Birger Madsen

Statistics for Non-Statisticians

Statistics for Non-Statisticians

Birger Madsen

Statistics for Non-Statisticians

Dipl.Math. Birger Madsen
Novozymes A/S
Project Management and Information
Krogshoejvej 36
2880 Bagsvaerd, Denmark
bsm@novozymes.com

Preface

Never have so many organizational decisions been taken based on statistics as today! Everything is supported by numbers. This applies to marketing, economics, social sciences, natural sciences, industry and administrative work within organizations, businesses and institutions. It is therefore important to have insight into basic statistical concepts, when assessing statistical data material, as well as when preparing an investigation, so that it produces useful statistical results.

There are several books on elementary statistics. So why write another? The simple answer is: Because it is needed! This book fills a gap in the existing literature about statistics. Most existing short introductions to statistics take one of the following approaches:

– An approach based primarily on descriptive statistics (charts, tables, etc.)
– A purely verbal approach without any mathematical formula, but also without practical guidelines
– An approach based primarily on probability theory

In contrast, this book is intended to be a "first course for the practitioner," giving a lot of useful details for e.g., planning of surveys. Comparing this book to standard 500–600 pages textbooks on statistics, you will actually find a lot of practical information in this book that is not available in the standard textbooks!

I have for some decades taught statistics at all levels. It is my experience that the most important concepts of statistics can be explained, so that "ordinary" people can understand it. I have experienced this through hundreds of courses for many different audiences. Now, I have put my words on paper!

Who Is This Book Written For?

The book is written for those, who need to know how to collect, analyze and present data. You may be working with administrative data, financial data or data from the social sciences or natural sciences. Maybe you plan to collect data through sample surveys, such as customer surveys or similar.

You do not know much about statistics. Maybe you have learned a little about the topic earlier, but forgotten most of it again. Maybe you never learned anything about the topic, but you are curious!

Although the book does not require knowledge of statistics, I assume that you are not totally unfamiliar with numbers! You are able to perform simple calculations with a calculator. And you don't panic, when you see a simple formula containing a square root! Don't worry: This book is not loaded with mathematical formulas. But it is unfortunately impossible to introduce statistical concepts without a minimum of mathematical calculations.

It is an advantage, if you have a basic understanding of spreadsheets. This book is not a course in the use of spreadsheets – the easiest way to learn spreadsheets is by reading a computer booklet or taking a course!

Neither is this a "How to do statistics with Excel" book – you can use the references in the literature list, if you need this. There are many books of this kind, often occupying hundreds of pages...

Yet it may be useful to know how the most important statistical calculations can be performed using the features of a spreadsheet. Spreadsheets have nowadays made numbers and graphs accessible to most people. This also applies to statistical calculations!

If you do not have access to a spreadsheet, I can recommend the spreadsheet Calc from Open Office (a free Office suite). See links to software in Appendix. Virtually everything discussed in this book can be done with this spreadsheet!

I recommend that, while reading the book, you work with some simple data, which you enter in a spreadsheet. It is easier to learn statistics if you work a little with the substance!

The beginner may be satisfied with a spreadsheet as a tool for statistical analysis. In professional work with statistics, however, you will very quickly discover the limitations of a spreadsheet. Then it is time to consider a better tool for the purpose! Therefore, the Appendix presents some of the major programs for statistics as well as links, where you can find more about them.

It is my hope that the book can be used for private study and as supplementary reading at business colleges, technical schools, high schools, and the initial training of statistics at business schools and social sciences at a university level. The book is not written for any specific education.

After reading this book, you should have the ability to dig further into some of the many other books on statistics available on the market. It is my hope that this book can ease the transition to the reading of the (many) more advanced books on the subject. The number of books on statistics grows dramatically as the professional level rises!

The mathematically oriented reader has to accept that the book does not achieve 100% mathematical precision everywhere. Focus is on an understandable, rather than mathematical precise presentation.

Some topics in the book are a bit more "technical" than the rest of the book. These issues can be skipped, without thereby losing coherence. Some of these items

are placed in a text frame and entitled "Technical note: ..." Some topics provide a clear indication that they may be skipped They are often put at the end of a chapter.

There are also many examples of using spreadsheets. If you do not use spreadsheets, you can just read the examples, without bothering about how the results were obtained in the spreadsheet.

Structure of the Book

The book is structured in such a way that what you learn in one chapter is used in the following chapters. This means that you should read it from the beginning, at least up to and including Chap. 5.

Chapters 1–2 are about the collection and presentation of data. These are crucial issues for most people working with statistics.

Chapters 3–5 are the core of the book. They introduce the basic statistical concepts, including descriptive statistics, the normal distribution and statistical tests.

When you have read Chap. 5, Chaps. 6–8 can be read independently of each other.

Chapter 6 supplements Chap. 1; it is about the planning of sample surveys and experiments.

Chapters 7–8 supplement Chap. 4 on the normal distribution. Chapter 8 is probably the "heaviest" material of the book, and appropriately placed at the end!

The Appendices of the book contain a lot of hopefully useful information: Review of probability theory, bibliography, glossary of statistical terms, list of statistical functions in spreadsheets, list of statistical software, useful links and various useful tables. All words in this book, which are marked with an *asterisk* (*), are explained in the glossary.

At the publisher's website, you will find additional material for the book: Useful worksheets, further explanation, examples, etc. Of course, there is also a spreadsheet with the example dataset "Fitness Club," which is used as recurrent example.

I wish you a pleasant reading!

Copenhagen, Denmark   Birger Madsen

Acknowledgements

Contents

Abbreviations

$\bar{x}$	Sample average
ANOVA	Analysis of variance
$B(n, p)$	Binomial distribution (n observations, probability p)
CV	Coefficient of variation
DF	Degrees of freedom
$E(X)$	Mean of X. $E=$"Expectation"
H_0	Null hypothesis
H_1	Alternative hypothesis
$N(0,1)$	Standardized normal distribution (mean 0, variance 1)
$N(\mu, \sigma^2)$	Normal distribution with mean μ and variance σ^2
R	Sample range
s	Sample standard deviation
s^2	Sample variance
$V(X)$	Variance of X
μ	Population mean
σ	Population standard deviation
Σ	Sum
σ^2	Population variance
χ^2	Chi-squared (distribution or test)

Chapter 1
Data Collection

This chapter explains some basic concepts within statistics. Also, we look at the most important ways to collect data in surveys.

> Statistics can be defined as a collection of techniques used when planning a data collection, and when subsequently analyzing and presenting data.

Dating back to ancient times people have needed knowledge about population size, to carry out a census of the armies or calculate expected taxes. The word statistics is derived from the word "status" (originally coming from Latin); and it was exactly the status of society, which was the subject of the first statistics! Later emerged probability theory (in connection with games!), demographics and insurance science as areas, in which statistical thinking was essential.

In today's digital age it is easy to collect as well as process and disseminate data, and therefore statistics is used for a variety of surveys throughout society.

> Most statistical surveys can be divided into the following phases:
>
> 1. Clarification of concepts
> 2. Planning of data collection
> 3. Data collection
> 4. Analysis and presentation of data
> 5. Conclusion
>
> Statistical methods (and statisticians!) are particularly useful in phases 2 and 4 of the survey.

> There are two kinds of statistics:
>
> – Descriptive statistics
> – Analytical statistics

Descriptive statistics means *describing data* using *tables*, *charts* and simple *statistical calculations* such as averages, percentages, etc. This is what many people understand by the word "statistics". It was also the kind of statistics that was produced in ancient times.

Analytical statistics is used to *assess differences and relationships* in data. For example, we could examine whether there is a relation between height and weight of a group of persons; or whether there is a difference between height of boys and height of girls, as well as provide an estimate of how large this difference is. Analytical statistics is a mathematical discipline, based on calculus of probability. It is a relatively new discipline that has been developed throughout the twentieth century.

This book is about descriptive statistics as well as analytical statistics. In practice, you need both. Analytical statistics is a very large topic, and here we can only scratch the surface (see especially Chaps. 5, 7, and 8). If you want to know more about analytical statistics, see some of the more advanced books in the literature list.

1.1   Sample Surveys

> In any *survey* (*), we collect information on the *individuals* of either the entire *population* (*) (a *total survey*) or a relatively small number of individuals of a *sample* (*) (a *sample survey*), in order to analyze and present data.

We are interested in the entire population of individuals. The advantage of investigating only a sample is that it is both faster and cheaper than investigating the whole population. In some situations, a carefully planned sample survey can even give more accurate results than a badly planned total survey!

We investigate the individuals in the sample in order to study the whole population! This means that the sample gives us an *estimate* (*) of the characteristics of the population (Fig. 1.1).

Examples of characteristics:

– *Average* (*) of a measurable attribute of individuals, e.g., height
– *Percentage* of individuals who belong to a particular category (e.g., who have a specific hobby)

The larger the sample, the better an estimate of the population!

It is also important that *the sample is representative* of the population. In practice, this means that the individuals in the sample are *selected at random*, in order to cover the whole population. We are dealing much more with *sampling* (*) in Chap. 6.

The sample (and the population) may consist of different types of individuals, depending on the context.

Some examples:

– People
– Companies

Sample and real world
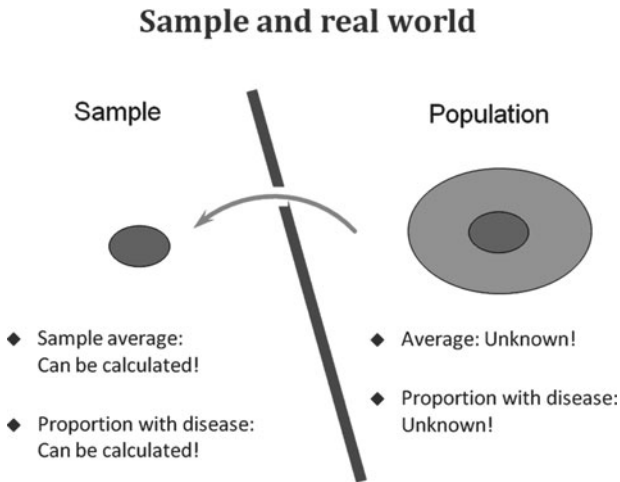
Fig. 1.1 Sample and population

- Public institutions
- Families
- Vouchers
- Houses
- Cars
- Trees
- Dogs
- Bacterial colonies
- Bottles or cans of beer
- Pills

The concepts in this book can be applied to all types of samples. The examples are mainly samples consisting of people. But the principles can be applied to all types of samples.

Typical applications with *samples consisting of people* are: analysis of attitudes, consumption, durable goods, interests and hobbies, eating and drinking habits, transportation, traffic, vacation, media (TV, radio, newspapers) and certain sensitive topics, such as alcohol consumption.

Sample surveys can provide *a high degree of flexibility*: One can in the same survey have questions on media consumption, traffic patterns and attitudes. Sample surveys are also widely used in commercial surveys, e.g., in connection with telephone interviews.

One of the essential applications of sampling is *sampling inspection* in the field of *statistical quality control*. If you are working with statistical quality control, most of this book will be relevant to you. The issues that are specific to this discipline, however, will not be dealt with here. See the literature list for books on statistical quality control.

1.2   Fitness Club: Example of a Sample Survey

This example is fictitious survey. It will be used as an example for the subsequent chapters.

Fitness Club has a number of sports facilities. This includes facilities for strength training, weight loss and cardiovascular workout.

Fitness Club wants to understand the needs of their young customers, kids of age 12–17 years. The club wants to know, how satisfied these kids are with the sports facilities. They also want to obtain information about their health in order to better customize the sports facilities for the various types of training.

Therefore, a sample survey is carried out among the kids using the sports facilities. We will later discuss how the survey can be organized. Moreover, we present some findings from the survey.

The population consists of kids using the sports facilities in Fitness Club. The individual is one kid. The sample consists of 30 kids.

Some data related to this example are found in the Appendices in Chap. 9 of this book.

1.3   Experiments

In certain situations, the information needed simply does not exist at all! In this case, one can plan (or design) an *experiment* (*), with the aim to provide the relevant data. In the experiment, we test the influence of one or more factors on a measured result. This approach is widely used in technical and industrial contexts.

In these situations the population is not well defined. The experiment can produce very different results, depending on how it is planned. Nevertheless, it also makes sense to talk about a sample in this situation. We can perceive the experimental results as a sample of all the (infinitely many) possible results we could have obtained in all similar experiments.

The statistical techniques used to analyze and present data are by and large the same, regardless of whether data are collected in a sample survey or an experiment.

The special techniques used in the planning of experiments will not be dealt with in this book; the space does not allow this. There are numerous books on this topic, see the literature list.

1.4   Experiments: An Example

To illustrate the difference between a *sample survey* and an experiment, we present an example of a *planned experiment*.

A producer of soft drinks wants to launch a new product with an entirely new blend of flavor ingredients. He is uncertain how to balance the flavors with

Table 1.1 Data from experiment

Flavor	Carbonate	Sugar	Rating
1	1	1	3.1
1	1	2	3.2
1	2	1	3.6
1	2	2	3.6
2	1	1	5.1
2	1	2	4.6
2	2	1	4.8
2	2	2	5.3

carbonate and sugar. He therefore performs an experiment, where he is testing various combinations of the ingredients.

The producer makes a list with the lowest respectively highest value added of the three ingredients, which he expects to use.

Then he sets an experimental plan. He makes a few bottles of product of each of the eight possible combinations of the three ingredients (Table 1.1).

Here "1" means add "a little" of the ingredient, while "2" means add "a lot".

A taste panel of ten persons tastes all bottles and rates them. They use a scale of 1–7, where 4 is "neither good nor bad," 1 is "unusually poor," 7 is "unusually good".

The average rating for each combination is shown in the last column of the table.

Afterwards the manufacturer studies the results of the experiment. He observes that the four combinations with a "high" value of the flavor have better rates than the four combinations with a "low" value of the flavor. The average of the four combinations with a "high" value is 4.95, and the average of the four combinations with a "low" value is only 3.38. It would therefore appear that a lot of the flavor ingredient must be added.

By calculating the other averages, he can see that it is not so important, how much carbonate and sugar is added. He therefore chooses to add an intermediate dose of carbonate and sugar.

Experiments of this type are widely used in industry, but they can also be used in marketing (as in this case) and social sciences.

1.5 Data Collection

We have now seen the difference between sample surveys and planned experiments. In the next chapter, we show how to present results from a sample survey or an experiment. First, however, we see how data are collected.

Most of the book describes general techniques that can be applied to all types of populations. Populations consisting of or involving humans are probably what will interest most readers of this book. That is why the rest of this chapter is on data collection methods, questionnaires and various sources of error for surveys of populations consisting of or involving people; for example, populations of enterprises, institutions or families.

If your population is not of this type, you can skip the rest of this chapter. The rest of the book will be useful, regardless of the type of population you are dealing with.

1.6   Registers

Before planning a sample survey, one should always check if the necessary information already exists in a database. A database consisting of people is often called a *register*. The vast majority of enterprises, institutions and organizations possess one or more databases with "business data"; for example member databases, customer databases, etc. It is varying from organization to organization, how much information on each individual these databases contain.

If a register contains the information you need, it is often relatively simple to produce the desired statistical results. This is simply reuse of data already collected for administrative purposes!

This will most often be done on the whole population, as it rarely is more difficult than making an investigation on only a sample of the individuals in the register. Therefore, *register surveys are nearly always total surveys*.

Briefly, you should *use registers*, when *the information needed is available* in an appropriate form, and you have access to it. On the other hand, you will use sample surveys, where registers do not exist, are inadequate, or you do not have access to them.

If the register does not contain the desired information, it can often be used as the *basis for selecting a sample*. Sometimes, a register survey is *combined with a sample survey*. The register provides part of the requested information, the rest you get through the sample survey.

Examples of register surveys are found in many research projects in the social sciences. One can use the information from the administrative registers for many purposes. The *national statistical institutes* in many countries have access to the registers, providing valuable information on the entire population of a nation.

1.7   Questionnaire Surveys

A *questionnaire survey* has three parties:

– *A researcher*, who formulates the questions.
– *An interviewer*, who asks the questions.
– *A respondent*, who answers the questions.

The usefulness of the results from a questionnaire survey depends on, whether *the researcher formulates the questions* in a way that ensures *effective communication* between the interviewer and the respondent!

There are two main types of questions:

1.7.1 Background Questions

This could be sex, age, marital status, type of accommodation, residence, education, employment, annual income, etc. These questions are used to *group results in tables and graphs*.

For example you create tables, where you study the distribution of the answer to a question across different groups (e.g., age groups). More can be found on this in Chaps. 2 and 5.

1.7.2 Study Questions

They constitute the body of the questionnaire.

A few examples:

(a) Are you in favor of or against nuclear power?
(b) Were you employed last week?
(c) How many nights in a hotel did you have last year?
(d) What was your annual income last year?
(e) Do you think the prime minister is doing a good job?

It is important to distinguish between closed and open questions:

1. *Closed questions* are of the type:
   What education do you have?

   – Elementary education
   – Apprentice/technical education
   – Secondary education
   – Higher education

   There are, in other words, a *limited number of categories* (often only two, such as yes/no). The advantage is that it is easy to handle the data processing afterwards, because it is not necessary to do any "recoding" of the responses. The disadvantage is that the respondent may not find the right answer option.
2. *Open questions* are of the type:
   What education do you have?
   Respondents have the opportunity to *write any free text*. The advantage is that the respondent always has an answer option. The disadvantage is that it is cumbersome and time consuming to handle the data processing afterwards: A person with subject matter knowledge must "code" the free text into well-defined categories.
3. *Semi-closed questions* are of the type:
   What education do you have?

   – Elementary education
   – Apprentice/technical education
   – Secondary education

– Higher education
– Other:

This is just a closed question, which has an "Other" category with the opportunity to write free text. This combines the best from the closed and open questions.

Many questions in questionnaires express an assessment. There are a number of *ordered response categories.*

An example:
What do you think about the course?

– Very good
– Good
– Neither good nor bad
– Bad
– Very bad

The number of categories in this kind of question is a topic of much discussion.
There is consensus that the number of categories *should not be too big* and *not too small either*! Approximately three to seven categories is probably the best, five answer categories is perhaps most commonly used.

On the other hand, many people prefer *an even number of categories.* It is based on an idea that an odd number of categories (for example 5) encourage the respondent to choose the middle category, because it is convenient. Therefore, many prefer four or maybe six answer categories.

Introduction questions are used now and then, particularly with complex issues, where the respondent needs to be "guided" in the topic. It is very important to be careful with the wording of these questions, otherwise you run the risk of being biased ("*leading questions*"). The risk of influencing the respondents is simply too big. There exist several examples on this in connection with, e.g., polls.

Deliberately influencing the respondent in order to provide a certain answer is (hopefully!) a rare exception, although it certainly occurs.

Control questions are used to ensure that the respondent has understood the questions and answered honestly, e.g., within a very complicated and sensitive issue.

You therefore ask the "same" question in a *new wording*. Or you ask a *supplementary question*, which may shed light on the responses to some of the key questions of the study.

An example from a survey, in which people were asked about their alcohol consumption the day before: The following control question was asked:
Do you normally drink more, the same or less?
The distribution of the answers is given in Table 1.2.

Table 1.2 Control question

More	App. 10%
The same	App. 50%
Less	App. 40%

The interviews were evenly spread over all weekdays and all months of 1 year! Thus, there should be a roughly equal number of people, who normally drink more respectively less. This was not the case, because the respondents want to appear not to drink too much. This may indicate that the respondents did not answer honestly!

It is very important not to ask too many questions!

There is a tendency to ask a lot of redundant questions just "out of curiosity". This bores the respondent and increases the risk that the respondent will skip the last questions. Maybe those questions were essential?

A basic requirement for the questionnaire is to *ensure a clear communication*, i.e., the use of simple and clear formulations. *Keep it simple!*

It is important that the researcher when formulating the questions tries to put himself in the respondent's place. Keep in mind that the respondent is not an expert!

If the results of the questionnaire are to be compared with results from other questionnaires, care must be taken to ensure *comparability*. This means that you pose *the same questions* and use *the same answer categories* as in the other surveys.

Remember to always *ask about one thing at a time*. It may be tempting to put two issues together in one question, but the result is rarely good.

An example from real life is the following question:

Are you shopping

– Alone?
– Together with others?
– Never?

Here, the researcher is dealing with two issues:

(a) Whether the respondent is responsible for a large or small proportion of purchases in the household.
(b) Whether the respondent is alone or together with others when shopping.

The result was, that many men answered "Alone" because they buy things for themselves, for example in a kiosk. On the other hand, many women answered "Together with others" because they bring their children when shopping in the supermarket. The purpose of the question was to identify the person responsible for most of the shopping in the household.

It is always a good idea to *test the questions on a few people* in order to see if they are understood and are answered without problems. The respondents might be "friends and relatives"! Such a study (known as a *pilot study*) can detect many problems before it is too late!

1.8 Sources of Errors in Surveys

Regardless of how data are collected in a questionnaire, problems may arise, which in the worst case may influence the results significantly. There may be errors that come from the question wording and errors arising from the data collection.

Many errors in surveys come from problems with the wording of the questions. This can be

- Unclear questions.
- People cannot or will not respond, they do not know, etc.
- Unconscious influence on the respondent, e.g., in connection with "introduction questions."

These errors can only avoided by a careful wording of the questions!

Non-response (*) *is another major problem.* This means that *some respondents do not participate* in the survey. This may be *due to problems in the data collection.*
Non-response may mean that the results of the survey are misleading!
For example, it is often difficult (both by telephone interview and by visit) to catch the "busy businessman," who is seldom at home (working more hours per week than the average). If you are not working hard to reach this group of respondents, it is obvious that the survey results will be "unbalanced".

The two most frequent causes of non-response are:

- *People are not at home, or not reached* (by telephone interview or visits).
- *Refusal: People may simply ignore any attempt of contact.*

The only solution to not reaching people is to *make numerous attempts of contact* (telephone calls, visits, etc.). Typically you should make attempts both in the weekend and on weekdays, both in daytime and evening hours, if you want to increase the response rate. This means at least four contact attempts, which of course is costly...
The problem with refusal is growing as the number of more or less dubious marketing companies increases, which again increases the number of phone calls to each household. People have simply had enough!
The problem is of course impossible to avoid completely. However, it is possible to reduce it:

- First of all, *proper training of the interviewers* is important, so that they can better cope with "troublesome" respondents.
- One can also *offer the respondents a reward* (e.g., a lottery ticket) for participating in the survey.
- *Combining multiple data collection methods* (for instance telephone interviews and visits) can also provide a higher response rate. This again increases the costs, however.

1.9   Comparing Methods of Data Collection

The main types of data collection in surveys are:

– *Internet*, e-mail or other electronic data collection (e.g., text messages on mobile phones)
– *Mail* questionnaire
– *Telephone* interviews
– *Personal* interview ("visit" or "face-to-face" interview)

When selecting a data collection method, you should obviously evaluate both cost and quality.

Usually, *there is a relationship between cost and quality*! The above list is roughly arranged in order of increasing cost and quality. Thus Internet or e-mail questionnaires are cheap, while personal interviews are expensive.

The quality depends largely on the non-response rate, i.e., proportion of respondents who do not participate. The non-response rate is considerable by Internet or e-mail; by personal interviews it is much smaller.

Errors in data are also a quality issue. The traditional mail questionnaire gives no opportunity to correct errors during the interviewing process, as do the other data collection types.

The first two methods of data collection are carried out without an interviewer, while there is an interviewer present with the last two methods. The *interaction between interviewer and respondent* in the last two data collection methods has both advantages and disadvantages that can affect quality:

– *Advantage*: The interviewer can help the respondent, if there is any doubt about what is meant by a question. This can reduce the number of questions that are not answered (or answered by "Do not know", etc.).
– *Disadvantage*: The interviewer may unconsciously affect the respondent's answer in a certain direction. This risk is largest in personal interviews, but can be minimized by careful training of interviewers.

Another key factor in the choice of data collection method is the questionnaire's length, i.e., number of questions.

The first two data collection methods should not be used for very long questionnaires, where the respondent's (lack of) patience will often mean "less serious" (or no!) replies.

The last two data collection methods involve a direct contact between interviewer and respondent. The talented interviewer can exploit this, if a respondent is becoming impatient.

The possibility of "*previewing*" is yet another factor to take into account. In personal interviews different visual effects, which can help the respondent to understand a difficult issue, can be shown. By telephone interview audio files can

be played during the interview. In mail questionnaires one can display pictures. Finally, in Internet interviews "almost everything" is possible by taking advantage of modern technology.

Often you will use a combination of two data collection methods.

One collection method (often telephone) is used for most interviews. For the reminders is used a second data collection method, e.g., personal interview. The skilled interviewer can better persuade a respondent to participate in the survey. This can be used to achieve a higher response rate.

Also, this enables an assessment of whether there is *a difference in the results* among those who answered respectively did not answer in the first round. This provides information on the *validity* of the survey results.

1.10   Example Continued

Fitness Club has a register of all customers. This register contains contact information (name, address, etc.) as well as information on sex and age of all customers; this makes it easy to extract a sample of customers in the age group 12–17 for a survey. We will discuss the process of selecting a sample in Chap. 6.

Fitness Club chooses to send a mail questionnaire to all the selected customers. It is a relatively inexpensive solution. A disadvantage is that for instance the questions about their health are answered subjectively. This is, indeed, a common feature of most questionnaire surveys!

Another drawback is that such surveys often lead to high non-response rates. To reduce this problem, Fitness Club encloses a stamped addressed envelope. Also, the customers responding will participate in a lottery, where they can win smart mobile phones and other electronic gadgets, which are particularly attractive to kids of this age group.

Chapter 2
Presentation of Data

In this chapter we show how to *present the results* of a questionnaire survey, using graphs and tables.

Graphs (charts, plots, etc.) are suited to get a feel of *patterns, structures, trends* and *relationships* in data and thus are an invaluable supplement to a statistical analysis. They are also useful tools to find *unlikely* (e.g., extremely large or extremely small) *data values* or *combinations of data values* (such as a very high person who does not weigh much), which may be *errors in the data*.

It is easy to create graphs with a spreadsheet, such as Microsoft Excel or Open Office Calc. Here, only the main types of graphs are covered. There are many other types of graphs than those shown here. See the "Help" menu in your spreadsheet to see the possibilities!

Tables are another way of presenting data, which is also discussed briefly here.

2.1 Bar Charts

Bar charts are familiar to most people. They conveniently summarize information from a table in a clear and illustrative manner.

Let us consider an example from the "Fitness Club" survey. As one focus of the young customers is weight loss, a table of average height and weight by sex is interesting. It may, in tabular form look as shown in Table 2.1.

A corresponding bar chart is shown in Fig. 2.1.

We see, at a glance, that the boys are both slightly taller and slightly heavier than the girls.

It is also well known that *this type of chart can be constructed to "cheat" with the axes*. If we consider for a moment only the (average) height, it could in graphical form look like the one shown in Fig. 2.2.

This chart contains the same information on the height of girls and boys, as the graph above. However, most people will *get a wrong impression* of the situation by considering this chart. The boys seem to be much taller than the girls, because the lower part of the bars is cut off!

It is therefore *important to be aware of the axes when studying bar charts*, as well as *when constructing them*.

Table 2.1 Average height
and weight

Sex	Average	
	Height (cm)	Weight (kg)
Girls	148.8	48.2
Boys	163.9	56.4
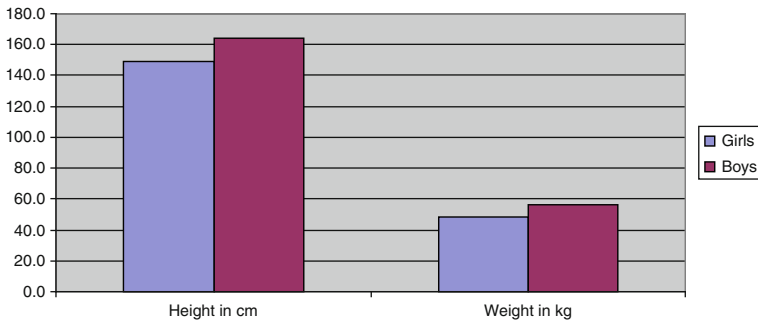
Fig. 2.1 Average height and weight
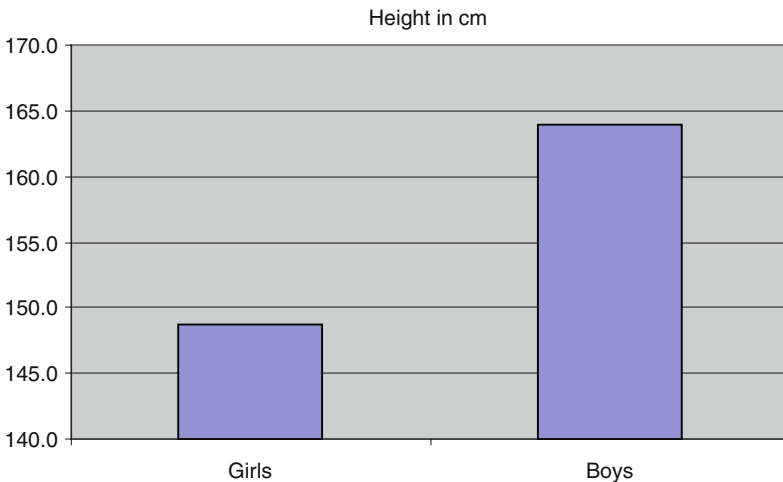
Fig. 2.2 Average height

2.2 Histograms

A *histogram* (*) is a special bar chart. It shows *the frequency* (*) *of the data values*: you can *visualize the distribution* of data values, for example, *where* "*the center*" *is* located (i.e., where there are many data values), *how large the* "*spread*" *is*, etc.

Table 2.2 Interval counts

In the interval from (but not including)	Up to and including	Number of kids	Interval center
100	120	3	110
120	140	3	130
140	160	10	150
160	180	10	170
180	200	4	190

Fig. 2.3 Histogram
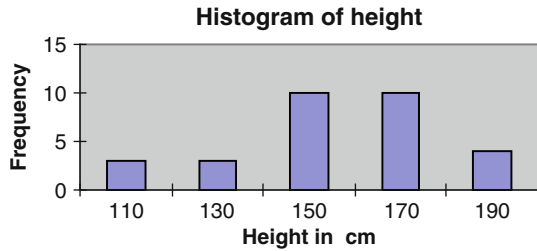
If you order and group the data values [see data in the (Chap. 9)] of height, you get Table 2.2.

This means: in the interval from (just over) 100 to 120 cm there are in total three kids. *The intervals must* of course *be constructed* so that *they do not overlap.* Therefore, only one endpoint belongs to the interval. The center of the first interval is 110 cm.

Counting of the frequencies can be done manually. Or you can let Microsoft Excel do it: Use the add-in menu "Data Analysis," which has a menu item "Histogram." This option is not available in Open Office Calc.

When the frequency data from the table are plotted in a bar chart, it looks like this as shown in Fig. 2.3.

The *general considerations* you should have in connection with determining the number of bars and the width of the bars are:

– The graph should *fit onto the paper* (or screen).
– The graph should be able to *accommodate all observations*. Be aware of the minimum value and the maximum value.
– The graph *does not* "*violate*" *the data material*. If, for example, there are two obvious "bulges" in the distribution, do not make so few bars that this important information disappears!
– The *intervals must be defined clearly*. There must be no doubt as to which interval an observation should belong. You must be sure to which interval the endpoints belong.
– You should be able to *compare the graph with other graphs*, e.g., from previous surveys.

Table 2.3 Number of bars

No. of values	No. of bars
10	3
100	7
1,000	10
10,000	13

The first decision is: How many bars should you use?

– The histogram should normally have 3–13 bars.
– The more the observations, the more the bars.

As a rough guide, you can use Table 2.3.

In the "Fitness Club" example, the number of values is 30. The number of bars should be between 3 and 7, so 5 is probably a fairly good choice.

Technical Note
To determine the number of bars, we can use the following formula:
No. of bars $= \log(n)/\log(2)$
Here n is the number of values and *log* is the logarithmic function (use calculator or spreadsheet for this calculation). You can use logarithms of base 10 or natural logarithms; the result of the formula will be the same.
In the "Fitness Club" example, $n = 30$. Here we use logarithms of base 10:
No. of bars $= \log(n)/\log(2) = \log(30)/\log(2) = 1.48/0.30 = 4.9 = $ app. 5.

The next question is: How wide should the bars be?
Having determined the number of bars, we can easily find out how wide each bar must be:

1. *Interval length = (Maximum value − minimum value)/(number of bars).*
2. *Round off the result to an appropriate number*, if necessary.

For the height data, maximum value $= 198$ cm and minimum value $= 112$ cm. This gives $(198 - 112)/5 = 17$. This might be rounded to 20. The previously proposed classification seems to provide a reasonably good description of data.

From time to time you see *histograms, where the bars are not equally wide*. You should absolutely avoid this, because it is difficult for the reader to interpret. It is, for example, not immediately clear how to scale the *Y*-axis. Should you take

into account the visual impression? Then the bar height should be smaller, if a bar is wider. This means, however, that the readings on the *Y*-axis cannot be readily interpreted.

You should only use histograms with equally wide bars!

2.3   Pie Charts

Pie charts are often used to show how large a part of the "pie" each group represents. This may be used with *frequency data* (equivalent to a histogram), but often the pie chart is used in connection with *economic quantities*, such as expenses or income.

A *frequency table* of the number of girls and boys in our sample looks like the tabular form shown in Table 2.4.

This information can be illustrated graphically as either a bar chart (Fig. 2.4) or a pie chart (Fig. 2.5).

The same information is given in the two graphs! If there are only a few groups like in this example, many feel that the pie chart is the most illustrative chart. The pie chart also has the advantage that you cannot "cheat" in the same manner as in

Table 2.4 Frequency of each sex

Sex	Number
Girls	13
Boys	17
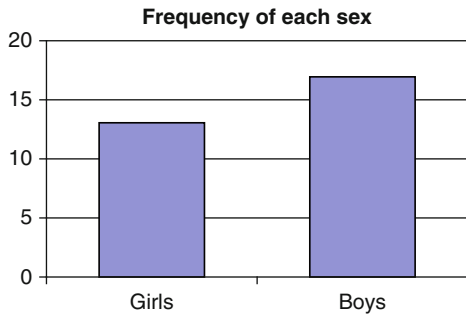
Fig. 2.4 Bar chart
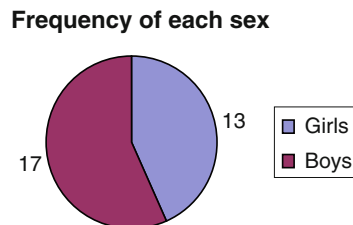
Fig. 2.5 Pie chart

the bar chart, where you can cut part of the axes. In case of more than six to seven groups, bar charts are probably most appropriate.

2.4   Scatter Plots

Scatter plots are well suited to show *relationships between two variables*.

