Quantitative Methods and Statistics

true

Version compiled 24 Jan 2021

Contents

Preface

Data are becoming ever more important, in all parts of society, including academia, and including the humanities. The availability of large amounts of digital data (such as text, speech, video, behavioural measurements) raises new research questions, which are typically and often investigated using quantitative methods. Aimed at humanities researchers and students, this book offers an overview of and introduction into the most important quantitative methods and statistical techniques used in the humanities. The book provides a solid methodological foundation for quantitative research, and it introduces the most commonly used statistical techniques to describe data and to test hypotheses. This will also enable the reader to critically evaluate such quantitative research.

This textbook is being used in the course *Methods and Statistics 1* at Utrecht University (Linguistics program). The book is also highly suitable for self-study at a basic level, for everybody who wishes to learn more about quantitative methods and statistics.

The main text has been kept free of mathematical derivations and formulas, which are typically not very helpful for humanities scholars and students. Our explanation is rather conceptual, and rich in examples. Where necessary we present derivations and formulas in separate sections.

This book also contains instructions on how to "do" the statistical analyses and visualisations, both in SPSS (version 22 or later) and in R (version 3.0 or later). These instructions too are in separate sections.

We would like to thank our co-teachers in various courses for the many discussions and examples that have been used in any shape or form in this textbook. We thank our students for their curiosity and for their sharp eyes in spotting errors and inconsistencies in previous versions.

We are also thankful to Gerrit Bloothooft, Margot van den Berg, Willemijn Heeren, Caspar van Lissa, Els Rose, Tobias Quené, Kirsten Schutter and Marijn Struiksma, for their advice, data, comments and suggestions.

We thank Aleksei Nazarov and Joanna Wall for translating this book from Dutch to English.

Utrecht, October 2020

Hugo Quené, https://www.hugoquene.nl

Huub van den Bergh, https://www.uu.nl/staff/HHvandenBergh

Notation

Following international usage we use the full stop (decimal point) as decimal separator; hence we write $\frac{3}{2} = 1.5$. Note that the decimal separator may vary between computers and between software packages on the same computer. Check which decimal separator is used by (each software package on) your computer.

License

This document is licensed under the *GNU GPL 3* license (for details see https://www.gnu.org/licenses/gpl-3.0.en.html).

Citation

Please cite this work as follows (in APA style):

Quené, H. & Van den Bergh, H. (2020). *Quantitative Methods and Statistics*. Retrieved 21 Oct 2020 from https://hugoquene.github.io/QMS-EN/ .

Technical details

All materials for this textbook are available at https://github.com/hugoquene/QMS-EN: this includes other versions of this textbook (EPUB, PDF, HTML), the source code (Rmarkdown and R) of the text including figures and examples, accompanying datasets used in the text, and figures as separate files.

The original Dutch version of this text was written in LaTeX, and was then converted to Rmarkdown, using `pandoc` (MacFarlane, 2020) and the `bookdown` (Xie, 2020) in Rstudio. The Dutch version is available at https://hugoquene.github.io/KMS-NL. The English translation is based on the Dutch LaTeX version (for Part I) and Rmarkdown version (for Parts II and III).

About the authors

Both authors work at the Faculty of Humanities at Utrecht University, the Netherlands. HQ is professor in the Quantitative Methods of Empirical Research in the Humanities, and he is also founding director of the Centre for Digital Humanities at Utrecht University. HvdB is professor in the Pedagogy and Testing of Language Proficiency, and he is also section chair in Dutch Language and Literature at the Dutch National Board of Tests and Examinations (CvTE).

Part I: Methodology

Chapter 1

Introduction

In this textbook, we will discuss the fundamental concepts, methods, and analytic techniques used in empirical scientific inquiry, both in general and as applied to the broad domain of language and communication. We will look at questions such as: What is a good research question? Which methodology is best for answering a given research question? How can researchers draw meaningful and valid conclusions from (statistical analyses of) their data? In this textbook, we will restrict ourselves to the most important fundamental concepts, and to the most important research methodologies and analytical techniques. In this first chapter, we will provide an overview of various types and forms of scientific research. In the following chapters, we will focus most of our attention on scientific research methodologies in which empirical observations are expressed in terms of numbers (quantitative), which may be analysed using statistical techniques.

1.1 Scientific research

To begin, we have to ask a question that refers back to the very first sentence above: what exactly is scientific research? What is the difference between scientific and non-scientific research (e.g., by investigative journalists)? Research conducted by a scholar does not necessarily have to be scientific research. Nor is research by journalists non-scientific by definition just because it is conducted by a journalist. In this textbook, we will follow this definition (Kerlinger and Lee, 2000, p.14):

> "Scientific research is systematic, controlled, empirical, amoral, public, and critical investigation of natural phenomena. It is guided by theory and hypotheses about the presumed relations among such phenomena."

Scientific research is systematic and controlled. Scientific research is designed such that its conclusions may be believed, because these conclusions are well-motivated. A research study can be repeated by others, which will (hopefully) lead to the same results. This demand that research be replicable also means that scientific research is designed and conducted in highly controlled ways (see Chapters **??** and **??**). The strongest form of control is found in a scientific experiment: we will therefore devote considerable attention to experimental research (§**??**). Any possible alternative explanations for the phenomenon studied are looked into one by one and excluded if possible, so that, in the end, we are left with one single explanation (Kerlinger and Lee, 2000). This explanation, then, forms our scientifically motivated conclusion on or theory of the phenomenon studied.

The definition above also states that scientific research is empirical. The conclusion a research draws about a phenomenon must ultimately be based on (systematic and controlled) observations of that phenomenon in reality – for example, on the observed content of a text or the behaviour observed in a test subject. If such observation is absent, then any conclusion drawn from such research cannot be logically connected to reality, which means that it has no scientific value. Confidential data from an unknown source or insights gained from a dream or in a mystical experience are not empirically motivated, and, hence, may not form the basis of a scientific theory.

1.1.1 Theory

The goal of all scientific research is to arrive at a theory of a part of reality. This theory can be seen as a coherent and consistent collection of "justified true beliefs" (Morton, 2003). These beliefs as well as the theory they form abstract away from the complex reality of natural phenomena to an abstract mental *construct*, which in its very nature is not directly observable. Examples of similar constructs include: reading ability, intelligence, activation level, intelligibility, active vocabulary size, shoe size, length of commute, introversion, etc.

When building a theory, a researcher not only defines various constructs, but also specifies the *relationships* between these constructs. It is only when the constructs have been defined and the relationships between these constructs have been specified that a researcher can arrive at a systematic explanation of the phenomenon studied. This explanation or theory can, in turn, form the basis of a *prediction* about the phenomenon studied: the number of spoken languages will decrease in the 21st century; texts without overt conjunctions will be more difficult to understand than texts with overt conjunctions; children with a bilingual upbringing will perform no worse at school than monolingual children.

Scientific research comes in many kinds and forms, which may be classified in various ways. In §**??**, we will discuss a classification based on paradigm: a

researcher's outlook on reality. Research can also be classified according to a continuum between 'purely theoretical' to 'applied'. A third way of classifying research is oriented towards the type of research, for instance, instrument validation (§1.3), descriptive research (§**??**), and experimental research (§**??**).

1.2 Paradigms

One criterion to distinguish different kinds of research is on the basis of the paradigm used: the researcher's outlook on reality. In this textbook, we have spent almost all of our attention on the empirical-analytical paradigm, because this paradigm has been written about the most and is the most influential. At present, this approach can be seen as 'the' standard approach, against the backdrop of which other paradigms try to distinguish themselves.

Within the *empirical-analytical* paradigm, we distinguish two variants: positivism and critical rationalism. Both schools of thought share the assumption that there exist lawful generalizations that can be 'discovered': phenomena may be described and explained in terms of abstractions (constructs). The difference between the two schools within the empirical-analytical tradition lies in the way generalizations are treated. Positivists claim that it is possible to make statements from factual observations towards a theory. Based on the observations made, we may generalize towards a general principle by means of induction. (All birds I have seen are also perceived by me to be singing, so all birds sing.)

The second school is critical rationalism. Those within this school of thought oppose the inductive statements mentioned above: even if I see masses of birds and they all sing, I still cannot say with certainty that the supposed general principle is true. But, say critical rationalists, we can indeed turn this on its head: we may try to show that the supposed general rule or hypothesis is not true. How would this work? From the general principle, we can derive predictions about specific observations by using deduction. (If all birds sing, then it must be true that all birds in my sample do sing.) If it is not the case that all birds in my sample sing, this means the general principle must be false. This is called the falsification principle, which we will discuss in more detail in **??**.

However, critical rationalism, too, has at least two drawbacks. The falsification principle allows us to use observations (empirical facts, research results) to make theoretical statements (regarding specific hypotheses). Strictly speaking, a supposed general principle should be immediately rejected after a single successful instance of falsification (one of the birds in my sample does not sing): if there is a mismatch between theory and observations, then, according to critical rationalists, the theory fails. But to arrive at an observation, a researcher has to make many choices (e.g., how do I draw an appropriate sample, what is a bird, how do I determine whether a bird sings?), which may cast doubt on the validity of the observations. This means that a theory/observation mismatch

could also indicate a problem with the observations themselves (hearing), or with the way the constructs in the theory (birds, singing) are operationalized.

