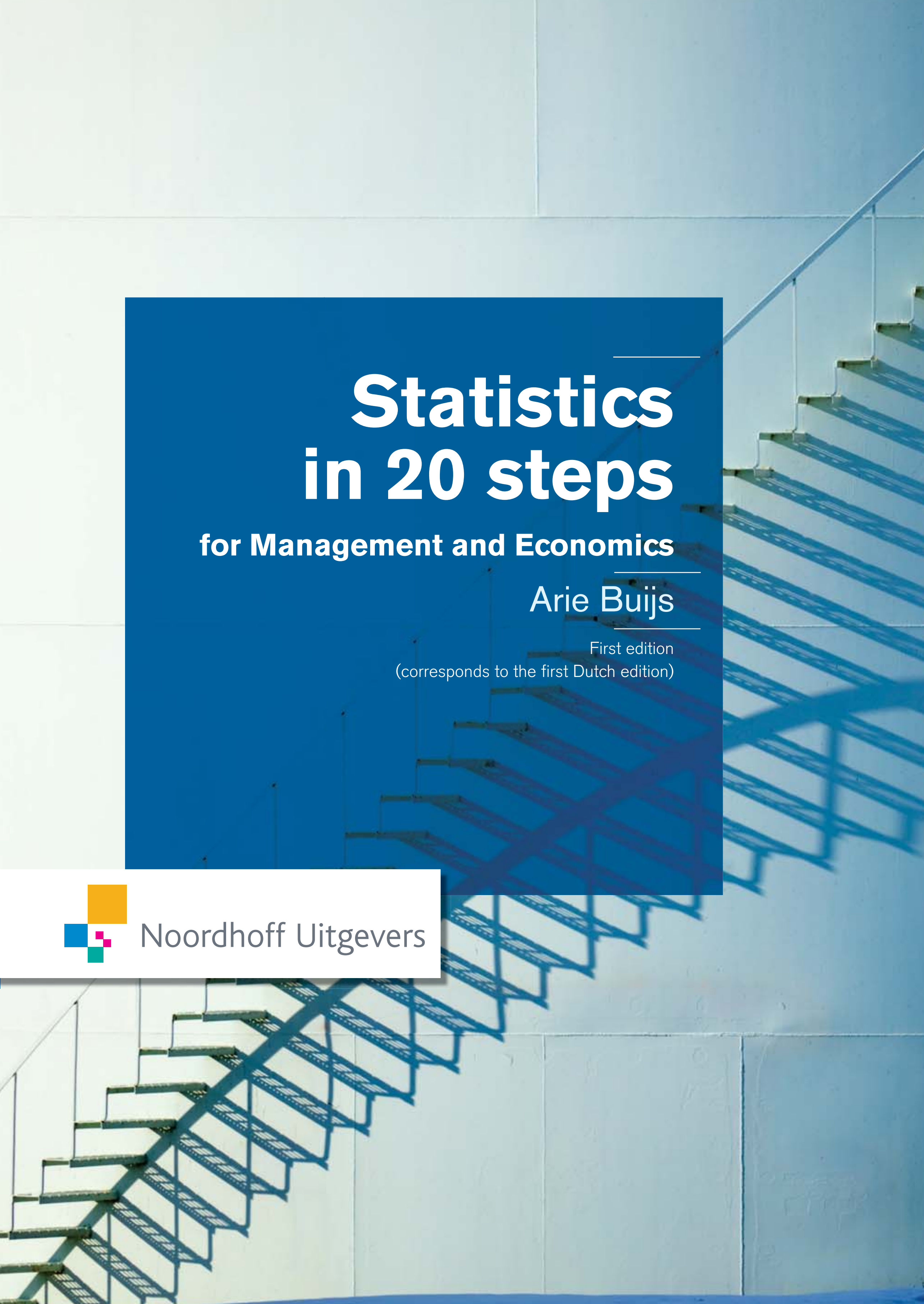

Statistics in 20 steps

for Management and Economics

Arie Buijs

First edition
(corresponds to the first Dutch edition)