In the "Fitness Club" example, we assume that there is *a relationship between height and weight*: the taller a kid is, the heavier it is. A scatter plot is illustrated in Fig. 2.6.

Weight is the *Y variable* or the *dependent* variable. Height is the *X-variable* or the *independent* variable. We imagine that weight depends on the height, i.e., there is a "*cause*" and an "*effect*."

In other cases, it is more arbitrary, as to which variable we choose as *X* and *Y*. We simply imagine that there must be a relationship (or *correlation*), without necessarily a "cause" and an "effect."

Chapter 7 gives tools to investigate *whether there is indeed a statistical correlation* between the two variables, height and weight.

On the other hand, *one cannot by statistical methods or by studying graphs determine whether there is a "cause" and an "effect."* Yet one can regularly in newspapers see examples of conclusions, where studying a plot leads to a conclusion, that *X* is the "cause" and *Y* is the "effect."

2.5   Line Charts

Line Charts are often used to illustrate *a trend*, where the *X*-variable is, e.g., time, age or seniority.
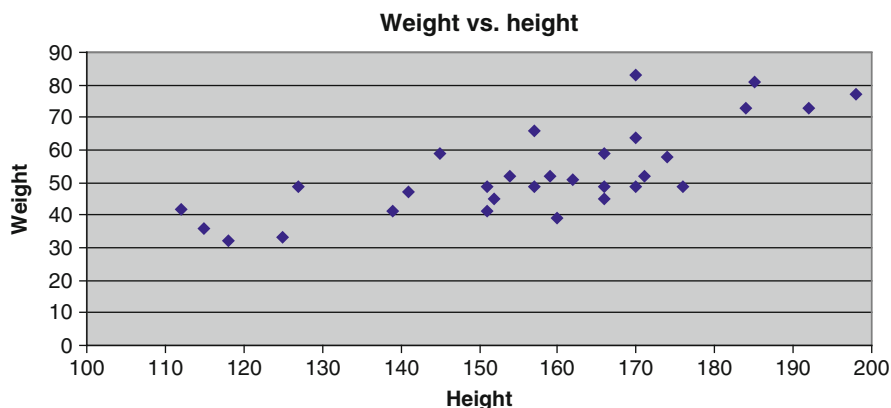
Fig. 2.6 Scatter plot

In the "Fitness Club" example, we would like to show how the average weight of the kids increases with age. We have the data as shown in Table 2.5.

We can illustrate this with a line chart as shown in Fig. 2.7.

This chart suggests that there might be some increase in weight with increasing age. As for bar charts, it is important to *be aware of the axes*. The same information can also be visualized in Fig. 2.8.

Table 2.5 Weight vs. age

Age	Average weight
12	48.40
13	50.00
14	49.50
15	61.75
16	56.50
17	66.00
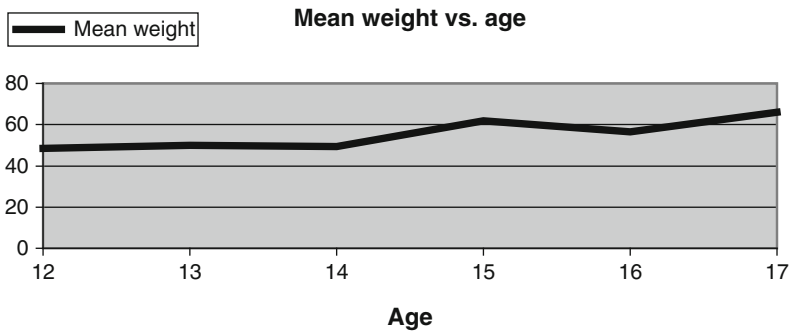
Fig. 2.7 Weight vs. age
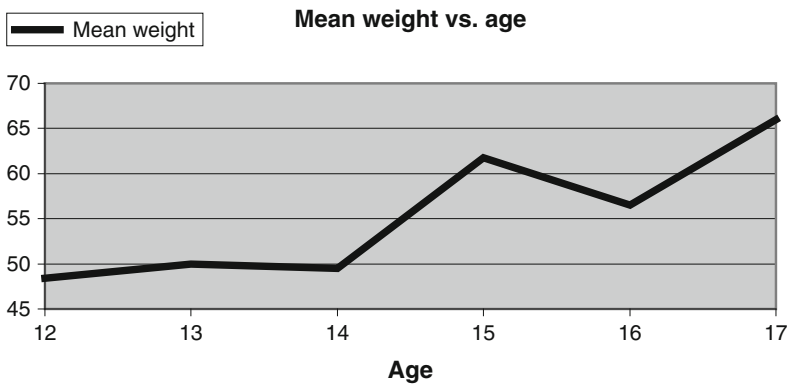
Fig. 2.8 Weight vs. age

The visual appearance is now quite different! It now seems as if there is a dramatic increase in weight with increasing age.

2.6 Bubble Plots

The *bubble plot* is a variant of the scatter plot. Instead of points, bubbles are plotted. The *size of each bubble* (either area or diameter) represents the *value of a third variable*.

It is probably most "fair" to let *the area of the bubble* be proportional to the value of the third variable, because the area is closely linked to the immediate visual impression.

However, if the third variable does not vary much (e.g., at the most by a factor of 2 between the minimum and maximum value), it is probably best to let *the diameter of the bubble* be proportional to this variable. Otherwise, it will be simply too hard to see the difference in size of the bubbles.

In the "Fitness Club" example we let the diameter of the bubble show the age of the kids. Large bubbles are the oldest kids and small bubbles are the youngest kids (Fig. 2.9).

In this plot, one can clearly see that the three kids at the left in the diagram, who are small in terms of both height and weight, are among the youngest, because the bubbles are relatively small.

2.7 Tables

Charts are an important way to present the results of an investigation. Other methods are tables, which we discuss here, and various statistical "key figures," which are the subject of the next chapter.
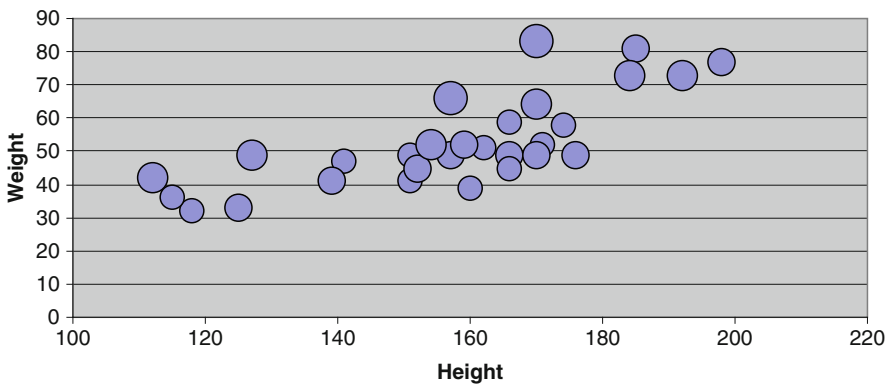
Fig. 2.9 Bubble plot

2.7.1 The Ingredients of a Table

Consider a typical table, such as Table 2.6.

The table ingredients are:

– *Table title*: "No. of kids by sex and age."
– *Column title*: It is a grouping of the variable "Age," supplemented with a "*Total*" column.
– *Row title*: It is a grouping of "Sex," supplemented with a "*Total*" row.
– *Cells*: This is the "core" of the table, for example, *frequencies, percentages* or an *average* of a variable. In Table 2.6 the cells contain a frequency.
– *Footnotes*: At the bottom of the table we find some information on the data sources, possibly supplemented by additional notes and comments.

So there are *two dimensions* of the table: *rows* and *columns*. Each dimension will often show *a grouping* of data. Or, one dimension can consist of *several different variables*, as in Table 2.7.

Here, the column dimension consists of the average for three variables: height, weight and age.

On other occasions, the column dimension could be *several calculations* of the same variable, as shown in Table 2.8.

Here, the column dimension consists of minimum, average and maximum values of weight.

Table 2.6 No. of kids by sex and age

No. of kids by sex and age				
Sex	Age			
	12–13	14–15	16–17	Total
Girls	5	6	2	13
Boys	6	8	3	17
Total	11	14	5	30

Source: Sample survey, Fitness Club
Other notes and comments

Table 2.7 Average, 3 variables

Sex	Average		
	Height (cm)	Weight (kg)	Age (years)
Girls	148.8	48.2	13.8
Boys	163.9	56.4	14.2

Table 2.8 Several statistics

Sex	Weight		
	Min	Average	Max
Girls	32	48.2	81
Boys	36	56.9	83

2.7.2 Percentages

The first table shown above gives the *sample frequencies* of kids in each combination of sex and age. Often, you will prefer to display the *sample percentages*. The sample frequencies are less interesting themselves.

The sample percentages are directly comparable to the population percentages, which may be known from a register. Thus, you can immediately assess *whether the sample is representative*; more about this in Chap. 5.

The percentage breakdown in the sample is shown in Table 2.9.

Our sample is not very large. The use of percentages in a small sample is of questionable value. For instance, 10% in the combination of boys aged 16–17 years covers only three kids... *Use percentages with caution, when the sample is small!*

Often percentages are given as *row percent* or *column percent*. This means that the percentages add up to 100% along rows, respectively, columns. This is shown as Row percent (Table 2.10) and Column percent (Table 2.11).

The reason for the use of row percent or column percent is often *a particular interest in the distribution of one dimension*, while the other dimension is only seen as a grouping.

It may also be that one perceives one dimension as a "*cause*" and the other as an "*effect*."

Table 2.9 Percentages

No. of kids by sex and age, percent				
Sex	Age			
	12–13	14–15	16–17	Total
Girls	16.7%	20.0%	6.7%	43.3%
Boys	20.0%	26.7%	10.0%	56.7%
Total	36.7%	46.7%	16.7%	100.0%

Table 2.10 Row percent

No. of kids by sex and age, row percent				
Sex	Age			
	12–13	14–15	16–17	Total
Girls	38.5%	46.2%	15.4%	100%
Boys	35.3%	47.1%	17.6%	100%

Table 2.11 Column percent

No. of kids by sex and age, column percent		
Sex	Age	
	12–13	14–15	16–17
Girls	45.5%	42.9%	40.0%
Boys	54.5%	57.1%	60.0%
Total	100%	100%	100%

Table 2.12 Row percent

Cardiovascular workout and physical fitness, row percent					
Cardiovascular workouts?	Physical fitness				Number
	Bad	Medium	Good	Total	
No	40%	40%	20%	100%	15
Yes	20%	40%	40%	100%	15

In the "Fitness Club" example the kids were asked whether they do cardiovascular workouts (row dimension). They have also been asked how they assess their physical fitness; here we use three categories: bad, medium and good. This is obviously a subjective assessment. The alternative is to measure their fitness rating through a physical test, which would be expensive.

In Table 2.12 we use row percent, because we expect that cardiovascular workouts may affect the physical fitness, not vice versa.

It is common, like here, to supplement with the column at the right, showing the number of individuals in each group.

In this table, one might suggest a trend that cardiovascular workouts do have an impact on the physical fitness. In Chap. 5 we return to this example.

Chapter 3
Description of Data

In Chap. 2, we discussed two different types of data: quantitative data and qualitative data.

– *Quantitative data*: These data are used for calculations and for defining the axes in graphs.
– *Qualitative data*: These data correspond to groupings of the sample, in tables or graphs.

In this chapter, we mainly discuss quantitative data. We present some important *sample statistics* (*); these are "key numbers" used to describe quantitative data from a sample, for example average and variance.

At the end of the chapter, we discuss in more detail various types of data and the statistics relevant for each type.

An important tool to describe quantitative data is the histogram, which gives an immediate visual impression of what is characteristic for data.

A histogram should be based on a sample of adequate size. In Fig. 3.1 we can see that a histogram based on a small sample is very "rough". As the sample size grows, one can gradually visualize the distribution of the data values.

However, it is often desirable to summarize the distribution by some "key numbers," i.e., the numbers that describe various characteristics of the distribution. These key numbers are the main topic of this chapter.

3.1 Systematic and Random Variation

Statistics is about describing the variation in data. It is important to distinguish between two different sources of variation (Fig. 3.2).

– *Systematic variation*: *The center of data*
– *Random variation*: *The spread of data*
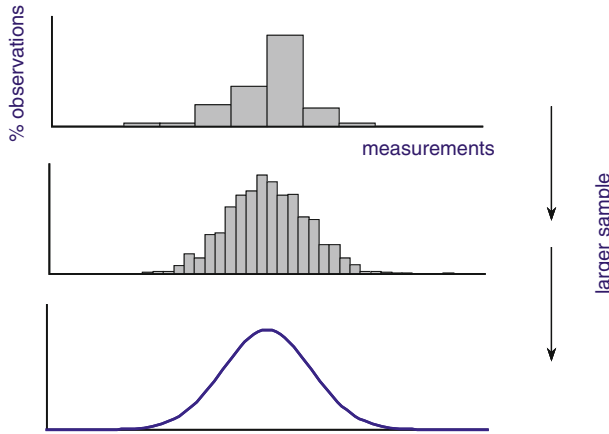
Fig. 3.1 Histogram for various sample sizes
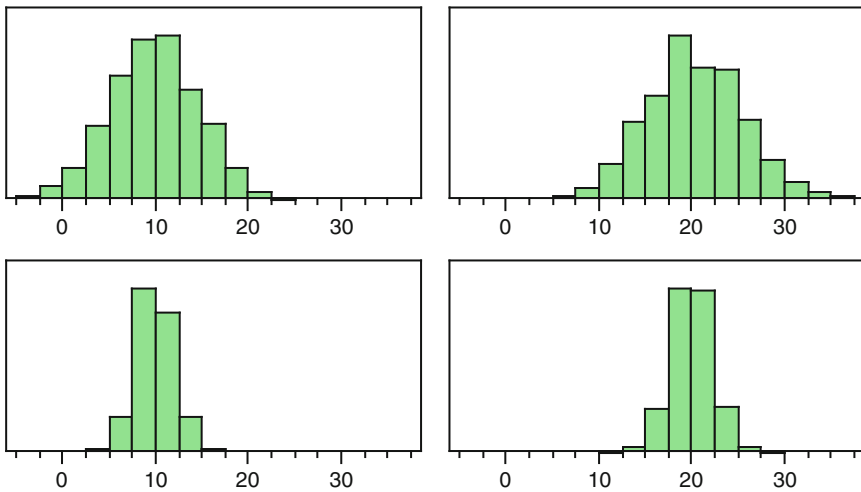
Fig. 3.2 Center and spread

Figure 3.2 illustrates the two different sources of variation. The figure contains four histograms with measurements of daily average temperatures (°C) at four different locations.

The top two distributions are characterized by having a *large spread*, i.e., *highly variable climate*. The two lower distributions have a *small spread*, i.e., a significantly more *stable climate*.

On the other hand, the two distributions to the left have a center of approx. 10°, and the two distributions to the right have a center of approx. 20°. This means that the two distributions to the left represent a *cold climate*, whereas the two distributions to the right represent a *warm climate*.

In statistical terminology, we often use the term *location* (*) rather than *center*. Also, we often use the term *dispersion* (*) rather than *spread*.

In this chapter, we present the most important sample statistics used to characterize the distribution of data:

– *Measures of location* (*center*)
– *Measures of dispersion* (*spread*)

3.2 Measures of Location

3.2.1 Average

3.2.1.1 Description

The *average* (*) is a measure of the center in the distribution of data values.
   The average is calculated as *the sum of all the data values divided by their number*.

As a symbol of the average, we often use the term $\bar{x}$, which is read as "*x-bar*".

Note: Often people use the word *mean* (*) rather than average.
   Strictly speaking, one should use the word *average of a sample*, and use the word *mean of a population*. Many use the two terms interchangeably.

The average is highly influenced by "extreme values" (i.e., very large or very small values). If for example, there are many very large data values, the average becomes "excessively" large.
   In a sample of income data, a single dollar billionaire can "destroy" the whole picture, "counting" of course just as much as several hundred "ordinary" people...
   In case of many extreme data values, an alternative is to use the *median* (*) rather than the average. See later.

3.2.1.2 Example

Data are the numbers 3, 5, 6, 4. The number of data values is obviously 4.
   The sum of all data values is $3 + 5 + 6 + 4 = 18$.

The average is:

$$\bar{x} = \frac{3 + 5 + 6 + 4}{4} = \frac{18}{4} = 4.5.$$

3.2.1.3 Spreadsheets

Most spreadsheets have many built-in *statistical functions*, including the average. This applies for example, to Microsoft Excel and Open Office Calc. Open Office is free!

To calculate the average, use the statistical function AVERAGE. See an example later.

3.2.1.4 Calculation Formula

With n data values x_1 up to x_n, the average can be calculated using this general formula:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}.$$

Here $x_1 + x_2 + \cdots + x_n$ is the sum of all data values.

This formula (and others) can be written shorter by using the "*sum*" *symbol* Σ, which corresponds to the "sum" button in a spreadsheet. We write Σx_i as a shorter way of writing the sum of all data values $x_1 + x_2 + \cdots + x_n$.

Then the formula for calculating the average can be written in a compact form:

$$\bar{x} = \frac{\sum x_i}{n}.$$

3.2.2 Median

3.2.2.1 Description

The median (*) is the data value "in the middle", i.e., a number that *divides the data values into two parts with an equal number of values.*

The median can be found by first sorting the data values in ascending order.

– In case of an *odd number* of data values, the median is *the middle value*.
– In case of an *even number* of data values, there is no single data value dividing data values into two equally large parts; we then define the median as the *average of the two middle values.*

Table 3.1 Median

No.	1	2	3	4	5	6	7	8	**9**	10	11	12	13	14	15	16	17
Value	12	12	13	13	13	13	14	14	**14**	14	14	15	15	15	16	17	17

The median is *not as sensitive to extreme values* as the average! Often, you will therefore complement the average with the median.

3.2.2.2 Example: Even Number of Data Values

Data are the numbers 3, 5, 6, 4.

First we sort the data in ascending order: 3, 4, 5, 6.

As there is an even number of data values, we take the average of the two middle values (in the sorted data), which are respectively 4 and 5.

Therefore, the median is:

$$M = \frac{4 + 5}{2} = 4.5.$$

In this example, the median and the average are the same. However, they need not be. In particular, they will be different in "skewed" distributions, see later.

Here we have used the symbol M for the median, as is common in many textbooks.

3.2.2.3 Example: Odd Number of Data Values

We consider the example "*Fitness Club*"; data values are the age of the boys. Data values (17 in total) are shown here, sorted in ascending order (see data in Chap. 9) (Table 3.1).

As there are 17 data values, the middle value is no. 9, i.e., the data value 14 as highlighted in bold. Out of the 17 data values, there are 8 data values on both sides of data value no. 9. Therefore, the median is 14.

3.2.2.4 Spreadsheets

The statistical function is called MEDIAN; see an example later.

3.2.3 Mode

3.2.3.1 Description

The mode (*) is simply the *most frequent data value!*

All that is required is a frequency count for each data value! No calculations are needed!

If you have a complete list of all data values and their frequency, you can immediately find the mode. In this situation it is easier to find the mode than calculating the average or the median. This is the only advantage of the mode compared with the average and the median!

In contrast, the mode has one very big disadvantage: If you have many different data values (perhaps with multiple digits), there will often be *only one occurrence of each value*. Should there by chance be 2 (or maybe even 3) occurrences of one single value, it may be just a statistical coincidence. The mode in this case is not a meaningful concept. Therefore, *the mode is not used very often in practice*!

3.2.3.2   Example

We use the example again "Fitness Club," age of the boys.

Here are the various ages and their frequency (Table 3.2).

We observe that the highest frequency is that of the age 14 years, which is thus the mode.

In a sample of 17 randomly selected people of all ages (kids as well as adults), data values from 0 to perhaps over 90 could occur. In this case, the mode will not be an informative number. If by chance there are two persons having the same age, we would merely consider this an uninteresting coincidence!

3.2.3.3   Spreadsheets

The function is called MODE. See an example later.

3.2.4   Choosing a Measure of Location

If the distribution is *symmetrical*, i.e., there are an equal number of large and small data values (see Fig. 3.3), you will usually use the average, but the median provides virtually the same result.

Table 3.2 Mode

Age	12	13	**14**	15	16	17
Frequency	2	4	**5**	3	1	2
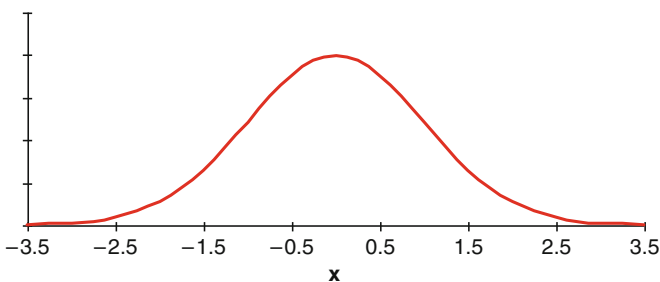
Fig. 3.3 Symmetrical distribution
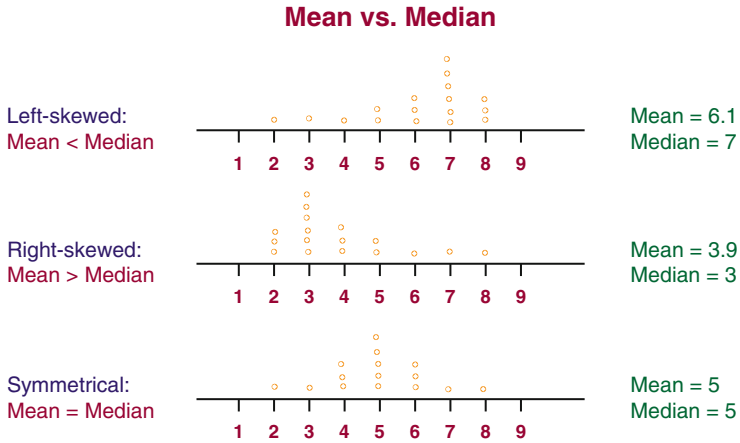
Fig. 3.4 Mean vs. median

In a *skewed* (i.e., nonsymmetrical) distribution, the average and the median are not identical:

– For a *right-skewed* distribution, i.e., a distribution with many large data values, *the average is larger than the median*.
– For a *left-skewed* distribution, i.e., a distribution with many small data values, *the average is smaller than the median*.

See Fig. 3.4.

3.2.4.1 Example: What Is the Average Salary?

Many economic and administrative data follow a right-skewed distribution, i.e., there are many large data values. An example might be the salary for a group of employees.

Figure 3.5 shows such a distribution. Also shown are the values of the mode, median and average salary.

We see that *in a right-skewed distribution, the mode is smaller than the median, and the median is smaller than the average!*

Now can we also see why most people find it so depressing to read a salary statistic? Most people have indeed a salary around the mode, but at the same time compare themselves to the average!
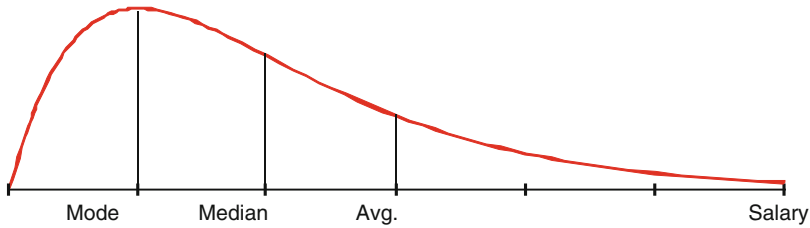
Fig. 3.5 Right-skewed distribution

Whether the actual data will follow a distribution like the one shown in the figure depends on how homogeneous the sample is.

The more the sample is divided into homogeneous groups, the less skewed the distribution will be.

If you have divided the data in many homogeneous groups, the average and the median (shown in most salary statistics) will be close to each other.

3.3 Measures of Dispersion

In this chapter, we have reviewed the main *measures of location*, i.e., the center of a distribution. Now we look at various *measures of dispersion*, i.e., the spread of a distribution.

3.3.1 Range

3.3.1.1 Description

The *range* (*) is simply the *width of the interval of the data values*.

The range is used if you need a measure of dispersion that is *easy to calculate* and *understand*!

The range is calculated as *the difference between the largest and smallest data values*, x_{max} and x_{min}, respectively:

$$R = x_{max} - x_{min}.$$

The range (denoted by the letter R) gives a numerical expression of the spread of data and of course has the advantage of being easy to calculate. The main advantage of range, however, is that it is easy to understand!

The range depends largely on the number of data values. If there are many data values, the range gets larger, because there are more small or large ("extreme") values.

The range is therefore used mainly for small samples.

A typical application is statistical quality control in the construction of *control charts*, where the number of data values in a sample is often in the order of magnitude 5. The exact purpose of the control chart is to quickly detect any changes in a production process by separating systematic and random variation!

3.3.1.2 Example

Data are the numbers 3, 5, 6, 4.

First we sort the data values in ascending order: 3, 4, 5, 6.

The smallest data value is $x_{min} = 3$, and the largest data value is $x_{max} = 6$.

The range then becomes:

$$R = x_{max} - x_{min} = 6 - 3 = 3.$$

3.3.1.3 Spreadsheets

The range does not exist as a function of its own in spreadsheets. However, there are functions for the *maximum* value (MAX) and the *minimum* value (MIN). See an example later.

3.3.2 Variance and Standard Deviation

3.3.2.1 Description

The most common *measure of dispersion* (spread) is the *standard deviation* (*), which can be interpreted as "*the average distance*" *between the data values and the average*. The larger the standard deviation, the larger is the spread of the distribution.

However, the standard deviation is not calculated as an ordinary average distance. Before explaining exactly how the standard deviation is calculated, we have to explain another concept:

The variance (*) is *the average of the squared distances between the data values and the average*.

The variance is not measured in the same units as the original data values, but in square units (e.g., square meters, if the data values are meters). Often, the term V is used for the variance.

> *The standard deviation (*) is the square root of the variance.*

The standard deviation is measured in the same units as the data values, e.g., meters. Often, the term s is used for the standard deviation (Fig. 3.6).

3.3.2.2 Example

Data are the numbers 3, 5, 6, 4.

The average of these figures was previously calculated to be 4.5.

The variance of these figures is:

$$V = \frac{(3-4.5)^2 + (5-4.5)^2 + (6-4.5)^2 + (4-4.5)^2}{n-1} = \frac{5}{3} = 1.67.$$

The standard deviation is:

$$s = \sqrt{\frac{(3-4.5)^2 + (5-4.5)^2 + (6-4.5)^2 + (4-4.5)^2}{n-1}} = \sqrt{\frac{5}{3}} = \sqrt{1.67} = 1.29.$$

Note: In the above formula, we divide by $n-1$, not by n. This is, for technical reasons and in practice, an unimportant detail, unless the sample is very small.

If $n = 1$ (i.e., only one data value), *the variance and the standard deviation cannot be calculated!* This is consistent with the fact that in this situation we cannot talk about the spread of the distribution!

Many spreadsheets and calculators have built-in functions to calculate the variance and the standard deviation. Often, there are *two versions of these functions*,
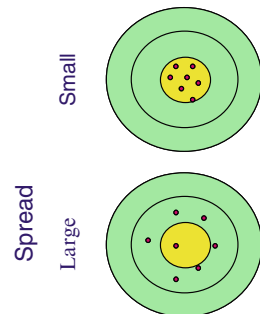
Fig. 3.6 Small and large spread

where division by n respectively division by $n - 1$ is used. This has been very confusing to many people.

Often, it is stated that the formulae with divisor n are used when calculating the variance or standard deviation of a population (as opposed to a sample). Most statisticians, however, *always use the formulae with divisor $n - 1$*!

3.3.2.3 Spreadsheets

For calculation of the variance we use the function VAR. To calculate the standard deviation, we use the function STDEV. See an example later.

3.3.2.4 Calculation Formulae

The variance is calculated using this general formula:

$$V = \frac{\Sigma(x_i - \bar{x})^2}{n - 1} = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n - 1}.$$

The standard deviation is the square root of the variance, i.e.,

$$s = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n - 1}}.$$

There is an *alternative calculation formula* for the standard deviation, see the text frame. It is particularly useful in cases where the calculations must be done on *a calculator without statistical functions*.

Technical note: alternative calculation formula for the standard deviation.

This formula for the standard deviation is useful when the calculations must be done on a calculator without statistical functions:

$$s = \sqrt{\frac{\Sigma(x_i)^2 - \Sigma(x_i)^2/n}{n - 1}}.$$

You need both the *sum of all data values* and the *sum of the squares*.
The principle is illustrated in Table 3.3.
We calculate the standard deviation as follows: $s = \sqrt{\dfrac{86 - 18^2/4}{4 - 1}} = \sqrt{1.67} = 1.29$.

Table 3.3 Sum and sum of squares

Data value no.	x	x^2
1	3	9
2	5	25
3	6	36
4	4	16
Sum	**18**	**86**

3.3.3 Interquartile Range

3.3.3.1 Description

Another important measure of dispersion is the *Interquartile Range* (*IQR*), explained below.

When the median is calculated, you can further divide the two parts of the data values into two parts each.

Thus the entire set of data values is divided into four parts, with (roughly) the same number of data values. The new points of division are called the *quartiles* (*).

The *difference between the quartiles* is called the *interquartile range*, often denoted by the abbreviation *IQR*. The interpretation of *IQR* is that it is *the length of the interval with the* "*middle* 50%" *of the data values* and it is often used when you use the median as a measure of location.

When we find the quartiles, we sort the data values in ascending order, as when calculating the median.

> The *lower quartile* (*or first quartile*) Q1 is a number that divides the data values into two parts so that *one-fourth of the data values are smaller than the lower quartile*, and three-fourths of the data values are larger.

Often, the median is considered to be the middle or second quartile and is sometimes denoted by Q2.

> The *upper quartile* (*or third quartile*) Q3 is a figure that divides the data values into two parts so that *three-fourths of the data values are smaller than the upper quartile* and one-fourth of the data values are larger.

> The *interquartile range* (*) *IQR* is then the difference between the upper and lower quartiles:
>
> $$IQR = Q3 - Q1.$$

Table 3.4 Quartiles

No.	1	2	3	**4**	**5**	6	7	8	**9**	10	11	12	**13**	**14**	15	16	17
Value	12	12	13	**13**	**13**	13	14	14	**14**	14	14	15	**15**	**15**	16	17	17

A few books, however, use half of the difference as the definition of the IQR.

3.3.3.2 Example

We consider again the example *Fitness Club*, age of the boys. Data values (17 in total) are shown in Table 3.4 in sorted order.

We have previously found the median to be the data value no. 9, i.e., the median is 14. This data value divides the data values in two equally large parts. Each of these two parts is now subdivided again:

The first half consists of data values no. 1–8. As there is an even number of data values, data values no. 4–5 form the point of division. Both data values are 13, i.e., the lower quartile Q1 equals 13.

The other half consists of data values no. 10–17. As there is an even number of data values, data values no. 13–14 form the point of division. Both data values are 15, i.e., the upper quartile Q3 equals 15.

Then the interquartile range is IQR $= 15 - 13 = 2$.

3.3.3.3 Spreadsheets

The quartiles are calculated using the function QUARTILE. See an example later.

Then the IQR is calculated by subtracting the quartiles from each other.

3.3.4 Choosing a Measure of Dispersion

The choice of a measure of dispersion will often reflect the measure of location that we have chosen.

– *If the distribution is symmetrical*, we often use the *average as a measure of location*. It is natural in this case to supplement with the *standard deviation as a measure of dispersion*. The standard deviation is, after all, based on the average.
– *If the distribution is skewed*, you will often use the *median as a measure of location*. In this case it is natural to supplement with the *IQR as a measure of dispersion*. This is also natural, because the median might be perceived as the "middle quartile".

3.3.5 Relative Spread (Dispersion)

3.3.5.1 Description

When comparing samples from several time periods (e.g., several years), *the average will often increase with time*; this is valid for many financial and administrative data during periods of growth, but also for many human and biological populations.

Often, *the spread increases with an increasing average*. Therefore, the spread is not interesting in itself. Instead, *the relative spread is more interesting*, i.e., the standard deviation divided by the average.

> As a measure of the relative spread (dispersion), we use the *coefficient of variation* (*), often denoted CV, defined as *the standard deviation as a percentage of average*:
>
> $$CV = \frac{s}{\bar{x}} \times 100\%.$$

Sometimes, this is also known as *the relative standard deviation* and denoted RSD.

Note: If the data values for instance are temperatures measured in °C, you cannot use the coefficient of variation! The average temperature (used as the denominator) may be 0°C or even negative! *If the CV has to be meaningful, there must therefore be a lower limit of* 0, i.e., negative values must not occur!

3.3.5.2 Example

Data are the numbers 3, 5, 6, 4.

We have previously found the average as $\bar{x} = 4.5$ and the standard deviation as $s = 1.29$.

This gives us the coefficient of variation:

$$CV = \frac{s}{\bar{x}} \times 100\% = \frac{1.29}{4.5} \times 100\% = 29\%.$$

3.3.5.3 Spreadsheets

The coefficient of variation does not exist as an independent function, but can be calculated manually using the standard deviation and the average.

3.4 Example: Statistical Functions in Spreadsheets

In most spreadsheets such as Microsoft Excel and Open Office Calc, *there is a wide range of statistical functions*. The most important functions are listed in Chap. 9.

Data values are once again the age of the boys from the *Fitness Club* survey, a total of 17 data values, which are entered in the first row of a spreadsheet in cells A1 up to Q1.

The data area should be specified when using all these statistical functions.

Figure 3.7 shows how all the above statistics are calculated using the statistical functions. Here the columns A to C in the spreadsheet are only shown, but the data values are located in the whole area A1:Q1.

The average is calculated by using the AVERAGE function, the standard deviation by using the STDEV function.

The coefficient of variation does not exist as a function, but it is calculated by dividing the standard deviation by the average, the result being displayed as a percentage (click on the *% key* in the spreadsheet).

The median is calculated by the function MEDIAN and the mode by using the function MODE.

Quartiles are calculated using the function QUARTILE. This function has an additional parameter to indicate which quartile we are calculating.

The value of this parameter is 3, when calculating the upper (third) quartile. It is 1 when calculating the lower (first) quartile. The median can be calculated as the second quartile by using the value 2, but of course the median also has its own function.

The IQR is calculated by subtracting the upper and lower quartiles.