A second drawback is that, in practice, there are very few theories that truly exclude some type of observation. When we observe discrepancies between a theory and observations made, the theory is adjusted such that the new observations still fit within the theory. In this way, theories are very rarely completely rejected.

One alternative paradigm is the critical approach. The *critical paradigm* is distinguished from other paradigms by its emphasis on the role of society; there is no one true reality: our image of reality is not a final one, and it is determined by social factors. Thus, insight into relationships within society, by itself, influences this reality. This means that our concept of science, as formulated in the definitions of research and theory given above, is rejected in the critical paradigm. Critical researchers claim that research processes cannot be seen as separate from the social context in which research is conducted. However, we must add that this latter viewpoint has lately been taken over by more and more researchers, including those that follow other paradigms.

1.3 Instrument validation

As stated above, research is a systematized and controlled way of collecting and interpreting empirical data. Researchers strive for insight into natural phenomena and into the way in which (constructs corresponding to) these phenomena are related to one another. One requirement for this is that the researcher be able to actually measure said phenomena, i.e., to express them in terms of an observation (preferable, in the form of a number). Instrument validation research is predominantly concerned with constructing instruments or methods to make phenomena, behaviour, ability, attitudes, etc. measurable. The development of good instruments for measurement is by no means an easy task: they truly have to be crafted by hand, and there are many pitfalls that have to be avoided. The process of making phenomena, behaviour, or constructs measurable is called *operationalization*. For instance, a specific reading test can be seen as an operationalization of the abstract construct of 'reading ability'.

It is useful to make a distinction between the abstract theoretical construct and the construct as it is used for measurements, which means: a distinction between the concept-as-intended and the concept-as-defined. Naturally, the desired situation is for the concept-as-defined (the test or questionnaire or observation) to maximally approach the concept-as-intended (the theoretical construct). If the theoretical construct is given a good approximation, we speak of an adequate or valid measurement.

When a concept-as-intended is operationalized, the amount of choices to be made is innumerable. For instance, the Dutch government institute that devel-

ops standardized tests for primary and secondary education, the CITO (Centraal instituut voor toetsontwikkeling, or Central Test Development Institute) must develop new reading comprehension tests each year to measure the reading ability exhibited by students taking the centralized final exams for secondary school students (eindexamens). For this purpose, the first step is to choose and possibly edit a text. This text cannot be too challenging for the target audience, but may also not be too easy. Furthermore, the topic of the text may not be too well-known – otherwise, some students' general background knowledge may interfere with the opinions and standpoints brought forward in the text. At the next step, questions must be developed in such a way that the various parts of the text are all covered. In addition, the questions must be constructed in such a way that the theoretical concept of 'reading ability' is adequately operationalized. Finally, exams administered in previous years must also be taken into consideration, because this year's exam may not differ too much from previous years' exams.

To sum up, a construct must be correctly operationalized in order to arrive at observations that are not only valid (a good approximation of the abstract construct, see Chapter **??**) but also reliable (observations must be more or less identical when measurement is repeated, see Chapter **??**). In each research study, the validity and reliability of any instance of measurement are crucial; because of this, we will spend two chapters on just these concepts. However, in instrument validation research, specifically, these concepts are absolutely essential, because this type of research itself is meant to yield valid and reliable instruments that are a good operationalization of the abstract construct-as-intended.

1.4 Descriptive research

Descriptive research refers to research predominantly geared towards describing a particular natural phenomenon in reality. This means that the researcher mostly aims for a description of the phenomenon: the current level of ability, the way in which a particular process or discussion proceeds, the way in which Dutch language classes in secondary education take shape, voters' political preferences immediately before an election, the correlation between the number of hours a student spent on individual study and the final mark they received, etc. In short, the potential topics of descriptive research are also be very diverse.

Example 1.1: Dingemanse et al. (2013) made or chose recordings of conversations in 10 languages. Within these conversations, they took words used by a listener to seek "open clarification": little words like *huh* (English), *hè* (Dutch), *ã?* (Siwu). They determined the sound shape and pitch contour of these words using acoustic measurements

> and phonetic transcriptions made by experts. One of the conclusions of this descriptive research is that these interjections in the various languages studied are much more alike (in terms of sound shape and pitch contour) than would be expected based on chance.

This example illustrates the fact that descriptive research does not stop when the data (sound shapes, pitch contours) have been described. Oftentimes, relationships between the data points gathered are also very interesting (see §1.1). For instance, in opinion polls that investigate voting behaviour in elections, a connection is often made between the voting behaviour polled, on the one side, and age, sex, and level of education, on the other side. In the same way, research in education makes a connection between the number of hours spent studying, on the one side, and performance in educational assessment, on the other side. This type of descriptive research, in which a correlation is found between possible causes and possible effects, is otherwise also referred to as *correlational research*.

The essential difference between descriptive and experimental research lies in the question as to cause and effect. Based on descriptive research, a causal relationship between cause and effect *cannot* be properly established. Descriptive research might show that there is a correlation between a particular type of nutrition and a longer lifespan. Does this mean that this type of nutrition is the cause of a longer lifespan? This is definitely not necessarily the case: it is also possible that this type of food is mainly consumed by people who are relatively highly educated and wealthy, and who live longer because of these other factors[1]. In order to determine whether there is a causal relationship, we must set up and conduct experimental research.

1.5 Experimental research

Experimental research is characterized by the researcher's systematically manipulating a particular aspect of the circumstances under which a study is conducted (Shadish et al., 2002). The effect arising from this manipulation now becomes central in the research study. For instance, a researcher suspects that a particular new method of teaching will result in better student performance compared to the current teaching method. The researcher wants to test this hypothesis using experimental research. She or he manipulates the type of teaching: some groups of students are taught according to the novel, experimental teaching method, and other groups of students are taught according to

[1] It is even possible that the nutrition habits under study cause people to live *shorter*, but that this negative effect is masked by the stronger positive effects of education and wealth.

the traditional method. The novel teaching method's effect is evaluated by comparing both types of student groups' performance after they have been 'treated' with the old vs. new teaching method.

The advantage of experimental research is that we may usually interpret the research results as the consequence or effect of the experimental manipulation. Because the research systematically controls the study and varies just one aspect of it (in this case, the method of teaching), possible differences between the performance observed in the two categories can only be ascribed to the aspect that has been varied (the method of teaching). Logically speaking, this aspect that was varied is the only thing that could have cause the observed differences. Thus, experimental research is oriented towards evaluating causal relationships.

This reasoning does require that test subjects (or groups of students, as in the example above) are assigned to experimental conditions (in our example, the old or the new method of teaching) at random. This random assignment is the best method to exclude any non-relevant differences between the conditions of treatment. Such an experiment with random assignment of test subjects to conditions is called a *randomized experiment* or *true experiment* (Shadish et al., 2002). To remain with our example: if the researcher had used the old research method only with boys, and the new research method only with girls, then any difference in performance can no longer just be attributed to the manipulated factor (teaching method), but also to a non-manipulated but definitely relevant factor, in this case, the students' sex. Such a possible disruptive factor is called a confound. In Chapter 6, we will discuss how we can neutralize such confounds by random assignment of test subjects (or groups of students) to experimental conditions, combined with other measures.

There also exists experimental research in which a particular aspect (such as teaching method) is indeed systematically varied, but in which test subjects or groups of students are not randomly assigned to the experimental conditions; this is called *quasi-experimental research* (Shadish et al., 2002). In the example above, this term would be applicable if teaching method were investigated using data from groups of students for which it was not the researcher, but their teacher who determined whether the old or new teaching method would be used. In addition, the teacher's enthusiasm or teaching style might be a confound in this quasi-experiment. We will encounter various examples of quasi-experimental research in the remainder of this textbook.

Within the type of experimental research, we can also make a further division: that between laboratory research and field research. In both types of experimental research, some aspect of reality is manipulated. The difference between both types of research lies in the degree to which the researcher is able to keep under control the various confounds present in reality. In laboratory research, the researcher can very precisely determine under which environmental conditions observations are made, which means that the researcher can keep many possible confounds (such as lighting, temperature, ambient noise, etc.) under control. In field research, this is not the case. When 'out in the field', the re-

searcher is not able to keep all (possibly relevant) aspects of reality fully under control.

Example 1.2: Margot van den Berg and colleagues from the Universities of Utrecht, Ghana and Lomé investigated how multilingual speakers use their languages when they have to name attributes like colour, size, and value in a so-called Director-Matcher task (Van den Berg et al., 2017). In this task, one research participant (the 'director') gave clues to another participant (the 'matcher') to arrange a set of objects in a particular order. This allowed the researchers to collect many instances of attribute words in a short period of time ("Put the yellow car next to the red car, but above the small sandal"). The interactions were recorded, transcribed, en subsequently investigated for language choice, moment of language switch, and type of grammatical construction. In this type of fieldwork, however, various kinds of non-controlled aspects in the environment may influence the sound recordings and, thus, the data, including "clucking chickens, a neighbour who was repairing his motorbike and had to start it every other second while we were trying to record a conversation, pouring rain on top of the aluminium roof of the building where the interviews took place." (Margot van den Berg, personal communication)

Example 1.3: When listening to spoken sentences, we can infer from a test subject's eye movements how these spoken sentences are processed. In a so-called 'visual world' task, listeners are presented with a spoken sentence (e.g., "Bert says that the rabbit has grown"), while they are looking at multiple images on the screen (usually 4 of them, e.g., a sea shell, a peacock, a saw, and a carrot). It turns out that listeners will predominantly be looking at the image associated with the word they are currently mentally processing: when they are processing *rabbit*, they will look at the carrot. A so-called 'eye tracker' device allows researchers to determine the position on the screen that a test subject is looking at (through observation of their pupils). In this way, the researcher can therefore observe which word is mentally processed at which time (Koring et al., 2012). Research of this kind is best conducted in a laboratory, where one can control background noise, lighting, and the position of test subjects' eyes relative to the computer screen.