Noordhoff Uitgevers

Statistics in 20 steps

for Management and Economics

Arie Buijs

First edition (corresponds to the first Dutch edition)

Noordhoff Uitgevers Groningen/Houten

Cover design: Rocket Industries, Groningen
Cover illustration: Getty Images
Translation: J. Hagoort

If you have any comments or queries about this or any other publication, please contact: Noordhoff Uitgevers bv, Afdeling Hoger Onderwijs, Antwoordnummer 13, 9700 VB Groningen, e-mail: info@noordhoff.nl

All due care has been taken in the preparation of this edition and the authors cannot accept any responsibility for any information that may be incorrect or incorrectly included. However, the authors would be pleased to consider any comments or suggestions for improvements.

MIX
Paper from responsible sources
FSC® C118189

0 / 15

ISBN (ebook) 978-90-01-86122-3
ISBN 978-90-01-86121-6
NUR 916

A Note from the Author

The Higher Education sector is changing at a fast pace. The government plays a huge role in this respect, for example through the financing and supervision of universities and polytechnics. Importance is attached to success rates, in terms of proportions of students who successfully complete study programs. On the other hand, concerns are raised about the level of knowledge of graduates, as reflected in the quality of graduation theses, for example. Various measures have been introduced to address these issues, such as the binding study advice, selection at the gate, and various forms of remedial education. In addition, students are expected to complete their program within a shorter time frame.

One of these concerns relates to students' level of knowledge of mathematics and statistics. This level varies widely between types of education. This is only logical, of course, as some professional areas demand a thorough ready knowledge of statistical methods and techniques, whereas in other areas a more limited set may suffice.
Still, discussions with colleagues in the education sector convinced me that every higher education graduate ought to leave with at least a basic range of knowledge and skills –irrespective of the field of study. This book has a specific focus on the field of statistics.

Statistics in 20 steps gives away that this book contains 20 compact chapters. The individual chapters can be dealt with within a one or two hours teaching module.
Each chapter elaborates on its main subject and encourages students to self-reflect on the guidance of short questions interspersed in the text. Several exercises are provided as well, and solutions are given at the end of the book. The companion website, www.statisticsin20steps.noordhoff.nl or www.statisticsintwentysteps.noordhoff.nl, offers additional computations and exercises for students.

When preparing this book I was privileged to brainstorm with Judith Boertjens of Noordhoff Publishers and with Gert-Jan Reus of *Fontys Hogeschool* in Tilburg. Mr. Reus in addition provided suggestions for case problems. Kai Strohmeyer advised on a number of statistical Excel applications. Dr. Bo van der Rhee of Nyenrode Business University provided useful tips. I am very grateful to all of them.

It is my wish that students will find this book easily readable and comprehensible. Moreover, I hope it will fire many students with enthusiasm for the field of statistics. Any comments or criticisms are very welcome.

Bilthoven, autumn 2014
Arie Buijs

Contents

Used cars for sale at Ben de Beun

A car magazine is investigating prices and characteristics of used cars. An assistant at the magazine inventoried the range of eight well-known models offered both by car dealers and private individuals via websites. Car dealer Ben de Beun appeared to have 30 of those cars in stock. The table below shows details of the first six cars on the list.

TABLE 1.1 Cars offered by car dealer Ben de Beun

No.	Brand	Colour	Year	Mileage (km)	Condition	Automatic yes/no	Asking price
1	Volvo	Blue	1999	345,000	Bad	Yes	1,800
2	BMW	Red	1996	320,000	Good	No	3,900
3	Mazda	Yellow	2010	56,000	Excellent	Yes	11,800
4	Opel	Gray	2003	130,000	Fair	Yes	3,500
5	Peugeot	White	2007	165,000	Average	No	4,700
6	Volkswagen	Black	2012	18,000	Good	No	17,900

Source: records De Beun BV, April 2014

After having collected those data the assistant wonders how to make this large amount of information more transparent? What kind of conclusions can be drawn?