	A	B	C
1	12	12	13
2	Value	Spreadsheet formula	Statistic
3	14.18	=AVERAGE(A1:Q1)	Average
4	1.510	=STDEV(A1:Q1)	Standard deviation
5	10.6%	=A4/A3	Coefficient of variation
6	14	=MEDIAN(A1:Q1)	Median
7	14	=MODE(A1:Q1)	Mode
8	15	=QUARTILE(A1:Q1;3)	Upper quartile
9	13	=QUARTILE(A1:Q1;1)	Lower quartile
10	2	=A8-A9	Interquartile range
11	17	=MAX(A1:Q1)	Maximum value
12	12	=MIN(A1:Q1)	Minimum value
13	5	=A11-A12	Range

Fig. 3.7 Statistical functions in spreadsheets

Age		*Height*		*Weight*	
Mean	14.03	Mean	157.10	Mean	53.17
Standard Error	0.27	Standard Error	4.03	Standard Error	2.50
Median	14	Median	159.5	Median	49
Mode	14	Mode	166	Mode	49
Standard Deviation	1.50	Standard Deviation	22.06	Standard Deviation	13.71
Sample Variance	2.24	Sample Variance	486.78	Sample Variance	187.87
Kurtosis	−0.31	Kurtosis	−0.21	Kurtosis	−0.17
Skewness	0.53	Skewness	−0.43	Skewness	0.70
Range	5	Range	86	Range	51
Minimum	12	Minimum	112	Minimum	32
Maximum	17	Maximum	198	Maximum	83
Sum	421	Sum	4713	Sum	1595
Count	30	Count	30	Count	30
Confidence Level(95.0%)	0.56	Confidence Level(95.0%)	8.24	Confidence Level(95.0%)	5.12

Fig. 3.8 Output from Data Analysis menu

The range does not exist as a function. On the other hand there are functions for the maximum (MAX) and minimum (MIN) data values. The range can then be calculated by subtraction.

The main measures for location and dispersion can also be calculated in Microsoft Excel using the *Add-in* menu *Data Analysis*, which has a menu item *Descriptive statistics*. The menu is not available in Open Office Calc; here you need the statistical functions.

As an example we use the data for age, height and weight of all 30 kids from the Fitness Club survey. The result of applying the menu *Data analysis* is shown in Fig. 3.8.

The concepts "Standard Error," "Kurtosis", "Skewness" and "Confidence Level" are explained in Chap. 4.

In this chapter, we have discussed the most important measures of location (center) and dispersion (spread).

The choice of measures depends on the distribution of the data values: Is it symmetrical or skewed?

In the next chapter, we primarily deal with symmetrical distributions. First, we give a more detailed discussion of various types of data, as well as the measures (of location and dispersion) to use for different types of data.

3.5   Data Type and Descriptive Statistics

If you work a lot with questionnaire data, you may want to read this section; otherwise, it can be skipped.

3.5.1   Data Types

The main types of data are:

Quantitative data are data such as weight, height, temperature, amounts of dollars, etc. These data are used for calculations and for defining the axes in a graph. They are subdivided into two types:

–   *Ratio data*: Here ratios are well defined; there is a natural zero, negative values do not occur. This applies to most physical measurements, e.g., length; a table might be twice as long as another.
–   *Interval data*: Here differences are well defined, and negative numbers may occur. An example is temperature measured in °C; an increase in temperature of 5° makes sense.

Qualitative data are data such as sex, education, occupation, etc. These data correspond to groupings of the sample (or the population), in tables or graphs. They are subdivided into three types:

–   *Ordinal data*: A number of categories with a natural ordering. This applies to many questionnaire data on, e.g., attitudes (on a scale of 1–5,...), grades in schools, etc.
–   *Nominal data*: A number of named categories. For instance, various types of fruit: apples, pears and bananas, etc.
–   *Alternative* (*binary*) *data*: Two categories (i.e., two alternatives). These data can optionally be viewed as ordinal or nominal data. Examples are: agree or disagree, good or bad, defective or non-defective, etc. We deal more with alternative data in Chap. 5

Integer data (counting data, i.e., the data values are 0, 1, 2, 3, etc.) is yet another type of data. They are really in-between quantitative (ratio) data and qualitative (ordinal) data.

If the counts result in very large numbers (such as strokes of lightning during a powerful thunderstorm), it nevertheless makes sense to consider integer data as quantitative (ratio) data.

3.5.2   Descriptive Statistics and Type of Data

Throughout this chapter, we have assumed that our data are quantitative. There it makes sense to calculate all statistics, except that *the coefficient of variation requires ratio data*.

For *integer data* we can calculate all the statistics, as for quantitative ratio data. However, *the actual value of the average might not be a real data value*! For instance, it is not possible that there can be 2.3 persons in a household. . .

However, an average of 2.3 persons in a household is actually very informative! It tells us that two persons in a household will be the most frequent situation;

Table 3.5 Descriptive statistics vs. data type

Sample statistic	Nominal data	Ordinal data	Integer data	Interval data	Ratio data
Mode	Yes	Yes	Yes	Yes	Yes
Median	No	Yes	Yes	Yes	Yes
Average	No	(No)	(Yes)	Yes	Yes
Quartiles	No	Yes	Yes	Yes	Yes
Interquartile range	No	(No)	Yes	Yes	Yes
Range	No	(No)	Yes	Yes	Yes
Standard deviation	No	No	(Yes)	Yes	Yes
Coefficient of variation	No	No	(Yes)	No	Yes

another frequent situation will be households with three people. With caution, we can also use the standard deviation and coefficient of variation as measures of spread.

For *ordinal data,* the situation is a bit more complicated! At least, the mode, the median, and the quartiles are well defined; here we are by definition using the fact that data can be ordered.

For typical *questionnaire data on a scale*, e.g., from 1 to 5, many people will also calculate the average. This is in actual fact not meaningful; however, as for integer data, the average may be informative, if used with caution.

Similarly, for this type of data we can calculate the range and IQR. If these statistics are to be meaningful, the difference between two numbers must be meaningful. This condition is usually not fulfilled. However, both statistics can be used as rough measures of spread, if used with caution.

For *nominal data*, there is only statistic that makes sense: This is *the mode as a measure of location*. The most frequent data value is well defined here. The concept of a spread does not make sense.

Table 3.5 shows the statistics that can be used in connection with different data types.

Chapter 4
The Normal Distribution

In Chap. 3, we explained how to calculate descriptive statistics such as the average and standard deviation of a sample. Now we will see what these measures can be used for.

Imagine that you buy a bag with 500 g of coffee. You are curious and empty the contents onto a weight to check whether the bag actually contains 500 g. If you have a very precise weight, you will hardly expect that the content weighs exactly 500 g and you are probably not surprised if it is a little more or less.

If you are repeating the experiment with many bags, you might expect that the weight of a bag will be very *close to* 500 g *on average*. You may also expect that there will *not be too much spread*. For instance, you do not expect to get less than 450 or 550 g, not even once. The weight of a bag can perhaps vary around 490–510 g, but it will rarely be more than 510 g or less than 490 g.

This variation can be described by a statistical distribution.

The most important statistical distribution is *the normal distribution* (*). Several statistical techniques require that data "follow" (i.e., can be described by) a normal distribution. If data do not follow a normal distribution, it becomes more difficult to analyze the data.

This chapter examines some important properties of the normal distribution. We also see how the fact that data follow a normal distribution can actually be verified. Finally, we see how to estimate the *statistical uncertainty* (*) of a sample average.

4.1 Characteristics of the Normal Distribution

The normal distribution curve is a symmetrical, "bell-shaped" curve similar to a histogram – in this case showing the weight of a very large number of coffee bags. It has been proven in practice that the normal distribution often gives a good description of many types of *measurement data*, such as weight, height, etc. But the normal distribution is very important also for *economic and administrative data*.

A normal distribution is completely described by its mean (average) and standard deviation.

The normal distribution in the example above describes the weight of all the coffee bags manufactured by the factory. Since we do not know the mean and standard deviation, they are often written in Greek letters:

- *Mean*: μ (read "mju") representing the "center"
- *Standard deviation*: σ (read "sigma") representing the "spread"

In Fig. 4.1, we see two normal distributions with small spread ($\sigma = 1$, above) and two normal distributions with large spread ($\sigma = 2$, below). The two distributions in each group have different means, $\mu = 10$ and $\mu = 24$.

Fig. 4.2 *shows the interpretation of the standard deviation in a normal distribution.* Here is shown a normal distribution representing the histogram of a population or a very large sample.

We observe that:

- 68% of the data values are in an interval around mean $\pm$ standard deviation
- 95% of the data values are in an interval around mean $\pm$ 2 standard deviations
- 99.7% of the data values are in an interval around mean $\pm$ 3 standard deviations

These percentages are unique to the normal distribution!

In a way, the normal distribution is "thinking" as if 0 corresponds to the mean and 1 unit corresponds to the standard deviation.
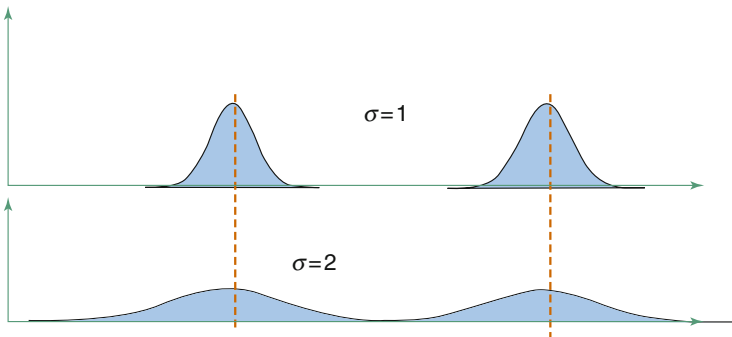
Fig. 4.1 Mean and spread

Fig. 4.2 Probabilities in the
normal distribution
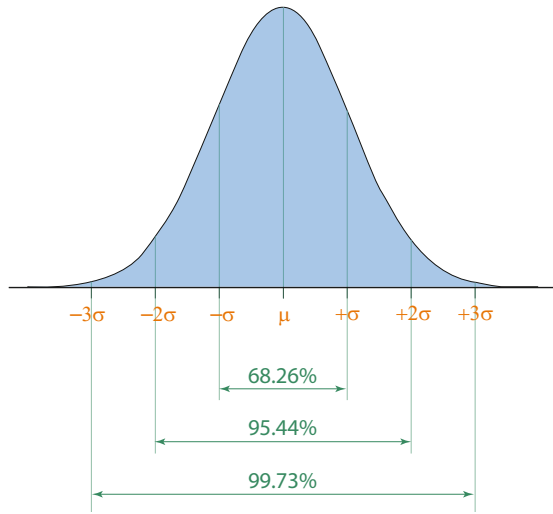
If *X* follows a normal distribution with mean μ and standard deviation σ, then

$$\frac{X - \mu}{\sigma},$$

follows a normal distribution with mean 0 and standard deviation 1.

There exists in a way, only one normal distribution!

The normal distribution with *mean* 0 and *standard deviation* 1 is therefore called the *standardized normal distribution*. See an example in the section "Calculations in the normal distribution".

4.2 Density Function and Distribution Function

In practice, it is therefore *areas under normal distribution curve*, which are interesting because they can be *interpreted as probabilities*. Therefore, we are usually interested in the curve showing areas under the normal distribution curve. The relationship between these two graphs is shown below.

The bell-shaped curve (Fig. 4.3) is called the *density function* (*), while *the curve showing areas* (probabilities) is called the *distribution function* (*).

Fig. 4.3 Density and
distribution function
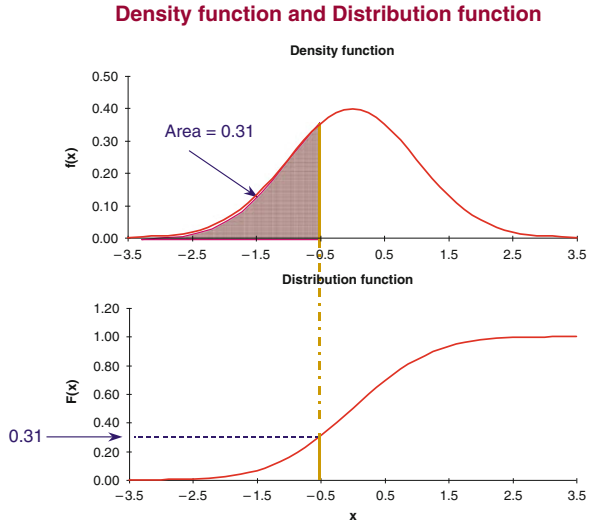
The distribution function is often written using the letter F. We can interpret the distribution function by noticing that $F(x)$ is the probability of observing data values up to and including x.

In real-world problems, we almost always need the distribution function. You only need the density function for constructing illustrations in a book.

4.3 Fractiles

Let us assume that the weight of the coffee in a bag of coffee follows a normal distribution with mean 500 g and standard deviation 5 g. We want to answer questions such as the following:

1. How many coffee bags are weighing at most 495 g?
   We use the distribution function: Find the value 495 on the x-axis and move vertically up to $F(495)$, i.e., find the corresponding value on the y-axis. This is precisely the probability of data values up to 495 g. On the graph (Fig. 4.4) below, we see that it is roughly 0.16, equivalent to 16%.
2. Which weight value separates the lightest 80% of the coffee bags from the rest?
   We now use the distribution function the opposite way: Find the value 0.80 (equivalent to 80%) on the y-axis (i.e., a probability of 80%) and move horizontally to the distribution curve, then find the corresponding value of the x-axis. On the graph below, we can see that it is roughly 504 g (Fig. 4.5).

We therefore need to use the normal distribution function in "both ways".

When we are using the distribution function in the "reverse" way, for example, from $0.80 = 80\%$ on the y-axis to a value on the x-axis, we are talking

Fig. 4.4 Distribution function
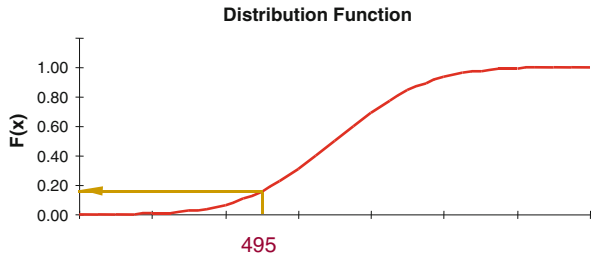
Fig. 4.5 Fractiles
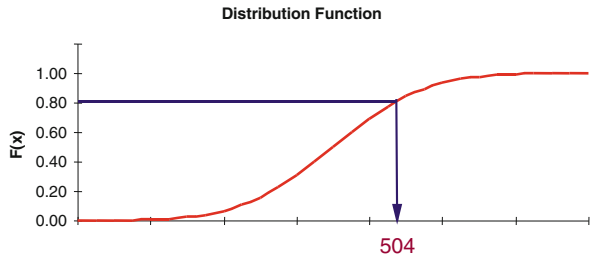
about finding a *fractile* (*) (also called a *quantile* or a *percentile*) in the distribution. The figure above shows the finding of 80% fractile in a normal distribution.

We have actually seen in Chap. 3 the most important fractiles: The quartiles are the 25 and 75% fractiles, and the median is the 50% fractile.

The fractiles corresponding to 10, 20, 30%, etc., are called the *deciles*.

4.4   Calculations in the Normal Distribution

We will now show how to do simple calculations in the normal distribution by using a table of the normal distribution.

The most important fractiles in the normal distribution are found in the table at the end of the book. More detailed tables on the normal distribution can be found in many books.

Example. Let us assume that the weight, X, of (the coffee in) a bag of coffee follows a normal distribution with a *mean* $\mu = 500$ g and *standard deviation* $\sigma = 5$ g.

As mentioned earlier:

$$\frac{X - \mu}{\sigma},$$

follows a normal distribution with mean 0 and standard deviation 1.

We say that we *standardize X* in this calculation. The standardized normal distribution function is tabulated in Chap. 9.

Now we can answer questions such as:

1. What is the probability that a random coffee bag weighs at most 510 g?
   We standardize the 510 g and obtain:

$$\frac{510 - \mu}{\sigma} = \frac{510 - 500}{5} = 2.$$

   By looking in a table of the standardized normal distribution, we find that the probability of a data value ≤ 2 is $0.0977 = 97.7\%$. This is then the probability that a random coffee bag weighs, at most, 510 g.

2. What is the 95% fractile in the distribution?
   In a table of the standardized normal distribution (at the end of the book), you find the 95% fractile to be 1.65. This represents a fractile in the distribution of

$$\frac{X - 500}{5},$$

   where X is the weight of a randomly selected coffee bag.
   By solving the equation

$$\frac{X - 500}{5} = 1.65,$$

   we obtain $X = 500 \times 1.65 = 508.25$.
   This means that the 95% fractile in the distribution of a randomly chosen coffee bag is 508.25.
   In other words, *the probability that a random coffee bag weighs less than 508.25 g is exactly* 95%.

4.5   The Normal Distribution and Spreadsheets

If you do not use spreadsheets, you can skip this section.

There are in Microsoft Excel and OpenOffice Calc two important functions for the normal distribution.

– NORMDIST: Provides the distribution function or density function for a normal distribution.
– NORMINV: Gives fractiles in a normal distribution.

4.5.1 NORMDIST (X; Mean; Stdev; Cumulative)

It is only if you have to make figures with the "bell-shaped" curve, that you need the density function. Therefore, you should *almost always use Cumulative* $= 1$ (Table 4.1).

The distribution function for the standardized normal distribution (with mean 0 and standard deviation 1) can also be obtained by using the function NORMS-DIST. This function only has one parameter: *X*.

4.5.2 NORMINV (Probability; Mean; Stdev)

For the standardized normal distribution (Table 4.2) (with mean 0 and standard deviation 1), you can use the function NORMSINV. This function only has one parameter: *probability*.

4.5.3 Example

Let us assume that the weight, *X*, of (the coffee in) a bag of coffee follows a normal distribution with a *mean* $\mu = 500$ g and *standard deviation* $\sigma = 5$ g.

1. What is the probability that a random coffee bag weighs, at most, 490 g?
   We use NORMDIST(490; 500; 5; 1) and get the result $0.023 = 2.3\%$.
2. What is the probability that a random coffee bag weighs at most 510 g?
   Similarly, we find the probability that a random coffee bag weighs at most 510 g:
   We use NORMDIST(510; 500; 5; 1) and get the result $0.977 = 97.7\%$.
3. What is the 5% fractile in the distribution?
   Remember that 5% equals 0.05.
   In the spreadsheet, we do not use percentages when writing probabilities.
   We therefore use NORMINV(0.05; 500; 5) and get the result 491.8.

Table 4.1 NORMDIST function

X	The number of which the value of the distribution function (or density function) is desired
Mean	The *mean* of a normal distribution
Stdev	The *standard deviation* of a normal distribution
Cumulative	*Cumulative* $= 0$ calculates the *density function*
	Cumulative $= 1$ calculates the *distribution function*

Table 4.2 NORMINV function

Probability	*The probability* for which a fractile in the normal distribution is wanted
Mean	The *mean* of a normal distribution
Stdev	The *standard deviation* of a normal distribution

	A	B	C	D
1	COFFEE BAGS EXAMPLE:			**Formula in spreadsheet**
2				
3	**Normal distribution**			
4	*Mean*	500		
5	*Standard deviation*	5		
6				
7	*Probability of weight ≤ x*		**Probability**	
8	x (490 g)	490	0.023	=NORMDIST(B8;B4;B5;1)
9	x (510 g)	510	0.977	=NORMDIST(B9;B4;B5;1)
10				
11	*In this normal distribution the fractiles are*		**Fractile**	
12	*Probability (5%)*	0.05	491.8	=NORMINV(B12;B4;B5)
13	*Probability (80%)*	0.80	504.2	=NORMINV(B13;B4;B5)

Fig. 4.6 Example in spreadsheet

This means that the probability that the weight of a random bag of coffee is ≤ 491.8 g is precisely 0.05, which is equivalent to 5%.

In other words, there is a 95% chance that a random coffee bag weighs more than 491.8 g.

4. What is the 80% fractile in the distribution?

Similarly, we find the 80% fractile as NORMINV(0.80; 500; 5) and get the result 504.2.

This means that the probability that the weight of a random bag of coffee is ≤504.2 g is precisely 0.80, which is equivalent to 80%.

In other words, there is a 20% chance that a random coffee bag weighs more than 504.2 g.

The spreadsheet shows how to make the calculations in a spreadsheet (Fig. 4.6).

4.6   Testing for the Normal Distribution

We have studied some key characteristics of the normal distribution, its density function and distribution function, and calculations in the normal distribution. In other words, the assumption has been that the data actually are following a normal distribution. There are several ways to check this. This is the topic of this section and the following section.

4.6.1 Simple Methods

1. *The histogram*

   It is always a good idea to study the histogram.

   This must show a symmetrical, "bell-shaped" appearance.

   Depending on the number of data values, the histogram can be more or less irregular; we discuss this later in this chapter.

2. *The average = the median*

   If data can be described by a normal distribution, the *average and median must be nearly identical*, because the normal distribution is symmetrical.

   This is very simple to check.

3. *Interquartile range larger than the standard deviation*

   In the normal distribution, the *interquartile range* (i.e., the difference between the upper and lower quartile) is somewhat *larger than the standard deviation*; actually, it is around $1.35\times$ the standard deviation, i.e.,

$$IQR = 1.35 \times s.$$

   This can be explained by the standardized normal distribution with mean 0 and standard deviation 1:

   Here, the upper quartile is 0.674 (see table in Chap. 9, 75% fractile). Because the normal distribution is symmetrical, the lower quartile is -0.674. The interquartile range, i.e., the distance between the quartiles, is therefore $2 \times 0.674 =$ approx. 1.35.

   By comparison, the standard deviation is precisely 1 in the standardized normal distribution.

4. *Number of data values in symmetric intervals around the mean*

   We have seen that around *68% of the data values* in a normal distribution are in an interval around *mean ± standard deviation*.

   If the data can be described by a normal distribution, the corresponding proportion for the data values therefore must be relatively close to 68%.

   If we have many data values (at least a couple of hundred), one can calculate the proportion of data values between the mean ± 2 standard deviations. This proportion should be relatively close to 95%.

Example A histogram of the height of all 30 kids from the *Fitness Club* survey is shown below (see also Chap. 2).

This histogram seems roughly symmetrical. When the sample is small, we must accept some deviation from the ideal appearance (Fig. 4.7).

In Chap. 3, we found a number of statistics for the height of the 30 kids. The most important statistics are shown in Table 4.3 with one decimal:

We see that the average and the median are roughly equal.

The interquartile range is slightly larger than the standard deviation, though not by a factor of 1.35.

The interval mean ± standard deviation corresponds to the interval from 135.0 to 179.2.

In this interval, you can count 21 of the 30 data values, equivalent to 70%, i.e., very close to 68%.

Fig. 4.7 Histogram of height
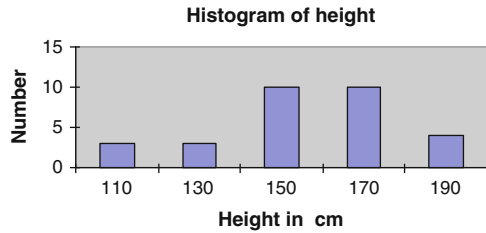
Table 4.3 Summary statistics of height

Height	
Mean	157.1
Median	159.5
Standard deviation	22.1
$Q1$	146.5
$Q3$	170
Interquartile range	23.5

Overall, we conclude that it appears reasonable that data can be described by a normal distribution.

Another issue is whether we should in fact use *two normal distributions*, one for each sex! We return to this issue in Chap. 8.

4.6.2 Skewness and Kurtosis

The two statistics skewness and kurtosis can be used to check whether data follow a normal distribution. They are, however, complicated to calculate and therefore require the use of a spreadsheet or other statistical software.

4.6.2.1 Description

These two statistics can be easily calculated in most spreadsheets or other statistical software. They provide an easy opportunity to check whether the data can be described by a normal distribution.

Skewness (*) is a measure of *how skewed the distribution is* compared to a symmetrical distribution:

– If data can be described by a symmetrical distribution, the skewness must be close to 0.
– Positive skewness indicates a right-skewed distribution.
– Negative skewness indicates a left-skewed distribution.

Table 4.4 Maximum deviation for skewness

n	Maximum deviation for skewness
25	1.00
100	0.50
400	0.25
1,600	0.12

As a very rough guide as to *how large deviations from 0 can be accepted for the skewness* for different sample sizes n, you can use the expression:

$$2\sqrt{\frac{6}{n}},$$

where n is the sample size.

This gives Table 4.4.

The smaller sample, the greater deviations from 0 you have to accept. When the sample size is multiplied by 4, the maximum acceptable deviation from 0 is divided by 2.

So, you check whether the skewness is within the maximum acceptable deviation from 0. If not, we do not have a symmetrical distribution.

If the distribution is symmetrical, you can supplement with evaluating another statistic:

Kurtosis (*) indicates how big "tails" the distribution has:

– *A normal distribution has kurtosis 0.*
– A *positive kurtosis* indicates *larger* "*tails*" than in the normal distribution.
– A *negative kurtosis* indicates *smaller* "*tails*" than in the normal distribution.

A distribution with positive kurtosis is often more "steep" in the top than the normal distribution.

Conversely, a distribution with negative kurtosis is often more "flat" in the top than the normal distribution.

However, these properties are not always true. For example, the *t-distribution* (*) has a positive kurtosis. Nevertheless, the *t*-distribution is more "flat" in the top than the normal distribution. See examples of the *t*-distribution later in this chapter.

We may accept larger deviations from 0 for the kurtosis than for the skewness.

For small sample sizes, Table 4.5 shows the minimum and maximum kurtosis that can be accepted if the distribution can be described by a normal distribution.

If kurtosis for a given sample size is outside the range above, data cannot be described by a normal distribution.

n	Min. Kurtosis	Max. Kurtosis
25	−1.2	2.3
100	−0.7	1.1
400	−0.4	0.5

Table 4.5 Min. and max. kurtosis

Assume, for example, that a sample of size $n = 100$ has kurtosis >1.1. This is a sign of a distribution with larger "tails" than the normal distribution.

Notice that for small sample sizes, the acceptable interval of kurtosis is not symmetrical.

For large samples sizes (about 1,000 or more), we can *accept twice as large deviations from 0* for kurtosis as for the skewness, i.e., *the maximum deviation from 0 for the kurtosis is*

$$4\sqrt{\frac{6}{n}},$$

4.6.2.2 Calculation

For calculation formula on these statistics, we refer to the Help in your spreadsheet (or other statistical software) or to textbooks, for example, *Montgomery: Introduction to Statistical Quality Control* (2005).

Note: There is an alternative formula for calculating kurtosis, where a normal distribution has the value 3.

In spreadsheets such as Microsoft Excel and Open Office Calc, and in most statistical software, the kurtosis of a normal distribution is 0.

4.6.2.3 Spreadsheets

Both the skewness and kurtosis exist as functions in spreadsheets:

- – *Skewness* can be obtained using the function SKEW.
- – *Kurtosis* can be obtained using the function KURT.

4.6.2.4 Example

In Chap. 3, we found using spreadsheet functions a number of statistics for the height of the 30 kids from the *Fitness Club* survey, including skewness and kurtosis, which are reproduced here (Table 4.6).

The skewness is close to 0. For a relatively small sample like this, we can accept values up to nearly ±1.

Table 4.6 Height statistics

Height	
Skewness	−0.43
Kurtosis	−0.21
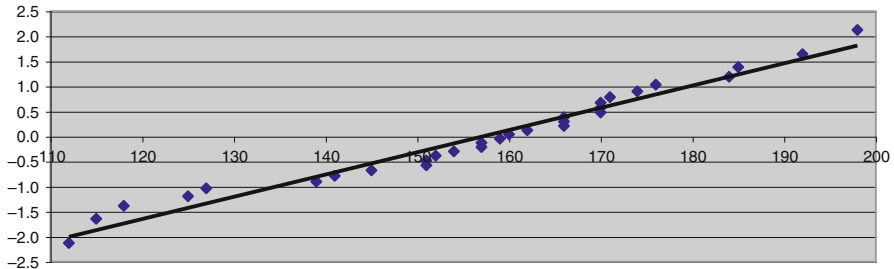
Fig. 4.8 Normal plot

Therefore, data can be described by a symmetrical distribution.

Kurtosis is very close to 0. This confirms that data can be described by a normal distribution.

There exist various statistical tests for the normal distribution, which are available in statistical software packages. These tests also confirm that the distribution of the height of the 30 kids can be described by a normal distribution. We will not cover such tests in this book; see the Help menu in your statistical software package.

4.6.3 Normal Plot

Finally, there exists a simple graphical tool to check if your data follow a normal distribution. This is called a *normal plot* (probability plot or quantile plot). It is a built-in feature of many statistical software packages, e.g., SAS, JMP, SPSS, Minitab, etc., see a list in Chap. 9.

If you have a statistical software package, you can produce a plot like the one shown above. It is also quite easy to do in a spreadsheet.

If data follow a normal distribution, the points should be randomly scattered around the straight line. This seems to be the case here. This confirms that data can be described by a normal distribution (Fig. 4.8).

If you do not have a statistical software package, you can construct the plot in a spreadsheet. The method is outlined in the text box.

Technical note: Construction of the normal plot in spreadsheets.

First, you sort the data values in ascending order. Here, we have used the height of the 30 kids from the *Fitness Club* survey; below we show only the two smallest data values, which are 112, respectively, 115 (Fig. 4.9):

(continued)

	A	B	C	D		
1	i	Height	(i-0.5)/30	Fractile	Formula column C	Formula column D
2	1	112	0.017	−2.13	= (A2-0.5)/30	= NORMSINV(C2)
3	2	115	0.050	−1.64		

Fig. 4.9 Calculations for normal plot

You make a column with consecutive numbers: this is column A. Data values are found in column B. Column C is used to calculate the expression $(i - 0.5)/n$, where i is the number of the data value (in column A) and n is the total number of values (here 30). Apart from 0.5 in the numerator (a technical correction), i/n is exactly the proportion of data values up to and including data value no. i. For the first data value, we get the result $(1 - 0.5)/30 = 0.017$.

A scatter plot with column C as Y and column B as X can be compared with the distribution function of a normal distribution.

However, it is difficult to check if the points follow a normal distribution curve. Therefore, the Y values in column C are transformed. For each value, we find the corresponding fractile in the standard normal distribution using the spreadsheet function NORMSINV, giving −2.13 for the first data value. This value is written in column D.

This corresponds to "twisting" the y-axis so that the curve becomes a straight line. A plot of column D as y and column B as x is shown above. Here, we have added a *regression line* (see Chap. 7).

4.7   Random Numbers

In evaluating how well a data set is consistent with the normal distribution, it can be a good benchmark to do the same calculations and charts for a similar number of *random numbers from the normal distribution*.

In Microsoft Excel, you can construct random numbers from a normal distribution using the add-in menu *Data Analysis*, which has a sub-item *Random number generation*. A similar option does not exist in Open Office Calc.

In this way, it is simple to construct random numbers from a normal distribution. In comparison with the histogram of the height of the 30 kids from the *Fitness Club* survey (see Chap. 2), we show a histogram based on 30 random data from a normal distribution (in this case with mean 0 and standard deviation 1, i.e., a standardized normal distribution).

We know that these numbers come from a normal distribution. Nevertheless, we can see that there are some irregularities in the histogram, which are due to the limited sample size (Fig. 4.10).

In fact, it is quite easy to create *many columns* with random data from a normal distribution. In this way, the author calculated the recommended limits for kurtosis

Fig. 4.10 Histogram, 30 data values

Histogram 30 data values
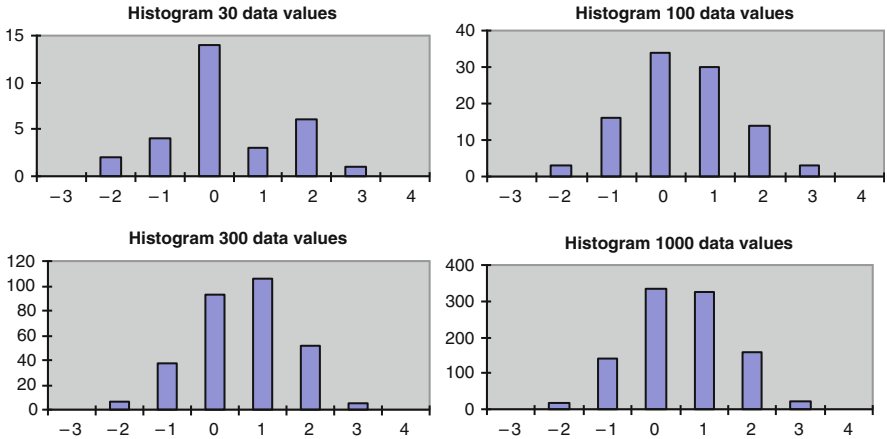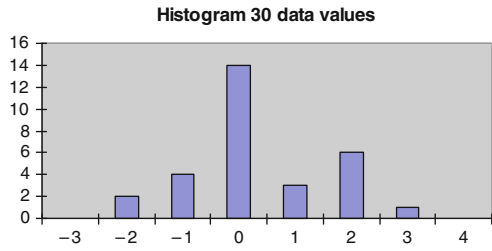
Fig. 4.11 Histograms, increasing sample size

in the table above; however, other statistical software has been used, which is more suitable for very large amounts of data.

Doing statistical calculations with random numbers is called simulation.

You can also use simulation to study *how the histogram gradually changes appearance when the sample size increases*. See the charts (Fig. 4.11).

It is evident that in a sample size of 30, you have to accept some irregularities in the histogram. When the sample size increases to, for example, 1,000, the histogram looks very similar to a normal distribution curve.

In these histograms, we have used the same number of bars for direct comparison. In practice, you will use more bars for large samples sizes, see Chap. 2.

4.8 Confidence Intervals

After studying the main characteristics of the normal distribution and how to control for the normal distribution, we now look at one of the main applications of the normal distribution: How to find the *statistical uncertainty* (*) associated with the average of a sample.

Fig. 4.12 Sample estimates

Sample and real world

Sample   Population ("real world")

- Sample average
  $\bar{x}$
- Sample standard deviation
  s

- Population mean
  μ
- Population standard
  deviation
  σ
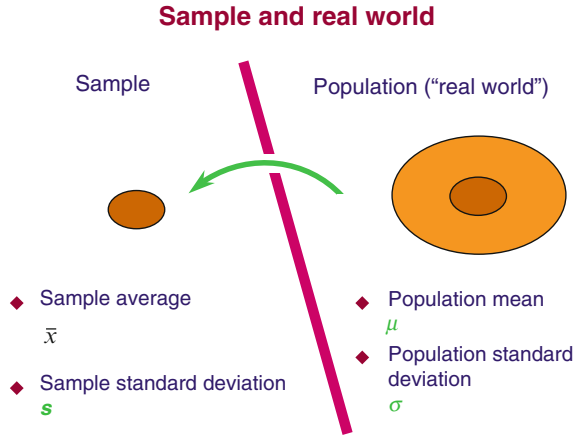
Assume the height of the kids follows *a normal distribution* with mean μ and standard deviation σ.

In practice, *we do not know μ and σ*, but we can calculate an estimate of μ and σ:

- As *an estimate of μ*, we use the *average from a sample*.
- As *an estimate of σ*, we use the *standard deviation from a sample*.

We do not anticipate that the average calculated from a sample corresponds completely to the unknown mean of the population.

But maybe we can find *an interval* that with a *high probability* (e.g., 95%) *contains the unknown mean*.