Both laboratory research and field research have advantages and disadvantages. The great advantage of laboratory research is, of course, the degree to which the researcher can keep all kinds of external matters under control. In a laboratory, the experiment is not likely to be disturbed by a starting engine or a downpour. However, this advantage of laboratory research also forms an important disadvantage, namely: the research takes place in a more or less artificial environment. It is not at all clear to what extent results obtained under artificial circumstances will also be true of everyday life outside the laboratory. Because of this, the latter forms a point to the advantage of field researcher: the research is conducted under circumstances that are natural. However, the disadvantage of field research is that many things can happen in the field that may influence the research results, but remain outside of the researcher's control (see example 1.2). The choice between both types of experimental research that a researcher has to make is obviously strongly guided by their research question. Some questions are better suited to being investigated in laboratory situations, while others are better suited to being investigated field situations (as is illustrated by the examples above).

1.6 Outline of this textbook

This textbook consists of three parts. Part I (Chapter 1 to 7) covers research methods and explains various terms and concepts that are important in designing and setting up a good scientific research study.

In part II (Chapters 8 to 12), we will cover descriptive statistics, and in part III (Capters 13 to 17), we will cover the basic methods of inferential statistics. These two parts are designed to work towards three goals.

Firstly, we would like for you to be able to critically evaluate articles and other reports in which statistical methods of processing and testing hypotheses on data have been used. Secondly, we would like for you to have the knowledge and insight necessary for the most important statistical procedures. Thirdly, these parts on statistics are meant to enable you to perform statistical analysis on your own for your own research, for instance, for your internship or final thesis.

These three goals are ordered by importance. We believe that an adequate and critical interpretation of statistical results and the conclusions that may be connected to these is of great importance to all students. For this reason, part I of this textbook devotes considerable attention to the 'philosophy' or methodology behind the statistical techniques and analyses we will discuss later.

We will also give you instructions on how you can perform these statistical analyses yourself in SPSS (a popular software package for statistical analysis) and in R (a slightly more challenging, but also much more powerful and versatile

software package that has been gaining popularity). For students and employees at Utrecht University, both packages are pre-installed in `MyWorkSpace`. SPSS is available at https://SurfSpot.nl for a small fee. R is freely available at https://www.R-project.org. A brief introduction to R can be found at https://hugoquene.github.io/emlar2020/. Longer introductions are available in the excellent free web books listed on https://statisticalhorizons.com/resources/free-web-books-for-learning-r, as well as in Dalgaard (2002).

Chapter 2

Hypothesis testing research

2.1 Introduction

Many empirical studies pursue the goal of establishing connections between (supposed) causes and their (supposed) effects or consequences. The researcher would like to know whether one variable has an influence on another. Their research tests the hypothesis that there is a connection between the supposed cause and the supposed effect (see Table 2.1). The best way to establish such a connection, and, thus, to test this hypothesis, is an experiment. An experiment that has been set up properly and is well executed is the 'gold standard' in many academic disciplines, because it offers significant guarantees concerning the validity of the conclusions drawn from it (see Chapter 5). Put differently: the outcome of a good experiment forms the strongest possible evidence for a connection between the variables investigated. As we discussed in Chapter 1, there are also many other forms of research, and hypotheses can also be investigated in other ways and according to other paradigms, but we will limit ourselves here to experimental research.

Table 2.1: Possible causes and possible effects.

Domain	Supposed cause	Supposed effect
trade	outside temperature	units of ice cream sold
healthcare	type of treatment	degree of recovery
eduction	method of instruction	performance on test
language	age at which L2 learning satrts	degree of proficiency
education	class size	general performance in school
healthcare	altitude	rate of malaria infection
language	age	speaking rate (speech tempo)

In experimental research, the effect of a variable manipulated by the researcher on some other variable is investigated. The introduction already provided an example of an experimental study. A novel teaching method was tested by dividing students between two groups. One group was taught according to the novel method, while the other group was taught as usual. The researcher hoped and expected that her or his novel teaching method would have a beneficial effect, meaning that it would lead to better student performance.

In hypothesis testing research, it is examined whether the variables investigated are indeed connected to one another in the way expected by the researcher. Two terms play a central role in this definition: 'variables' and 'in the way expected'. Before we consider experimental research in more detail, we will first take a closer look at these terms.

2.2 Variables

What is a variable? Roughly speaking, a variable is a particular kind of property of objects or people: a property that may vary, i.e., take different values. Let us look at two properties of people: how many siblings they have, and whether their mother is a woman or a man. The first property may vary between individuals, and is thus a (between-subject) variable. The second property may not vary: if there is a mother, she will always be a woman by definition [at least, traditionally]. Thus, the second property is not a variable, but a constant.

In our world, almost everything exists in varying quantities, in varying manners, or to various extents. Even a difficult to define property, like a person's popularity within a certain group, may form a variable. This is because we can rank people in a group from most to least popular. There are ample examples of variables:

- regarding *individuals*: their length, their weight, shoe size, speaking rate, number of siblings, number of children, political preference, income, sex, popularity within a group, etc.
- regarding *texts*: the total number of words ('tokens'), the number of unique words ('types'), number of typos, number of sentences, number of signs of interpunction, etc.
- regarding *words*: their frequency of use, number of syllables, number of sounds, grammatical category, etc.
- regarding *objects* such as cars, phones, etc.: their weight, number of components, energy use, price, etc.
- regarding *organizations*: the number of their employees, their postal code, financial turnover, numbers of customers or patients or students, number

of surgeries or transactions performed or number of degrees awarded, type of organization (corporation, non-profit, …), etc.

2.3 Independent and dependent variables

In hypothesis testing research, we distinguish two types of variables: dependent and independent variables. The *independent* variable is whatever is presumed to bring about the supposed effect. The independent variable is the aspect that a research will manipulate in a study. In our example where an experiment is conducted to evaluate the effects of a new teaching method, the teaching method is the independent variable. When we compare performance between the students that were taught using the new method and those whose writing instruction only followed the traditional method, we can see that the independent variable takes on two values. In this case, we can give these two values (also called *levels*) that the independent variable can take the names of "experimental" and "control", or "new" and "old". We might also express the independent variable's values as a number: 1 and 0, respectively. These numbers do not have a numerical interpretation (for instance, we might as well give these values the names 17 and 23, respectively), but are used here solely as arbitrary labels to distinguish between groups. The manipulated variable is called 'independent' because the chosen (manipulated) values of this variable are not dependent on anything else in the study: the researcher is independent in their choice of this variable's values. An independent variable is also called a *factor* or a *predictor*.

The second type of variable is the dependent variable. The *dependent* variable is the variable for which we expect the supposed effect to take place. This means that the independent variable possibly cause an effect on the dependent variable, or: it is presumed that the dependent variable's value depends on the independent variable's value - hence their names. An observed value for the dependent variable is also called a *response* or *score*; oftentimes, the dependent variable itself may also be given these names. In our example where an experiment conducted to evaluate the effect a new teaching method has on students' performance, the student's performance is the dependent variable. Other examples of possible dependent variables include speaking rate, score on a questionnaire, or the rate at which a product is sold (see Table 2.1). In short, any variable could be used as the dependent variable, in principle. It is mainly the research question that determines which dependent variable is chosen, and how it is measured.

This being said, it must be stressed that independent and dependent variables themselves must not be interpreted as 'cause' and 'effect', respectively. This is because the study has as its goal to convincingly demonstrate the existence of a (causal) connection between the independent and the dependent variable. However, Chapter 5 will show us how complex this can be.

The researcher varies the independent variable and observes whether this results

in differences observed in the dependent variable. If the dependent variable's values differ before and after manipulating the independent variable, we may assume that this is an effect that the manipulation has on the independent variable. We may speak of a relationship between both variables. If the dependent variable's value does not differ under the influence of the independent variable's values, then there is no connection between the two variables.

Voorbeeld 2.1: Quené et al. (2012) investigated whether a smile or frown influences how listeners process spoken words. The words were 'pronounce' (synthesized) by a computer in various phonetic variants - specifically, in such a way that these words sounded as if pronounced neutrally, with a smile, or with a frown. Listeners has to classify the words a 'positive' or 'negative' (in meaning) as quickly as possible. In this study, the phonetic variant (neutral, smile, drown) takes the place of the independent variable, and the speed with which listeners give their judgment is the dependent variable.

2.4 Falsification and null hypothesis

The goal of scientific research is to arrive at a coherent collection of "justified true beliefs" (Morton, 2003). This means that a scientific belief must be properly motivated and justified (and must be coherent with other beliefs). How may we arrive at such a proper motivation and justification? For this, we will first refer back to the so-called induction problem discussed by Hume (1739). Hume found that it is logically impossible to generalize a statement from a number of specific cases (the observations in a study) to a general rule (all possible observations in the universe).

