponds to the identity line, that is the diagonal line of equality, for no concentration. When there is some concentration, then the curve deviates from this line. The amount of deviation depends on the strength of concentration. Suppose we want to design a measure of concentration which is 0 for no concentration and 1 for perfect (i.e. extreme) concentration. We can simply measure the area between the Lorenz curve and the identity line and multiply it by 2. For no concentration, the area will be zero and hence the measure will be zero. If there is perfect concentration, then the curve will coincide with the axes, the area will be close to 0.5, and twice the area will be close to one. The measure based on such an approach is called the Gini coefficient:

$$G = 2 \cdot F. \tag{3.36}$$

Note that F is the area between the curve and the bisection or diagonal line.

The Gini coefficient can be estimated by adding up the areas of the trapeziums F_i as displayed in Fig. 3.6:

$$F = \sum_{i=1}^{n} F_i - 0.5,$$

where

$$F_i = \frac{u_{i-1} + u_i}{2}(v_i - v_{i-1}).$$

It can be shown that this corresponds to

$$G = 1 - \frac{1}{n}\sum_{i=1}^{n}(v_{i-1} + v_i), \tag{3.37}$$

but the proof is omitted. The same formula can be used for grouped data except that $\tilde{v}$ is used instead of v. Since

$$0 \leq G \leq \frac{n-1}{n}, \tag{3.38}$$

one may prefer to use the standardized Gini coefficient

$$G^{+} = \frac{n}{n-1}G, \tag{3.39}$$

which takes a maximum value of 1.

Example 3.4.3 We return to our farmland example. Suppose we have 7 farmers with farms of different sizes:

Farmer	1	2	3	4	5	6	7
Farmland size x_i	20	14	59	9	36	23	3

Using the ordered values, we can calculate u_i and v_i using (3.32) and (3.33):

i	$x_{(i)}$	u_i	v_i
1	3	$\frac{1}{7} = 0.1429$	$\frac{3}{164} = 0.0183$
2	9	$\frac{2}{7} = 0.2857$	$\frac{12}{164} = 0.0732$
3	14	$\frac{3}{7} = 0.4286$	$\frac{26}{164} = 0.1585$
4	20	$\frac{4}{7} = 0.5714$	$\frac{46}{164} = 0.2805$
5	23	$\frac{5}{7} = 0.7143$	$\frac{69}{164} = 0.4207$
6	36	$\frac{6}{7} = 0.8571$	$\frac{105}{164} = 0.6402$
7	59	$\frac{7}{7} = 1.0000$	$\frac{164}{164} = 1.0000$

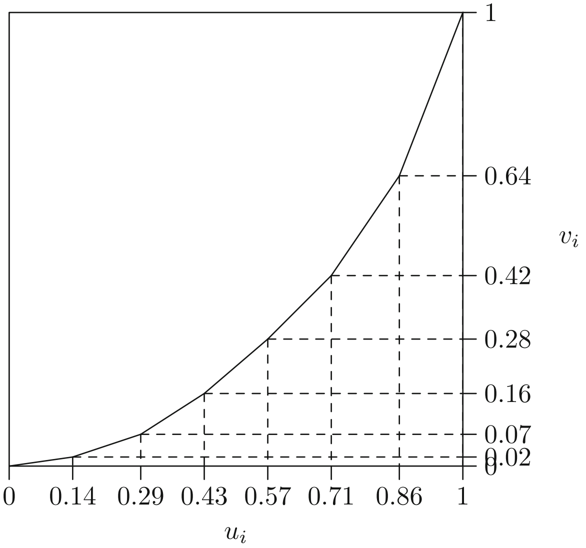


Fig. 3.7 Lorenz curve for Example 3.4.3*

The Lorenz curve is displayed in Fig. 3.7. Using this information, it is easy to calculate the Gini coefficient:

$$G = 1 - \frac{1}{7}(0.0183 + [0.0183 + 0.0732] + [0.0732 + 0.1585] + [0.1585 + 0.2805] \\ +[0.2805 + 0.4207] + [0.4207 + 0.6402] + [0.6402 + 1]) = 0.402$$

We know that $G = 0.4024 \le \frac{6}{7} = \frac{n-1}{n}$. To standardize the coefficient, we therefore have to use (3.39):

$$G^{+} = \frac{7}{6}G = \frac{7}{6} \cdot 0.4024 = 0.4695\,.$$

In *R*, we can obtain the non-standardized Gini Coefficient using the `ineq` function in the library `ineq`.

```
library(ineq)
farm <- c(20,14,59,9,36,23,3)
ineq(farm)
```

3.5 Key Points and Further Issues

Note:

✓ A summary on how to descriptively summarize data is given in Appendix D.1.