Such an interval is known as a 95% *confidence interval* (*) for the mean in the population. We will now show how to find such an interval (Fig. 4.12).

4.8.1 Confidence Interval for the Mean

4.8.1.1 Description

The technique in this section requires that *data can be described by a normal distribution*. The purpose is to calculate an estimate of the mean and find a confidence interval for it.

Furthermore, we assume that

- *We know the standard deviation in advance* or
- *The sample is sufficiently large*

Knowing the standard deviation in advance is not the usual situation.

However, *if the sample is large enough*, we can use a calculated standard deviation, as if it is known.

With a sample size of more than 30, we are on the safe side. If the sample size is just over 10, we will not make very big mistakes using the technique in this section.

4.8.1.2   Calculation

As an estimate of the mean μ in the population, we use the average $\bar{x}$ of the sample.

An average is more precisely determined, the larger the sample size. More specifically, we have the following rule:

The standard deviation for an average is obtained by dividing the original standard deviation σ with the square root of the number of data values n.

This is called the *standard error* (*) *of the mean* and sometimes abbreviated SE.

$$\text{Standard error} = \frac{\sigma}{\sqrt{n}}.$$

We have seen that slightly more than 95% of all the data values in a normal distribution are in an interval around mean $\pm$ 2 standard deviations. If we want exactly 95% of all data values in the interval, we must multiply the standard deviation by 1.96 instead of 2.

A 95% *confidence interval* (*) for the mean is therefore

$$\bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}}.$$

The number after $\pm$ can be interpreted as the *statistical uncertainty* (*) of the average.
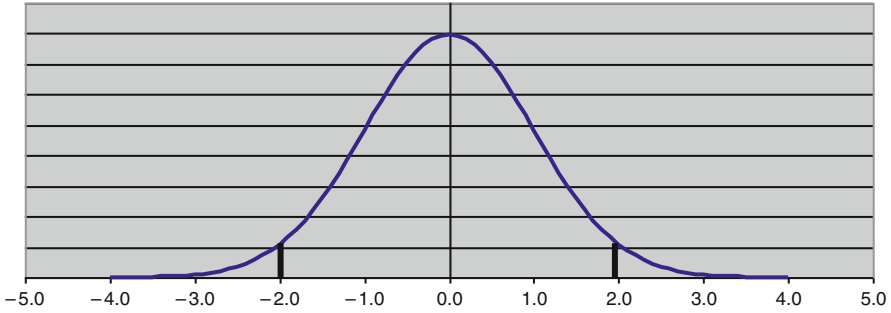
Fig. 4.13 Normal distribution

The precise reason why the 95% confidence interval is calculated in this way, is fairly technical. See, for example, G.E.P. Box, W.G. Hunter, and J.S. Hunter (Wiley 2005, 2nd ed.): *Statistics for experimenters*.

Unfortunately, the term statistical uncertainty is not given a name in most books on statistics! It is referred to as "The half-length of a confidence interval for the mean" or just "The number after ±".

In fact, 1.96 is just the 97.5% fractile of the standardized normal distribution, with 2.5% of the data values larger than 1.96. This means that 95% of the data values are between the −1.96 and 1.96. In Chap. 9, a table of the main fractiles of the standardized normal distribution can be found (Fig. 4.13).

4.8.1.3   Example Height of kids in *Fitness Club* survey

We want to calculate a 95% confidence interval for the mean height of all the kids in the population.

We do not know the mean μ, but we have data from a sample of $n = 30$ kids to estimate it.

In the sample, we have an average height of $\bar{x} = 157.10$ cm and a standard deviation $s = 22.06$ cm. As the sample size is 30, we can consider the standard deviation known, i.e., we can put $\sigma = 22.06$ cm.

We calculate a 95% confidence interval for the mean using the formula:

$$\bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}} = 157.10 \pm 7.90 \text{ cm.}$$

The 95% confidence interval for the mean height can be written as 157.10 ± 7.90 cm. The endpoints of the interval can be calculated as 149.2 and 165.0 cm. This interval will with 95% probability include the unknown mean of the population.

Sometimes we want an interval that with *a probability of* 99% contains the unknown mean. Then, we shall instead multiply the standard error with 2.576. The endpoints of the interval are then calculated as 146.7 cm and 167.5 cm. So, *the confidence interval is wider*, if we want a 99% probability.

Actually, the number 2.576 is just the 99.5% fractile in the standardized normal distribution, i.e., exactly 99% of all data values are between -2.576 and 2.576. Thus, 0.5% of the data values are larger than 2.576.

Technical note: The statistical uncertainty on an average in a finite population.

Usually, we take samples from a population with a finite number of individuals, and *the sample is relatively small compared to the population*, at most 10% of the population. In this case, the above formulas for standard error and confidence interval are valid.

If the sample is larger than 10% of the population, we must modify the formulas.

The correct formula for standard error is then:

$$\text{Standard error} = \frac{\sigma}{\sqrt{n}} \times \sqrt{1 - \frac{n}{N}}.$$

Here, N = number of individuals in the population. The fraction n/N is called the *sampling fraction* (*).

Similarly, the formula for the confidence interval is modified: A 95% confidence interval for the mean is

$$\bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}} \sqrt{1 - \frac{n}{N}}.$$

When the sample is small, n/N is close to 0, and thus the square root of $1 - n/N$ is very close to 1. This means that the simpler formula is valid.

4.8.1.4   Spreadsheets

This section can be skipped if you do not use spreadsheets.

– First, calculate the average of the sample with the function AVERAGE.
– Then, calculate the *statistical uncertainty* using the function CONFIDENCE.

For the function CONFIDENCE, we need to specify the parameters (Table 4.7).

This allows you to calculate the end points of the confidence interval. This is shown below (Table 4.8):

We imagine that data are located in the area B2: B31. The average is calculated using AVERAGE. The standard deviation is calculated using the STDEV.

Then, calculate the statistical uncertainty using function CONFIDENCE.

For this function, we must specify the "rest probability" 0.05 (equivalent to 5%), the standard deviation (calculated using the STDEV), and the sample size (30).

Table 4.7 CONFIDENCE function

Alpha	"Rest probability" (e.g., 0.05 = 5%, if you need a 95% confidence interval)
Stdev	Standard deviation (considered to be known)
Size	Sample size, n

Table 4.8 Example, CONFIDENCE function

	A	B	C
33	Average	157.10	=AVERAGE(B2:B31)
34	Standard deviation	22.06	=STDEV(B2:B31)
35	Statistical uncertainty	7.90	=CONFIDENCE(0.05;B34;30)

The 95% confidence interval for the mean height then can be written as 157.10 ± 7.90. The endpoints of the confidence interval thus become 149.0 and 165.0.

4.8.2 Confidence Interval for the Mean in Case of a Small Sample

If your sample size is typically larger than 20, you can skip this section.

4.8.2.1 Description

Let us assume that the weight of coffee in a bag of coffee follows a normal distribution with mean μ and standard deviation σ.

We know neither the mean nor the standard deviation in this normal distribution.

– We estimate both the mean and the standard deviation from a sample.
– The purpose is to estimate the mean μ and find a confidence interval for it.
– We do not know the standard deviation σ, but estimate it by the sample standard deviation s.

The sample is relatively small, less than 30 (or maybe even less than 10). Therefore, the standard deviation from the sample cannot be considered to be known.

4.8.2.2 Calculation

By analogy with the last section, we construct a 95% confidence interval for the mean:

$$\bar{x} \pm t \frac{s}{\sqrt{n}}.$$

Because the sample is not big enough for us to consider the standard deviation to be known, it appears that the multiplier t of the standard error becomes (maybe even much) larger than 1.96.

The constant 1.96 previously used was the 97.5% fractile of a normal distribution.

Instead of fractiles in the normal distribution, we must now apply *fractiles* in the *t-distribution* (*), also known as "Students *t*-distribution". This is not a single distribution, but a whole family of distributions. If there are n data values (at least 2), we say that the *t*-distribution has $n - 1$ *degrees of freedom* (*).

Note: "*Degrees of freedom*" is often abbreviated DF.

If we want a *95% confidence interval* for the mean μ, we must use *the 97.5% fractile* in the *t*-distribution with $n - 1$ degrees of freedom.

If we want a *99% confidence interval* for the mean μ, we must use the *99.5% fractile* in the *t*-distribution with $n - 1$ degrees of freedom.

The confidence interval, which we calculate in this situation, is wider than when the standard deviation σ is known. If the sample size is small, the confidence interval is much wider.

The most important fractiles in the *t*-distribution can be found in a table in Chap. 9.

In the table, we see that the 97.5 and 99.5% fractiles in the *t*-distribution are larger than the corresponding fractiles in normal distribution; this is particular noticeable, when the number of degrees of freedom is less than 10.

When the number of degrees of freedom is at least 30, there is practically no difference between the *t*-distribution and the normal distribution. Therefore, the table only shows fractiles for up to 30 degrees of freedom.

The following figure shows the probability density function of the *t*-distribution with 1, 2, and 5 degrees of freedom as well as the normal distribution. Observe that even a *t*-distribution with 5 degrees of freedom at first glance does not seem very different from the normal distribution; however, there is still a big difference in the "tails" of the distribution.

Note: The *t*-distribution has "heavier tails" than the normal distribution, i.e., it has positive kurtosis (Fig. 4.14).
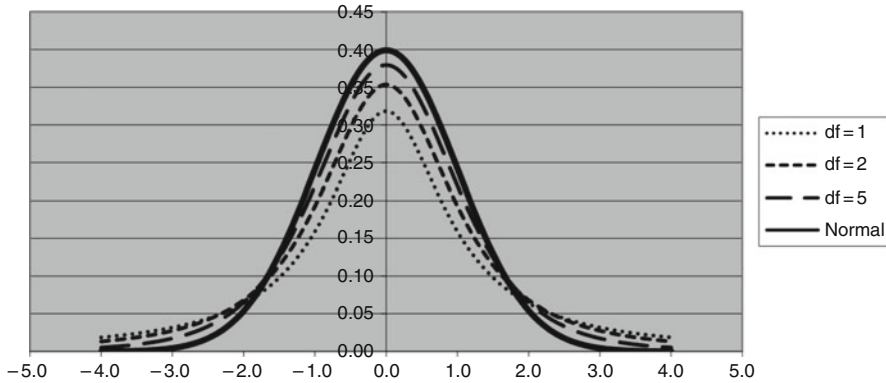
Fig. 4.14 Normal distribution and *t*-distribution

4.8.2.3 Example

Let us assume that the weight of the coffee in bags of coffee follows a normal distribution. We do not know the mean μ, but we take a sample of $n = 4$ coffee bags to estimate it.

Suppose that in the sample we have an average $\bar{x} = 505.8$ g and a standard deviation of $s = 5.30$ g.

We calculate a 95% confidence interval for the mean from the formula

$$\bar{x} \pm t \frac{s}{\sqrt{n}}.$$

We need a confidence interval, which with probability 95% contains the unknown value of the mean μ, so we must use the 97.5% fractile of the *t*-distribution.

Since the sample consists of $n = 4$ coffee bags, the number of degrees of freedom is $df = n - 1 = 3$. In the table of the *t*-distribution in Chap. 9, we find the 97.5% fractile for a *t*-distribution with 3 degrees of freedom as 3.182. Notice that this fractile is substantially larger than 1.96.

The formula now gives us the *statistical uncertainty* of the average as 8.4. The endpoints of the interval can now be calculated as 497.4 and 514.2 g.

4.8.2.4 Spreadsheets

This section can be skipped if you do not use spreadsheets.

If the sample is small, you need the following to calculate the confidence interval for the mean:

– The average of the sample calculated using the AVERAGE.
– The standard deviation calculated using the STDEV.

(continued)

- The sample size n. We need the square root of n obtained using the function SQRT.
- A fractile in the t-distribution with $n - 1$ degrees of freedom, for example, the 97.5% fractile in case of a 95% confidence interval: Calculated using the function TINV.

In the example with coffee bags, there are four data values, i.e., the number of degrees of freedom is 3. If we want a confidence interval that with probability 95% contains the unknown value of the mean μ, we use the 97.5% fractile. This fractile can be calculated in spreadsheets, use the function TINV.

Note: The special way the probability must be specified in the spreadsheet function:

- *First parameter* of TINV: Find the "rest probability" and multiply by 2.
- *Second parameter* of TINV: Number of degrees of freedom.

An example: For the 97.5% = 0.975 fractile, the "rest probability" is 2.5% = 0.025. We multiply by 2 and obtain 5% = 0.05. With 3 degrees of freedom we get the fractile TINV(0.05; 3) = 3.182.

All the information needed is found in the output below, including the calculation formula using spreadsheet functions. The 95% confidence interval for mean is found as 505.8 $\pm$ 8.4 or the interval from 497.4 to 514.2 (Fig. 4.15).

If you have Microsoft Excel, another option is to use the add-in menu *Data Analysis*, which has a menu item *Descriptive statistics*. Remember to tick the *Confidence level for the mean* box.

Below is the (partial) output from the Microsoft Excel menu *Descriptive statistics* (Table 4.9).

	A	B	C
1	Weight of 4 bags of coffee		Formula in spreadsheet
2	Bag 1	504.3	
3	Bag 2	499.6	
4	Bag 3	507.0	
5	Bag 4	512.3	
6	Average	505.8	= AVERAGE(B2:B5)
7	Standard deviation	5.30	= STDEV(B2:B5)
8	Sample size	4	
9	Rest probability	0.05	
10	Statistical uncertainty	8.4	= TINV(B9;B8-1)*B7/SQRT(B8)
11	Lower limit of interval	497.4	= B6-B10
12	Upper limit of interval	514.2	= B6+B10

Fig. 4.15 Example: Statistical functions

Table 4.9 Example: Data Analysis menu

Weight of four coffee bags	
Mean	505.8
Standard error	2.65
Standard deviation	5.30
Confidence level (95.0%)	8.4

4.8.3 Confidence Interval for the Standard Deviation

This section can be skipped without loss of continuity.

4.8.3.1 Description

Let us assume that the weight of coffee in bags of coffee follows a normal distribution with mean μ and standard deviation σ. We know neither the mean nor the standard deviation in this normal distribution.

The purpose of this section is to estimate the standard deviation σ and find a confidence interval for it.

4.8.3.2 Calculation

The confidence interval for the standard deviation σ is found by *specifying the lower and upper endpoint* of the interval directly.

We need fractiles of the *Chi-squared distribution* (*). We also here need to specify the number of *degrees of freedom*, which is again $n - 1$.

A *95% confidence interval for the standard deviation σ* is:

$$\sqrt{\frac{(n-1)s^2}{\chi^2_{97.5\%}}} \leq \sigma \leq \sqrt{\frac{(n-1)s^2}{\chi^2_{2.5\%}}}$$

The *denominator* under the square root is:

- *97.5% fractile* (lower limit)
- *2.5% fractile* (upper limit) in a Chi-squared distribution with $n - 1$ degrees of freedom

The letter χ is the Greek letter "Chi".

For a *99% confidence interval*, we use the *99.5% fractile* (lower limit) and the *0.5% fractile* (upper limit) in a Chi-squared distribution with $n - 1$ degrees of freedom.

4.8.3.3 Example

Let us assume that the weight of the coffee in bags of coffee follows a normal distribution. We do not know the standard deviation σ, but we take a sample of $n = 4$ coffee bags to estimate it.

We need a confidence interval, which with probability 95% contains the unknown value of the standard deviation σ.

In the sample, we have a standard deviation of $s = 5.30$ g.

We calculate a 95% confidence interval for the standard deviation from the formula above.

Since the sample consists of $n = 4$ coffee bags, the number of degrees of freedom is $df = n - 1 = 3$.

In the table of the *t*-distribution in Chap. 9, we find the 97.5% fractile for a Chi-squared distribution with 3 degrees of freedom as 9.35 and the 2.5% fractile as 0.22.

Thus, we get from the formula above the following 95% confidence interval for the standard deviation:

– Lower limit is 3.0.
– Upper limit is 19.8.

This means, that with 95% probability, the standard deviation is between 3.0 and 19.8.

This interval might seem rather wide. This is of course due to the small sample size. *If we need a narrower interval, we must increase the sample size.*

4.8.3.4 Spreadsheets

You need the following for a 95% confidence interval for the standard deviation:

– The variance (calculated using the function VAR).
– The sample size n.
– The number of degrees of freedom: $df = n - 1$.
– The 2.5% fractile and the 97.5% fractile in the Chi-squared distribution with $n - 1$ degrees of freedom.

The fractiles can be calculated in Microsoft Excel/Open Office Calc using the function CHIINV.

Note that you should specify the "rest" probability rather than the probability itself.

Example: For the 97.5% = 0.975 fractile, the "rest" probability is 2.5% = 0.025.

With 3 degrees of freedom, you get the fractile CHIINV(0.025;3) = 9.35 (Fig. 4.16).

	A	B	C
1	**Weight of 4 bags of coffee**		**Formula in spreadsheet**
2	Bag 1	504.3	
3	Bag 2	499.6	
4	Bag 3	507.0	
5	Bag 4	512.3	
6	Average	505.8	=AVERAGE(B2:B5)
7	Standard deviation	5.30	=STDEV(B2:B5)
8	Variance	28.13	=VAR(B2:B5)
9	Sample size	4	
10	Lower limit of interval	3.00	=SQRT((B9-1)*B8/CHIINV(0.025;B9-1))
11	Upper limit of interval	19.77	= SQRT ((B9-1)*B8/CHIINV(0.975;B9-1))

Fig. 4.16 Example using statistical functions

4.9   More About the Normal Distribution

Often, you hear people say that "you need at least a certain sample size in order that data should follow a normal distribution". This is nonsense!

Data from a sample of a population will follow the same distribution, no matter the size of the sample!

However, one can show the following:

> *When data do not follow a normal distribution:*
> The calculation of the confidence interval for the mean as shown in this chapter can still be used.
> *The sample average will follow a normal distribution, if the sample size is large.*
> This is just one of the reasons why the normal distribution is so important: *The normal distribution can be used as an approximation to describe the sample average*, regardless of the distribution of data in the population.
> The literature refers to this as "*The central limit theorem*".

The requirement to the sample size is not huge: *A sample size of 5 will often be enough*. However, if the distribution of data in the population is extremely skewed, a larger sample size (e.g., 20) may be necessary.

Furthermore:

Even if data follow a normal distribution, *we need a certain sample size in order to use fractiles from the normal distribution* when constructing a confidence interval for the mean.

We have seen that for small sample sizes we must use fractiles from the *t*-distribution. This is the case, when the sample size is up to approx. 20.

In short:

> *If the sample size is larger than 20*, you are assured that:
>
> 1. The sample average will follow a normal distribution.
> 2. You can use fractiles from the normal distribution to construct confidence intervals for the mean.

Many statistical methods (e.g., the methods described in Chaps. 7 and 8) *require that data follow a normal distribution*. This is another reason why the normal distribution is so important.

If data do not follow a normal distribution, one solution (which often can be used for right-skewed distributions) is to *transform the data values* (e.g., taking the logarithm of data values), so that they can be described by a normal distribution, and then doing the calculations on the transformed data. In case of transforming with the logarithm, we say that the original data follow a *lognormal distribution*.

Another approach is to use *nonparametric* statistical techniques (also referred to as *distribution free* techniques). These techniques are described in more advanced books on statistics, and they are implemented in many statistical software packages.

In this chapter, we have primarily dealt with quantitative data, i.e., data values are numbers that can be used for calculations.

We have seen how to describe quantitative data using the normal distribution, and how we check for the normal distribution. Also, we have seen how we can find the statistical uncertainty associated with a sample average.

In the next chapter, we deal with the statistical methods for qualitative data corresponding to groups in the population.

Chapter 5
Analysis of Qualitative Data

In this chapter we look at *qualitative data*, i.e., data values correspond to groups in the population. One particularly important type of qualitative data is *alternative (binary) data* with only two groups ("alternatives").

Some examples of alternative data:

– *Sample surveys*: For example, questionnaire surveys: the answers "Yes"/"No" to a question.
– *Statistical quality control*: For instance, classification of items in the categories "Good"/"Defective."
– *Games*: For example, "Heads"/"Tails" in tossing a coin.

Alternative data are described by a statistical distribution called the binomial distribution (*). It is used very often when analyzing data from surveys, but also in many other contexts, such as social sciences, economics, administration, science and technology.

5.1 The Binomial Distribution

The situation, in which the binomial distribution is used, can be characterized as follows:

– *Each observation ("trial") can be classified into two categories.* Often, we call them "*success*" and "*failure*" regardless of whether one of the categories can be said to be "better" than the other.
– *The probability that an observation is classified as "success" is constant.* For example, in statistical quality control there must not be a trend that defective items become more frequent.
– *The observations are independent.* This means, for example, that two respondents do not affect each other's answers in a questionnaire survey.

Using probability theory [see Chap. 9], one can calculate *the probability of getting exactly x successes out of n observations*, given that *the constant probability of success is p*.

The probabilities of the binomial distribution are tabulated in many books (for small values of *n*). They can also be calculated in most spreadsheets using a statistical function, see later.

Also, one can show the following:

> *In a binomial distribution with the number of observations = n and probability of success = p:*
>
> $$\text{Mean} = n \times p$$
>
> $$\text{Variance} = n \times p(1 - p)$$
>
> $$\text{Standard deviation} = \sqrt{n \times p(1 - p)}.$$

5.1.1 Example

Let us consider a dice where the probability of getting six eyes is 1/6 (approx. 16.7%). We throw the dice 48 times in total. We think of six eyes as "success" and everything else as a "failure."

In other words, the number of throws with six eyes follows a binomial distribution with

- $n = 48$ = number of throws
- $p = 1/6 = 16.7\%$ = probability of six eyes in each throw.

In this distribution, we have:

$$\text{Mean} = 48 \times \frac{1}{6} = 8.$$

That is, on average we will have 8 out of 48 throws with six eyes, which probably is not surprising.

Furthermore:

$$\text{Variance} = 48 \times \frac{1}{6} \times \frac{5}{6} = 8 \times \frac{5}{6} = \text{app. } 6.67.$$

The standard deviation is the square root of the variance, i.e.:

Binomial distribution with n=48, p=1/6
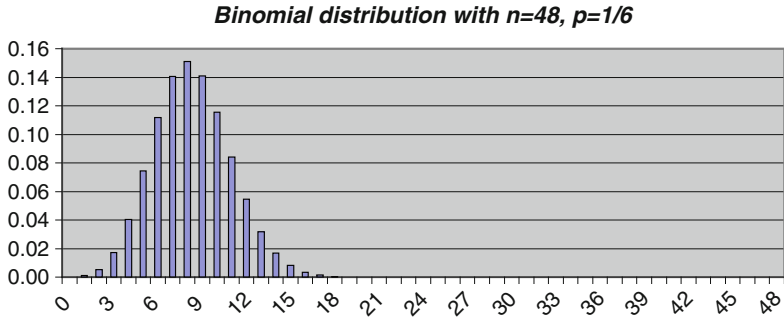
Fig. 5.1 Binomial distribution
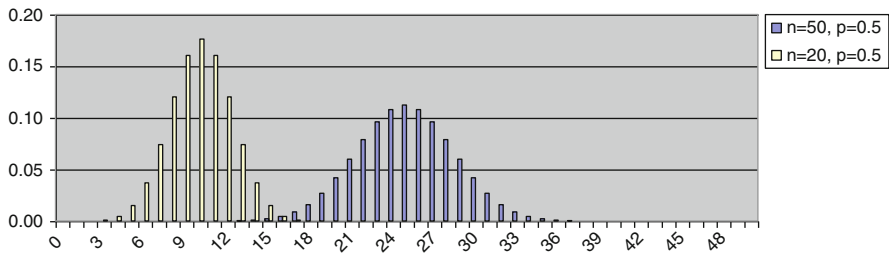
Fig. 5.2 Two binomial distributions

$$\text{Standard deviation} = \sqrt{48 \times \frac{1}{6} \times \frac{5}{6}} = \sqrt{8 \times \frac{5}{6}} = \text{app.} \, 2.58.$$

Fig. 5.1 shows the probabilities of this binomial distribution.

It can be seen that it is very unlikely to get more than (about) 20 throws out of 48 with six eyes.

As expected, the distribution is concentrated around 8 (the mean).

5.2   The Binomial Distribution and the Normal Distribution

Fig. 5.2 illustrates that the binomial distribution shows some similarity with the normal distribution.

The figure shows two binomial distributions. Both distributions have $p = 0.5$, equivalent to tossing a coin. One distribution is with $n = 20$ tosses and the other is with $n = 50$ tosses.

The chart shows that the similarity with a normal distribution is improving, the larger the n is.

Sometimes we have a binomial distribution, where p is not 0.5 (as in the example with throwing a dice).

If p is close to 0 or 1, n must be very large before the binomial distribution becomes similar to a normal distribution.

As a rule of thumb, we have the following:
The normal distribution can be used as an approximation to the binomial distribution, if
$n \times p > 5$ *and* $n(1 - p) > 5$

Note that $n \times p$ is precisely the *mean of the number of* "*successes*" (e.g., throws with six eyes).

Similarly, $n(1 - p)$ is the mean of the number of "failures" (e.g., throws with less than six eyes).

Instead of the binomial distribution, we can *use a normal distribution with the same mean and standard deviation*. See an example later in this chapter.

Technical note: Sampling with or without replacement?
Often, we take samples from a specific population with a (finite) number of individuals. Ideally, *sampling* (*) should be carried out *with replacement* in order to use the binomial distribution to describe the distribution of the number of individuals with a certain characteristic (e.g., people with a specific hobby).

Replacement means that each individual is "put back" before selecting the next individual of the sample. This way you can, in theory, select an individual twice (or maybe several times) in the sample, which means *constant probability* of getting individuals with that characteristic.

Usually, the sample is relatively small compared to the population, for example, *less than* 10% of the population. If we use *sampling with replacement* in this situation, the probability of selecting an individual twice (or several times) will be very small. Therefore, it corresponds roughly to *sampling without replacement*, i.e., each individual can only be selected once.

In practice, the vast majority of samples are selected *without replacement*. Also, the vast majority of samples are *small compared to the population*, typically less than 10% of the population. In this situation it is therefore possible to use the *binomial distribution*, although in principle it can only be used in conjunction with sampling with replacement.

If the sample is larger than 10% of the population and sampling is without replacement, one must use *the hypergeometric distribution*, which is considerably more complicated than the binomial distribution. If the sample is small compared to the population, the hypergeometric distribution is very similar to the binomial distribution. We refer to more advanced books on statistics.

5.3 The Binomial Distribution in Spreadsheets

This section can be skipped if you do not use spreadsheets.

In Microsoft Excel and Open Office Calc, the following function determines probabilities of the binomial distribution:

BINOMDIST (Value; Sample size; Probability; Cumulative).

The function can calculate *probabilities* (e.g., probability of *exactly* two successes) and *cumulative probabilities* (such as the probability of *maximum* two successes, i.e., 0, 1 or 2 successes) (Table 5.1).

5.3.1 Example

As an example, we throw a dice four times and count the number of throws with six eyes.

In other words, the number of throws with six eyes follows a binomial distribution with:

- $n = 4$ throws
- $p = 1/6 = 16.7\% = $ probability of six eyes in each throw

We wish to find:

- The probability of maximum two throws with six eyes
- The probability of exactly two throws with six eyes

We enter the information in a spreadsheet as shown below (Fig. 5.3).

We find that the probability of max. 2 (i.e., 0, 1 or 2) throws with six eyes is approx. 98%, i.e., it is very unlikely to have 3 or 4 throws (out of 4) with six eyes,

Table 5.1 BINOMDIST function

Number	Number of successes (e.g., throws with six eyes)
Sample size	Number of observations (or trials) (e.g., throws of a dice)
Probability	Probability of success in each observation (e.g., 1/6)
Cumulative	Cumulative $= 0$ calculates the probability of an exact number of successes
	Cumulative $= 1$ calculates a cumulative probability.

	A	B	C
1	*n*	4	Formula in spreadsheet:
2	*p*	0.16667	=1/6
3	*x*	2	
4	*Max. 2 throws with 6 eyes*	0.9838	= BINOMDIST(B3;B1;B2;1)
5	*Exactly 2 throws with 6 eyes*	0.1157	= BINOMDIST(B3;B1;B2;0)

Fig. 5.3 BINOMDIST function example

which is hardly surprising. The probability of getting exactly 2 (out of 4) throws with six eyes is almost 12%.

Note: The function BINOMDIST can be used only when the sample size, *n*, is up to approximately 1,000. If *n* is much larger than in 1,000, we must *approximate with the normal distribution* with the same mean and standard deviation. See an example later in this chapter.

5.4   Statistical Uncertainty in Sample Surveys

We have now studied the main characteristics of the binomial distribution. In this and the next section, we study the most important applications in sample surveys.

Let us imagine that we want to find an estimate of the *relative frequency* (*) of a particular activity among the kids in the *Fitness Club* survey. For instance, we want to find out *which proportion* of the kids is doing strength training. We can get this information through a *sample survey*. There will be some *statistical uncertainty* (*) connected with our estimate, and we also want to estimate this statistical uncertainty.

The population consists of kids in the *Fitness Club*. The *relative frequency in the population* corresponds to *the probability p that a randomly selected kid* does strength training.

The number of kids in the sample, who do strength training, can be described by a binomial distribution with

- n = sample size
- p = relative frequency in the population

Assume that in a sample of *n* kids, there are *x* kids who do strength training. The *estimate of the relative frequency p in the population* is the *relative frequency in the sample x/n*.

Note: Here the term *relative frequency* is interpreted as a *proportion* and used in much the same way as the term incidence (e.g., of a disease), which is often *expressed as a percentage*. In opposition to this is the (absolute) *frequency* (*), which is *expressed as a number* (of occurrences, individuals, etc.).

> *Two types of probability.*
>   You can use the term *probability* in (at least) two different ways:
>
> - As an expression of a *relative frequency (proportion)* that can be found from a sample survey or an experiment. This is exactly the case here. We have a proportion (of kids doing strength training) in the population. We expect that this proportion is corresponding to the probability that a randomly chosen kid does strength training. This is the reason why we use the binomial distribution in this situation.
>
> *(continued)*

– As an expression of an *expectation*. We may say that the probability that a certain football match will end with a home win is 40%. This is not necessarily based on knowledge about previous results of matches between the two teams; maybe, there are no previous matches! Rather, we are using knowledge about the latest series of results of both teams and their position in the table, injured players, etc.

5.4.1 Example

Let us assume that out of $n = 30$ kids in the sample, $x = 12$ do strength training. We can estimate the relative frequency p in the population by the relative frequency in the sample:

$$\frac{x}{n} = \frac{12}{30} = 0.40 = 40\%.$$

That is, we estimate that 40% of the kids in the population do strength training.

If the sample is large enough, we can *approximate the binomial distribution* with the normal distribution.

The estimate x/n of the relative frequency p can also be approximated by a normal distribution.

We can show that *an estimate of standard deviation in this normal distribution* is

$$\sqrt{\frac{p(1-p)}{n}} = \text{app.} \sqrt{\frac{0.40 \times 0.60}{30}} = \sqrt{0.008} = 0.09 = 9\%.$$

The *standard deviation of our estimate x/n* of the relative frequency p is thus $0.09 = 9\%$.

We can now construct a *95% confidence interval of p*, which with probability 95% contains the relative frequency in the population.

This is done in the same way as when constructing a 95% confidence interval for the mean of a normal distribution (see Chap. 4):

$$p = \frac{x}{n} \pm 1.96 \times 0.09 = 0.40 \pm 0.18.$$

This means that with 95% probability the relative frequency p of kids doing strength training in the population is somewhere between 22% and 58%.

It seems, perhaps, that we really do not know much about the proportion of kids doing strength training! The reason is, of course, that the sample is not very large.

If we increase the sample size, the statistical uncertainty will become smaller, and more about this is given below.

The term after $\pm$ is the *statistical uncertainty* (*) u of the estimate of the relative frequency in the population.

The general formula for the statistical uncertainty of a relative frequency is:

$$u = 1.96\sqrt{\frac{p(1-p)}{n}}.$$

When this formula is used in calculations, the relative frequency in the sample x/n is substituted for p.

The above formula can only be used if the sample is substantially smaller than the population, for example, less than 10% of the population.

For example, if the population consists of 100 kids, we cannot use the formula with a sample size of $n = 30$. See textbox p. 79–80 for the situation where the sample is larger than 10% of the population.

In other words:

As long as the sample is substantially smaller than the population, the statistical uncertainty is independent of the population size.

For example, a random sample of 1,000 persons in China is statistically just as good (or bad!) as a sample of 1,000 persons in, for example, Sweden, although the size of the populations is dramatically different. This comes as a surprise to many people!

The above formula also has the interesting consequence that *the largest statistical uncertainty is obtained, when $p = 0.5 = 50\%$.* See Fig. 5.4, where we have plotted the statistical uncertainty vs. p for a sample size of $n = 100$.

It is clear from the figure that as long as p is not too close to 0 or 1 (for example, if p is somewhere in the interval from $0.2 = 20\%$ to $0.8 = 80\%$), the statistical uncertainty is roughly the same as when $p = 0.5$.
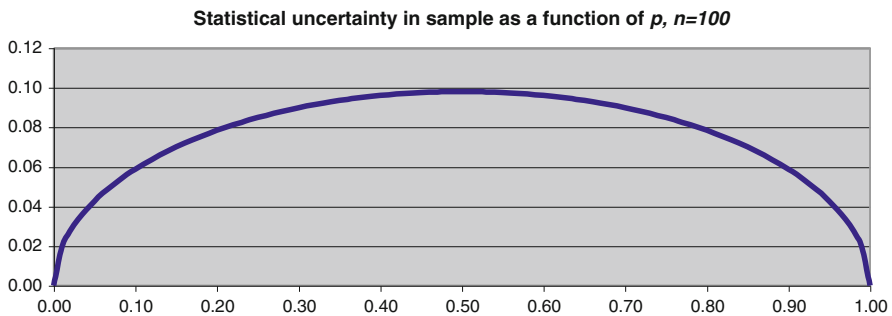
Fig. 5.4 Statistical uncertainty vs. p
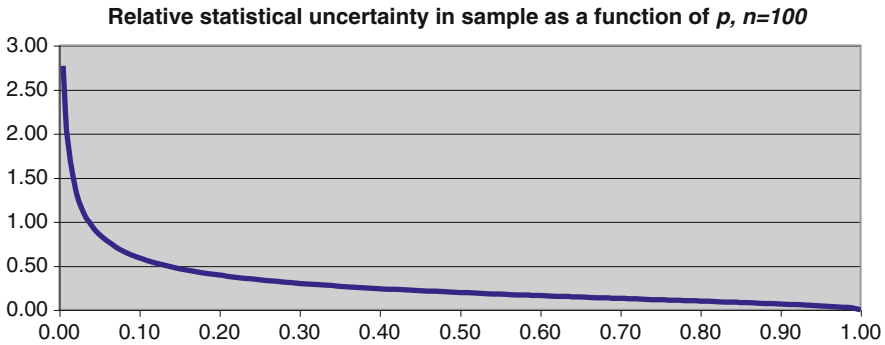
Fig. 5.5 Relative statistical uncertainty

In other words:

> *The statistical uncertainty is roughly constant as long as the relative frequency is not too close to the extremes.*

The *relative statistical uncertainty* is u/p; this number *gets larger, the smaller p is*, see Fig. 5.5 which is still based on a sample size of $n = 100$.