We will illustrate the problem inherent in this generalization or induction step with the belief that 'all swans are white'. If I had observed 10 swans that are all white, I might consider this as a motivation for this belief. However, this generalization might be unjustified: perhaps swans also exist in different colours, even if I might not have seen these. The same problem of induction remains even if I had seen 100 or 1000 white swans. However, what if I had seen a single black swan? In that case, I will know immediately and with completely certainty that the belief of all swans' being white is false. This principle is also used in scientific research.

Let us return to our earlier example in which we presumed that a new teaching method will work better than an older teaching method; this belief is called H1.

Let us now set this reasoning on its head, and base ourselves on the complementary belief that the new method is *not* better than the old one[1]; this belief is called the null hypothesis or H0. This belief that 'all methods have an equal effect' is analogous to the belief that 'all swans are white' from the example given in the previous paragraph. How can we then test whether the belief or hypothesis called H0 is true? For this, let us draw a representative sample of students (see Chapter **??**) and randomly assign students to the new or old teaching method (values of the independent variable); we then observe all participating students' performance (dependent variable), following the same protocol in all cases. For the time being, we presume that H0 is true. This means that we expect no difference between the student groups' performance. If, despite this, the students taught by the new method turn out to perform much better than the students taught by the old method, then this observed difference forms the metaphorical black swan: the observed difference (which contradicts H0) makes it unlikely that H0 is true (provided that the study was valid; see Chapter 5 for more on this). Because H0 and H1 exclude each other, this means that it is very likely that H1 is indeed true. And because we based our motivation upon H0 and not H1, sceptics cannot accuse us of being biased: after all, we did try to show that there was indeed no difference between the performance exhibited by the students in each group.

The method just described is called falsification, because we gain knowledge by rejecting (falsifying) hypotheses, and not by accepting (verifying) hypotheses. This method was developed by philosopher of science Karl Popper (Popper, 1935, 1959, 1963). The falsification method has interesting similarities to the theory of evolution. Through variation between individual organisms, some can successfully reproduce, while many others die prematurely and/or do not reproduce. Analogously, some tentative statements cannot be refuted, allowing them to 'survive' and 'reproduce', while many other statements are indeed refuted, through which they 'die'. In the words by Popper (1963) (p.51, italics removed):

> " … to explain (the world) … as far as possible, with the help of laws and explanatory theories …there is no more rational procedure than the method of trial and error — of conjecture and refutation: of boldly proposing theories; of trying our best to show that these are erroneous; and of accepting them tentatively if our critical efforts are unsuccessful."

Thus, a proper scientific statement or theory ought to be falsifiable or refutable or testable (Popper, 1963). In other words, it must be possible to prove this statement or theory wrong. A testable statement's scientific motivation, and, therefore, its plausibility increase with each time this statement proves to be immune to falsification, and with each new set of circumstances under which this

[1] Two beliefs are complementary when they mutually exclude each other, like H1 and H0 in this example.

happens. 'Earth's climate is warming up' is a good example of a statement that is becoming increasingly immune to falsification, and, therefore, is becoming increasingly stronger.

Voorbeeld 2.2: 'All swans are white' and 'Earth's climate is warming up' are falsifiable, and therefore scientifically useful statements. What about the following statements?
a. Gold dissolves in water.
b. Salt dissolves in water.
c. Women talk more than men.
d. Coldplay's music is better than U2's.
e. Coldplay's music sells better than U2's.
f. If a patient rejects a psychoanalyst's reading, then this is a consequence of their resistance to the fact that the psychoanalyst's reading is correct.
g. Global warming is caused by human activity.

2.5 The empirical cycle

So far, we have provided a rather global introduction to experimental research. In this section, we will describe the course of an experimental study in a more systematic way. Throughout the years, various schemata have been devised that describe research in terms of phases. The best known of these schemata is probably the empirical cycle by De Groot (1961).

The empirical cycle distinguishes five phases of research: the observation phase, the induction phase, the deduction phase, the testing phase, and the evaluation phase. In this last phase, any shortcomings and alternative interpretations are formulated, which lead to potential new studies, each of which once again goes through the entire series of phases (hence the name, 'cycle'). We will now look at each of these five phases of research one by one.

2.5.1 observation

In this phase, the researcher constructs a problem. This is to say, the researcher forms an idea of possible relationships between various (theoretical) concepts or constructs. These presumptions will later be worked out into more general hypotheses. Presumptions like these may come about in myriads of different ways – but all require for the researcher to have sufficient curiosity. The researcher

may notice an unusual phenomenon that needs an explanation, e.g., the phenomenon that the ability to hear absolute pitch occurs much often in Chinese musicians than in American ones (Deutsch, 2006). Systematic surveys of scientific publications may also lead to presumptions. Sometimes, it turns out that different studies' results contradict each other, or that there is a clear gap in our knowledge.

Presumptions can also be based on case studies: these are studies in which one or several cases are studied in depth and extensively described. For instance, Piaget developed his theory of children's mental development based on observing his own children during the time he was unemployed. These observations later (when Piaget already had his own laboratory) formed the impetus for many experiments that he used to sharpen and strengthen his theoretical insights.

It is important to realize that purely unbiased and objective observation is not possible. Any observation is influenced by theory or prior knowledge to a greater or smaller extent. If we do not know what to pay attention to, we also cannot observe properly. For instance, those that specialize in the formation of clouds can observe a far greater variety of cloud types than the uninitiated. This means that it is useful to first lay down an explicit theoretical framework, however rudimentary, before making any observations and analysing any facts.

A researcher is prompted by remarkable phenomena, case studies, studying the literature, etc. to arrive at certain presumptions. However, there are no methodological guidelines on how this process should come about: it is a creative process.

2.5.2 induction

During the induction phase, the presumption voiced in the observation phase is generalized. Having started from specific observations, the researcher now formulates a hypothesis that they suspect is valid in general. (**Induction** is the logical step in which a general claim or hypothesis is derived from specific cases: my children (have) learned to talk → all children (can) learn to talk.)

For instance, from the observation made in their own social circle that women speak more than men do (more minutes per day, and more words per day), a researcher may induce a general hypothesis: H1: women talk more than men do (see Example 2.2; this hypothesis may be further restricted as to time and location).

In addition, the hypothesis' empirical content must be clearly described, which is to say: the type or class of observations must be properly described. Are we talking about all women and men? Or just speakers of Dutch (or English)? And what about multilingual speakers? And children that are still acquiring their language? This clearly defined content is needed to test the hypothesis (see the subsection on testing below, and see Chapter **??**).

Finally, a hypothesis also has to be logically coherent: the hypothesis has to be consistent with other theories or hypotheses. If a hypothesis is not logically coherent, it follows by definition that it cannot be unambiguously related to the empirical realm, which means that it is not properly testable. From this, we can conclude that a hypothesis may not have multiple interpretations: within an experiment, a hypothesis, by itself, must predict one single outcome, and no more than one. In general, three types of hypotheses are distinguished (De Groot, 1961):

- Universal-deterministic hypotheses.
  These take the general shape of *all As are B*. For example: all swans are white, all human beings can speak. If a researcher can show for one single A that it is not B, then the hypothesis has, in principle, been falsified. A universal deterministic hypothesis can never be verified: a researcher can only make statements about the cases they have observed or measured. If we are talking about an infinite set, such as: all birds, or all human beings, or all heaters, this may lead to problems. The researcher does not know whether such a set might include a single case for which 'A is not B'; there is one bird that cannot fly, et cetera. Consequently, no statement can be made about these remaining cases, which means that the universal validity of the hypothesis can never be fully 'proven'.

- Deterministic existential hypotheses.
  These take the general shape of *there is some (at least one) A that is B*. For example: there is some swan that is white, there is some human being that can speak, there is some heater that provides warmth. If a researcher can demonstrate that there exists one A that is B, the hypothesis has been verified. However, deterministic existential hypotheses may never be falsified. If we wanted to do that, it would be necessary to investigate all units or individuals in an infinite set for whether they are B, which is exactly what is excluded by the infinite nature of the set. At the same time, this makes it apparent that this type of hypotheses does not lead to generally valid statements, and that their scientific import is not as clear. One could also put it this way: a hypothesis of this type makes no clear predictions for any individual case of A; a given A might be the specific one that is also B, but it might also not be. In this sense, deterministic existential hypotheses do not conform to our criterion of falsifiability.

- Probabilistic hypotheses.
  These take the general shape of *there are relatively more As that are B compared to non-As that are B*. In the behavioural sciences, this is by far the most frequently occurring type of hypothesis.
  For example: there are relatively more women that are talkative compared to men that are talkative. Or: there are relatively more highly performing students for the new teaching method compared to the old teaching method. Or: speech errors occur relatively more often at the beginning

rather than at the end of the word. This does not entail that all women speak more than all men, nor does this entail that all students taught by the new method perform better than all students taught by the old method.

2.5.3 deduction

During this phase, specific predictions are deduced from the generally formulated hypothesis set up in the induction phase. (**Deduction** is the logical step whereby a specific statement or prediction is derived from a more general statement: all children learn to talk → my children (will) learn to talk.)