1 A first step

In this chapter we will answer the following questions:
- What exactly is the field of statistics?
- Why is it important to distinguish between various types of variables?
- What is the proper way to prepare a table and how can we draw conclusions from the tabular information?

In many situations we come across information in the form of data sets. We collect data, for example, by asking people's opinions via surveys, or by searching the Internet. How to best collect data is in itself already an important question. Still, how to make good use of these data is even more important. This book is concerned with the field of statistics.
This introductory chapter presents an overview of this field in section 1.1. In section 1.2 we address the differences between various types of variables, and in section 1.3 we show how tables can be valuable tools to present observed outcomes.

1.1 What is statistics?

We are familiar with the term statistics in various idiomatic expressions. For example: according to the statistics this or that must be true. Or: according to the statistics the number of beachgoers is always highest on Sundays. Or: according to the statistics Ajax often wins the competition in the final match of the season.
Statistics is also the name of a broad discipline encompassing elements such as the theory of probability, but also censuses or opinion polls. It is sometimes said that with clever use of statistics you can make a person believe anything, be it true or not. Still, the data are the central point and form the basis for all insights. This book aims to demonstrate how you can collect, organize, summarize and analyze your data, with the ultimate objective to let you understand more of the world around you. Let us first look at the different types of statistics and familiarize ourselves with some fundamental principles.

Types of statistics
Statistics can be applied for different purposes. One of these purposes is discovering certain patterns in data you have collected. Or in other words: analyzing a data set. To describe the data you have the option to draw tables and graphs, or to calculate specific measures such as the mean outcome, or the extent to which outcomes appear to vary. You are describing a large collection of data in a purposeful manner in order to be able to sketch a concise picture. This application is called *descriptive statistics*.
You may even aim higher. Once you have gained solid knowledge of a particular topic, you may want to make predictions for the future, or use your analysis to make decisions. The concept of probability plays an important role in this matter. But in the end you still want to reach a valid conclusion – in spite of all kinds of uncertainties.

There is a range of statistical application in the fields of economics and management. This is a selection:

- A company would like to have insight in the levels of sickness absence among its employees. After having collected the data, all kinds of questions can be explored. Is the current absence rate lower than in the previous year? Are levels among older employees higher than among younger ones? Are there any departments with a strikingly high absence rate?
- The authorities are considering levels of employee pension contributions. In order to correctly estimate future payments they will need data on the life expectancies of men and women. In addition, numbers of people in employment now and in the future must be estimated, as those are the ones who pay the contributions. It follows that they will have to collect and analyze a huge amount of data.
- A supermarket is planning what they will need to take in stock for the coming Christmas period. This requires insight in the expected number of customers and possibly shifting preferences of customers.
- The Minister of Economic Affairs has asked to be informed about the level of inflation. To this aim prices of a great many products are collected. These are converted into a consumer price index number.
- A manufacturer has specified a highest and a lowest allowable limit for the dimensions of the parts that are produced. Using statistical methods the quality of the products can be measured periodically.

These are just a few examples. It will be clear that mastery of a number of statistical techniques is very important in many disciplines and professions – ranging from marketing tot operations management and from accountant to government policy maker. Moreover, knowledge of statistical methods is often a prerequisite for successful completion of graduation projects.

Some basic concepts

To gain more knowledge about a topic of interest you will need to do some research. Everything is possible, for example an opinion poll among a group of people, quality control of products in a factory, or establishing the percentage of companies that export goods. Before setting out, you would do well to decide about what elements you would like to draw a conclusion on in the end.

Are you planning an opinion poll among all adults in the Netherlands? Are you looking at the export figures of all companies, or only of those companies that employ at least 25 people?

The total collection of elements in the research is called the *population*. For example, if we should like to measure satisfaction with the student finance system, the population is likely to consist of all Dutch students who are eligible to receive funding. And in the case of a political survey the population consists of Dutch citizens who are eligible to vote.

Often it will be problematic to involve the total population in your research. It may be too time-consuming to survey everybody and, in addition, it may happen that not all information is available. For this reason you will often make use of a sample, which implies studying a selection of the population. In this case, it is hoped that the sample provides a good impression of the bigger picture.