When the relative frequency approaches 0, the relative statistical uncertainty can get infinitely large.

This can be translated to the opinion polls: The largest political parties, which come close to 50%, have the largest statistical uncertainty. In contrast, the smallest political parties have the largest relative statistical uncertainty.

It appears from the formula that the statistical uncertainty u is inversely proportional to the square root of the sample size n. For example, if the sample gets four times larger, the statistical uncertainty is halved. Conversely, if the sample gets four times smaller, the statistical uncertainty is doubled.

This is illustrated in Fig. 5.6, which shows the statistical uncertainty vs. n, where $p = 0.5$. In other words, this is *the maximum statistical uncertainty for a given sample size*.

See also the table in section 9.4.5 showing the statistical uncertainty for different values of n and p.

> *Technical note: Statistical uncertainty of the relative frequency for a large sample.*
>
> If the sample is *larger than 10%* of the population and sampling is carried out *without replacement*, the formula for the statistical uncertainty of a relative frequency should be modified. The correct formula for the statistical uncertainty in this case is as follows:
>
> *(continued)*

$$1.96\sqrt{\frac{p(1-p)}{n}} \times \sqrt{1-\frac{n}{N}}.$$

Here N = number of individuals in the population. The fraction n/N is called the *sampling fraction* (*).

When the sample is small, n/N is close to 0, and thus the square root of $1 - n/N$ is very close to 1. If the sample is larger than 10% of the population, $1 - n/N$ is noticeably smaller than 1. Therefore, the statistical uncertainty can never be larger than:

$$1.96\sqrt{\frac{p(1-p)}{n}}.$$

In other words, you can see this number as *an upper limit for the statistical uncertainty*. If the sample is very large compared to the population, the actual statistical uncertainty will be considerably smaller.
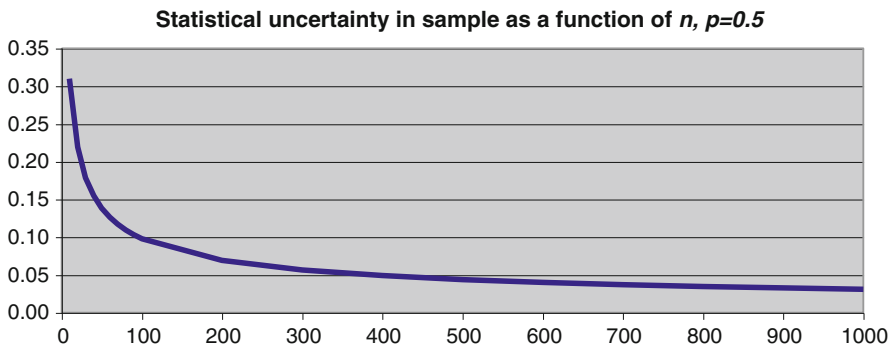
Fig. 5.6 Statistical uncertainty vs. n

5.5 Is the Sample Representative?

Here, we illustrate an important application of the above calculations.

We conducted a survey of kids in the *Fitness Club*. In the sample, there are 17 boys and 13 girls. We are interested in knowing whether *the sample is representative* with respect to sex or whether there are *systematic errors* or *bias* (*) related to the *sampling* (*). See more in Chaps. 1 and 6.

An example of systematic errors or bias: Suppose that the sample is selected by visiting the *Fitness Club* a given day and asking the first 30 kids, we meet, to participate in the survey. Assume that the boys are more frequent users of the *Fitness Club* than the girls. Then we would have a *biased* composition of the sample, i.e., it is not a representative sample with respect to sex.

Table 5.2 Data for example

$x = $ number of boys in the sample	17
$n = $ sample size	30
$p = x/n = $ proportion of boys in the sample	0.567

The population consists of all kids in the *Fitness Club*. The club might know from its registers that there are 65% boys and 35% girls among the kids, who are customers.

> If you do not have knowledge about the relative frequencies in the population, you cannot use this approach to judge whether the sample is representative.

Now we will ask ourselves the question: Is the sample representative with respect to sex?

This question can be answered by using data from the sample to *calculate a confidence interval for the proportion of boys in the population*. If this confidence interval contains the known proportion of boys (in the population), the sample is representative with respect to sex.

With (Table 5.2) the formula for the statistical uncertainty of p is:

$$u = 1.96\sqrt{\frac{p(1-p)}{n}}.$$

The statistical uncertainty is then $u = 0.177$ and the confidence interval for p has the endpoints $0.567 - 0.177 = 0.389$ and $0.567 + 0.177 = 0.744$, i.e., the confidence interval goes from 38.9% to 74.4%. As we can see, the confidence interval is very wide. The reason is, of course, that the sample is not very large!

The confidence interval contains the known value of the proportion of boys in the population, which is 65%. Therefore, we consider the sample to be *representative with respect to sex*.

We have so far in this chapter studied the binomial distribution, including confidence intervals. In the rest of the chapter, we look at *statistical tests* which are used for surveys as well as for experiments.

5.6 Statistical Tests

Sometimes, you have a hypothesis, you want to confirm or reject. A simple example is examining whether a dice or coin is "genuine." That is, the probability of, for example, six eyes when throwing a dice is 1/6 or the probability of heads is 0.5 when tossing a coin.

You then set up *a hypothesis*, for instance, in this case the assumption that the probability of heads is $p = 0.5$ when tossing a coin. Generally, this is a *hypothesis that a parameter in the population* (e.g., a probability) *equals a certain value*. In statistical literature, the hypothesis is often called the *null hypothesis* (*).

Statistical test of a hypothesis:
    The objective is to decide *whether the hypothesis is supported by data* from a sample (or an experiment).

–   *The hypothesis can be either true or false.*
–   *We consider the hypothesis true, unless data indicate that it is false.*

    The practical approach is as follows:

1. Assume that the hypothesis is true.
2. Calculate the probability of outcomes at least as "rare" as the observed outcome.
3. If this probability is small (typically less than 5%), reject the hypothesis. Otherwise, accept it.

5.6.1 Example

This approach is best illustrated with an example.

Let us assume that we toss a coin $n = 20$ times and observe the outcome heads $x = 5$ times.

We are now asking the question: Is the coin genuine? The general approach is as follows:

1. *We assume that the coin is genuine, i.e., $p = 0.5$.*

    We assume that the coin is genuine, i.e., we can use a binomial distribution with $n = 20$, $p = 0.5$. Thus, we would expect around 10 times heads out of 20 tosses with the coin.

    Figure 5.7 shows the probabilities of this binomial distribution.

2. *We calculate the probability of getting an outcome that is at least as rare as the observed outcome.*

    We have observed five times heads in 20 tosses, i.e., somewhat less than expected. An outcome at least as rare will be at most five times heads in 20 tosses.

    Usually you will add the probability of getting at least 15 (i.e., $20 - 5$) times heads, which is just as rare to the "other extreme"; it is "just as bad" for the hypothesis. This is called a *two-sided test* (*).

    We can easily calculate the probability of at most five times heads in 20 coin tosses by using a spreadsheet. This is precisely the sum of the probabilities in the bars from 0 to 5 in the chart above.
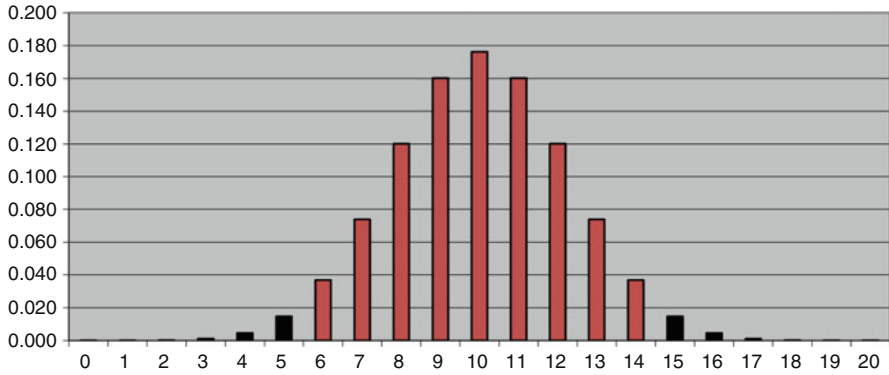
Fig. 5.7 Binomial distribution for example

The function BINOMDIST (5; 20; 0.5, 1) gives us the result 0.0207, i.e., 2.07%. (This can also be looked up in a table of binomial probabilities.) Similarly, the probability of at least 15 times heads in 20 coin tosses is the same, i.e., 0.0207.

The total probability of an outcome at least as rare as the observed outcome is therefore: $2 \times 0.0207 = 0.0414 = 4.14\%$.

This probability is called the *p-value* (*) in statistical "jargon."

3. *If this probability is small (typically less than 5%), reject the hypothesis. Otherwise, accept it.*

The probability of an outcome at least as rare as the observed outcome is 4.14% <5%.

The conclusion is, therefore, that *we reject the hypothesis that p = 0.5!* This means that there is statistical evidence that the coin is false!

The philosophy behind the approach outlined above is as follows: If the probability of a more "rare" outcome is small, there are two options:

– *The hypothesis is true, but we have observed a rare event.*
– *The hypothesis is actually false.*

Of course it is conceivable that the first option is correct. If so, we have observed a rare event! However, statisticians do not believe in miracles. Therefore, we prefer to believe the second option.

5.6.2 Approximation with the Normal Distribution

In some situations, you need to use the approximation of the binomial distribution with a normal distribution:

- The sample size is very large, e.g., more than 1,000.
- You have to perform the calculations using only a calculator and a table of the normal distribution.

In the example above, we should use a normal distribution with mean $n \times p = 10$ and variance $n \times p(1 - p) = 5$. The standard deviation is $\sqrt{5} = 2.236$.

In this normal distribution, we must calculate the probability of data values up to (and including) 5.

Since the normal distribution *covers the whole axis* (not only integer values), we should actually find the probability of data values up to 5.5 rather than 5. We therefore *add 0.5 to the value of x*.

We must therefore find the value NORMDIST (5.5; 10; 2.236; 1). We use 1 for the last parameter, because we need the distribution function (not the density function). The result is 2.2%. Again, you multiply by 2 and get 4.4%, i.e., still below 5%.

That it is indeed permissible to use the normal distribution can be demonstrated as follows:

With $n = 20$ and $p = 0.5$, we obtain $n \times p = 10 > 5$ and $n(1 - p) = 10 > 5$.

5.6.3 Significance Level

How small should the probability be, before we reject the hypothesis?

This limit is called the *significance level* (*). Usually, we choose the significance level 0.05, i.e., 5%.

If the hypothesis is indeed true, there is a small probability (5%) that we commit an error, i.e., *reject a true hypothesis*. This kind of error is called a *type I error* (*).

If not explicitly stated, the significance level is 5%.

If the consequences of committing the error of rejecting a true hypothesis *are very serious*, we could choose *the significance level 1%*.

The price we pay for choosing a 1% significance level is that it becomes more difficult to detect differences that actually exist, for example, to detect a false coin.

In the above example with tossing a coin, the *p*-value is just below 5%, but well over 1%. If we use a significance level of 1%, the hypothesis $p = 0.5$ is accepted. We need more convincing data to reject the hypothesis $p = 0.5$, if we use the significance level of 1%.

Note: *The significance level has to be decided before doing the statistical test.*

The example with tossing a coin shows why: Depending on the choice of significance level, you can either accept or reject the hypothesis.

5.6.4 Statistical Test or Confidence Interval

Let us in the example above construct a confidence interval for the probability of heads, p. We have observed a relative frequency of heads $x/n = 5/20 = 0.25 = 25\%$. This estimate for p is inserted in the formula for the statistical uncertainty u:

$$u = 1.96\sqrt{\frac{p(1-p)}{n}}.$$

The result is $u = 0.22$. The confidence interval is thus $p = 0.25 \pm 0.22$, i.e., the confidence interval goes from 0.03 to 0.47. That is, a 95% confidence interval for p does not contain 0.5. So based on our data we are (at least) 95% certain that p is not 0.5, as it should be for a genuine coin.

We thus get the same conclusion when using a confidence interval as when using a statistical test.

> To perform a statistical test at 5% significance level for the hypothesis $p = 0.5$ corresponds to constructing a 95% confidence interval for p and check whether the confidence interval contains the value 0.5.

What are the advantages and disadvantages of the two approaches?

– *A confidence interval* gives a *yes/no conclusion*. On the other hand, a confidence interval also gives us *a set of values* of p, which can be *considered likely*.
– *A statistical test* gives both a *yes/no conclusion* (accept/reject) and a more *graduated answer* with *the p-value*. In the example with tossing a coin, the probability of a rarer event is below 5%, but only slightly below. That is, the hypothesis is only just rejected, which is not a very convincing conclusion. We would have felt more convinced, if the *p*-value was less than 1%.

In this section, we have dealt with a statistical test of a hypothesis in one distribution (the binomial distribution). The next section describes situations where two or more distributions are to be compared.

5.7 Frequency Tables

5.7.1 Introduction to Chi-Squared Test

Let us take a look at some data from the *Fitness Club* sample survey. We are still interested in the proportion of kids doing strength training, but now *we want to group data* according to sex. This is given in Table 5.3.

Table 5.3 Observed frequencies

Observed frequency of kids			
	Does strength training	No strength training	Total
Boys	10	7	17
Girls	2	11	13
Total	12	18	30

Table 5.4 Row percentages

Percent			
	Does strength training (%)	No strength training (%)	Total
Boys	59	41	17
Girls	15	85	13
Total	40	60	30

We call this a *frequency table*. This is a *2 × 2 table* (read "2 by 2 Table"), i.e., two rows (Boys/Girls) and two columns (Does strength training/No strength training).

It looks as if (perhaps not very surprising) strength training is a boy's thing. We can illustrate this by calculating the proportion (in percent) of kids who does strength training for each sex. This is given in Table 5.4.

We see that as many as 59% of the boys do strength training. By contrast, only 15% of the girls do strength training. In total in the sample, 40% of the kids do strength training.

This is a "subjective" evaluation based on comparing the two proportions; it is not an objective statistical test of whether the proportions differ.

In the example above, it is probably quite obvious that *there actually is a difference* between the proportion of boys and girls doing strength training; we hardly need a statistical test. But in other cases there might be only minor differences in the proportions; then the conclusion is less obvious, and there is a need for an objective criterion like a statistical test, so we are not left in doubt.

Here we present a statistical test that can be used in situations where you have to *compare two relative frequencies (proportions)*.

The hypothesis is that there is the same proportion of boys and girls doing strength training. This means that there is *independence between rows and columns*, i.e., doing strength training is independent of sex.

We use a statistical test of the hypothesis. We use the general approach:

1. *We assume that the hypothesis is true*.
2. *We calculate the probability of getting an outcome that is at least as rare as the observed outcome*.

First, let us consider how the frequencies would be if the hypothesis was correct. Then the proportion of kids doing strength training would be the same for boys and girls, i.e., equal to 40%.

This does not necessarily lead to integers as shown in Table 5.5.

Table 5.5 Expected frequencies

Expected frequency of kids			Total
	Does strength training	No strength training	
Boys	**6.80**	10.20	**17**
Girls	5.20	7.80	13
Total	**12**	18	**30**

For example, we calculate the expected 6.80 boys doing strength training as 40% of 17 boys: $17 \times 0.40 = 6.80$.

The 40% can be written $12/30 = 0.40 = 40\%$. This means that we can *write the calculation of the "upper-left corner" of the table as:* $\dfrac{17 \times 12}{30} = 6.80$.

Note which frequencies are used in this formula. They are highlighted in bold in the table above.

Similarly, the other frequencies are calculated. We call these frequencies the *expected frequencies*. The original frequencies are called the *observed frequencies*.

The hypothesis must be considered false if the observed frequencies are far from the expected frequencies. Before showing the calculation of the probability, it should be emphasized that a necessary condition is that *all expected frequencies should be at least 5*. This is the case here.

We need a measure of the magnitude of the difference between the observed and expected frequencies.

A measure of the difference between the observed and expected frequencies can be calculated using the following formula:

$$\chi^2 = \sum \frac{(O - F)^2}{F}.$$

Here:

- O = *observed frequency*
- F = *expected frequency*

Σ is the "sum symbol," i.e., we must add all 4 numbers of this type, as there are 4 observed frequencies and 4 expected frequencies.

The result is called χ^2. χ is the Greek letter Chi (pronounced "Ki"), which corresponds to the letters "ch". χ^2 is read "*chi-squared.*"

The calculation of χ^2 is as follows:

$$\chi^2 = \frac{(10 - 6.80)^2}{6.80} + \frac{(7 - 10.20)^2}{10.20} + \frac{(2 - 5.20)^2}{5.20} + \frac{(11 - 7.80)^2}{7.80} = 5.79.$$

If all the observed frequencies are the same as the expected frequencies, we will have $\chi^2 = 0$.

Small values of χ^2 are "good" for our hypothesis – this is evidence that the observed and expected frequencies are close. Large values of χ^2 are "bad" for the hypothesis – this is evidence that the observed and expected frequencies are far apart.

Now we need to find the probability of getting a value of χ^2 that is larger (i.e., "worse" for the hypothesis) than the value calculated above. One can show that we have to use a so-called *chi-squared distribution* (*) (or χ^2-distribution). In this case, we say that we use a chi-squared distribution with one *degree of freedom* (*) (more on this later).

We can get a rough estimate of the probability by looking at the table of chi-squared distribution (with one degree of freedom) in Chap. 9. Here it is found that

- *The 97.5% fractile is 5.02.*
- *The 99% fractile is 6.63.*

This means that the probability of getting a value greater than 5.79 is somewhere between 1% and 2.5%.

We can show (see section 5.7.4) that the probability is actually in 0.016 or 1.6%.

3. *If this probability is small (typically less than 5%), reject the hypothesis. Otherwise, accept it.*

We have seen that the probability of getting a value of χ^2 above 5.79 is between 1% and 2.5%. As this is below 5%, we reject the hypothesis, that the proportion doing strength training is independent of sex. This was what we had expected.

Here, we see the density function of chi-squared distribution with one degree of freedom. It can be seen that the value 5.79 is pretty "extreme" in the distribution (Fig. 5.8).
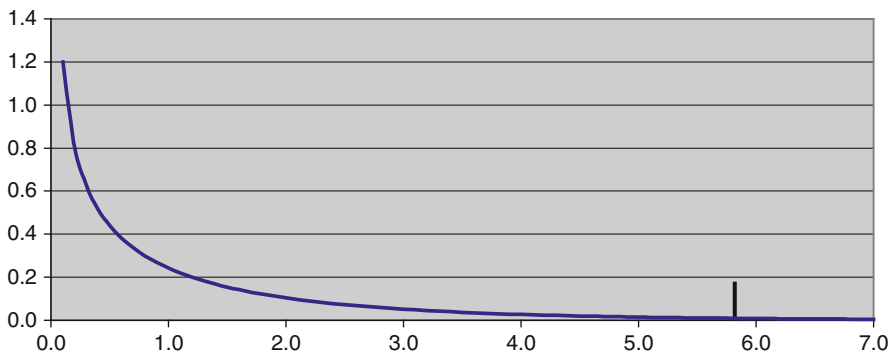
Fig. 5.8 Chi-square fractile in example

5.7.2 Confidence Interval for Difference Between Two Proportions

This section can be omitted without loss of continuity.

We have seen above that there is statistical evidence that the proportions of girls and boys doing strength training are indeed different.

It seems natural to ask the question: *How large is the difference between the two proportions?*

The estimated proportions of boys and girls doing strength training are $59\% = 0.59$ and $15\% = 0.15$, so we estimate the difference between the two proportions to be $0.59 - 0.15 = 0.44$.

The statistical uncertainty of the estimated difference between the two proportions is:

$$u = 1.96\sqrt{p(1-p) \times \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}.$$

Here p is the *proportion in the whole sample* doing strength training, and n_1 and n_2 are the total frequencies of boys and girls in the sample.

From the tables earlier in the chapter, we obtain Table 5.6.

Using the formula above, we obtain the statistical uncertainty for the difference between the two proportions as $u = 0.35$.

The difference between the two proportions is 0.44. Thus, a 95% confidence interval for the difference is 0.44 ± 0.35.

The confidence interval goes from $0.09 = 9\%$ to $0.79 = 79\%$. The confidence interval does not contain 0, in accordance with the fact that the hypothesis, that the two proportions are identical, is rejected.

It may seem that we do not know much about the size of this difference (except that it is not 0). If we need a narrower confidence interval, we must increase the sample size.

5.7.3 Several Rows and/or Columns

The statistical test we have described above can be used generally for the comparison of several distributions of a grouping of several categories. There can therefore be more than two rows and/or columns in the table.

Table 5.6 Two proportions

P	0.40
n_1	17
n_2	13

Table 5.7 Observed
frequencies in example

Cardiovascular workouts?	Physical fitness			
	Bad	Medium	Good	Total
No	6	6	3	**15**
Yes	3	6	6	**15**
Total	**9**	**12**	**9**	**30**

Table 5.8 Expected
frequencies in example

Cardiovascular workouts?	Physical fitness			
	Bad	Medium	Good	Total
No	4.5	6.0	4.5	**15**
Yes	4.5	6.0	4.5	**15**
Total	**9**	**12**	**9**	**30**

As an example, let us return to table 2.12, which was shown in Chap. 2. The
Fitness Club kids were asked whether they do cardiovascular workouts or not.
Moreover, they were asked how they perceive their physical fitness. Here is the
frequency of kids in all combinations of the two questions (Table 5.7).

We will examine whether physical fitness is independent of cardiovascular
workouts. The hypothesis is that *there are the same relative frequencies (proportions) among the rows*, i.e., cardiovascular workouts do not affect physical fitness.

You can use the same procedure as above, i.e., first we calculate the expected
frequencies. They are shown in Table 5.8. For example, the expected frequency
of kids doing cardiovascular workouts, who have a bad physical fitness, is calculated as:

$15 \times 9/30 = 4.5$.

We observe that several expected frequencies are smaller than 5, but none are
smaller than 4.5. Strictly speaking, it is not allowed to use the method described
above; however, it will hardly be a very big mistake.

We can now calculate χ^2 using the formula:

$$\chi^2 = \sum \frac{(O - F)^2}{F}.$$

This gives the result 2.00.

How many degrees of freedom should be used for the chi-square distribution?
It turns out that the general formula is:
Degrees of freedom $=$ (Number of rows $-$ 1) $\times$ (Number of columns $-$ 1)

In this example, there are two rows (No/Yes) and three columns (Bad/Medium/
Good). Therefore, the number of degrees of freedom is DF $= (2 - 1) \times (3 - 1) =
1 \times 2 = 2$.

In the table in Chap. 9, we find that the 95% fractile of the chi-squared distribution with two degrees of freedom is 5.99. We have calculated a value $\chi^2 = 2.00$.
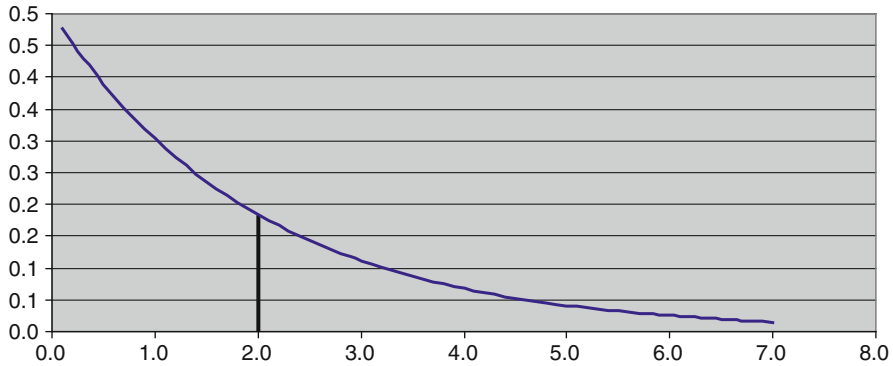
Fig. 5.9 Chi-square fractile in example

This is far smaller than the 95% fractile, i.e., the probability of getting a larger value is (probably a lot) larger than 5%.

Therefore, *we accept the hypothesis* that physical fitness is independent of doing cardiovascular workouts. This does not necessarily mean that the hypothesis is true. We just *do not have statistical evidence to reject it*.

We can also calculate the probability directly using the *CHITEST* function in a spreadsheet (see the next section) and get approx. 37%, much larger than 5%.

The following is the density function for a chi-squared distribution with two degrees of freedom. It is evident that the value 2.00 by no means is "extreme" in the distribution (Fig. 5.9).

From the table of frequencies, one might suggest a trend that kids not doing cardiovascular workouts have a worse physical fitness compared to kids doing cardiovascular workouts. However, *there is no statistical evidence in the data to support this assumption*!

> *Note:* The hypothesis that *there is independence between the row and column variables* is accepted.
>
> In the case, when *the hypothesis is rejected*, this is itself not a guarantee that there is a *causal relationship* between the two variables! And if there is a causal relationship, *the statistical test cannot tell which variable is cause and which variable is effect*.
>
> It might also be that there exists an "*indirect relationship*" between the two variables, i.e., both variables are related to a third variable. We return to this issue in Chap. 7 in relation to quantitative variables.

In this chapter we have seen some techniques, which are useful in the *statistical analysis* of sample surveys and experiments. In the next chapter, we look at some issues within *planning* of sample surveys and experiments. First, we show how the calculations above can most easily be done by calculator or spreadsheet.

5.7.4   Calculations in Spreadsheets

If you do not use spreadsheets, you can skip this section.

We use the table of observed frequencies of *Fitness Club* kids doing strength training grouped according to sex (Fig. 5.10).

First an explanation regarding the expected frequencies: For instance, cell B9 is calculated as $17 \times 12/30$.

The formula in cell B9 can thus be programmed like this:

```
=D3 × B5/D5
```

However, it is an advantage, if you program cell B9 like this:

```
=$D3 × B$5/$D$5
```

The dollar signs are "*absolute references*"; see more in the help of your spreadsheet. If you copy cell B9 over the area B9:C10 (i.e., all the expected frequencies), the references will be kept properly. They will always be referred to the correct cells.

When you have calculated the expected frequencies, the rest is easy. See cell B13, how to use the spreadsheet function CHITEST.

For the function *CHITEST*, you specify the cells with the observed frequencies (cell B3:C4) and the cells with the expected frequencies (cell B9:C10).

The result is the p-value, i.e., the probability of getting an outcome that is at least as rare as the observed outcome. It is approximately 0.016 or 1.6%. Previously, we found that the *p*-value is between 1% and 2.5%.

We do not get the value of χ^2, but we really do not need it for anything! If we want it after all, it can be calculated as shown in cell A14. We do an "*inverse calculation*" from the *p*-value using the function *CHIINV*. For the function

	A	B	C	D
1		**Observed frequency of kids**		
2		**Does strength training**	**No strength training**	**Total**
3	**Boys**	10	7	17
4	**Girls**	2	11	13
5	**Total**	12	18	30
6				
7		**Expected frequency of kids**		
8		**Does strength training**	**No strength training**	**Total**
9	**Boys**	6.80	10.20	17
10	**Girls**	5.20	7.80	13
11	**Total**	12	18	30
12				
13	0.016	=CHITEST(B3:C4;B9:C10)		
14	5.79	=CHIINV(A13;1)		

Fig. 5.10 Chi-squared test in spreadsheet

Table 5.9 Data for
calculations using calculator

Observed frequency of kids			
	Does strength training	No strength training	Total
Boys	10	7	17
Girls	2	11	13
Total	12	18	30

CHIINV, you specify *the p-value* and the number of *degrees of freedom*, in this case
1. We get the same result 5.79, as was obtained previously.

5.7.5 Calculations by Calculator

If you do not use a calculator, you can skip this section. You need a mathematical
calculator with logarithms.

   We use the table of observed frequencies of *Fitness Club* kids doing strength
training grouped according to sex (Table 5.9).

   There is another formula for calculating χ^2, which is much easier to use on a
calculator. It will give practically the same result as the above formula. The larger
the sample, the better the agreement will be.

   In the table, *we calculate a contribution from each number*, including the sub-
totals and the total.

   Every contribution is of the form $x \cdot \ln(x)$, where x is one of the numbers of the
table and ln is the *natural logarithm function*, which is available on most mathe-
matical calculators.

   The contribution corresponding to the *observed frequencies* in the 2×2 table
and the *total* is *added*, the remaining contributions (corresponding to *subtotals* of
each row and each column) are *subtracted*. Finally, we *multiply by* 2.

   In the example, the calculations look like this:

$$\chi^2 = 2(10 \times \ln(10) + 7 \times \ln(7) + 2 \times \ln(2) + 11 \times \ln(11) + 30 \times \ln(30) - 12$$
$$\times \ln(12) - 18 \times \ln(18) - 17 \times \ln(17) - 13 \times \ln(13))$$
$$= 6.18.$$

   We get a value of χ^2, which is pretty close to the one obtained previously. The
conclusion remains the same.

   The advantage of this formula is that *we do not need to calculate the expected
frequencies*. This can be pretty cumbersome if there are many rows and columns.

   With this approach, we cannot control the necessary condition that all expected
frequencies are at least 5. One easy solution is to *calculate the minimum expected
frequency* using the minimum row total and the minimum column total and this
should be >5.

   Here the row totals are 17 and 13; the smallest value is 13 from row 2. The column
totals are 12 and 18; the smallest value is 12 from column 1. Thus, in row 2, column 1,
we get the minimum expected frequency. This was previously calculated to be $5.20 > 5$.

Chapter 6
Error Sources and Planning

6.1 Two Kinds of Errors

We saw in Chap. 3 that there are *two types of variation* in data: *systematic* variation and *random* variation.

> There are therefore *two sources of errors* in sample surveys as well as in planned experiments:
>
> – *Systematic errors*, usually called *Bias* (*): The difference between the true value and the mean.
> – *Random errors* (*): The dispersion (spread) around the mean.

This is illustrated in Fig. 6.1.
Increasing the size of the sample (or experiment) *can reduce the random errors*. This can be used in the planning stages to *determine an appropriate size of the sample* (or experiment). We discuss this in the first part of the chapter.

After that follows some topics exclusively related to sample surveys. First we discuss various sources of systematic errors in sample surveys, in continuation of the discussion in Chap. 1. We also give an overview of the principles of *sampling* (*) or sample selection.

6.2 Random Error and Sample Size

The *random errors* (*) are coming from *general causes of variability*, which give rise to a natural variability. The natural variability is reflected in dispersion (spread) around the mean. This *dispersion is always present* to some extent. The individuals in a sample survey (or an experiment) will never be completely identical.

Fig. 6.1 Systematic and Random Variation
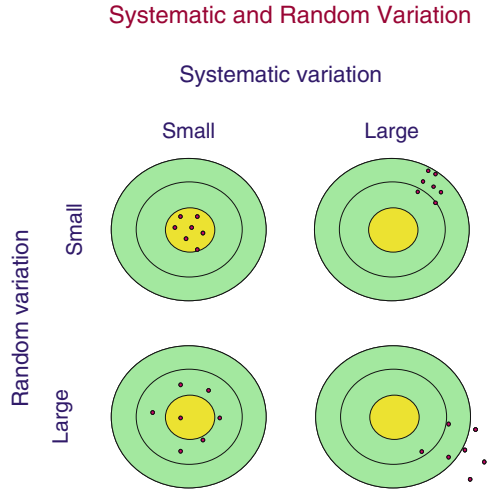
The following considerations are based on a necessary condition:

The experiment or the sample must be organized through randomization (*). This means that *an experiment should be conducted in random order*, or that *sampling should be done by a random selection mechanism* (more on this last in the chapter).

Randomization is necessary in order to be able to estimate the statistical uncertainty!

When planning a sample survey or an experiment we often face the question: How large a sample size should we choose?

The first question one must consider is *what to record on each individual* in the sample or experiment.

Roughly speaking, there are two situations:

1. We record a *qualitative* variable, often an alternative variable, e.g., "yes/no".
2. We record a *quantitative* variable, i.e., a number for each individual.

6.2.1 A Qualitative Variable

Qualitative variables are usually associated with sample surveys, but may also arise in connection with planned experiments.

We saw in Chap. 5 that *the statistical uncertainty of a relative frequency* is given by

$$u = 1.96\sqrt{\frac{p(1-p)}{n}}.$$

Here n is the sample size and p the relative frequency of one answer category (e.g., the answer "Yes" to a question). In this formula, we use *the relative frequency x/n from the sample* as estimate of p.

The formula requires that *the sample size n is large enough* for the conditions $n \times p > 5$ and $n(1-p) > 5$ to be fulfilled. On the other hand, the *sample size should be at most 10% of the population* (if sampling is without replacement), see Chap. 5.

In practical calculations one can safely replace 1.96 with 2. There is no important difference in the result, and it makes the calculations much simpler. Therefore, we use the simpler formula

$$u = 2\sqrt{\frac{p(1-p)}{n}}.$$

As shown in Chap. 5 the formula gives the maximum value for $p = 0.5 = 50\%$.

If we put $p = 0.5$, the formula can be reduced to the following expression for the maximum statistical uncertainty of a relative frequency:

$$u = \frac{1}{\sqrt{n}}.$$

This formula is remarkably simple!! Yet, I have found it in no other books on statistics!

A few examples:

– The maximum statistical uncertainty for $n = 100$ is $u = 1/\sqrt{100} = 1/10 = 0.1 = 10\%$.
– The maximum statistical uncertainty for $n = 10,000$ is $u = 1/\sqrt{10,000} = 1/100 = 0.01 = 1\%$.

The formula can be used to *estimate the sample size* to achieve a given maximum statistical uncertainty. If *the maximum statistical uncertainty must be u*, we get:

(continued)

$$n = \frac{1}{u^2}.$$

This is the *minimum sample size*, we can use. This formula is extremely useful!

Note: The formula above assumes that the sample size is less than 10% of the population. If the sample size found from the formula is larger than 10% of the population, the statistical uncertainty becomes smaller. Thus, *we are on the safe side*, if we use the formula above to determine the sample size.

Technical note: If the sample is large compared to the population.

In this case, the statistical uncertainty can be considerably less than the value determined from the above formula. See end of Chap. 5, where the correct formula for the statistical uncertainty is indicated.

The required sample size then becomes much smaller. The easiest way is to program the formula from the Chap. 5 in a spreadsheet and try different sample sizes, until you get the desired maximum statistical uncertainty. As a starting value, you can use $n = 1/u^2$.