✓ The median is preferred over the arithmetic mean when the data distribution is skewed or there are extreme values.

✓ If data of a continuous variable is grouped, and the original ungrouped data is not known, additional assumptions are needed to calculate measures of central tendency and dispersion. However, in some cases, these assumptions may not be satisfied, and the formulae provided may give imprecise results.

✓ QQ-plots are not only descriptive summaries but can also be used to test modelling assumptions, see Chap. 11.9 for more details.

✓ The distribution of a continuous variable can be easily summarized using a box plot.

3.6 Exercises

Exercise 3.1 A hiking enthusiast has a new app for his smartphone which summarizes his hikes by using a GPS device. Let us look at the distance hiked (in km) and maximum altitude (in m) for the last 10 hikes:

Distance	12.5	29.9	14.8	18.7	7.6	16.2	16.5	27.4	12.1	17.5
Altitude	342	1245	502	555	398	670	796	912	238	466

(a) Calculate the arithmetic mean and median for both distance and altitude.
(b) Determine the first and third quartiles for both the distance and the altitude variables. Discuss the shape of the distribution given the results of (a) and (b).
(c) Calculate the interquartile range, absolute median deviation, and standard deviation for both variables. What is your conclusion about the variability of the data?
(d) One metre corresponds to approximately 3.28 ft. What is the average altitude when measured in feet rather than in metres?
(e) Draw and interpret the box plot for both distance and altitude.
(f) Assume distance is measured as only short (5–15 km), moderate (15–20 km), and long (20–30 km). Summarize the grouped data in a frequency table. Calculate the weighted arithmetic mean under the assumption that the raw data is not

known. Determine the weighted median under the assumption that the values within each class are equally distributed.

(g) What is the variance for the grouped data when the raw data is known, i.e. when one has knowledge about the variance in each class? How does it compare with the variance one obtains when the raw data is unknown?

(h) Use *R* to reproduce the results of (a), (b), (c), (e), and (f).

Exercise 3.2 A gambler notes down his wins and losses (in €) from playing 10 games of roulette in a casino.

Round	1	2	3	4	5	6	7	8	9	10
Won/Lost	200	600	−200	−200	−200	−100	−100	−400	0	

(a) Assume $\bar{x} = -$ €90 and $s =$ €294.7881. What is the result of round 10?

(b) Determine the mode and the interquartile range.

(c) A different gambler plays 33 rounds of roulette. His results are $\bar{x} =$ €12 and $s =$ €1000. Is it meaningful to compare the variability of results of the two players by using the coefficient of variation? If yes, determine the coefficients of variation; if no, why is a comparison not possible?

Exercise 3.3 A fashion boutique has summarized its daily sales of designer socks in different groups: men's socks, women's socks, and children's socks. Unfortunately, the data for men's socks was lost. Determine the missing values.

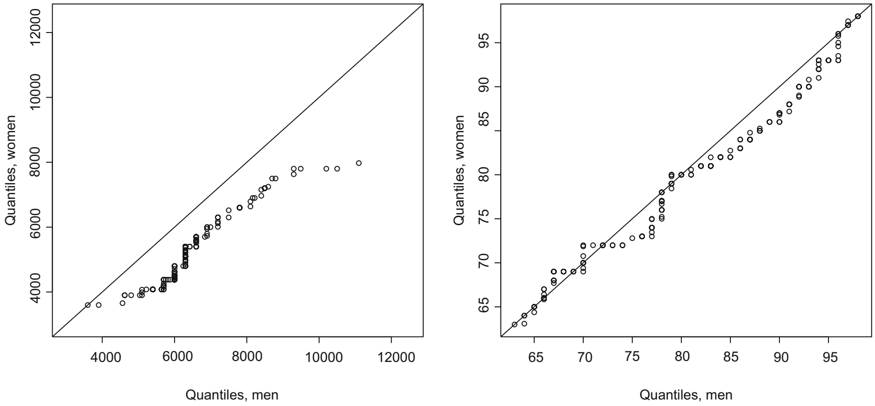
	n	Arithmetic mean	Standard deviation
		in €	in €
Women's wear	45	16	$\sqrt{6}$
Men's wear	?	?	?
Children's wear	20	7.5	$\sqrt{3}$
Total	100	15	$\sqrt{19.55}$