An element of the population can have all kinds of characteristics. These characteristics are called *variables*. These variables can often take different values.

?? Please look at the information about the cars for sale in tables 1.1 and 1.5. What do you think is the population? And the sample? What variables have been collected?

Table 1.1 at the beginning of this chapter lists the first six cars. Table 1.5 on page 16 provides the whole range of 30 cars for sale at Ben de Beun. The car magazine was interested in a number of characteristics of used cars. The population is basically made up of all second-hand cars on offer on the Dutch market. From this population, a sample of cars offered by Ben de Beun's car dealership was selected, limited to eight popular models. The characteristics of interest are price, brand, mileage, condition, year of manufacture, colour, and whether or not the car has an automatic gearbox. Note that these variables quite differ in character.

?? What are the possible values the variable 'brand' can take? And what are the values the variable 'mileage' can take?

The possible outcomes for the variable 'brand' are Fiat, BMW, Mazda, Opel, Volkswagen, and so on. These are all names, also known as *qualifications*. The possible outcomes for the variable mileage are all numerical values. And, that is a good thing, for we can use these to make calculations.

Variables can be classified as either *qualitative* or *quantitative*. If the outcome of the variable is a numerical value, we are dealing with a quantitative variable. On the other hand, if the outcomes are qualifications, and no numerical values, we are dealing with a qualitative variable. If the variable is qualitative, the statistical analysis is rather limited. Quantitative variables, however, allow for statistical calculations.

?? Is it possible to convert a qualitative variable into a quantitative variable, and is it possible to convert a quantitative variable into qualitative variable?

You might consider representing qualitative data in a computer database by a numeric code. Regarding the variable 'brand', for example, the outcome Fiat might be indicated by code 1, BMW by code 2, Mazda by code 3, and so on. Please note, however, that after converting qualifications into numeric codes, it is still not possible to perform calculations that provide meaningful results. Assume that the mean value of the codes would amount to, for example 3.27. This is an outcome which clearly cannot be interpreted meaningfully.
It is also possible the other way round. A quantitative variable can indeed be converted into a qualitative variable. For example, in the above example mileages from 0 up to and including 79,999 could be qualified as 'very low'; 80,000 up to and including 119,999 as 'low'; 120,000 up to and including 199,999 as 'high'; and 200,000 and more as 'very high'.
Such a conversion can be useful to give others a quick impression of the observed mileages. Caution is required, however, as this approach in principle gives much less detailed information than when using the original data.

1.2 Types of variables

Different types of variables can be distinguished. We have already come across the concepts of quantitative and qualitative variables. Within the category of quantitative variables we can distinguish between *discrete* and *continuous* variables. A discrete variable in principle has a limited number of possible outcomes. How many passengers are on a bus? The outcome is likely to be a number between 0 and the maximum capacity of, let's say, 80 passengers. What are the prices per bottle of four types of Cola? Assume these four types cost €0,95, €1,28, €1,45, and €1,68, respectively. This goes to show that outcomes can be expressed as numbers with decimal places. Basically, it is possible to list all possible outcomes for a discrete variable.
In contrast, a continuous variable in principle can have an infinite number of outcomes. Think of the time you have to wait before the next train comes in, the weight of a pumpkin, or the distance between two trees alongside a road. Provided you have the right measurement instruments, measurements can be very precise and in theory may result in a huge amount of possible outcomes. As an aside, you may find it useful to remember that, when dealing with a continuous variable, for any two chosen outcomes it will always be possible to imagine an additional outcome lying between these two.

Please be aware that the contrast between a discrete and a continuous variable in practice is less rigorous than in theory, because precision of measurement instruments in fact is limited. Mileage, for example, is a continuous variable, but the car's speedometer usually indicates mileage as a number with no more than one decimal place. In this way, continuous variables are actually converted into discrete variables.

Four scales of measurement
Variables are measured on different scales. This is also referred to as the level of measurement. Four different scales of measurement are distinguished: nominal, ordinal, interval or ratio.