If we presume (H0) that "women talk more than men", we can make specific predictions for specific samples. For example, if we interviewed 40 female and 40 male school teachers of Dutch, without giving them a time limit, then we predict that the female teachers in this sample will say more than the male teachers in the sample (including the prediction that they will speak a greater number of syllables in the interview).

As explained above (§2.4), most scientific research does not test H1 itself, but its logical counterpart: H0. Therefore, for testing a H1 (in the next phase of the empirical cycle), we use the predictions derived from H0 (!), for instance: "women and men produce equal numbers of syllables in a comparable interview".

In practice, the terms "hypothesis" and "prediction" are often used interchangeably, and we often speak of testing hypotheses. However, according to the above terminology, we do not test the hypotheses, but we test predictions that are derived from those hypotheses.

2.5.4 testing

During this phase, we collect empirical observations and compare these to the worked-out predictions made "under H0", i.e., the predictions made if H0 were to be true. In Chapter 13, we will talk more about this type of testing. Here, we will merely introduce the general principle. (In addition to the conventional "frequentist" approach described here, we may also test hypotheses and compare models using a newer "Bayesian" approach; however, this latter method of testing is outside the scope of this textbook).

If the observations made are extremely unlikely under H0, there are two possibilities.

- (i) The observations are inadequate, we have observed incorrectly. But if the researcher has carried out rigorous checks on their work, and if they take themselves seriously, this is not likely to be true.

- (ii) The prediction was incorrect, meaning that H0 is possibly incorrect, and should be rejected in favour of H1.

In our example above, we derived from H0 (!) the prediction that, within a sample of 40 male and 40 female teachers, individuals will use the same amount of syllables in a standardized interview. However, we find that men use 4210 syllables on average, while women use 3926 on average (Quené, 2008, p.1112). How likely is this difference if H0 were true, assuming that the observations are correct? This probability is so small, that the researcher rejects H0 (see option (ii) above) and concludes that women and men do *not* speak *equal* amounts of syllables, at least, in this study.

In the example above, the testing phase involves comparing two groups, in this case, men and women. One of these two groups is often a neutral or control group, as we saw in the example given earlier of the new and old teaching methods. Why do researchers often make use of a control group of this kind? Imagine that we had only looked at the group taught by the new method. In the testing phase, we measure students' performance, which is a solid B on average (7 in the Dutch system). Does this mean that the new method is successful? Perhaps it is not: if the students might have gotten an A or A- (8 in the Dutch system) under the old method, the new method would actually be worse, and it would be better not to add this new method to the curriculum. In order to be able to draw a sensible conclusion about this, it is essential to compare the new and old methods between one another. This is the reason why many studies involve components like a neutral condition, null condition, control group, or placebo treatment.

Now that we know this, how can we determine the probability of the observations we made if H0 were to be true? This is often a somewhat complex question, but, for present purposes, we will give a simple example as an illustration: tossing a coin and observing heads or tails. We presume (H0): we are dealing with a fair coin, the probability of heads is 1/2 at each toss. We toss the same coin 10 times, and, miraculously, we observe the outcome of heads all 10 times. The chance of this happening, given that H0 is true, is $P = (1/2)^{10} = 1/1024$. Thus, if H0 were to be true, this outcome would be highly unlikely (even though the outcome is not impossible, since $P > 0$); hence, we reject H0. Therefore, we conclude that the coin most likely is not a fair coin.

This leads us to an important point: when is an outcome unlikely enough for us to reject H0? Which criterion do we use for the probability of the observations made if H0 were to be true? This is the question of the level of significance, i.e., the level of probability at which we decide to reject H0. This level is signified as α. If a study uses a level of significance of $\alpha = 0.05$, then H0 is rejected if the probability of finding these results under H0[2] is smaller than 5%.

[2]More accurately: If the probability to find either these results or other results that would differ even more from those predicted by H0 is smaller than 5%, then H0 is rejected.

In this case, the outcome is so unlikely, that we choose to reject H0 (option (ii) above), i.e., we conclude that H0 is most probably not true.

If we thus reject H0, there is a small chance that we are actually dealing with option (I): H0 is actually true, but the observations happen *by chance* to strongly diverge from the prediction under H0, and H0 is falsely rejected. This is called a Type I error. This type of error can be compared to unjustly sentencing an innocent person, or undeservedly classifying an innocent email message as 'spam'. Most of the time, $\alpha = 0.05$ is used, but other levels of significance are also possible, and sometimes more prudent.

Note that significance is the probability of finding the extreme data that were observed (or data even more extreme than that) given that H0 is true:

$$\text{significance} = P(\text{data}|\text{H0})$$

Most importantly, significance is *not* the probability of H0 being true given these data, $P(\text{H0}|\text{data})$, even though we do encounter this mistake quite often.

Each form of testing also involves the risk of making the opposite mistake, i.e., not rejecting H0 even though it should be rejected. This is called a Type II error: H0 is, in fact, false (meaning that H1 is true), but, nevertheless, H0 is not rejected. This type of mistake can be compared to unjustly acquitting a guilty person, or undeservedly letting through a spam email message (see Table 2.2).

Table 2.2: Possible outcomes of the decision procedure.

Reality	Decision	
	Reject H0	**Maintain H0**
H0 is true (H1 false)	Type I error (α)	correct
H0 is false (H1 true)	correct	Type II error (β)
	Convict defendant	**Acquit defendant**
defendant is innocent (H0)	Type I error	correct
defendant is guilty	correct	Type I error
	Discard message	**Allow message**
message is OK (H0)	Type I error	correct
message is spam	correct	Type II error

If we set the level of significance to a higher value, e.g., $\alpha = .20$, this also means that the chance of rejecting H0 is much higher. In the testing phase, we would reject H0 if the probability of observing these data (or any more extreme data) were smaller than 20%. This would mean that 8 times heads within 10 coin tosses would be enough to reject H0 (i.e., judging the coin as unfair). Thus, more outcomes are possible that lead to rejecting H0. Consequently, this higher level of significance entails a greater risk of a Type 1 error, and, at the same

time, a smaller risk of a Type II error. The balance between the two type of error depends on the exact circumstances under which the study is conducted, and on the consequences that each of the two types of error might have. Which type of error is worse: throwing away an innocent email, or letting a spam message through? The probability of making a Type I error (the level of significance) is controlled by the researcher themselves. The probability of a Type II error depends on three factors and is difficult to gauge. Chapter 14 will discuss this in more detail.

2.5.5 evaluation

At the end of their study, the researcher has to evaluate the results the study yielded: what do they amount to? The question posed here is not merely whether the results favour the theory that was tested. The goal is to provide a critical review of the way in which the data were collected, the steps of reasoning employed, questions of operationalization, any possible alternative explanations, as well as what the results themselves entail. The results must be put in a broader context and discussed. Perhaps the conclusions will also lead to recommendations, for example, recommendations for clinical applications or for educational practice. This is also the appropriate moment to suggest ideas for alternative or follow-up studies.

During this phase, the aim is primarily to interpret the results, a process in which the researcher plays an important and personal role as the one who provides the interpretation. Different researchers may interpret the same results in widely different ways. Finally, in some cases, results will contradict the outcome that was predicted or desired.

2.6 Making choices

Research consists of a sequence of choices: from the inspirational observations during the first phase, to the operational decisions involved in performing the actual study, to interpreting the results during the last stage. Rarely will a researcher be able to make the best decision for every choice point, but they must remain vigilant of the possibility of making a bad decision along the way. The entire study is as strong as the weakest link: the entire study is as good as the worst choice in its sequence of choices. As an illustration, we will provide an overview of the choices a researcher has to make throughout the empirical cycle.

The first choice that has to be made concerns the formulation of the problem. Some relevant questions that the researcher has to answer at that moment include: how do I recognize a certain research question, is research the right choice in this situation, is it possible to research this idea? The best answers to such

questions depend on various factors, such as the researcher's view of humankind and society, any wishes their superiors or sponsors might have, financial and practical (im)possibilities, etc.

The research question does have to be answerable given the methods and means available. However, within this restriction, the research question may relate to any aspect of reality, regardless of whether this aspect is seen as irrelevant or important. There are many examples of research that was initially dismissed as irrelevant, but, nevertheless, did turn out to have scientific value, for instance, a study on the question: "is 'Huh?' a universal word?" (Dingemanse et al., 2013) (Example 1.1). In addition, some ideas that were initially dismissed as false later did turn out to be in accordance with reality. For instance, Galilei's statement that Earth revolved around the Sun once was called unjustified. In short, research questions should not be rejected too soon for being 'useless', 'platitudes', 'irrelevant', or 'trivial'.

If the researcher decides to continue their study, the next step is usually studying the literature. Most research handbooks recommend doing a sizeable amount of reading, but how is an appropriate collection of literature found? Of course, the relevant research literature on the area of knowledge in question must be looked at. Fortunately, these days, there are various resources for finding relevant academic publications. For this, we recommend exploring the pointers and so-called "libguides" offered by the Utrecht University Library (see http://www.uu.nl/library and http://libguides.library.uu.nl/home_en). We would also like to warmly recommend the guide by Sanders (2011), which contains many extremely helpful tips to use when searching for relevant research literature.