6.2.1.1 Example

We take an example of a situation from a sample survey. In Chap. 5, we found that among the $n = 30$ kids in the *Fitness Club* survey, we find that $x = 12$ do strength training. The relative frequency in the sample is thus $x/n = 12/30 = 0.4 = 40\%$. We also found that the statistical uncertainty is 0.18 or 18%.

If we find this statistical uncertainty too large, we can use the formula for n. If we want to have a maximum statistical uncertainty of $0.1 = 10\%$, we find the minimum sample size $n = 1/u^2 = 1/0.1^2 = 100$.

The above formula can obviously be used in *subgroups* of the population. The formula then finds the value of n for each subgroup separately.

For example, we want to be sure of getting a fairly accurate estimate for both boys and girls separately. We then apply the formula for each sex separately. We specify the maximum statistical uncertainty acceptable for the relative frequency of boys doing strength training. This gives us the necessary number of boys to be included in the sample. The same calculation can be performed with the girls.

6.2.2   A Quantitative Variable

Quantitative data occur often in sample surveys as well as planned experiments.

We saw in Chap. 4 that *the statistical uncertainty of an average*, if the standard deviation is known, is approximately

$$u = 2 \times \frac{\sigma}{\sqrt{n}},$$

where σ is the standard deviation and n is the sample size.

Strictly speaking, we should replace 2 by 1.96. This does not mean much in practice, however.

> If we know the standard deviation σ and *want a maximum statistical uncertainty u* of the average, we find the *necessary sample size* as:
>
> $$n = \left(2 \times \frac{\sigma}{u}\right)^2.$$
>
> We cannot determine the sample size without knowledge about the dispersion of what we measure!

Note: We often use the term *sample size* even when planning *an experiment*!

6.2.2.1   Example

We show an example from a planned experiment, but it might as well be a sample survey.

We want to perform an experiment with baking of breads with a special wheat variety. We want to assess the impact on bread volume of the addition of a special additive, aimed at increasing the volume.

We know from previous experiments that we can expect a standard deviation of app. $\sigma = 10$ ml of the bread volumes. We want a maximum statistical uncertainty on both averages (bread baked with and without additive) of $u = 5$ ml.

If we use the formula for n, we get $n = (2 \times 10/5)^2 = 4^2 = 16$. If we are to be on the safe side, we should choose $n = 16$ breads in each group (with/without additive).

In total, in the experiment we should bake $2 \times 16 = 32$ breads.

Strictly speaking, we do not need to specify both the standard deviation σ and the statistical uncertainty u. It is enough that *we specify the desired ratio between σ and u*.

Maybe we do not have precise knowledge of the standard deviation.

However, we may want a maximum statistical uncertainty, which is half of the standard deviation obtained. This means that we require that σ/u is at least 2.

We now use the formula with $\sigma/u = 2$. This results again in the required sample size $n = 16$ for each group.

6.2.2.2   Notes

1. *If the value of n determined from the formula is small* (e.g., smaller than 10), we must be careful. In practice, the necessary size of the experiment may be larger. We have used the factor 2 to construct a confidence interval for the mean. Strictly speaking, we should use the 97.5% fractile of a *t*-distribution with $n - 1$ degrees of freedom (see Chap. 4). If n is smaller than 10, the 97.5% fractile is larger than 2, see table of the *t*-distribution at the end of the book.
2. *In most sample surveys, we often have only one group*. Thus, we get directly the necessary sample size from the entire population using the formula for *n*.
3. In most experiments, we are often interested in comparing two or more groups ("treatments"). Therefore, you might prefer to determine the necessary sample size to obtain a given statistical uncertainty of the difference between two means. See more in Chap. 8.

So far, we have been dealing with the necessary size of a sample (or an experiment). This depends on how large a random error, we can accept.

The rest of this chapter focuses on sample surveys. We look in more detail at *bias* (systematic errors). Also, we discuss *how to do the sampling* (sample selection), so that bias is avoided as much as possible.

6.3   Bias (Systematic Errors)

In Chap. 1, we discussed some errors associated with sample surveys, in particular questionnaires.

Bias (*) or systematic error comes from *specific causes, which often can be identified*. By removing these specific causes, the bias can in principle be avoided.

The main causes of bias in sample surveys (questionnaires) are:

1. Errors caused by the interviewing process and wording of the questions.
2. Errors caused by nonresponse.
3. Errors in the sampling (sample selection).
4. Errors in the definition of the sample.

In Chap. 1 we discussed 1 and 2. Now, we discuss 3 and 4.

6.3.1 Errors in the Sampling (Sample Selection)

The sampling (*) or sample selection should preferably be done using randomization (*). We discuss this in the next section.

The *danger of not using randomization*, but instead using some kind of "convenience sample" is that there *may be a bias*, because we get too few of one type of people and too many of another type.

Typical examples are Internet polls and telephone polls during television programs.

– Who are using a certain website at any given day and bother to vote?
– Who are viewing a television program and bother to make a phone call to express an opinion?

This need not be a sample representative of the population!

6.3.1.1 Example

The *Fitness Club* survey could be organized by letting interviewers visit the club and "haphazardly" selecting some of the kids, who are present.

The disadvantage of organizing the survey in this way is that we do not know what type of kids, we select. We will probably get many kids, who are frequent users of the club! Maybe we are interested the other kids also... Perhaps we want to find out, why some of the kids are less frequent users of the club!

6.3.2 Errors in the Definition of the Sample

The ideal situation is that you have a database (a register) of the whole population. This makes it easy to provide a *sampling frame*, from which you select the individuals included in the sample. The sampling frame can be a separate copy of the database, as it looks when the sample is selected.

The individuals that are selected from the sampling frame are called sampling units (*). A sampling unit may be the same as *an individual* in the population. It may also be *a group of individuals* from the population.

The population will often consist of persons. The sampling frame may consist of households. A number of households are selected from the sampling frame. From each household one or more persons are selected. In this context we call a person the *analysis unit*.

An *incomplete sampling frame* is a frequent source of bias in sampling surveys. This means that the sampling frame does not correspond exactly to the population. This can be because:

– Some individuals from the population *cannot be part of the sample*. For instance, people living in institutions cannot be part of the sample, if the sampling frame consists of private households only.
– The sampling frame is not up-to-date. There is often a time lag from the time of selection to the time of the interview. During this period *the population may change*! For example, a person from the sampling frame may die before the interview.
– The sampling frame is incomplete or incorrect for other reasons. A list of households could for example be based on an incomplete or incorrect list of roads and road numbers.

6.3.2.1   Example

Once every month *Fitness Club* prints a list of all the kids, who are using the club. The sample survey is organized by taking a number of kids from this list.

Problems may arise in this context:

– A kid from the list has stopped in the club at the time of interview.
– A new kid has started in the club after the list has been printed.

6.3.3   What Is a Representative Sample?

The term *representative sample* is used in many ways, without being defined. Use of this term should preferable be avoided. If we need a definition, it might be something like this:

A sample can be called representative, if there are only random errors, i.e., no bias.

Under this definition, representative sample surveys do not exist! A representative sample can therefore be seen as an ideal! There will always be bias, but we can do much to reduce it!

> *Important note:*
>
> – *Only the random errors will become smaller, when the sample size increases!*
> – *The bias will not become smaller when the sample increases!*

If, for example, the sampling unit is a household, and we select the first available person from each household, we will often get too many women and too few men,

because women on the average have shorter working hours than men. This will not change by increasing the sample size! We will still have too many women and too few men, regardless of sample size...

6.4   Sampling (Sample Selection)

In this section we describe the main principles of sampling (*) or sample selection. As mentioned earlier, sampling should be based on randomization (*); otherwise it may cause systematic errors. Therefore, the most important methods of sampling are based on randomization.

We also give a short description of some other methods, which are not based on randomization.

6.4.1   Simple Random Sampling

This is the basic method. Simple random sampling (*) is in fact a gigantic lottery!

> The formulas for the determination of the statistical uncertainty from the start of this chapter assume simple random sampling. If the sample has been selected by some other mechanism, the formulas are not correct!

We refer to specialized books on survey sampling, if you need to calculate the statistical uncertainty associated with other methods of sampling.

Nowadays, simple random sampling is done using random numbers in statistical software or a spreadsheet.

In a spreadsheet, you can use the function RAND. This function is used without any parameters:

=RAND()

This provides a random number between 0 and 1.

> The general approach for selecting n sampling units from a sampling frame can be summarized as follows:
>
> 1. Use the function RAND for each sampling unit in the sampling frame.
> 2. Sort all sampling units according to the value of the random numbers.
> 3. Select the first n sampling units, where n is the required sample size.

6.4.2   Stratified Sampling

Stratified sampling (*) is a method of dividing the population into homogeneous groups, called *strata* (singular: stratum). Within each group simple random sampling is used!

> The main reason to do *stratification* is that we can reduce the statistical uncertainty significantly. Conversely, *we can reduce the sample size* significantly without increasing the statistical uncertainty! This happens, *if the strata are homogeneous*, i.e., the spread within strata is small, while on the other hand, the spread between strata is large.

Let us as a hypothetical (?) example imagine that men and women are 100% divided with respect to the opinion on a particular issue. For instance, all men will answer "Yes" to a certain question, while all women will answer "No". In this situation the sample need not be very large in order to cover the population! In fact, a sample size of 2 is enough (one man and one woman)... There is no further information in a larger sample!

This example is of course entirely hypothetical! However, if the situation has a certain similarity with the hypothetical example, there will still be a huge gain by stratification. This is true if e.g., *most men* will answer "Yes" to the question and *most women* will answer "No".

If we have data from an earlier sample survey, we can use these data to plan a new sample survey. Often, there is one particular variable, which is the most important variable. Then we can carry out some statistical calculations to determine the factors that have the greatest influence on this variable. The statistical techniques used for this purpose are extensions of the methods discussed in Chaps. 5, 7, and 8.

Other reasons for stratification may be:

1. *Administrative reasons*. This might be stratification according to geographical criteria. In this situation, the statistical uncertainty will rarely be reduced much compared to simple random sampling; it may even be larger.
2. *Groups (strata) have an interest in itself*. We want to ensure that all strata are adequately represented in the sample. This is particularly important, if there are small groups in the population, which have our particular interest.
3. *Sampling is conducted according to different principles* in different groups of the population. For example, some people live in institutions rather than ordinary households. The practical problems involved in sampling are very different.

6.4.2.1 Example

In the *Fitness Club* sample survey, stratifying by age will probably be a good idea. There is in virtually every aspect very big differences between a 12-year-old and a 17-year-old... We could have two strata: kids 12–14 years old and kids 15–17 years old. It might also make sense to stratify by sex. It depends on the main purpose of the survey. You can also stratify after both age and sex, for example, using four strata in total.

How many individuals should be selected from each stratum?

Normally, we will select a number of individuals in each stratum, corresponding to its size in the population.

If one group is twice as large as another, we must select twice as many individuals in this group compared to the other.

In some cases we will *over-sample some groups* and under-sample others. For example, some groups are of particular interest. In these cases, we must calculate a *weighted average*, when calculating an average from the sample. That is, we multiply each group average with its weight from the population. If one group in the population is twice as large as another, this average should count twice as much as the other average.

6.4.3 Cluster Sampling

Cluster sampling (*) is based on a sampling frame, consisting of sampling units, which in turn contain several analysis units.

The classic example is *a household, which consists of several people*. We select a number of households by simple random sampling. We can then select one person, all persons or for instance half of the persons from the household. In this context the sampling unit (a household) is often called a *cluster*.

Cluster sampling is used mainly for administrative reasons. You may have access to a sampling frame consisting of households, but not a sampling frame of persons.

The statistical uncertainty associated with cluster sampling will usually be larger than when simple random sampling is used. This happens when the individuals in a cluster are similar.

Assume that we are interested in the television viewing of adult persons. In most households consisting of two adults, they will watch a television program together. If you ask them, which television programs they saw the day before, you

will get the same response from both adults in the household! There is, in other words, no additional information asking both adults as compared to asking only one of them!

Of course, this is a simplified description. For example one of them may be shopping, while the other watches television at home. As long as the population consists of only adults, this description is, however, reasonably correct.

> *In some cases cluster sampling has smaller statistical uncertainty than when simple random sampling is used. This happens if there are large differences between the individuals in a cluster.*

Assume that we still restrict the population to adults, and we are interested in their consumption of sanitary towels. This situation is just the reverse: Households with two adults will most often be one person of each sex; therefore, there will be a very large difference in their consumption of sanitary towels!

6.4.3.1   Example

The sampling of school children in a specific school may be carried out in two stages:

–   First, we select a number of classes by simple random sampling from a sampling frame of all classes.
–   Then we select a number of children from each selected class by simple random selection.

We use cluster sampling, because it only requires lists of children in the classes, which have been selected!

> *How to select sampling units and analysis units?*
> We select a number of sampling units by simple random sampling. This step has (at least) two options:
>
> –   Sampling units are selected with the *same probability*.
> –   Sampling units are selected with a *probability proportional to their size*. The larger a sampling unit (e.g., household or school class), the larger the probability of selection.
>
> From each sampling unit, one or more analysis units are selected. This step has several options:
>
> –   To *select one* analysis unit.
> –   To *select all* the analysis units.
>
> *(continued)*

– To select a number of analysis units *proportional to the size* of the sampling unit. The larger a sampling unit, the more analysis units must be selected.

The topic is very large. We refer to specialized books on the topic.

6.4.4 Systematic Sampling

If you do not have a sampling frame in the form of a database, you can use *systematic sampling*, which despite the name is based on randomization!

Systematic sampling can be seen as a practically feasible approach in situations where simple random sampling is not feasible! The method is best illustrated by an example.

6.4.4.1 Example

In the *Fitness Club* survey, we have a list of all kids in the club. There are 300 in total; we select 30 kids for the sample. We must therefore select every tenth kid from the list.

The only randomization in this sampling approach is the selection of the first kid!

– We choose a random number between 1 and 10, for example 7. We choose kid no. 7 from the list.
– Then we select every tenth kid from the list. This means, that we will select no. 7, 17, 27, etc.

The method can be used to select customers entering a shop, who should participate in a questionnaire survey. If every tenth customer is to be selected, we can use exactly the same approach.

The statistical uncertainty may be larger or smaller, than when simple random sampling is used. It is very difficult to tell a priori, what will be the case.

6.4.5 Quota Sampling

We now briefly discuss some methods of sampling, which are not based on a randomization mechanism. This should be avoided if possible, because we are not able to assess the size of the random error!

Having said that, there are situations where randomization is not feasible. *Quota sampling* is typically used in situations where the interviewer is in e.g., a shopping mall. Simple random sampling of customers is not feasible. *It is not possible to prepare a sampling frame.*

Instead, the interviewer will have a number *quota* to be "filled". Often we use a number of age groups for each sex: e.g., men aged 15–29, 30–44, etc. The interviewer needs a number of persons in each group.

The interviewer has opportunity to "spot" potential candidates for each group, as they appear. In this way it seems that we cover the population in a reasonable "representative" manner. However, we do not have any opportunity to demonstrate, that this is true...

The sample may be "representative" in terms of sex and age. But there may be a number of other criteria, where the sample is biased ("unbalanced"), and we do not know anything about it! At the same time we have no idea about the size of the statistical uncertainty!

6.4.6 Purposive Sampling

This technique is used as an easy way to get *a quick sample*, which is similar to the population. We may have detailed knowledge about the population, so that we can sample a few "typical" sampling units. This is done quite purposively, without any kind of randomization.

If we are lucky (and clever!) we can thereby obtain *a sample, which is very similar to the population*.

The sampling unit will often be an administrative unit consisting of several analysis units.

The drawbacks of this method are the same as for quota sampling.

6.4.6.1 Example

Early exit polls after elections can be produced by selection a few "typical" municipalities, which "reflect" the nation quite precisely. In each municipality we select a few polling stations, using our knowledge about the municipality to "cover" the municipality "representatively". In each polling station we ask a number of voters, how they voted. The selection of these voters could be done by e.g., systematic sampling.

6.4.7 Convenience Sampling

The sample is selected "haphazardly". It may be volunteers, "friends and relatives," etc.

Such sample surveys have no statistical value!

See discussion on Internet polls and telephone polls, etc. earlier in this chapter.

However, this type of samples need not totally be condemned: They can be useful to test (parts of) a questionnaire in order to assess whether the wording of one or more questions needs to be changed. This is often called a *pilot survey*. We are only interested in a qualitative (not a statistical) evaluation of the questionnaire!

In Chap. 5, we discussed how to analyze qualitative data from a sample survey or an experiment.

In this chapter we have discussed various aspects in connection with planning a sample survey or an experiment. In the next two chapters, we describe, how we can analyze quantitative data from a sample survey or an experiment.

Chapter 7
Assessment of Relationship

In many different disciplines you need to assess, *whether there is a relationship* between two variables. This can be in administration, social sciences, economics, industry, and science.

The purpose could be one of the following:

– To get a basic understanding of a subject area
– To find reasons or explanations of phenomena
– To try to predict future developments

We study some techniques to assess a relationship and assess *whether an apparent relationship is real* or just a statistical coincidence.

The technique is called *regression analysis* (*). We will only consider the basic technique, *linear regression*, which assumes that there is *a linear relationship* between two variables, i.e., a plot of *Y* against *X* shows a number of points scattered around a straight line.

One of the variables is the *Y*-variable or *dependent variable*. The other variable is the *X*-variable or the *independent variable*.

The subject is treated fairly briefly. We refer to the literature list, if you want to study the issue thoroughly.

The calculations are quite complicated. You can use an advanced calculator with built-in regression analysis, but you are better off with a spreadsheet or statistical software.

There are also more advanced types of regression analysis, for instance *nonlinear regression* or *multiple regression* (multiple *X*-variables). You can read more in many of the books from the literature list or study the Help facility in your spreadsheet or other statistical software.

This chapter discusses the fundamental concepts of linear regression, with a practical example. We do not show the calculation formulas for the different

statistics. These formulas are only important, if you do not have a spreadsheet (or other software) or an advanced calculator to do the calculations!

Here, we use statistical functions of Microsoft Excel, Open Office Calc, and several other spreadsheets. Microsoft Excel also has additional options in the add-in menu "Data Analysis", under the item "Regression".

It is important to use these techniques critically. Ask questions such as:

– Is there a relationship?
– Is it a linear relationship?
– Is there causality?

Note: Statistics can tell, whether there is *a statistical relationship* between two variables, whether it is linear or possibly more complicated (nonlinear). But statistics cannot tell, whether one phenomenon is the cause of another, i.e., whether there is *a causal relationship*. Here professional knowledge is needed.

Many variables in the social and natural sciences *increase with time*. In this situation, a plot of any variable against any other variable will show a reasonably clear relationship.

However, this is a statistical relationship, not a causal relationship. The real relationship, which is hidden in this way, is *an increase with time for both variables*. This is a fairly common error in connection with interpretation of statistical results.

Assume that the number of storks and the number of children in a given area increase in a certain period of time; you cannot conclude that the storks are coming with the children! This is a kind of false conclusion, we often find in newspaper articles.

The underlying (third) variable need not be time, but often it is.

7.1   Example

Let us consider the height and weight of the 17 boys in the *Fitness Club* survey.

We ask the following question: Is there a (possibly linear) relationship between height (*X*) and weight (*Y*)?

We assume that the weight is dependent on the height. Therefore, we put height as *X* and weight as *Y*.

One purpose could be to identify boys, who weigh too much compared to their height; these boys might be interested in an intensive weight loss program!

The first step is always to plot data.

The following is a plot of weight vs. height. In addition, we can see the straight line that makes the best fit to data. This line is established by the *method of least squares* (*). See the literature list for books with a detailed review of the

method. The line is called the *regression line* (*). See the Help of your spreadsheet how to do this (Fig. 7.1).

A statistical model can describe these data. It describes the weight Y of a randomly chosen boy, knowing his height X, through the equation

$$Y = a + b \cdot X + e$$

If we ignore the term e that is precisely the equation for a straight line. Here

$X =$ Height of a boy
$Y =$ Weight of a boy
$a = $ *Intercept* (on Y-axis) of the regression line
$b = $ *Slope* of the regression line
$e = $ Random variation of Y (for a given value of X), often called *residual*

This is illustrated in Fig. 7.2.
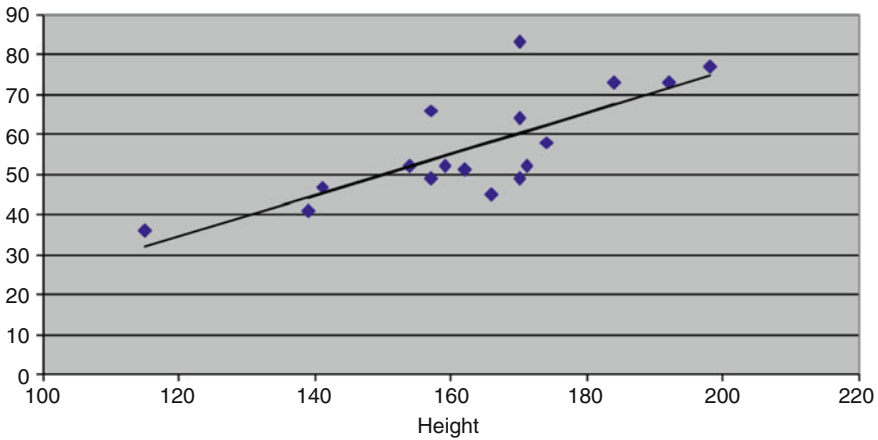

Fig. 7.1 Weight vs. height
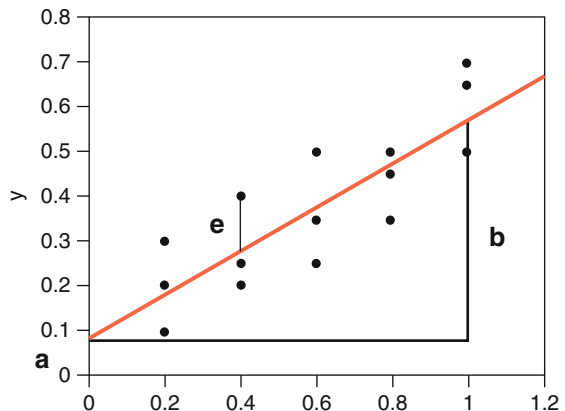
Fig. 7.2 Regression line

Sometimes, *data need to be transformed* in order to obtain a linear relationship. If the points tend to be grouped around a (nonlinear) curve, a seemingly nonlinear relationship can become linear by *transforming Y and/or X*, e.g., with the logarithm function. In the graph with weight plotted against height, there is no immediate sign of a nonlinear relationship.

To describe the degree of (linear) relationship between X and Y we use the *correlation coefficient* (*), often simply called the *correlation* and labeled with the letter r (Table 7.1).

This is illustrated in Fig. 7.3.

Practical situations are usually not that clear-cut! The graph with weight plotted against height is equivalent to a situation where $r > 0$. By visual inspection, most people will probably find that r may be closer to 1 than to 0. We will determine r below.

Sometimes we use r^2, i.e., *the squared value of r*, which is a number between 0 and 1, as a measure of the degree of (linear) relationship between X and Y. It is often written R^2 (i.e., with capital "R") or R-SQUARE. It is also called the *coefficient of determination*. You can say that R^2 expresses how large a part of the variation in Y *is*, which is "explained" by X.

This can be used to compare different models. It can sometimes be difficult to see from a graph, whether we should transform, e.g., the Y-variable or not (typically with the logarithm). If it is difficult to see from the graph, we can choose the model that gives the highest value of R-SQUARE.

Please note that a relatively high value of R-SQUARE is not a guarantee that a linear relationship is an adequate description of data. Always study the plot also!

Table 7.1 Correlation coefficient, r, is a number between -1 and 1. The interpretation is as follows:

$r = -1$	A perfect linear relationship, where the line is tilted downwards
$r = 1$	A perfect linear relationship, where the line is tilted upwards
$r = 0$	No (linear) relationship between X and Y
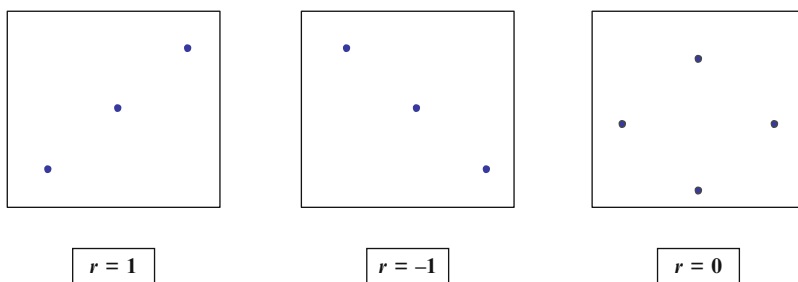
Fig. 7.3 Correlation

7.2 Linear Regression with Spreadsheets

Data for the boys are shown below, but only the first few rows. Data continue until row 18 in the spreadsheet (Fig. 7.4).

To perform linear regression, we use the following spreadsheet functions:

INTERCEPT
SLOPE
CORREL
RSQ
FORECAST

Note: There is another worksheet function called PEARSON to calculate the correlation coefficient. This is just another name for the same function as CORREL; the reason is that the correlation coefficient is often called the *Pearson correlation coefficient*.

Input parameters to the first four functions are the data cells for the *Y* and *X* variables. See column F for the formulas; the result of applying the formulas is given in column E.

Column C shows the *forecasts* of weight, as *predicted by the model*. They correspond to points on the regression line, i.e., we move vertically from a point (up or down), until we hit the regression line.

The predicted values are calculated using the function FORECAST.

Input parameters for this function are the following: First the value of *X*, for which you want a prediction (forecast) of the *Y*-value. Then the relevant range of data cells for *Y* and *X*. Here is, for example, how the content of the cell C2 is programmed (Table 7.2):

	A	B	C	D	E	F
1	Height	Weight	Forecast	Residual	Results	Formula in spreadsheet
2	157	66	53.61	12.39	−27.3148	=INTERCEPT(B2:B18;A2:A18)
3	115	36	31.96	4.04	0.5154	=SLOPE(B2:B18;A2:A18)
4	174	58	62.37	-4.37	0.7645	=CORREL(A2:A18;B2:B18)
5	171	52	60.82	-8.82	0.5845	=RSQ(B2:B18;A2:A18)

Fig. 7.4 Spreadsheet example

Table 7.2 Forecasts
=FORECAST(A2;$B$2:$B$18;$A$2:$A$18)

Note: We have used "absolute references" (dollar signs) to display the range of data cells of *Y* and *X* (see Chap. 5, section on frequency tables). This means that you can copy the contents of cell C2 down over the whole area C2: C18, only the reference to the actual value of *X* will change! See also the Help to your spreadsheet.

We see that the intercept (on *Y*-axis) is negative. We can calculate a 95% confidence interval of this statistic; this confidence interval ranges from -66.7 to 12.1. (This could, for example, be calculated in Microsoft Excel using the add-in menu "Data Analysis", menu-item "Regression".) This means that 0 is in the confidence interval, which makes sense. This corresponds to the fact that a boy of 0 cm weighs 0 kg!

The slope (SLOPE) is about 0.5, representing an increase of 0.5 kg in body weight for each additional cm in height.

The correlation coefficient is about 0.76 (positive, and as expected closer to 1 than 0).

The value of *R*-SQUARE is 0.58; this is the same as the squared value of the correlation coefficient.

A plot of weight against height is shown in the beginning of this chapter, with the regression line.

Column D contains the *residuals*, i.e., the vertical distances between each point and the regression line. They can be calculated as the difference between the columns B (weight) and C (forecast), i.e., they are calculated as Weight-Forecast.

Residuals are very useful for model control. You can examine, if the residuals follow a normal distribution using some of the methods from Chap. 4. This is skipped here.

Moreover, we can plot the residuals against the *X*-variable or other variables. In Figs. 7.5 and 7.6 we show diagrams with the residuals plotted against height, and the predicted value of weight (the "forecast").

Looking at these graphs, there is no obvious "pattern". That is exactly what we hope for: The residuals should show only random variation!
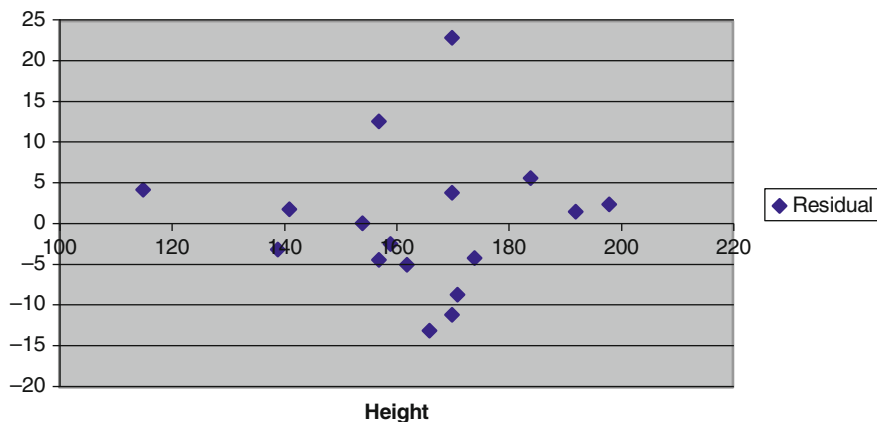

Fig. 7.5 Residual vs. height
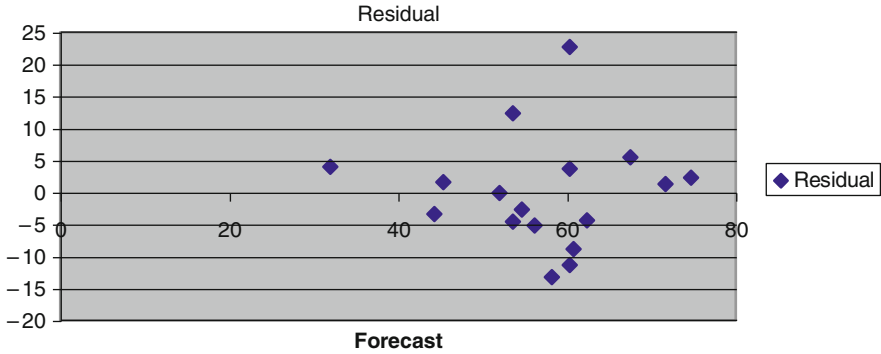
Fig. 7.6 Residual vs. forecast

The residuals can also be used to *identify extreme observations*, for instance boys weighing too much. This could be done by calculating the standard deviation of the residuals; the average of the residuals will always be 0. The standard deviation of the residuals can easily be found to be 8.68; this is the standard deviation for random variation.

The 95% fractile of the normal distribution is 1.645. This can be seen from the table of the normal distribution in Appendices (Chap. 9). Thus, if the residuals roughly follow a normal distribution, values larger than 1.645 times the standard deviation for random variation $= 14.3$ should occur with 5% probability.

This could also be used to identify future boys weighing too much. A future boy customer giving information about his height and weight could be offered an intensive weight loss program, if his weight exceeds the expected weight by more than 14.3 kg. His expected weight can be calculated using the intercept and slope found above as $-27.31 + 0.515 \cdot \text{Height}$.

Similar calculations can of course be done using data for the girls in the survey. It is not a priori certain that the intercept and slope for the girls will be similar to those for the boys.

7.3 Is There a Relationship?

We accept from the above that the relationship between height (as X) and weight (as Y) can be described by a linear relationship. Now we ask the next question: Is the regression line different from a simple average, i.e., could the regression line in fact just as well be horizontal? If so, there is no (linear) relationship between the two variables.

We therefore formulate the following hypothesis:
The line is horizontal, i.e., $b = 0$.
This is the same as: The correlation coefficient between the two variables is 0.

We use the general approach from Chap. 5 for testing a hypothesis:

1. We assume that this hypothesis is true.
2. Calculate the p-value (*), i.e., the probability of getting a more "rare" result.
   It may in this case be shown that one should calculate

$$t = r\sqrt{\frac{n-2}{1-r^2}}.$$

Here, n = number of observations = 17 in the example and r = correlation coefficient = 0.7645. Inserting these values in the formula, we get $t = 4.59$.

This statistics can be described by a t-distribution with $n-2$ degrees of freedom. We subtract 2 from the sample size instead of 1, due to the fact that we must calculate two parameters, intercept and slope.

Here $n = 17$, i.e., there are 15 degrees of freedom.

If $r = 0$, we get $t = 0$. If $r > 0$, we get $t > 0$. If $r < 0$, we get $t < 0$. If r is far from 0, t will also be far from 0.

The hypothesis $r = 0$ is rejected for both negative and positive values of r, which are far from 0. In other words: We reject the hypothesis in "both ends" of the t-distribution.

We can now look up in the table of the t-distribution with 15 degrees of freedom [see Appendices (Chap. 9)] and find that the 99.5% fractile is 2.947. Thus, the probability of getting a larger value than 4.59 is (probably a lot) less than 0.5%. The probability of getting values below -4.59 is also less than 0.5%. In total, the probability of values of t "more extreme" than ± 4.59 is less than 1%.

The following is the density function for a t-distribution with 15 degrees of freedom. It is evident that the value in 4.59 is rather "extreme" in this distribution (Fig. 7.7).
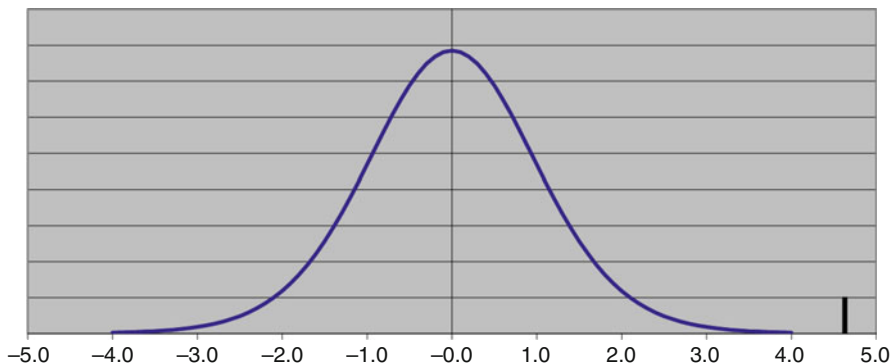

Fig. 7.7 Fractile in t-distribution

3. If this probability is small, reject the hypothesis, otherwise accept it.

Here we have found a probability of more "extreme" values of t less than 1%, i.e., the hypothesis $b = 0$ is rejected. We conclude that there is statistical evidence of a (linear) relationship between height and weight.

Using the Excel add-in menu "Data Analysis", menu-item "Regression", you can create a 95% confidence interval for the slope. This goes from 0.28 to 0.75. Once again, we conclude that the slope is not 0, i.e., the line cannot be horizontal.

7.3.1   Note

The t-test above is exactly the same as the t-test for $b = 0$, which is presented in many other books on statistics, and calculated in the Excel menu "Data Analysis" under "Regression" or in other statistical software.