1 Nominal scale
Using a nominal scale, it is not possible to logically arrange the conceivable outcomes in a in a fixed order. True, characteristic of a variable can be converted into a numeric code, but this does not necessarily imply that these codes have a logical order.

?? What characteristics of the used cars are measured on a nominal scale? See Table 1.1.

In Table 1.1, car brand and colour are examples of characteristics measured on a nominal scale.

2 Ordinal scale
An ordinal scale, on the other hand, permits to arrange outcomes in a fixed order. Here, a clear method is available to determine what the logical order should be.

?? What characteristics of the used cars case are measured on an ordinal scale?

A car's condition is an example of a characteristic measured on an ordinal scale. It is easy to see that 'excellent' comes first, followed by 'good', 'fair', 'average' and 'bad'. Clearly a logical sequence. Shuffling these qualifications around is likely to result in a non-logical series.
Other familiar examples of ordinal scale are military ranks (from private third class up to and including general) and response categories in a questionnaire (from very bad to very good).

3 Interval scale
When measuring on an interval scale, differences between two outcomes are well-defined. An interval scale has no natural zero point. In fact it is a rather arbitrarily chosen characteristic. A typical example is temperature. The zero point is the freezing point of water, but one could just as well have chosen something else.

?? What variables in the table are characterized by an interval scale?

Year of manufacture is a typical example of a characteristic measured on an interval scale. The interval between years 1995 and 1998 is equal to that between 2009 and 2012, that is to say three years. Thus the intervals are of the same size.
Other well-known examples of an interval scale are temperature and time indication.

1

4 Ratio scale
In the case of a ratio scale we do have a natural zero point. Thus, referring to the ages of the cars, a ten-year-old car is twice as old as a five-year-old car.

?? What car characteristics are measured on a ratio scale?

The characteristics 'mileage' and 'asking price' are examples of characteristics measured on a ratio scale. Zero is a logical starting point of the scale in this case.
Other familiar instances of a ratio scale are weight and elapsed time.

?? Why is it important to distinguish between the different measurement scales?

It is important to choose the correct scale because this defines the arithmetic calculations to be applied.

1.3 Preparing tables

A data set can be quite voluminous and may have grown so unmanageable that it requires some order. The information is easily presented in a transparent manner in tables and graphs. As a basic principle, the data are classified into groups. These groups are called *classes*. For each class we are going to determine how many observations can be assigned to this class. These numbers are called *frequencies*. A proper table needs to meet some requirements:
- Provide a title, so as to inform the reader what it is about.
- Provide headings, indicating what information is provided in the columns.
- Choose the classes such that all observations can be included. Whenever possible these classes should be listed in a logical order.
- Provide a total row at the bottom.
- If relevant, provide the source of the data.

Regarding the characteristics of the 30 cars (Table 1.5), we could provide an overview of, for example, conditions, as shown in Table 1.2.

TABLE 1.2 Conditions of 30 cars

Condition	Number
Excellent	5
Good	9
Fair	5
Average	7
Bad	4
Total	30

Source: sample April 2014

In Table 1.2 all five response categories represent a different class. In some other case it may be necessary to first define classes.

?? How would you group the mileages into classes?

The researcher is free in choosing what classes to use for ratio or interval variables. Classes of similar width are preferable, but occasionally, if there are few observations in a part of the distribution, it may be more useful to have some wider intervals. This is done in Table 1.3 for the mileages, classified as mentioned in section 1.1.

1

TABLE 1.3 Mileages of 30 cars

Mileage	Number
0 - < 80,000	6
80,000 - < 120,000	4
120,000 - < 200,000	11
200,000 - < 400,000	9
Total	30

Source: sample April 2014

Note the format given for the 0 - < 80,000 class. This signifies that outcomes of at least 0 to a maximum of 79,999 fall in this class. In other words: this format tells us that the lower boundary belongs to this class whereas the upper boundary does not.

?? Assume that a car has exactly 120,000 kilometres on it. In what class do we categorize this?