During the next phase, the first methodological problems start appearing: the researcher has to formulate the problem more precisely. One important decision that has to be made at that point is whether the problem posed here is actually suited for research (§2.4). For instance, a question like "what is the effect of the age of onset of learning on fluency in a foreign language?" cannot be researched in this form. The question must be specified further. Crucial concepts must be (re)defined: what is the age of onset of learning? What is language fluency? What is an effect? And how do we define a foreign language? How is the population defined? The researcher is confronted with various questions regarding definitions and operationalization: Is the way concepts are defined theoretical, or empirical, or pragmatic in nature? Which instruments are used to measure the various constructs? But also: what degree of complexity should this study have? Practically speaking, would this allow for the entire study be completed? In which way should data be collected? Would it be possible at all to collect the desired data, or might respondents never be able or willing to answer such questions? Is the proposed manipulation ethically sound? How great is the distance between the theoretical construct and the way in which it will be measured? If anything goes wrong during this phase, this will have a direct effect upon the rest of the study.

If a problem has been successfully formulated and operationalized, a further ex-

ploration of the literature follows. This second bout of literature study is much more focussed on the research question that has been worked out by this point, compared to the broad exploration of the literature mentioned earlier. On the grounds of earlier publications, the researcher might reconsider their original formulation of the problem. Not only does one have to look at the literature in terms of theoretical content, but one should also pay attention to examples of how core concepts are operationalized. Have these concepts been properly operationalized, and if there might be different ways of operationalizing them, what is the reason behind these differences? In addition, would it be possible to operationalize the core concepts in such a way that the distance between the concept-as-intended and the concept-as-defined become (even) smaller (§**??**)? The pointers given above with regard to searching for academic literature are useful here, as well. After this, the research is to (once again) reflect upon the purpose of the study. Depending on the problem under consideration, questions such as the following should be asked: does the study contribute to our knowledge within a certain domain, does the study create solutions for known stumbling blocks or problems, or does the study contribute to the potential development of such solutions? Does the research question still cover the original problem (or question) identified by superiors or sponsors? Are the available facilities, funds, and practical circumstances sufficient to conduct the study?

During the next step, the researcher must specify how data will be collected. This is an essential step, which influences the rest of the study; for this reason, we will devote an entire chapter to it (Chapter **??**). What constitutes the population: language users? Students? Bilingual infants? Speech errors involving consonants? Sentences? And what is the best way to draw a representative sample (or samples) from this population (or populations)? What sample size is best? In addition, this phase involves choosing a method of analysis. Moreover, it is advisable to design a plan of analysis at this stage. Which analyses will be performed, what ways of exploring the data are envisioned?

All the choices mentioned so far are not yet sufficient for finishing one's preparations. One must also choose one's instruments: which devices, recording tools, questionnaires, etc., will be used to make observations? Do suitable instruments already exist? If so, are these easily accessible and does the researcher have permission to use them? If not, instruments must be developed first (§**??**). However, in this latter case, the researcher must also take the task upon themselves to first test these instruments: to check whether the data obtained with these instruments conform to the quality standards that are either set by the researcher or that may be generally expected of instruments used in scientific research (in terms of reliability and validity, see Chapters 5 and 12).

It is only when the instruments, too, have been prepared that the actual empirical study begins: the selected type of data is collected within the selected sample in the selected manner using the selected instruments. During this phase, also, there are various, often practical problems the researcher might encounter. An example from actual practice: three days after a researcher had sent out their

questionnaire by mail, a nationwide mail workers' strike was set in motion and lasted two weeks. Unfortunately, the researcher had also given the respondents two weeks' notice to respond by mail. This means that, once the strike was over, the time frame the subjects were given to respond had already passed. What was the researcher to do? Lacking any alternatives, our protagonist decided to approach each of the 1020 respondents by phone, asking them to fill out the questionnaire regardless and return it at their earliest convenience.

For the researcher who has invested in devising a plan of analysis in advance, now is the time of harvest. Finally, the analyses that were planned can be performed. Unfortunately, reality usually turns out to be much more stubborn than the researcher might have imagined beforehand. Test subjects might give unexpected responses or not follow instructions, presumed correlations turn out to be absent, and unexpected (and undesirable) correlations do turn out to be present to a high degree. Later chapters will be devoted to a deeper exploration of various methods of analysis and problems associated with them.

Finally, the researcher must also report on their study. Without an (adequate) research report, the data are not accessible, and the study might as well *not* have been performed. This is an essential step, which, among other things, involves the question of whether the study may be checked and replicated based on the way it is reported. Usually, research activity is reported in the form of a paper, a research report, or an article in an academic journal. Sometimes, a study is also reported on in a rather more popular journal or magazine, targeted towards an audience broader than just fellow researchers.

This concludes a brief overview of the choices researchers have to make when doing research. Each empirical study consists of a chain of problems, choices, and decisions. The most important choices have been made before the researcher starts collecting data.

Chapter 3

Integrity

3.1 Introduction

Scientific research has brought humanity immeasurably great benefits, such as reliable computing technology, high-quality medical care, and an understanding of languages and cultures that are not our own. All these assets are based on scientifically motivated knowledge. Researchers produce knowledge, whose progress and growth comes about because researchers build upon their predecessors' experience and insights.

Example 3.1*: Sir Isaac Newton wrote of his scientific work: "If I have seen further it is by standing on [the] shoulders of Giants" (in a letter addressed to Robert Hooke, dated 5 Feb 1675)[1]. This metaphor can be traced back to the medieval scholar, Bernard de Chartres: "nos esse quasi nanos gigantum umeris insidentes" [that we are like dwarfs sat on giants' shoulders] compared to scholars from the times of Antiquity. The aforementioned quote by Newton has also become Google Scholar's motto (scholar.google.com).

In this chapter, we will discuss the ethical and moral aspects of scientific research. Science is done by human beings, and requires a well-developed sense

[1] A copy of this letter may be found at https://digitallibrary.hsp.org/index.php/Detail/objects/9792; for some background information, see http://www.bbc.co.uk/worldservice/learningenglish/movingwords/shortlist/newton.shtml.

of judgment on the part of the researchers. The *Netherlands Code of Conduct for Research Integrity* (VSNU, 2018) (https://www.vsnu.nl/en_GB/research-integrity) describes how researchers (and students) are to behave. According to this code of conduct, scientific research and teaching should be based on the following principles:

- honesty,
- diligence,
- transparency,
- independence, and
- responsibility.

The following sections will go over how these principles are to be implemented in our actions during the various phases of scientific research. How are we to set up a study, collect and process data, and report on our study in a way that is honest, diligent, transparent, independent, and responsible? This is something we have to think about even before we start working on our project, which is why these topics are discussed at the beginning of this reader, even though we will also refer to terms and concepts that will be worked out in more detail in subsequent chapters.

3.2 Design

To be sure, scientific research does bring us immeasurably great benefits, but this is balanced by considerable cost. This includes direct expenses, such as setting up and maintaining laboratories, equipment, and technical support, but also researchers' salaries, financial compensation for informants and test subjects, travel expenses for access to libraries, archives, informants, and test subject, etc. These direct expenses are usually subsidized by public funds held by universities and other academic institutions. In addition, there is an indirect cost, which is partially borne by informants and test subjects: time and effort that can no longer be spent on something else, loss of privacy, and possibly other risks we are not yet aware of. One often forgotten type of cost is loss of naïveté: a test subject who has participated in an experiment learns from it, and, because of this, will possibly respond differently in a subsequent experiment (see §5.4, under History). This means that any results obtained in this subsequent experiment will generalize less well to other subjects who have a different history, and have not yet participated in a study.

Given its great cost, research has to be thought through and designed in such a way that its expected benefits are reasonably balanced by its expected cost (Rosenthal and Rosnow, 2008, Ch.3). If the chance that a study will yield valid

conclusions is very low, it is better *not* to go ahead with this study, which will save on both direct expenses and indirect cost.

Example 3.2: Suppose that we would like to examine whether 4-year-old bilingual children might have a cognitive advantage over monolingual children of the same age. Based on earlier research, we expect a different of at least 2 points (on a 10 point scale) between both groups (with a "pooled standard deviation" $s_p = 4$, hence $d = 0.5$, §**??** and §**??**).

We then compare two group of $n = 4$ children each. Even if there were actually a difference of 2 points between both groups (meaning, if the hypothesis were true), this study would still have a mere 51% chance of yielding a significant difference: the power of the experiment is only .51 (Chapter 14), because the two groups contain so few test subjects. It would be better for the four-year-olds and their parents to do engage in other activities (at school, at home, or at work) instead of participating in this study.

However, if $n = 30$ children would participate in each of the two groups, and if there were indeed a 2 point difference between both groups (meaning, if the research hypothesis were true), then the power of the experiment would be .90. This means that bigger groups lead to a much better chance of confirming our study's hypothesis. This elaborate research design will cost more (for the researchers and the children and their parents), but presumably it will also yield much more: a valid conclusion with great impact on society.

A study's design (see Chapter **??**) has to be as efficient as possible, and the researcher has to start thinking about it at an early stage. First of all, efficiency depends on choices regarding how the independent variables are manipulated. Is there a separate group of test subjects for each condition of the independent variable (meaning we have "between-subjects" conditions, like in example 3.2 above)? In a between-subjects design that involves two groups, we need about $n = (5.6/d)^2$ subjects in each group (for further explanation of this, see Gelman and Hill (2007), and see §13.8). Or are all test subjects involved in all conditions (meaning we have "within-subjects" conditions)? A within-subjects design with two conditions requires only $n = (2.8/d)^2$ subjects in each condition, and the study will therefore also have lower expenses and indirect cost

for a much smaller number of test subjects. In general, this means that, if possible, it is much better to manipulate independent variables within subjects than it is between subjects. However, this is not always possible, firstly because individual characteristics only differ between subjects by definition (for example: female/male sex, multilingual/monolingual youth, aphasia/no aphasia, etc.). Secondly, we must take proper care to recognize effects of so-called transfer between conditions, which threaten our study's validity (for example: experience, learning, fatigue, maturation). We will return to this in §**??**.