The advantage of the above formula for t-test is that it is much easier to calculate. It requires no special statistical software. It requires only calculating the correlation coefficient.

The correlation coefficient is easy to calculate in almost all spreadsheets, including Open Office Calc. There are also many advanced calculators, which can calculate the correlation coefficient.

If you have Excel (add-in menu "Data Analysis", menu-item "Regression") or more advanced statistical software, you do not need to calculate the t-statistics manually.

7.4   Multiple Linear Regression

This section can safely be omitted without loss of continuity.

We give a brief example of *multiple linear regression* with two or more X-variables. This can be performed using statistical software or the Microsoft Excel add-in "Data Analysis", item "Regression".

We have seen that the height of the boys has a significant influence on their weight. Suppose we want to investigate the possible relationship of both height and age of kids with their weight. This is exactly, what multiple linear regression does.

In this example, we use all 30 kids, data values for age and height (as X) and weight (as Y). You might want to do this statistical analysis on boys and girls separately first, to ensure that the results are similar; this is actually found to be the case.

If you use the Microsoft Excel add-in "Data Analysis" item "Regression" with this input, you will get the following output. Output from other statistical software packages will look similar (Fig. 7.8).

Regression Statistics						
Multiple R	0.787					
R Square	0.619					
Adjusted R Square	0.591					
Standard Error	8.768					
Observations	30					
ANOVA						
	DF	SS	MS	DF	Significance F	
Regression	2	3372.39	1686.20	21.9	0.0000022	
Residual	27	2075.77	76.88			
Total	29	5448.17				
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	−47.654	17.728	−2.69	0.0122	−84.03	−11.28
Age	2.330	1.108	2.10	**0.0449**	0.06	4.60
Height	0.434	0.075	5.77	**0.0000**	0.28	0.59

Fig. 7.8 Multiple regression

The p-values of age and height are highlighted. It will be seen that both age and height have a statistical significant relationship with weight (though the p-value of age is just below 0.05).

Our interpretation is that age has an influence on weight *in addition to* the influence of height.

Again, residuals from this model could be used for model control, as well as for identifying extreme observations, e.g., kids weighing too much.

This topic is a huge topic, and we refer to more advanced books on statistics for more details.

7.5   Final Remarks

In this chapter, *we have considered the X-variable as a "random" variable* like the *Y*-variable. Therefore, it makes sense to talk about a relationship (i.e., the correlation coefficient) between the two variables.

Sometimes, X is a variable considered "given", such as time. We do not imagine time to vary in a random way. In this situation, there is no sense in talking about "correlation" between *X* and *Y*.

However, *all the calculations above can be performed in exactly the same way*. The *t*-test is now interpreted merely as a test that the line is horizontal, i.e., $b = 0$. It is just the interpretation of the *t*-test that is different!

In this chapter we have discussed how to assess the relationship between two variables. In the next chapter, we discuss another important issue: Comparing two groups.

Chapter 8
Comparing Two Groups

We may for example want to evaluate the effect of physical exercise on the weight of kids. This could be evaluated in two different ways:

- *In a planned experiment*: We select a group of subjects (i.e., kids) and measure their weight. Then they must exercise daily for a period, after which we measure their weight again. We compare the weight before and after the experiment.
- *In a sample survey*: We consider two groups of kids: one group of kids who do not exercise regularly and another group of kids who do. We compare the weight of the kids in the two groups.

These two approaches, which are introduced in this chapter, illustrate the two main techniques of comparing the two sets of data.

The techniques are reviewed on the basis of specific examples. Finally, we mention some extensions of these statistical techniques.

8.1 Matched Pairs: The Paired *t*-Test

8.1.1 *Example*

The girls among kids in the *Fitness Club* survey are selected for an experiment, where they must exercise at least 1 h daily over a period of 4 weeks. Otherwise, they do not change their lifestyle.

The purpose of this experiment is to investigate the potential for intensive weight loss programs among the girl customers.

In this context, we need more precise numbers than can be obtained from asking the kids about their weight. Therefore, their weight before and after the experiment is measured; see data in Table 8.1.

Table 8.1 Data from
experiment

| Before	After	Difference
42	42	0
58	57	1
58	56	2
40	41	−1
49	48	1
80	77	3
50	49	1
48	49	−1
49	47	2
34	33	1
33	32	1
43	43	0
44	42	2
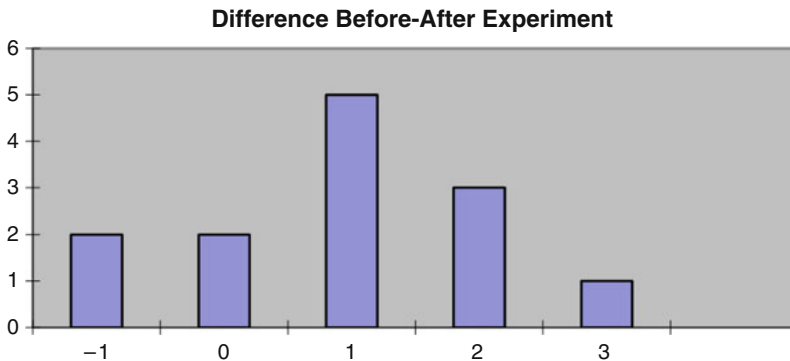
Fig. 8.1 Histogram of differences

At first glance, the difference in weight before and after the experiment seems small. The table does, however, also provide the difference in weight for each kid before and after the experiment. It is seen that in most cases (9 out of 13) there is a small weight loss, but there are also some kids who weigh the same or even slightly more as before the experiment. This is illustrated in Fig. 8.1 showing the histogram of differences.

It is evident from the graph that the "center" of the distribution is to the right of 0.

8.1.2 Description

> We have a number of *pairs of data values*. The two data values in a pair *belong to two different groups*. We are interested in whether there is a difference between the two groups.

The most common application of this technique is in statistical analysis of data from planned experiments. The situation could be the following: we have n individuals,

each of whom has been subjected to two "treatments". We want to examine *whether there is a difference between the two treatments* and possibly *find the average difference*.

This situation is referred to as *matched pairs*.

8.1.3   Calculation

The hypothesis is that *the mean of the differences is 0*, i.e., there is *no difference between the two groups*.

We use the general approach:

1. We assume that the hypothesis is true.
2. Calculate the p value, i.e., the probability of getting a more "rare" result.

    The mean difference is estimated by the average of the differences, which is calculated to be:

$$\bar{d} = 0.92 \text{ kg.}$$

We also calculate the standard deviation of the differences to be $s = 1.19$.

It is natural to relate the average difference to $s/\sqrt{n}$, the estimate of the standard error (see Chap. 4).

We therefore calculate:

$$t = \frac{\bar{d}}{s/\sqrt{n}}.$$

This is called the paired *t*-test.

In the example we get $t = 2.803$.

This statistic follows a *t*-distribution. The *number of degrees of freedom is $n - 1$ because we have n differences* (see Chap. 4, section "Confidence interval for the mean in case of a small sample"). Once we have calculated the differences, the original data are unimportant, i.e., *it is the number of differences that count*.

In this example, there are 13 differences, i.e., the number of degrees of freedom is 12.

If all differences are 0, we get $t = 0$. Values of t close to 0 are "good" for our hypothesis.

Values of t far from 0 are "bad" for the hypothesis. If t is far from 0, we therefore reject the hypothesis. This corresponds to an average difference far from 0.

From the table at the end of the book, we get:

- The 99% fractile of a *t*-distribution with 12 degrees of freedom is 2.681.
- The 99.5% fractile in a *t*-distribution with 12 degrees of freedom is 3.055.

The probability of getting a larger value of *t* is thus between 0.5% and 1%. Normally, we add the probability of getting a value of *t*, which is at least as "far out" to the opposite side. This is just as "bad" for the hypothesis!

The probability of a rarer result is thus between 1% and 2%.

The graph in Fig. 8.2 shows a *t*-distribution with 12 degrees of freedom. It is evident that the value 2.803 is quite "far out" in the distribution.

3. If this probability is small, we reject the hypothesis.

As the probability is less than 2%, we reject the hypothesis. This means that there is statistical evidence that the mean difference in weight before and after the experiment is not 0. In this example, the mean difference is positive, i.e., there is a weight loss.

Now we have demonstrated that there is indeed a difference in weight. The question that follows is: how big is the mean difference?

This question can be answered by calculating a 95% confidence interval for the mean difference; see Chap. 4 about this. The confidence interval is calculated as follows:

$$\bar{d} \pm ts/\sqrt{n}.$$

For *t* we use the 97.5% fractile in a *t*-distribution with $n - 1$ degrees of freedom. This gives us a 95% confidence interval, i.e., with probability 95% the interval contains the true value of the mean difference.

In the table with the *t*-distribution at the end of the book we find the 97.5% fractile in a *t*-distribution with 12 degrees of freedom as 2.179.

If we insert this in the formula, we get the confidence interval 0.923 ± 0.718, i.e., with probability 95% the mean difference is somewhere between 0.205 and 1.641.
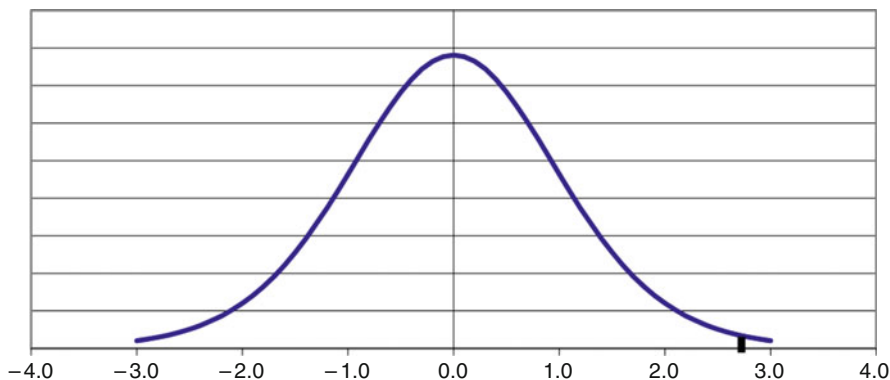

Fig. 8.2 Fractile in *t*-distribution

8.1.4 Spreadsheets

With a spreadsheet, we can directly calculate the *p* value, i.e., the probability of a value of *t* rarer than the above calculated value 2.803, using the TTEST function:

TTEST (Data1; Data2; Tails; Type)

Calculates the *p* value of a *t*-test (Table 8.2).
These data could be in cells A2:A14 ("Before" data), and B2:B14 ("After" data). We now use the function as follows:

=TTEST(A2:A14;B2:B14;2;1)

We use:

– Tails = 2, since we reject the hypothesis in both sides of the distribution.
– Type = 1, since we perform a *t*-test for matched pairs.

The result is $0.016 = 1.6\%$. The probability of a rarer value of *t* is thus app. 1.6%. By comparison, we found out in the above test that the *p* value is between 1% and 2%.

When using the TTEST function, we do not need to calculate the differences! We get the *p* value directly and can compare this with 0.05.

8.2 Comparing Two Groups Means

8.2.1 Example

We want to examine whether there is a difference between the physical fitness of boys and girls in the *Fitness Club* survey. The purpose is to investigate whether potential boy customers and potential girl customers should be addressed in the same way when recruiting new customers for intensive weight loss programs.

Several different parameters are relevant in this context: one could for example compare their weight. However, this would not be appropriate, as a difference may be due to differences in height and/or age.

Table 8.2 TTEST function

Data1	Data cells for group 1
Data2	Data cells for group 2
Tails	Tails = 1 means reject the hypothesis on only one side of the *t*-distribution
	Tails = 2 means reject the hypothesis on both sides of the *t*-distribution, i.e., for both small and large values of *t*; this is the normal situation
Type	Type = 1 means matched pairs (this section)
	Type = 2 means comparing the means of two groups; group standard deviations are required to be identical
	Type = 3 means comparing the means of two groups in general (next section)

Table 8.3 BMI data

| Girls	18.0	28.1	21.4	15.2	17.8	23.7	15.8	21.5	30.4	21.1	23.0	33.5	19.5			
Boys	26.8	19.2	17.8	19.6	23.6	16.3	19.8	22.1	19.4	21.6	19.9	21.2	27.2	20.6	17.0	21.9	28.7
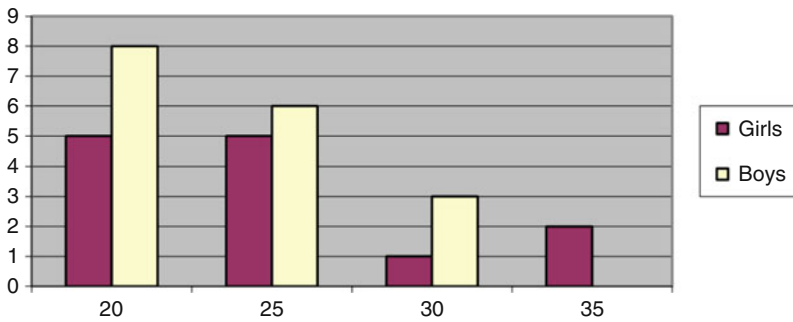
Fig. 8.3 Histogram of BMI data

Therefore, we calculate their body mass index (BMI), i.e.,

$$\text{BMI} = \frac{\text{Weight in kg}}{(\text{Height in meters})^2}.$$

This is an internationally accepted measure. For instance, a person who is 2.00 m tall and weighs 100 kg has a BMI of $100/2^2 = 100/4 = 25$.

– A BMI below 20 is considered to be under normal.
– A BMI of 20–25 is considered normal.
– A BMI of 25–30 is considered overweight.
– A BMI over 30 is considered heavily overweight.

The values here are shown to one decimal place for all kids. We used the questionnaire data on height and weight; see the table with data at the end of the book (Table 8.3).

Figure 8.3 is a combined histogram of BMI for boys and girls.

Immediately, there seems to be no major differences in the distribution of BMI for girls and boys. We want to confirm this using a statistical test.

We start by calculating the average and standard deviation in each group, i.e., for girls and boys separately (Table 8.4).

We might, e.g., name girls "Group 1" ($i = 1$) and boys "Group 2" ($i = 2$).

The average difference (i.e., the difference between the averages) is $22.22 - 21.34 = 0.88$.

8.2.2 Description

This technique can be used for statistical analysis of data from sample surveys and planned experiments.

Table 8.4 Data from two groups

Girls	Boys	Calculation	Notation
22.22	21.34	Mean	$\bar{x}_i$
5.53	3.52	Standard deviation	s_i
13	17	Number of values	n_i
12	16	Degrees of freedom	$n_i - 1$

We have two groups of data values. We are interested in whether there is a difference between the mean of the two groups (and if there is, we want to estimate the mean difference).

The two groups may be two different groups of individuals in a population that we want to compare using a sample survey. Or they might be two groups of individuals, which have been subject to two different treatments in a planned experiment.

8.2.3 Calculation

The hypothesis is that *the mean difference is 0*, i.e., that the two means are identical.

1. *We assume that the hypothesis is true.*
2. *We calculate the p value, i.e., the probability of getting a more "rare" result.*
   Note: Calculation of the *p* value is easy in a spreadsheet; see the next section. We now calculate the following:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}.$$

This statistic contains the average, standard deviation and number of data values for each group.

This is called a *t*-test for two samples with unequal variances. This *t*-test allows the variances in the two groups to be unequal, in contrast to a *t*-test for two samples with equal variances (see later).

In the example we get $t = 0.50$. This value of *t* is to be compared to a fractile in a *t*-distribution. So what is the number of degrees of freedom?

- The number of degrees of freedom can *never be smaller* than the number of degrees of freedom in *the smallest group*.
- The number of degrees of freedom can *never be larger* than the *sum of the number of degrees of freedom* in each group. (This will be the case when both the standard deviation and number of data values are identical in both groups.)

Technical note: Degrees of freedom in t-test for two samples with unequal variances.

There is a fairly complicated formula to determine the precise number of degrees of freedom:

$$f = \frac{\left(\left(s_1^2/n_1\right) + \left(s_2^2/n_2\right)\right)^2}{\left(\left(s_1^2/n_1\right)^2/(n_1 - 1)\right) + \left(\left(s_2^2/n_2\right)^2/(n_2 - 1)\right)}.$$

In the example, we get $f = 19.2$, which is rounded to 19.

In the example, we have minimum 12 degrees of freedom and maximum 28 degrees of freedom. This is in agreement with the value $f = 19$, found using the formula in the box above.

If the two averages are identical, we get $t = 0$! Values of t close to 0 are "good" for the hypothesis.

Values of t far from 0 are "bad" for the hypothesis. If t is far from 0, we will therefore reject the hypothesis.

From the table at the end of the book we get:

The 90% fractile of a t-distribution with 19 degrees of freedom is 1.328.

We have calculated the value of t to be 0.50, which is smaller than 1.328. The probability of getting a larger value of t is therefore (probably much) more than 10%. Normally, we add the probability of getting a value of t, which is at least as "far out" to the opposite side. This is just as "bad" for the hypothesis! The probability of a rarer result is thus more than 20%.

3. If this probability is small, we reject the hypothesis.

Since the observed probability is larger than 20%, we accept the hypothesis.

We can also calculate a 95% confidence interval for the mean difference using the following formula:

$$\bar{x}_1 - \bar{x}_2 \pm t\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

For t we use the 97.5% fractile in a t-distribution. This gives us a 95% confidence interval, i.e., with probability 95% the interval contains the value of the mean difference.

The number of degrees of freedom is determined as shown in the text box; in the example, we get 19 degrees of freedom. In the table with the t-distribution at

the end of the book we find the 97.5% fractile in a *t*-distribution with 19 degrees of freedom to be 2.093.

   If we insert this in the formula, we get the confidence interval 0.88 ± 3.68. With probability 95% the mean difference is somewhere between -2.79 and 3.68. This interval contains 0, in agreement with the fact that the hypothesis is accepted.

Note: If both the groups are large, e.g., more than ten data values, we can without too much error use fractiles in the normal distribution instead of the *t*-distribution, i.e., the 97.5% fractile is app. 2.
   In the example above, the *t*-fractile is 2.09 instead of 1.96, which is not a very large difference.

8.2.4 Spreadsheets

With a spreadsheet, we can directly calculate the *p* value, i.e., the probability of a value of *t* rarer than the above calculated value 0.50 using the TTEST function; see the description earlier in this chapter.

   The data could be in cells B1:N1 (girls), and B2:R2 (boys).
   We now use the function as follows:
   =TTEST(B1:N1;B2:R2;2;3)
   We use:

   – Tails $= 2$, since we reject the hypothesis in both sides of the distribution.
   – Type $= 3$, since we perform a *t*-test for comparing the mean of two groups.

   The result is $0.62 = 62\%$. The probability of a rarer value of *t* is thus app. 62%. By comparison, we found out in the above test that the *p* value is more than 20%.
   When using the TTEST function, we do not need to do any calculations! We get the *p* value directly and can compare this with 0.05.

8.2.5 Size of an Experiment

Let us assume that the variances in both the groups are identical and equal to σ, and the sample sizes are identical and equal to *n* (larger than 10). Then the statistical uncertainty of the difference between the two group means is

$$u = 2\sqrt{\frac{2\sigma^2}{n}}.$$

   This number is just "the number after $\pm$" in the formula for the confidence interval for the difference between two means; see above.

This can be used to determine *the necessary sample size* in order to obtain a *given statistical uncertainty u* of the *difference between two means*. The necessary sample size is

$$n = 8\left(\frac{\sigma}{u}\right)^2.$$

Here, *n* is the necessary sample size *in each group*, i.e., the total sample size is $2 \times n$. In general, if we have more than two groups, *n* is multiplied with the number of groups.

This formula can be used in the same way, as the formula given in Chap. 6. It will most often be used in connection with planning of experiments. Experiments usually involve comparing two or more groups; in most sample surveys, on the contrary, there is usually just one group.

8.3 Other Statistical Tests for Two Groups

8.3.1 Test for the Same Variance in the Two Groups

Sometimes you will be interested in whether the variance (or standard deviation) is the same in both groups. This can be examined by an *F*-test, which we will not cover in detail here.

There is a function FTEST for this purpose in most spreadsheets.

The *F*-test uses an *F*-distribution. This distribution is relatively complicated because it requires two numbers of degrees of freedom (one for each group). For the FTEST function, however, you only need to specify the areas with data for both the groups.

In the example, we have BMI data from girls in cells B1:N1 and from the boys in cells B2:R2. We want to test the hypothesis that the variance (or standard deviation) is the same for girls and boys. We could then use the FTEST function as given in Table 8.5.

The result is the *p* value for the hypothesis that the two variances are equal. This *p* value is found to be 0.093 or 9.3%. Therefore, we accept the hypothesis that the two variances are equal.

8.3.2 Comparing Two Group Means: Two Samples with Equal Variances

There is a third kind of *t*-test: a *t*-test for comparing the mean of two groups, assuming that there is *the same variance* (or standard deviation) in both the groups. This might be examined first by an *F*-test.

Table 8.5 FTEST function

=FTEST(B1:N1;B2:R2)

You can do this *t*-test in a spreadsheet, by selecting Type = 2 in the TTEST function.

The situation where this *t*-test should be considered is the following:

– The two variances (or standard deviations) are virtually identical. Test this with the *F*-test.
– One sample is substantially smaller than the other, with less than ten data values.

In this situation you get more degrees of freedom for this *t*-test, than when using the *t*-test from the last section. This is an advantage, because it will be easier to detect differences that actually exist!

In general, however, there is not much need for this *t*-test!

In the example with the BMI data, we have accepted the hypothesis that the two variances are equal. Thus, we could use this *t*-test (i.e., Type = 2) in this case. Then we get a *p* value of 0.60, i.e., practically the same as before.

8.4   Final Remarks

We have in this chapter studied the two main techniques to compare the two groups. There is another technique that can be seen as an extension of the *t*-test: *Analysis of variance* (*), often abbreviated ANOVA.

ANOVA is, despite the name, used to compare means.

A simple ANOVA example could be comparing several group means (i.e., more than two groups). This is available in Microsoft Excel using the add-in menu "Data Analysis," menu-item "ANOVA: Single factor". Single factor ANOVA is useful when analyzing data from sample surveys as well as planned experiments.

Some situations involve two or several factors, which define the groups. This is an ANOVA with two or more factors and is available in statistical software. See the list of statistical software at the end of the book. In Microsoft Excel, you can do ANOVA with two factors, but only in the simple case, where there is the same number of data values (or just 1) in all of the groups defined by the two factors.

ANOVA with two or more factors is mostly used when analyzing data from planned experiments!

We cannot cover ANOVA in this book. See more advanced books from the literature list. At this point, you have sufficient background to read more advanced books on statistics, if needed.

You have now sufficient knowledge about statistics to go out and use it in practice! I wish you good luck in your further work with statistics!

Chapter 9
Appendices

9.1 Probability Theory

The reader, who is mainly interested in applying statistical methods, can safely skip this appendix.

Probability theory gives the mathematically oriented reader *a better understanding of fundamental statistical concepts*, such as statistical distributions (e.g., the binomial distribution and the normal distribution), i.e., the concepts explained in Chaps. 4 and 5. In particular, one purpose of this appendix is to obtain a better understanding of the binomial distribution.

Only the fundamental concepts of probability theory, which are relevant for explaining the statistical concepts in this book, are explained here. Other books must be consulted for a more thorough introduction to probability theory.

Probability theory was historically founded in medieval times when *analyzing problems in games*, e.g., throwing dice. And even today, most introductions to probability theory use examples from games. This also has the advantage that these examples are (relatively) simple compared to other (maybe more practically relevant) examples.

In this appendix, the basic terms of probability are explained intuitively by examples using only a minimum of mathematical notation.

For a more complete explanation of probability theory, see other books, e.g., Sincich TL, Levine DM, Stephan D, Sincich T and Berenson M (2002) *Practical statistics by example – using Microsoft excel and Minitab*. 2nd ed. Prentice Hall, NJ.

9.1.1 Sample Space, Events, and Probability

When recording an *observation* (for example in a survey) or a *measurement* (for example in an experiment) there are a number of *outcomes* (*).

The set of all possible outcomes is called the *sample space* (*).

A *subset* of the sample space is called an *event* (*).

The *probability* (*) of an event (or an outcome) is *a number between 0 and 1*, which indicates *the likelihood that the event will occur.*

If all outcomes in a finite sample space have the same probability (are equally likely), we have the following:

Probability of event = (# of outcomes of the event)/(# of outcomes of the sample space).

Other words for probability are *risk*, in case of an undesirable outcome, and *chance*, in case of a desirable outcome.

In this appendix, we consider only *finite sample spaces*.

An event may consist of just one outcome or an event may even be empty, i.e., contain no outcomes. On the other hand, an event may cover the whole sample space.

9.1.1.1   Example 1

When throwing a dice, the possible *outcomes* are 1, 2, 3, 4, 5, or 6 eyes. The *sample space* consists of all these six outcomes.

Examples of *events* are:

1. The dice shows an *even* number of eyes, i.e., 2, 4, or 6.
2. The dice shows an *odd* number of eyes, i.e., 1, 3, or 5.
3. The dice shows *at most* 3 eyes.
4. The dice shows *exactly* 3 eyes.

   As the last example shows, an event can consist of just one outcome.
   The events in example 1 and 2 are *complementary events*. This means that

– The two events are *mutually exclusive*, i.e., the two events have *no outcomes in common*.
– The two events together contain *all outcomes* of the sample space.

The sum of the probabilities of two complementary events is 1.

If we can calculate the probability of an event, the probability of the complementary event can thus be calculated by subtraction from 1. Sometimes it is easier to calculate the probability of the complementary event!

9.1.1.2   Example 2

Throwing one dice twice: 36 outcomes; all of them are considered equally likely. Below the total number of eyes is indicated for all outcomes (Fig. 9.1).

Fig. 9.1 Example 1

1. \ 2. throw	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

Fig. 9.2 Example 2

1. \ 2. throw	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

In this example, we study the event: "*The total number of eyes is more than 3.*"
What is the probability of this event?

The complementary event is "The total number of eyes is at most 3." It is easily seen that the number of outcomes in this event is 3 (*shaded area*, see upper left corner of the figure), i.e., the probability of the event is 3/36 = 1/12.

Hence, the probability of the original event is 1 − 1/12 = 11/12.

In this case, calculation of the probability of the complementary event was easier.

9.1.1.3 Example 3

Throwing one dice twice: 36 outcomes; all of them are considered equally likely (Fig. 9.2):

In this example, we study the event: *"The total number of eyes is a multiple of 6."*

It is seen that the total number of outcomes in this event is 6 (*shaded area*). The probability of the event "The total number of eyes is a multiple of 6" is thus 6/36 = 1/6.

The event is equivalent to occurrence of *either* the event "The total number of eyes is 6" (5 outcomes) *or* the event "The total number of eyes is 12" (1 outcome). Their probabilities are seen to be 5/36 and 1/36, respectively. Mathematically, this event is *the union of two mutually exclusive events.*

Another way to find the probability of the event is to add the probabilities of the two separate events; this again gives 5/36 + 1/36 = 6/36 = 1/6.

This illustrates the "*rule of addition*":

The probability of the occurrence of either one or the other of two mutually exclusive events is the sum of the probabilities of the individual events.

9.1.1.4 Example 4

Throwing one dice twice: 36 outcomes; all of them are considered equally likely (Fig. 9.3):

In this example, we study the event: *"The total number of eyes is 12."*

This event contains just 1 outcome (*lower right corner, shaded*). The probability of this event is 1/36.

It can be seen that the event is equivalent to the occurrence of *both* the event "6 eyes in throw 1" *and* the event "6 eyes in throw 2." Both of these events have a probability of $6/36 = 1/6$, as they consist of 6 outcomes (row 6, respectively, column 6, shown in *bold*). Mathematically, this event is the *intersection* of the two separate events.

It is seen that the probability of the event also can be found as the product of the probabilities of the two separate events, i.e., as $1/6 \times 1/6 = 1/36$.

> If *the probability of the intersection* of two events is exactly the *product of the probabilities* of the individual events, the two events are said to be *independent* (*).

This means that *the probability of obtaining six eyes in the second throw does not depend on whether or not we obtained six eyes in the first throw*.

This is also expressed by stating that the *conditional probability* of obtaining six eyes in the second throw, *given* that we obtained six eyes in the first throw is 1/6, i.e., the same as the *unconditional* probability.

No matter what the result of the first throw is, the probability of six eyes in the second throw will still be 1/6.

9.1.1.5 Example 5

Throwing one dice twice: 36 outcomes; all of them are considered equally likely (Fig. 9.4):

In this example, we study the events:

Event A: "Number of eyes in first throw is 1."

Event B: "Total number of eyes is at least 7."

1. \ 2. throw	1	2	3	4	5	6
1	2	3	4	5	6	**7**
2	3	4	5	6	7	**8**
3	4	5	6	7	8	**9**
4	5	6	7	8	9	**10**
5	6	7	8	9	10	**11**
6	**7**	**8**	**9**	**10**	**11**	**12**

Fig. 9.3 Example 4

Fig. 9.4 Example 5

1. \ 2. throw	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

It is seen that event A consists of six outcomes (first row, shown in *bold*), hence the probability of A is $P(A) = 6/36 = 1/6$ (the letter P is used as abbreviation for "Probability").

It is seen that event B consists of 21 outcomes (*shaded area*, diagonal plus lower right half), hence the probability of B is $P(B) = 21/36 = 7/12$.

The *intersection $A \cap B$ of event A and B* consists of just one outcome, the upper right corner. This event has the probability $P(A \cap B) = 1/36$.

As

$$P(A \cap B) = 1/36,$$

and

$$P(A) \times P(B) = (1/6) \times (7/12) = 7/72,$$

we see that

$$P(A) \times P(B) \neq P(A \cap B),$$

Thus, A and B are not independent by the definition in example 4.

Another way to see that A and B are not independent is to calculate the conditional as well as unconditional probabilities:

Two events A and B are *independent*, if *the conditional probability* of event B occurring given that event A occurred is equal to *the unconditional probability* of B occurring.

In other words, *knowledge of occurrence of A has no influence on the probability of B*.

The *conditional probability of event B given that A occurred* is denoted by $P(B|A)$. In the first row, there are six outcomes (event A). Out of these, one outcome is included in B. Hence, we see that $P(B|A) = 1/6$.

The probability of B was found to be 7/12.

As

$$P(B|A) \neq P(B),$$

we once again see that A and B are not independent.

Actually, in the table above, we can easily see how the probability of getting at least seven eyes in total depends on the number of eyes obtained in the first throw.

The larger the number of eyes in the first throw, the larger the probability of getting at least seven eyes in total.

If the number of eyes in the first throw is 6, the probability of at least 7 eyes in total is actually 100%.

If two events are *not independent*, they are said to be *dependent*.

Dependent events occur frequently in real life!

For instance, the probability (or risk) of a person developing lung cancer is dependent on whether or not that person is a smoker: For a nonsmoker, the risk might be 1%; for a smoker, the risk might be as high as 10%.

9.1.2   *Random Variables; the Binomial Distribution*

A *random variable* (*) is a mathematical function on the sample space.

The mathematical function will often be the identity! For instance, in example 1, the sample space consists of the outcomes 1, 2, 3, 4, 5, or 6 (eyes on a dice). The random variable is simply *the number itself*, i.e., the number of eyes shown; *no mathematical operation is done!* This number will vary randomly, hence the term random variable.

In example 2, the total number of eyes in two throws with a dice is a random variable: The sample space consists of 36 pairs of numbers of eyes in each throw. For each of the possible outcomes, the total number of eyes can be calculated by *adding the number of eyes in each throw*.

The result is a mathematical function of the sample space. As the sample space in the example is *finite* (36 outcomes), the random variable is a *discrete random variable*.

A discrete random variable need not have a finite sample space. One example is the number of flashes of lightning in a thunderstorm. There is no upper limit to this random variable; however, the sample space is still discrete, as only *integer values* are possible (0, 1, 2, 3, 4, etc.).

In contrast to a discrete random variable, a *continuous random variable* may take any fractional value; this will be the case with many *measurement data*, where data values can be *any real number* (or any non-negative number). Such data are often described by the *normal distribution*.

In this appendix, we will only cover discrete random variables. The binomial distribution is the most important distribution used to describe discrete random variables.

The *binomial distribution* (*) is used when the following conditions are satisfied:

- *Each observation (or "trial") can be classified into two categories.* Often, we call them "*success*" and "*failure*" regardless of whether one of the categories can be said to be "better" than the other.
- *The probability that an observation is classified as "success" is constant.* For example, in statistical quality control there must not be a trend that defective items become more frequent.
- *The observations are independent.* This means, for example, that two respondents do not affect each others answers in a questionnaire survey.

Notation:

- n is the sample size, i.e., number of observations (trials)
- p is (the constant) probability of "success" in each trial
- $1 - p$ is probability of "failure" in each trial
- *X is a random variable* indicating the number of successes out of n trials
- x is *the actual number* of successes *in a specific sample* of n trials

---

$P(X = x)$ is *the probability of obtaining exactly x successes out of n observations or trials.* In Chap. 5, we showed graphs of this probability. We also showed how to calculate this probability using a spreadsheet.

Here, we will derive the mathematical expression of this probability.

Step 1:

The probability of *x successes in the first x trials* is p^x (where p is the probability of success in each trial), as *the probabilities should be multiplied.* This follows by the fact that the trials are independent.

The probability of obtaining $n - x$ failures in the remaining $n - x$ trials can be calculated in the same way and is found to be $(1 - p)^{n-x}$.

The expression $p^x(1 - p)^{n-x}$ is thus the probability of a certain combination of x successes (each having probability p) and $n - x$ failures (each having probability $1 - p$).

Step 2:

However, there are several different combinations of x successes and $n - x$ failures. And all of them will have the same probability $p^x(1 - p)^{n-x}$.

As *these events are mutually exclusive* (they cannot occur at the same time), *we can use the rule of addition.*

Thus, the probability of x successes in n trials will be $p^x(1-p)^{n-x}$ multiplied by the number of different combinations of x successes and $n-x$ failures.

Step 3:
We need an expression for the number of different combinations of x successes and $n-x$ failures.

The number of combinations (*) of n individuals, when we take a sample of x is often written $\binom{n}{x}$, reading "n over x."

This is also called the *binomial coefficient*. It is tabulated in many textbooks for small values of n. It can also be found in a spreadsheet using the function COMBIN.

A mathematical formula for the binomial coefficient can be found (see technical note at end of this appendix):

$$\binom{n}{x} = \frac{n(n-1)\cdots(n-x+1)}{x(x-1)\cdots 2 \times 1}.$$

For instance, with $n=4$ and $x=2$:

$$\binom{4}{2} = \frac{4 \times 3}{2 \times 1} = \frac{12}{2} = 6.$$

And entering

=COMBIN(4;2)

in a spreadsheet cell gives the result 6.
This gives us the desired formula for the probability of x successes in n trials:

$$P(X=x) = \binom{n}{x}p^x(1-p)^{n-x} = \frac{n(n-1)\cdots(n-x+1)}{x(x-1)\cdots 2 \times 1}p^x(1-p)^{n-x}.$$

9.1.3 Random Variables: Mean and Variance

Let X be a *discrete random variable*.