Exercise 3.4 The number of members of a millionaires' club were as follows:

Year	2011	2012	2013	2014	2015	2016
Members	23	24	27	25	30	28

(a) What is the average growth rate of the membership?

(b) Based on the results of (a), how many members would one expect in 2018?



(a) for the salary (b) for length of service

Fig. 3.8 QQ-plots

(c) The president of the club is interested in the number of members in 2025, the year when his presidency ends. Would it make sense to predict the number of members for 2025?

In 2015, the members invested €250 million on the stock market. 10 members contributed 16% of the investment sum, 8 members contributed €60 million, 8 members contributed €70 million, and another 4 members contributed the remaining amount.

(d) Draw the Lorenz curve for this data.
(e) Calculate and interpret the standardized Gini coefficient.

Exercise 3.5 Consider the monthly salaries Y (in Swiss francs) of a well-reputed software company, as well as the length of service (in months, X), and gender (Z). Figure 3.8 shows the QQ-plots for both Y and X given Z. Interpret both graphs.

Exercise 3.6 There is no built-in function in *R* to calculate the mode of a variable. Program such a function yourself. Hint: type `?table` and `?names` to recall the functionality of these functions. Combine them in an intelligent way.

Exercise 3.7 Consider a country in which 90 % of the wealth is owned by 20 % of the population, the so-called upper class. For simplicity, let us assume that the wealth is distributed equally within this class.

(a) Draw the Lorenz curve for this country.
(b) Now assume a revolution takes place in the country and all members of the upper class have to give away their wealth which is then distributed equally across the remaining population. Draw the Lorenz curve for this scenario.
(c) What would the curve from (b) look like if the entire upper class left the country?

Exercise 3.8 A bus route in the mountainous regions of Romania has a length of 418 km. The manager of the bus company serving the route wants his buses to finish a trip within 8 h. The bus travels the first 180 km with an average speed of 48 km/h, the next 117 km with an average speed of 37 km/h, and the last section with an average speed of 52 km/h.

(a) What is the average speed with which the bus travels?
(b) Will the bus finish the trip in time?

Exercise 3.9 Four friends have a start-up company which sells vegan ice cream. Their initial financial contributions are as follows:

Person	1	2	3	4
Contribution (in €)	800	10300	4700	2220

(a) Calculate and draw the Lorenz curve.
(b) Determine and interpret the standardized Gini coefficient.
(c) Does G^+ change if each of the friends contributes only half the amount of money? If yes, how much? If no, why not?
(d) Use *R* to draw the above Lorenz curve and to calculate the Gini coefficient.

Exercise 3.10 Recall the pizza delivery data which is described in Appendix A.4. Use *R* to read in and analyse the data.

(a) Calculate the mean, median, minimum, maximum, first quartile, and third quartile for all quantitative variables.
(b) Determine and interpret the 99 % quantile for delivery time and temperature.
(c) Write a function which calculates the absolute mean deviation. Use the function to calculate the absolute mean deviation of temperature.
(d) Scale the delivery time and calculate the mean and variance for this variable.
(e) Draw a box plot for delivery time and temperature. The box plots should not highlight extreme values.
(f) Use the `cut` command to create a new variable which summarizes delivery time in steps of 10 min. Calculate the arithmetic mean of this variable.
(g) Reproduce the QQ-plots shown in Example 3.1.6.

→ Solutions to all exercises in this chapter can be found on p. 333

**Source* Toutenburg, H., Heumann, C., *Deskriptive Statistik*, 7th edition, 2009, Springer, Heidelberg