Being multilingual or being female are characteristics that may only vary between individuals. But other conditions may also vary within individuals, for instance, the day on which a cognitive measurement is taken. Suppose that we expect a difference of $D = 2$ points between cognitive measurements taken on Monday and on Friday, respectively (with $s = 4$ and $d = 0.5$, see example 3.2). If we manipulate the day of measurement between subjects, meaning we make separate groups for children tested on Monday and those tested on Friday, this entails that we need $n = (5.6/0.5)^2 = 126$ children in each group, yielding a total of $N = 252$. However, if we manipulate the day of measurement within subjects, meaning that we observe each test subject on Monday and also on Friday, this entails that we need a total of just $N = (2.8/0.5)^2 = 32$ children. The within-subjects design means that far fewer children's routines will need to be disturbed for our cognitive measurements. However, we must be properly aware of learning effects between the first and second measurement, and take appropriate precautionary measures. For instance, we can no longer use the same questionnaires in both conditions.

A study's efficiency also depends on the dependent variable, in particular, on the observations' level of measurement (Chapter **??**), accuracy, and reliability (Chapter 12). The lower the level of measurement, the lower also the study's efficiency. As accuracy goes down, the study's efficiency also goes down, and more subjects and observations will be needed to be able to draw valid conclusions.

Example 3.3: Suppose that we would like to examine a difference between two within-subjects conditions, and suppose that the actual difference between them is 2 points (which yields $s_D = 4$ and $d = 0.5$, see example 3.2). However, suppose that we decide to look only at the *direction* and not at the size of the difference between the two observations for each subject: does the subject have a positive or negative difference between the first and second condition? This binomial dependent variable contains less information than the original point score (it contains just the direction and not the size of the difference), making the study less efficient. For this specific example, this means we would need 59 instead of 34 test subjects.

Thus, researchers are responsible for diligently and honestly considering and balancing their study's cost and benefits, and they need to have a sufficient methodological background to be able to choose a proper research design, taking in account time constraints, the available test subjects and instruments of measurement, etc.

3.3 Participants and informants

Scientific research is done by human beings: researchers are but human. In the realm of humanities, these researchers themselves study (other) human beings' behaviour and intellectual products. These activities are governed by laws, rules, guidelines, and codes of conduct that researchers (and students!) must follow, stemming from the aforementioned principles of diligence and responsibility. A study and the data collected for it may not lead to any kind of harm or significant loss of privacy for the parties involved.

For research in the humanities in the Netherlands, two laws are relevant:

- The General Data Protection Regulation (GDPR), see https://autoriteitpersoonsgegevens.nl/nl/onderwerpen/avg-europese-privacywetgeving (in Dutch) or https://ec.europa.eu/info/law/law-topic/data-protection/data-protection-eu_en,
- Wet Medisch-wetenschappelijk Onderzoek met mensen (WMO; English: Medical Research Involving Human Subjects Act), see https://wetten.overheid.nl/BWBR0009408/2019-04-02 (in Dutch) or https://english.ccmo.nl/investigators/legal-framework-for-medical-scientific-research/laws/medical-research-involving-human-subjects-act-wmo

It is compulsory to ask participants (or their legal guardians) for their explicit informed consent. This means that participants are fairly informed about the study, about its cost and benefits, and about their remuneration, and that, after this, they explicitly consent to participate. For researchers and students at Utrecht University, helpful examples of informed consent (information letters and consent statements) can be found on the website of the Faculty Ethics Review Committee for the Humanities (FETC-GW, discussed in more detail below), via https://fetc-gw.wp.hum.uu.nl.

All data that may be used to identify an individual are considered to be *personal data*, which may only be collected and processed according to the GDPR. It is advisable to separate one's research data from any personal data as early as possible, which means anonymizing the data. Any information that links personal data and research data (e.g., a list with test subjects' names and their corresponding anonymous personal code) is, itself, confidential and must be

saved and stored with care. Do not keep personal data any longer than necessary. Research data may only be used for the (scientific) goal for which they were collected. Make sure that participants and informants are not recognizable in reports and publications on the study (i.e., use anonymous codes).

Photos and recordings of individuals (including audio, video, physiological data, and EEG) are subject to what we call *portrait rights*. This means that photos and other identifying recordings are considered on a par with portraits. When such a photo or recording is published, the person shown or represented may appeal to their portrait rights and claim damages for the harm done to them by this publication. This means that, if you might be interested in publishing a recording from which someone could be recognized, you must ask the individual who was recorded or their legal guardian for explicit consent beforehand (see above for the notion of informed consent). This also applies if you intend to demonstrate or show a fragment of such a recording at a conference presentation or on a website.

The Dutch WMO law (see above) states that any research involving human subjects must first be approved by a special committee; for the Faculty of Humanities at Utrecht University, this is the Medical Ethics Assessment Committee (Medisch-Ethische Toetsingscommissie or METC), which is administered by the University Medical Centre (UMC). This committee assesses whether the possible benefits of a study are reasonably balanced against the costs and possible harm done to test subjects.

Most research in languages and communication at Utrecht University is exempt from review by the METC, which would otherwise be time-consuming, but it must be submitted to the Faculty Ethics Review Committee for the Humanities (Facultaire Ethische Toetscommissie - Geesteswetenschappen or FETC-GW, see https://fetc-gw.wp.hum.uu.nl/en/). However, this does not apply to research done by students, provided that some conditions apply. You can find more information on the FETC-GW website. When in doubt, always consult with your supervisor or teacher. This ethics assessment is also compulsory for students and researchers in other fields (literature, history, media & culture) who plan to do research with human subjects.

3.4 Data

The data collected form the motivation and empirical basis for the conclusion drawn from scientific research. These data therefore have an essential importance: no data means no valid conclusions. Moreover, as we saw above (§**??**), these data are very costly (in terms of time, money, privacy, etc.). This means that we should treat them very diligently. We must be able to convince others of our conclusions' validity based on these data, and we must be able to share the underlying data with other researchers, if asked.

Thus, diligence requires, at the very least, making a sufficient number of backup copies as soon as possible. Think of what might happen if a fire or flood would completely destroy the place where you work or live, or if your laptop would be stolen during your thesis project (this actually happened to one of our students!). If so, would proper and recent copies of the data be stored in other locations? For storing backup copies, a sufficiently secured cloud service[2] is a good option.

Diligence also requires a proper record of what the data stand for, and how they were collected. Data without a matching description are practically useless for scientific research. Charles Darwin carefully noted down which bird found on which of the Galapagos Islands had which beak shape, and these observations later formed (a part of) the motivation for his theory of evolution. In the same way, we strongly encourage you to keep a log (on paper or digitally) of all steps of your research study, including motivations for these steps, if needed. Also note the brand, type, and settings used for any equipment you use, and note the version and settings for any software used. Keep a record of which processing steps were applied to the data, and why, and which file contains which data.

If you are working with digitized data (e.g., in Excel, or SPSS, or R), make sure to carefully keep track of which variable is stored in which column, using which unit of measurement and which coding scheme.

Example 3.4: The file found at <http://tinyurl.com/nj4pjaq. contains data from 80 speakers of Dutch, partially taken from the Corpus of Spoken Dutch (Corpus Gesproken Nederlands or CGN). The first line contains the variable names. Each subsequent line corresponds to one speaker. The pieces of data on each line are separated by spaces. The first column contains the anonymized speaker ID code, as used in the CGN. In the fifth column, the speaker's region of origin is coded with a single letter: `W` for Western region (Randstad), `M` Central (Mid), `N` North, `S` South) (Quené, 2008). Because of the careful annotation, these data may still be used with no problem, even if they were collected over 20 years ago by fellow researchers.

Data remain the intellectual property of those who collected them. Use of other researchers' data with no citation may be seen as theft or plagiarism.

Data fraud (fabricating data, meaning, coming up with data out of thin air, instead of observing them) is obviously at odds with multiple principles in the code of conduct mentioned above (VSNU, 2018). Fraud harms the mutual trust

[2]Students and employees of most Dutch educational institutions can use SurfDrive (https://www.surfdrive.nl) for easy data storage on secured servers.

on which science is based. It misleads other researchers who might be building on the fictional results, and any research funds allotted to a fraudulent line of research are taken away from other, non-fraudulent research – in other words, it is a mortal sin of academia. If you would like to discuss any questions or dilemmas around this topic, please contact prof.dr. Christoph Baumgartner, confidential advisor on academic integrity for the Humanities at Utrecht University (`c.baumgartner@uu.nl`).

3.5 Writing

Scientific research only really reaches its purpose once its results are being divulged. Research that is not reported on could as well *not* have been conducted at all, and the cost associated with this research was, basically, spent in vain. For this reason, reporting research results is an important part of academic work. Publications (as well as patents) form a very important part of the "output" of scientific research. Researchers are measured by the number of their publications and these publications' "impact" (the number of times these publications are cited by others who build upon them). This great importance is one of the reasons we ought to be diligent in treating others' writings, as well as our own.