The *mean* (or *expectation*) of X is defined as

$$E(X) = \sum xP(X = x) = \mu.$$

The *variance* of X is defined as

$$V(X) = \sum (x - \mu)^2 \times P(X = x).$$

The *standard deviation* of X is defined as the square root of the variance, i.e.,

$$\sigma = \sqrt{V(X)}.$$

In the expressions for the mean and variance, *summation is over all possible values x* of the random variable X.

For continuous random variables, the concepts mean and variance can also be defined.

However, *summation should be replaced by integration*, which is defined in the mathematical discipline *calculus*. We will not go further into this; see advanced textbooks on probability theory.

9.1.3.1 Example 6

Let the random variable X be *the number of successes out of n trials*, which follows a binomial distribution. In this case, summation in the expressions for $E(X)$ and $V(X)$ is over *all values from 0 to n*.

We have derived the probabilities $P(X = x)$ above. It can be shown mathematically that inserting

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

in the above expressions for $E(X)$ and $V(X)$ gives the result.

$$E(X) = np,$$
$$V(X) = np(1 - p).$$

These are exactly the expressions for the mean and variance of a binomial distribution shown in Chap. 5.

9.1.4   Technical Note: The Binomial Coefficient

We want to determine the number of *combinations* (or *groups*) of n individuals when taking a sample of x.

First, determine the number of *permutations* (*ordered groups*) of n individuals, when taking a sample of x. In this way, (A, B) and (B, A) are considered two different groups.

For instance, with $n = 4$ persons labeled A, B, C, D, we select ordered groups of $x = 2$ persons. In order to find out how many ordered groups exist, we *first* select person no. 1. This can be done in $n = 4$ ways. *Then* we select person no. 2. With three persons left, this can be done in $n - 1 = 3$ ways.

In total, we can select $4 \times 3 = 12$ ordered groups.

> *Generally, the number of permutations of n objects, when taking a sample of x objects, is equal to $n(n - 1) \cdots (n - x + 1)$.*
> *In this expression, the total number of factors is x.*

This number can be found in spreadsheets using the function PERMUT. For instance, entering

=PERMUT(4;2)

in a spreadsheet cell gives the result 12.

From the number of permutations, we can find the number of combinations.

In the example, we have counted (A, B) and (B, A) as two different groups; we determined the number of permutations to be 12. In order to find the number of combinations of four objects, when taking a sample of 2, we divide 12 by 2, obtaining 6.

In the general case, we divide the number of permutations $n(n - 1) \cdots (n - x + 1)$ by the number of permutations of x individuals, which is $x(x - 1) \cdots 2 \times 1$.

Thus, we obtain *the number of combinations of n individuals, when taking a sample of x*:

$$\binom{n}{x} = \frac{n(n - 1) \cdots (n - x + 1)}{x(x - 1) \cdots 2 \times 1}.$$

With $n!$ (read "n factorial") meaning all the numbers 1, 2, etc., up to n multiplied together, the binomial coefficient can also be written as

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}.$$

9.2 Summary of Statistical Methods

Important points to clarify when doing a statistical analysis:

1. Quantitative or qualitative data?
2. One group or two groups?
3. Confidence interval or statistical test?

9.2.1 Quantitative Data

9.2.1.1 Descriptive Statistics

See Chap. 3 (Table 9.1).

9.2.1.2 The Normal Distribution

See Chap. 4.

Probabilities (Table 9.2).

Table 9.1 Descriptive statistics

	Symmetric distribution	Skewed distribution
Location (center)	Average $\bar{x} = \dfrac{\Sigma x_i}{n}$	Median $Q2 =$ Data value "in the middle"
Dispersion (spread)	Standard deviation $s = \sqrt{\dfrac{\Sigma(x_i - \bar{x})^2}{n-1}}$	Interquartile range IQR $= Q3 - Q1$ $Q3 =$ Upper quartile $Q1 =$ Lower quartile

Table 9.2 Normal distribution

Interval	Probability
Mean $\pm$ 1 standard deviation	68% of data values
Mean $\pm$ 2 standard deviations	95% of data values
Mean $\pm$ 3 standard deviations	99.7% of data values

Table 9.3 Confidence intervals: One group

95% confidence interval for the mean: Large sample or standard deviation known.	$\bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}}$
95% confidence interval for the mean: Small sample and standard deviation unknown. Use 97.5% fractile from *t*-distribution, Degrees of Freedom $= n - 1$.	$\bar{x} \pm t \frac{s}{\sqrt{n}}$
95% confidence interval for the standard deviation: Use fractiles from Chi-squared distribution, Degrees of Freedom $= n - 1$.	$\sqrt{\dfrac{(n-1)s^2}{\chi^2_{97.5\%}}} \leq \sigma \leq \sqrt{\dfrac{(n-1)s^2}{\chi^2_{2.5\%}}}$

Testing for the Normal Distribution

1. Simple methods

    – The histogram
    – The average $=$ the median
    – Interquartile range larger than the standard deviation
    – Number of data values in symmetric intervals around the mean

2. Skewness and kurtosis

    – Calculate in spreadsheet
    – Reasonably close to 0? Compare with min. and max. limits in Chap. 4.

3. Normal plot

9.2.1.3 Confidence Intervals: One Group

See Chap. 4 (Table 9.3).

9.2.1.4 Confidence Intervals: Two Groups

See Chap. 8 (Table 9.4).

9.2.1.5 Sample Size

See Chap. 6.

Table 9.4 Confidence intervals: Two groups

Matched pairs (paired t-test): 95% confidence interval for mean difference. Use 97.5% fractile from *t*-distribution, Degrees of Freedom $= n - 1$.	$\bar{d} \pm ts/\sqrt{n}$
Comparing two group means: 95% confidence interval for difference between two group means. Use 97.5% fractile from *t*-distribution.	$\bar{x}_1 - \bar{x}_2 \pm t\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}$
Degrees of Freedom: Formula available.	

Table 9.5 Statistical tests: Two variables or two groups

Two variables: Test, that correlation $= 0$ (or slope $= 0$). Correlation coefficient (*r*) calculated with spreadsheet. Compare with 97.5% fractile from *t*-distribution, Degrees of Freedom $= n - 2$.	$t = r\sqrt{\dfrac{n-2}{1-r^2}}$
Matched pairs (paired t-test): Test, that mean difference $= 0$. Compare with 97.5% fractile from *t*-distribution, Degrees of Freedom $= n - 1$.	$t = \dfrac{\bar{d}}{s/\sqrt{n}}$
Comparing two group means: Test, that two means are equal. Compare with 97.5% fractile from *t*-distribution, Degrees of Freedom: Formula available.	$t = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$

Table 9.6 Confidence intervals: One group

Confidence interval for proportion (p) of binomial distribution: *x* is number of observations (out of *n*) of interest, $p = x/n$.	$\dfrac{x}{n} \pm 1.96\sqrt{\dfrac{p(1-p)}{n}}$

If *we know the standard deviation σ* and *want a maximum statistical uncertainty u* of the average, we find the *minimum necessary sample size n* as

$$n = \left(2\frac{\sigma}{u}\right)^2.$$

If there are two (or more) groups, this is the necessary sample size *in each group*.

9.2.1.6 Statistical Tests: Two Variables or Two Groups

See Chap. 7 and 8 (Table 9.5).

9.2.2 Qualitative Data

9.2.2.1 Confidence Intervals: One Group

See Chap. 5 (Table 9.6).

9.2.2.2 Confidence Intervals: Two Groups

See Chap. 5 (Table 9.7).

9.2.2.3 Sample Size

See Chap. 6.

If the *maximum statistical uncertainty of a proportion* is *u*, we find the *minimum necessary sample size n* as

$$n = \frac{1}{u^2}$$

The above formula can obviously be used in subgroups of the population. The formula then finds the value of *n* for each subgroup separately.

9.2.2.4 Statistical Tests: Two Groups or Two Variables

See Chap. 5 (Table 9.8).

9.3 Statistical Functions in Spreadsheets

This is an overview of the most important statistical (and a few mathematical) functions in Microsoft Excel and the free spreadsheet OpenOffice Calc. All these functions have the same name and work in the same way in both spreadsheets (and in some other spreadsheets).

Note: The names of these functions are translated when using spreadsheets in other languages than English!

Table 9.7 Confidence intervals: Two groups

Confidence interval for difference between proportions:
  (p_1 and p_2) from two binomial distributions: p is $p_1 - p_2 \pm 1.96\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$
  the proportion in the whole sample.

Table 9.8 Statistical tests: Two groups or two variables

Test, that proportions (p_1 and p_2) from two binomial distributions are equal: *O* Observed frequency, *E* Expected frequency. Compare to fractiles from $\chi^2 = \sum \frac{(O-F)^2}{F}$ Chi-squared distribution, Degrees of Freedom $= 1$.

Frequency table:
Test of independence between rows and columns. O Observed frequency, $\chi^2 = \sum \frac{(O-F)^2}{F}$
  E Expected frequency. Compare to fractiles from Chi-squared distribution,
  Degrees of Freedom $= (\#\text{Rows} - 1) \times (\#\text{Columns} - 1)$.

Table 9.9 Statistical functions

Function	Short description
AVERAGE	Gives the average of the data values.
BINOMDIST	Gives a probability of the binomial distribution.
CHIDIST	Gives the distribution function of a Chi-squared distribution.
CHIINV	Gives fractiles in a Chi-squared distribution.
CHITEST	Gives the p-value of a test of independence in a frequency table.
CONFIDENCE	Gives a confidence interval of the mean, when the standard deviation is known.
CORREL	Gives the correlation coefficient between two variables.
CRITBINOM	Gives the critical value of a binomial distribution.
FORECAST	Predicts a y value based on an x value and a linear regression model of the data values.
FTEST	Gives the p-value of an F-test for comparing two variances.
INTERCEPT	Gives the intercept on the y-axis in linear regression.
KURT	Gives the kurtosis.
LOG	Gives the logarithm to base 10 of a number.
MAX	Gives the largest data value.
MEDIAN	Gives the median.
MIN	Gives the smallest data value.
MODE	Gives the mode, i.e., the data value with the largest frequency.
NORMDIST	Gives the distribution function of a normal distribution.
NORMINV	Gives fractiles in a normal distribution.
NORMSDIST	Gives the distribution function of a standardized normal distribution.
NORMSINV	Gives fractiles in a standardized normal distribution.
PERCENTILE	Gives any fractile in a set of data values.
QUARTILE	Gives quartiles in a set of data values.
RSQ	Gives the square of the correlations coefficient.
SKEW	Gives the skewness.
SLOPE	Gives the slope in linear regression.
SQRT	Gives the square root of a number.
STANDARDIZE	Standardization: Subtraction of the mean followed by division by the standard deviation.
STDEV	Gives the standard deviation.
TDIST	Gives the distribution function of a t-distribution.
TINV	Gives fractiles in a t-distribution.
TTEST	Gives the p-value in a t-test for comparing two means. Includes paired t-test.
VAR	Gives the variance.
ZTEST	Gives the p-value in a test for comparing the mean with a known value, when the standard deviation is known.

We refer to the "Help" menu of your spreadsheet for more details (Table 9.9).

9.4 Statistical Tables

9.4.1 Fractiles in the Normal Distribution

Fractiles in the (standard) normal distribution are calculated in Microsoft Excel/ OpenOffice Calc using the function NORMSINV.

Table 9.10 Fractiles in the normal distribution

50%	60%	70%	75%	80%	90%	95%	97.5%	99%	99.5%	99.9%	99.95%
0.0000	0.2533	0.5244	0.6745	0.8416	1.2816	1.6449	1.9600	2.3263	2.5758	3.0902	3.2905

Table 9.11 Probabilities in the normal distribution

−3	−2.5	−2	−1.5	−1	−0.5	0	0.5	1	1.5	2	2.5	3
0.0013	0.0062	0.0228	0.0668	0.1587	0.3085	0.5000	0.6915	0.8413	0.9332	0.9772	0.9938	0.9987

Example: For the 97.5% = 0.975 fractile, we obtain NORMSINV(0.975) = 1.960 (Table 9.10).

9.4.2 Probabilities in the Normal Distribution

Probabilities in the (standard) normal distribution are calculated in Microsoft Excel/ OpenOffice Calc using the function NORMSDIST.

Example: NORMSDIST(2) = 0.9772. Thus, the probability of values ≤ 2 is 97.72% (Table 9.11).

9.4.3 Table of the t-Distribution

Note: The number of degrees of freedom in a sample is $n - 1$, where n is the sample size.

If the number of degrees of freedom is more than 30, you can with very good approximation use fractiles of the normal distribution instead.

Fractiles in the *t*-distribution are calculated in Microsoft Excel or OpenOffice Calc using the function TINV. Notice the peculiar way to specify the probability: Find the "rest" probability and multiply by 2.

Example: For the 97.5% = 0.975 fractile, the "rest" probability is 2.5% = 0.025. When multiplied by 2, we get 5% = 0.05. With, for example, 9 degrees of freedom we obtain the fractile as TINV(0.05;9) = 2.262 (Table 9.12).

9.4.4 Table of the Chi-Squared Distribution

Fractiles in the Chi-squared distribution are calculated in Microsoft Excel or OpenOffice Calc using the function CHIINV.

Notice that you should specify the "rest" probability rather than the probability itself.

Table 9.12 Fractiles in the *t*-distribution

Degrees of freedom	90%	95%	97.5%	99%	99.5%
1	3.078	6.314	12.706	31.821	63.656
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
30	1.310	1.697	2.042	2.457	2.750

Example: For the 97.5% = 0.975 fractile, the "rest" probability is 2.5% = 0.025. With, for example, 9 degrees of freedom we obtain the fractile as CHIINV (0.025;9) = 19.02 (Table 9.13).

9.4.5 Statistical Uncertainty in Sample Surveys

This table can be used for questionnaire data with two answer categories, e.g., "Yes/No."

The table shows the statistical uncertainty of the result of a sample survey.

The number in the table is "the number after ±." It is used to construct a 95% confidence interval.

Simple random sampling is assumed.

By stratified sampling, the statistical uncertainty will often be smaller.

Table 9.13 Fractiles in the chi-squared distribution

Degrees of freedom	0.5%	1.0%	2.5%	5.0%	95.0%	97.5%	99.0%	99.5%
1	0.00	0.00	0.00	0.00	3.84	5.02	6.63	7.88
2	0.01	0.02	0.05	0.10	5.99	7.38	9.21	10.60
3	0.07	0.11	0.22	0.35	7.81	9.35	11.34	12.84
4	0.21	0.30	0.48	0.71	9.49	11.14	13.28	14.86
5	0.41	0.55	0.83	1.15	11.07	12.83	15.09	16.75
6	0.68	0.87	1.24	1.64	12.59	14.45	16.81	18.55
7	0.99	1.24	1.69	2.17	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	15.51	17.53	20.09	21.95
9	1.73	2.09	2.70	3.33	16.92	19.02	21.67	23.59
10	2.16	2.56	3.25	3.94	18.31	20.48	23.21	25.19
11	2.60	3.05	3.82	4.57	19.68	21.92	24.73	26.76
12	3.07	3.57	4.40	5.23	21.03	23.34	26.22	28.30
13	3.57	4.11	5.01	5.89	22.36	24.74	27.69	29.82
14	4.07	4.66	5.63	6.57	23.68	26.12	29.14	31.32
15	4.60	5.23	6.26	7.26	25.00	27.49	30.58	32.80
16	5.14	5.81	6.91	7.96	26.30	28.85	32.00	34.27
17	5.70	6.41	7.56	8.67	27.59	30.19	33.41	35.72
18	6.26	7.01	8.23	9.39	28.87	31.53	34.81	37.16
19	6.84	7.63	8.91	10.12	30.14	32.85	36.19	38.58
20	7.43	8.26	9.59	10.85	31.41	34.17	37.57	40.00
21	8.03	8.90	10.28	11.59	32.67	35.48	38.93	41.40
22	8.64	9.54	10.98	12.34	33.92	36.78	40.29	42.80
23	9.26	10.20	11.69	13.09	35.17	38.08	41.64	44.18
24	9.89	10.86	12.40	13.85	36.42	39.36	42.98	45.56
25	10.52	11.52	13.12	14.61	37.65	40.65	44.31	46.93
26	11.16	12.20	13.84	15.38	38.89	41.92	45.64	48.29
27	11.81	12.88	14.57	16.15	40.11	43.19	46.96	49.65
28	12.46	13.56	15.31	16.93	41.34	44.46	48.28	50.99
29	13.12	14.26	16.05	17.71	42.56	45.72	49.59	52.34
30	13.79	14.95	16.79	18.49	43.77	46.98	50.89	53.67
31	14.46	15.66	17.54	19.28	44.99	48.23	52.19	55.00
32	15.13	16.36	18.29	20.07	46.19	49.48	53.49	56.33
33	15.82	17.07	19.05	20.87	47.40	50.73	54.78	57.65
34	16.50	17.79	19.81	21.66	48.60	51.97	56.06	58.96
35	17.19	18.51	20.57	22.47	49.80	53.20	57.34	60.27
36	17.89	19.23	21.34	23.27	51.00	54.44	58.62	61.58

By cluster sampling, the statistical uncertainty will often be larger (Table 9.14). Example:

A result (e.g., percentage answering "Yes" to a question) in a sample survey is 25%; the sample size is 500. The statistical uncertainty of the result is found in the table to be ±3.8%.

This means that if interviewing the whole population, the result would with 95% probability be in the interval 25% ± 3.8%, i.e., an interval from 21.2% to 28.8%.

Note: You get the same statistical uncertainty, if the result in the sample survey is 75%.

9.5   Fitness Club: Data from the Sample Survey

Data from the example used throughout the book. Data are sorted by sex and age (Table 9.15).

Table 9.14 Statistical uncertainty in sample surveys

Sample size	Result in percent											
	1%	3%	5%	10%	15%	20%	25%	30%	35%	40%	45%	50%
	99%	97%	95%	90%	85%	80%	75%	70%	65%	60%	55%	50%
50	2.8%	4.7%	6.0%	8.3%	9.9%	11.1%	12.0%	12.7%	13.2%	13.6%	13.8%	13.9%
100	2.0%	3.3%	4.3%	5.9%	7.0%	7.8%	8.5%	9.0%	9.3%	9.6%	9.8%	9.8%
150	1.6%	2.7%	3.5%	4.8%	5.7%	6.4%	6.9%	7.3%	7.6%	7.8%	8.0%	8.0%
200	1.4%	2.4%	3.0%	4.2%	4.9%	5.5%	6.0%	6.4%	6.6%	6.8%	6.9%	6.9%
300	1.1%	1.9%	2.5%	3.4%	4.0%	4.5%	4.9%	5.2%	5.4%	5.5%	5.6%	5.7%
400	1.0%	1.7%	2.1%	2.9%	3.5%	3.9%	4.2%	4.5%	4.7%	4.8%	4.9%	4.9%
500	0.9%	1.5%	1.9%	2.6%	3.1%	3.5%	3.8%	4.0%	4.2%	4.3%	4.4%	4.4%
600	0.8%	1.4%	1.7%	2.4%	2.9%	3.2%	3.5%	3.7%	3.8%	3.9%	4.0%	4.0%
700	0.7%	1.3%	1.6%	2.2%	2.6%	3.0%	3.2%	3.4%	3.5%	3.6%	3.7%	3.7%
800	0.7%	1.2%	1.5%	2.1%	2.5%	2.8%	3.0%	3.2%	3.3%	3.4%	3.4%	3.5%
900	0.7%	1.1%	1.4%	2.0%	2.3%	2.6%	2.8%	3.0%	3.1%	3.2%	3.3%	3.3%
1000	0.6%	1.1%	1.4%	1.9%	2.2%	2.5%	2.7%	2.8%	3.0%	3.0%	3.1%	3.1%
1250	0.6%	0.9%	1.2%	1.7%	2.0%	2.2%	2.4%	2.5%	2.6%	2.7%	2.8%	2.8%
1500	0.5%	0.9%	1.1%	1.5%	1.8%	2.0%	2.2%	2.3%	2.4%	2.5%	2.5%	2.5%
1750	0.5%	0.8%	1.0%	1.4%	1.7%	1.9%	2.0%	2.1%	2.2%	2.3%	2.3%	2.3%
2000	0.4%	0.7%	1.0%	1.3%	1.6%	1.8%	1.9%	2.0%	2.1%	2.1%	2.2%	2.2%
3000	0.4%	0.6%	0.8%	1.1%	1.3%	1.4%	1.5%	1.6%	1.7%	1.8%	1.8%	1.8%
4000	0.3%	0.5%	0.7%	0.9%	1.1%	1.2%	1.3%	1.4%	1.5%	1.5%	1.5%	1.5%
5000	0.3%	0.5%	0.6%	0.8%	1.0%	1.1%	1.2%	1.3%	1.3%	1.4%	1.4%	1.4%
6000	0.3%	0.4%	0.6%	0.8%	0.9%	1.0%	1.1%	1.2%	1.2%	1.2%	1.3%	1.3%
7000	0.2%	0.4%	0.5%	0.7%	0.8%	0.9%	1.0%	1.1%	1.1%	1.1%	1.2%	1.2%
8000	0.2%	0.4%	0.5%	0.7%	0.8%	0.9%	0.9%	1.0%	1.0%	1.1%	1.1%	1.1%
9000	0.2%	0.4%	0.5%	0.6%	0.7%	0.8%	0.9%	0.9%	1.0%	1.0%	1.0%	1.0%
10000	0.2%	0.3%	0.4%	0.6%	0.7%	0.8%	0.8%	0.9%	0.9%	1.0%	1.0%	1.0%

Table 9.15 Fitness Club data

30 randomly chosen kids from fitness club

No.	Sex	Age (years)	Height (cm)	Weight (kg)
6	F	12	145	59
20	F	12	151	49
26	F	12	118	32
7	F	13	166	59
10	F	13	160	39
2	F	14	151	41
12	F	14	166	49
15	F	14	185	81
18	F	14	176	49

(*continued*)

Table 9.15 (continued)

No.	Sex	Age (years)	Height (cm)	Weight (kg)
— 30 randomly chosen kids from fitness club

No.	Sex	Age (years)	Height (cm)	Weight (kg)
25	F	14	125	33
30	F	15	152	45
24	F	16	127	49
28	F	17	112	42
1	M	12	157	66
21	M	12	115	36
3	M	13	174	58
4	M	13	171	52
8	M	13	141	47
9	M	13	166	45
14	M	14	162	51
17	M	14	157	49
19	M	14	139	41
22	M	14	159	52
23	M	14	170	49
5	M	15	198	77
11	M	15	192	73
27	M	15	154	52
13	M	16	170	64
16	M	17	184	73
29	M	17	170	83

9.6   Where to Go from Here

9.6.1   Literature

The following books can be recommended:

Darrell Huff (1991). *How to Lie with Statistics*. Penguin Books
A classic. Put it in your pocket and read it in the bus or in the train. . .

Larry Gonick and Woolcott Smith (1993). *The Cartoon Guide to Statistics*.
Harperperennial
Statistics as a cartoon!

Sincich TL, Levine DM, Stephan D, Sincich T and Berenson M (2002) 2nd ed
Practical Statistics by Example – using Microsoft Excel and Minitab. Prentice Hall.
Plenty of examples in virtually all disciplines.
Both for users of Microsoft and Excel and the widely used statistics software Minitab

Agresti and Finlay B (1997). *Statistical methods for the social sciences*. 3rd ed.
Prentice Hall, NJ
Detailed book on statistics for the social sciences.

Vic Barnett (2003). *Sample survey principles and methods*, 3rd ed. Arnold Publishers, New Delhi
An excellent book on sample surveys.

Groves RM (2004). *Survey errors and survey costs*. Wiley, NJ
Thorough book on the practical aspects of sample surveys. Not mathematically advanced.

Cochran WG (1978): *Sampling techniques*, 3rd ed. Wiley, NJ
Still the bible on survey sampling!

Cox DR (1992) *Planning of experiments*. Wiley, NJ
Elementary, yet thorough book on planned experiments. Focus is on applications, not on the theory.
Not mathematically advanced.

Cochran WG and Cox GM (1992) 2nd ed. *Experimental design*. Wiley, NJ
Planning of experiments for practitioners. Tables with specific experimental designs.

Box GEP, Hunter WG and Hunter JS (2005) *Statistics for experimenters*, 2nd ed. Wiley, NJ
Excellent book on statistics with an emphasis on planning of experiments and analyzing the results, but it is useful for most people working with statistics. A legendary book!

Douglas Montgomery (2005). *Introduction to statistical quality control*, 6th ed. Wiley, N.
Basic statistics with a thorough introduction to statistical quality control.

Douglas Montgomery (2005). *Design and analysis of experiments*, 6th ed. Wiley, NJ
Design and statistical analysis of experiments, analysis of variance, regression analysis, etc.
Somewhat higher level of mathematics than the other books in this list.

9.6.2 Useful Links

Table 9.16 Useful links

Statistics about society	
Eurostat European statistics	http://epp.eurostat.ec.europa.eu/
OECD statistics	http://www.oecd.org/statsportal/
UN Statistics Division UN/ECE Stat. Division	http://unstats.un.org/unsd http://www.unece.org/stats

Table 9.16 (continued)

Statistical organizations	
International Statistical Institute Many useful links, e.g.: Glossary of Statistical Terms	http://www.isi-web.org/ http://www.isi-web.org/glossary.htm
European Network for Business and Industrial Statistics	http://www.enbis.org
American Statistical Association	http://www.amstat.org
American Society for Quality	http://www.asq.org
Other useful links	
Consortium for the advancement of undergraduate statistics education	http://www.causeweb.org Many links, articles, data, etc.
Statpages	http://statpages.org Overview of statistical software, books, demos, links etc.
Statistical software providers	http://www.stata.com/links/stat_software.html Comprehensive overview of statistical software.
Electronic Statistics Textbook	http://www.statsoft.com/textbook/stathome.html Electronic textbook in statistics. Includes Statistics Glossary.
Computer-Assisted Statistics Textbooks	http://cast.massey.ac.nz/collection_public.html
Java Applets for visualization of statistical concepts	http://ucs.kuleuven.be/java/index.htm Visualizing statistical concepts.
Environments to experiment	http://lstat.kuleuven.be/env2exp/index.html

9.6.3 Overview of Statistical Software

Table 9.17 Statistical software

Open Office Free Office suite! With the spreadsheet Calc.	http://www.openoffice.org
SAS Several modules, very comprehensive system. *JMP* General statistical software. Particularly for industrial applications.	http://www.sas.com http://www.jmp.com
SPSS Several modules, very comprehensive system. Particularly for questionnaires, etc.	http://www.spss.com
Minitab General statistical software. Particularly for industrial applications.	http://www.minitab.com
Stata Particularly for questionnaires, etc.	http://www.stata.com Comprehensive list of statistical software!
Statistica Several modules, very comprehensive system.	http://www.statsoft.com Online textbook in statistics!

Table 9.17 (continued)

Genstat General statistical software.	http://www.vsni.co.uk
S-Plus General statistical software.	http://www.insightful.com
Statistical Solutions: Several programs, e.g., − BMDP − nQuery − SOLAS	http://www.statsol.ie General statistical software. Calculation of the necessary sample size. Handling and estimating missing data values.
Systat General statistical software.	http://www.systat.com Relatively cheap program.
Instat	http://www.graphpad.com Very cheap. Only small amounts of data.
Free statistical software	http://statpages.org/javasta2.html

9.7   Glossary

Term	Explanation
Alternative hypothesis	The opposite hypothesis of the null hypothesis. Is true, when the null hypothesis is false.
Analysis of Variance ANOVA	A technique that partitions the total variation in components caused from one or more groupings of data. It can, for instance, be used to test, if several group means are identical.
Average	The average is calculated as the sum of all data values divided by their number and it is calculated for the data values of a sample or an experiment. The average of the population is called the mean.
Bias Systematic error	The difference between the true value and the mean due to specific (known or unknown) causes, e.g., nonresponse in a sample survey.
Binomial distribution	A statistical distribution used to describe the probability that x individuals in a sample of size n have a certain characteristic. The probability that one single individual has the characteristic is constant. The observations from several individuals are independent.
Chi-squared distribution	A statistical distribution that takes on positive values only. The number of degrees of freedom must be specified and it is used, for example: − Test for independence in frequency tables. − Confidence interval for a variance.
Cluster sampling	The population is divided in clusters, each consisting of several sampling units. A number of clusters are selected at random. Large clusters may be selected with larger probability than small clusters. Within each cluster, one or more (or all) sampling units are selected.

Term	Explanation
Coefficient of correlation Correlation	The degree of (linear) relationship between two variables. A number between -1 and 1. A value of -1, and 1 respectively corresponds to a linear relationship (with negative, and positive slope respectively). 0 corresponds to no (linear) relationship.
Coefficient of variation CV	The standard deviation expressed as a percentage of the average. Also called the relative standard deviation (RSD).
Confidence interval	An interval, which with a given probability, e.g., 95% or 99%, contains the (true) population value of a parameter, e.g., a mean.
Critical value	Fractile (typically 95% or 97.5%) of a distribution (e.g., Chi-squared distribution or t-distribution) used to compare with a sample statistic in a statistical test to determine whether to accept or reject the null hypothesis.
Degrees of freedom DF	Parameter of a t-distribution or a Chi-squared distribution. Examples: $-$ In a sample: number of data values minus one. $-$ In a frequency table: (No. of rows $-$ 1)$\times$ (No. of columns $-$ 1).
Density function	Can be considered an idealized histogram of a (possibly fictitious) population. The area under the density curve is 1 corresponding to 100% (all data values).
Dispersion	The spread of the data values of a distribution or a population.
Distribution function	The area under the density curve for data values up to a given value x. Corresponds to the probability (in a given distribution) of data values $\leq x$.
Estimate	Estimate of a population parameter, e.g., an average, calculated in a sample.
Experiment	A systematic investigation to determine which factors influence a product or a process. The various factor combinations are tested on a number of individuals (units), and a result (response) is measured.
Event	A subset of the sample space.
Fractile Quantile	In a distribution, the p-fractile is a value which separates the fraction p of the smallest data values from the largest.
Frequency	Number of occurrences of a given value in a distribution. Used for qualitative data or grouped data values of quantitative data.
Histogram	Bar chart showing the frequency of grouped data values of a quantitative variable.
Independent events	If the probability of the intersection of two events is exactly the product of the probabilities of the individual events, the two events are said to be independent.
Inter-Quartile Range	The difference between the upper and the lower quartile. (In some books defined as half of the difference).
Kurtosis	A parameter, indicating how "big tails" a distribution has compared to the normal distribution. Positive values indicate a distribution with "large tails." Negative values indicate a distribution with "small tails." Values around 0 indicate a distribution with tails like the normal distribution.
Location	The center of the data values of a distribution or population.

Term	Explanation
Mean Expectation	The average of a population. Usually unknown. An estimate is obtained by calculating the average in a sample.
Median	A number that divides the data values into two parts with an equal number of data values. The data value "in the middle" or 50% fractile (or the 2. quartile).
Method of least squares	A technique used to find the best model for data, e.g., a straight line. The idea is to choose the model that minimizes the sum of the squared residuals (i.e., the distances between the data values and the data values predicted by the model).
Mode	Data value with the largest frequency.
Nonresponse	The fact that some respondents do not participate in a sample survey. Can be caused by problems with the data collection.
Normal distribution	A symmetric distribution for description of quantitative (continuous) data.
Null hypothesis	A statistical hypothesis (assumption) about a population parameter, e.g., an average. The hypothesis can be true or false. The null hypothesis is considered true, unless sample data indicate that it is false (i.e., that the alternative hypothesis is considered true).
One-sided test	A test of a hypothesis, where you only reject for *either* small *or* large values (but not both) of a sample statistic due to specific subject matter knowledge.
Outcome	Result of an observation or measurement.
Population	The total set of individuals to be considered.
Probability of event	A number between 0 and 1 indicating the likelihood that the event will occur.
P-value	The probability of more extreme values (than observed) of a sample statistic, either extreme values in both sides (two-sided test) or only in one side of the distribution (one-sided test).
Quartile	Lower (1.) quartile: The 25% fractile in a distribution. Upper (3.) quartile: The 75% fractile in a distribution.
Random error	Differences between the average of a sample and the mean of a population due to the general variation ("natural variability") between individuals in the population and due to sampling variability.
Random variable	A mathematical function on the sample space.
Randomization	Sorting of the individuals in random order and it is used for − (Simple) random selection of sampling units in a sample. − Conduction of an experiment with combinations of several factors.
Range	The difference between the largest and smallest data value.
Regression analysis	A statistical technique used to assess the (e.g., linear) relationship between a (dependent) Y-variable and one or more (independent) X-variable(s).
Regression line	A statistical model, where the mean of Y depends linearly on X. The graph is a straight line determined by the method of least squares.
Relative Frequency	A frequency expressed as a proportion or percentage of the total frequency.
Sample	A number of individuals in a population, which are selected (at random) to give information about the population.
Sample space	The set of all possible outcomes.
Sample statistic	A function of the data values in a sample, e.g., an average.

Term	Explanation
Sampling	The process to select (draw) a sample.
Sampling fraction	Number of individuals in the sample divided by number of individuals in the population.
Sampling unit	An individual in the population, which can be selected for a sample.
Significance level	The probability to commit a type I error. Often the significance level is chosen to be 5% or (in some cases) 1%.
Simple random sampling	Selection of a sample in such a way that all individuals in the population are selected randomly and have the same probability of being selected.
Skewness	A measure of the departure from symmetry of a distribution. — Positive value indicates a "right skewed" distribution (too many data values to the right). — Negative value indicates a "left skewed" distribution. — A value around 0 indicates a symmetric distribution.
Standard deviation	Square root of the variance. The most used measure of spread (dispersion).
Standard error	The standard deviation of a mean. The standard deviation divided by $\sqrt{n}$.
Statistical uncertainty	The statistical uncertainty (e.g., of an average) is half the length of a (typically 95%) confidence interval (e.g., for the mean). This is the number after $\pm$.
Stratified sampling	The population is divided in homogeneous groups (strata). Simple random sampling is used within each group. The sampling fraction can be different from group to group.
Survey	A total investigation of the population or a (representative) sample.
t-distribution Students t-distribution	A symmetric distribution with larger tails than the normal distribution. The number of degrees of freedom needs to be specified. Applications, for example: — Confidence interval for a mean. — Test that two means are identical.
Type I error	To reject the null hypothesis, when it is true.
Type II error	To accept the null hypothesis, when it is false.
Two-sided test	A test of a hypothesis, where you reject for *both* small and large values of a sample statistic; this is the common situation.
Variance	Expresses the average of the squared distances between the data values and their average and it is calculated as: $$V = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}$$

Index