The value 120,000 is the lowest number that belongs to the 120,000 - < 200,000 class.

Tables can take a variety of forms. A well-known form is the so-called *cross-table*. In a cross-table the observations are classified on the basis of two characteristics simultaneously. We will show this for the variable ‘mileage’ in relation to the variable ‘condition’. See Table 1.4.

TABLE 1.4 Condition versus mileage

Kilometres × 1000					
Condition	**0 -< 80**	**80 -< 120**	**120 -< 200**	**200-< 400**	**Total**
Excellent	3	1	1	0	5
Good	3	2	2	2	9
Fair	0	0	2	3	5
Average	0	1	5	1	7
Bad	0	0	1	3	4
Total	6	4	11	9	30

Source: sample April 2014

The table shows the frequencies of combinations of the two variables. Note that the total column to the right and the total row at the bottom display, perhaps surprisingly, the same information as presented separately in tables 1.2 and 1.3.

TABLE 1.5 Cars on sale at car dealer Ben de Beun

No.	Brand	Colour	Year of manufacture	Mileage (km)	Condition	Automatic yes/no	Asking price
1	Volvo	Blue	1999	345,000	Bad	Yes	1,800
2	BMW	Red	1996	320,000	Good	No	3,900
3	Mazda	Yellow	2010	56,000	Excellent	Yes	11,800
4	Opel	Gray	2003	130,000	Fair	Yes	3,500
5	Peugeot	White	2007	165,000	Average	No	4,700
6	Volkswagen	Black	2012	18,000	Good	No	17,900
7	Volvo	Gray	2004	163,000	Average	No	7,400
8	BMW	Gray	2002	225,000	Fair	No	5,300
9	Fiat	White	2007	92,000	Good	No	4,900
10	Mazda	Gray	2003	197,000	Average	Yes	3,500
11	BMW	Black	2005	173,000	Good	No	6,800
12	Peugeot	Red	2004	144,000	Average	No	2,900
13	Opel	Blue	2006	88,000	Excellent	Yes	10,500
14	Opel	Gray	2007	146,000	Fair	Yes	8,200
15	Toyota	White	1997	267,000	Bad	No	1,200
16	Volkswagen	Blue	2000	279,000	Bad	No	1,800
17	Toyota	White	2001	162,000	Average	No	1,900
18	Opel	Black	2004	210,000	Fair	Yes	4,700
19	Fiat	Red	2008	63,000	Excellent	No	4,800
20	Volkswagen	Gray	2005	173,000	Good	Yes	5,200
21	Opel	Gray	2004	228,000	Good	Yes	2,900
22	Volvo	Red	2009	135,000	Excellent	Yes	14,600
23	BMW	Gray	2002	293,000	Average	No	4,600
24	Volvo	Blue	2003	255,000	Fair	No	4,800
25	Volkswagen	Black	1998	173,000	Bad	No	1,300
26	Volkswagen	White	2009	46,000	Good	No	9,800
27	Toyota	Blue	2006	92,000	Good	Yes	4,100
28	Volkswagen	Red	2002	116,000	Average	No	2,300
29	Fiat	White	2010	35,000	Excellent	No	5,900
30	BMW	Gray	2011	46,000	Good	Yes	16,200

Source: record De Beun BV, April 2014

Exercises

1.1 Consider the 30 cars for sale (Table 1.5) and prepare a table which shows numbers of the different brands.

1.2 Prepare a table which shows the distribution of the cars on the basis of the asking price. Create €5,000 wide classes, with class €15.000 and higher as the last class.

1.3 Draw a cross-table showing the observations on the basis of the characteristics 'asking price' and 'automatic'. Use the same classes as in exercise 1.2.

1.4 Explore the following statements:

a Condition of the black cars typically is not that good.
b German cars tend to have a higher asking price than the rest.
c If you were to classify these cars into 'old' and 'young', on the basis of year of manufacture, the younger ones are more likely to have an automatic gearbox.

Chapter 20 presents more exercises dealing with a larger data set of 200 cars.