The researchers involved in a study must discuss amongst each other who will be listed as authors of a report or publication, and in which order. Those listed as co-authors of a research report have to satisfy three conditions (Office of Research Integrity, 2012, Ch.10). Firstly, the must have made a substantial academic contribution to one or more phases of the study: think of the original idea, setting up and designing the study, collecting the data, or analysing and interpreting the data. Secondly, they must have been a part of writing up the report, either by doing part of the writing or by providing comments on it. Thirdly, they must have approved the final version of the report (most often implicitly, sometimes explicitly), and they must also have consented to being a co-author. It is best practice for the researchers to come to a mutual agreement on the order in which their names are listed. Usually, names are ordered by decreasing importance and extent of each author's contribution. If the lead researcher is the main investigator and also a co-author, this person is often listed last.

Example 3.5: A, a student research assistant, helped collect data, but has made no other contributions, and is not entirely sure what the research is about. This means that A need not be listed as a co-author on the report, but the authors do have to describe and acknowledge A's contribution in their report.

> B, another student, conducted one of the parts of the research project supervised by researcher C. Supervisor C thought of the entire project, but B has collected literature, set up and conducted one part of the study, collected, analysed, and interpreted data, and reported on this all in a paper. Because of this, B and C are both co-authors of a publication on B's part of the research project. They come to an agreement on the order in which authors are listed. Because student B was the most prominent person in the work, while C was the main investigator, they agree that B will be first author and C will be second (and last) author.

Researchers build upon their predecessors' work (see example 3.1). This may also involve building upon their arguments and even their writing, but these cases do require that we always correctly refer to the appropriate source, i.e., to these predecessors' work. After all, if we did not do this, we could no longer distinguish who is responsible for which thought or which fragment of writing. Plagiarism is "copying others' documents, thoughts, arguments, and passing them off as one's own work" (Van Dale, 12th edition [our translation]). This form of fraud is also a mortal sin of academia that may lead to substantial sanctions. The Faculty of Humanities at UU has the following to say about it:

> Plagiarism is the appropriation of another author's works, thoughts, or ideas and the representation of such as one's own work. The following are some examples of what may be considered plagiarism:
>
> - Copying and pasting text from digital sources, such as encyclopaedias or digital periodicals, without using quotation marks and referring to the source;
> - Copying and pasting text from the Internet without using quotation marks and referring to the source;
> - Copying information from printed materials, such as books, periodicals or encyclopaedias, without using quotation marks and referring to the source;
> - Using a translation of the texts listed above in one's own work, without using quotation marks and referring to the source;
> - Paraphrasing from the texts listed above without a (clear) reference: paraphrasing must be marked as such (by explicitly linking the text with the original author, either in text or a footnote), ensuring that the impression is not created that the ideas expressed are those of the student;
> - Using another person's imagery, video, audio or test materials without reference and in so doing representing them as one's own work;

- Resubmission of the student's own earlier work without source references, and allowing this to pass for work originally produced for the purpose of the course, unless this is expressly permitted in the course or by the lecturer;
- Using other students' work and representing it as one's own work. If this occurs with the other student's permission, then he or she may be considered an accomplice to the plagiarism;
- When one author of a joint paper commits plagiarism, then all authors involved in that work are accomplices to the plagiarism if they could have known or should have known that the other was committing plagiarism;
- Submitting papers provided by a commercial institution, such as an internet site with summaries or papers, or which have been written by others, regardless of whether the text was provided in exchange for payment.

https://students.uu.nl/en/practical-information/policies-and-procedures/fraud-and-plagiarism

In the case of self-plagiarism, the fragments or writing or thoughts in question are not taken from others, but from one of the authors. There are various schools of thought on self-plagiarism; however, it is advisable to be sure to cite the relevant source if one is to take ideas from one's own work, building on the principles of diligence, reliability, transparency, and responsibility.

A reference or citation is a shortened mention of a source in the body of the text; you might have seen these quite a few times in this syllabus already. At the end of the report or text, a full list of sources follows, which is usually given the heading, "Sources", "Sources consulted", "References", "Literature", or "Bibliography". A mistake in the references may be seen as a form of plagiarism (Universiteitsbibliotheek, Vrije Universiteit Amsterdam, 2015) because the reader is directed towards an incorrect source. For this reason, it is imperative that researchers cite their sources correctly. Various conventions, depending on the area of study, have been developed for this. Usually, instructors will indicate which style or convention is to be used for citing one's sources. In this textbook, we have intended to follow the style described by the American Psychological Association (2010), a style commonly used in the social sciences and some disciplines within the humanities. (For technical reasons, references may deviate slightly from the APA style.)

The rules for citing sources may sometimes be complex. In addition, authors must make sure that the citations in the body of the text correspond to the list of full references at the end. These tasks are best performed by a so-called "reference manager", a program that collects references or citations, and correctly inserts them into the body of the text. An overview of such programs can be found at https://en.wikipedia.org/wiki/Comparison_of_

reference_management_software. In writing this textbook we have used Zotero (https://www.zotero.org), combined with BibTeX (https://www.bibtex.org).

Chapter 4

Levels of measurement

4.1 Introduction

In Chapter 2, we were already introduced to variables: properties that can take different values. As we know, a variable's value is a way of indicating a property or quality of an object or person. If we are dealing of a dependent variable, this value may also be called a *score* or *response*, often represented with the symbol Y. The way in which a property is expressed in a measurable value is called the variable's *level of measurement*; thus, level of measurement is an inherent property of the variable itself! We distinguish four levels of measurement, in order of increasing informativeness: nominal, ordinal, interval, ratio. For the former two levels of measurement, only discrete categories are distinguished, with or without ordering. The latter two levels of measurement use numerical values, with or without a zero point. We will discuss the levels of measurement in more detail below. Insight into a variable's level of measurement is important for interpreting scores for that variable, and – as we will see later – for choosing the correct statistical test to answer a research question.

4.2 Nominal

We speak of a nominal variable (or a nominal level of measurement) when a property is categorized into separate (discrete) categories that have no order between them. Well-known examples include a test subjects nationality, a car's make, the colour of someone's eyes, the flavour of a tub of ice cream, one's living arrangements (with one's family, with housemates, living independently, living with a partner, other), etc. Scores may only be used to distinguish between the categories (a statement like, "vanilla is different from strawberry" does make sense). We can, indeed, count how often each category occurs, but there is no

interpretable order (the statement, "vanilla is larger than strawberry" does not make sense), and we can also not do any arithmetic on the values measured for a nominal variable. For instance, we can determine the most frequently occurring nationality, but we cannot calculate the average nationality.

4.3 Ordinal

We speak of an ordinal variable (or an ordinal level of measurement) when a property is categorized into separate categories that do have an *order* or ranking between them. However, in the case of an ordinal variable, we do not know anything about the distance between the various categories. Well-known examples include level of education (primary education, secondary education, bachelor's degree, master's degree/PhD, …), answer on a scale question (*agree, do not know, disagree*), position within a ranking, order of elimination in a talent show, clothing size (XS, S, M, L, XL, …), or military rank (soldier, major, general, …). Here, as well, we can count how often each category occurs, and we can even sensibly interpret the rank order (whoever is eliminated last has performed better than whoever is eliminated first, size L is greater than size M, a general outranks a major). However, we still can do no arithmetic on the values measured for an ordinal variable. We may determine the bestselling clothing size, but we cannot calculate the average clothing size sold[1].

4.4 Interval

We speak of an interval variable (or an interval level of measurement) when a property is expressed as a number on a continuous scale for which there is *no zero point*. Because of the scale, we know what the distances or intervals are between the various values of an interval variable. Well-known examples include temperature in degrees Celsius (the zero point is arbitrary) or calendar year (ditto for this zero point). We can count how often each category occurs, we can sensibly interpret the rank order (in our Gregorian calendar, the year 1999 preceded the year 2000), and we can also sensibly interpret the intervals (the interval between 1918 and 1939 is just as long as that between 1989 and 2010). We may, indeed, do arithmetic on the values of an interval variable, but the only operations that make sense are addition and subtraction. These are enough to calculate an average, e.g., the average year in which the individuals in the sample obtained their first mobile phone.

[1] If half of our respondents answers *agree*, and the other half answers *disagree*, it does not make sense to conclude that the responses are *neutral* on average.

4.5 Ratio

The fourth and highest level of measurement is the ratio level. We speak of a ratio variable (or a ratio level of measurement) when a property is expressed as a number on a continuous scale for which there is, indeed, a *zero point*. Because of the scale, we know what the distances or intervals are between the various values of a ratio variable. In addition, because of the zero point, we know what the proportions or ratios are between the various values (hence the name of this level). Well-known examples include temperature in degree Kelvin (measured from absolute zero), response time[2] in thousandths of a second (ms), your height in centimetres (cm), your age in years, the number of errors made on a test, etc. When we are dealing with a ratio variable, we can count how often each category occurs, we can sensibly interpret the rank order (someone whose height is 180 cm is taller than someone whose height is 179 cm), we can sensibly interpret intervals (the increase in age between 12 and 18 is two times as large as that between 9 to 12), and we can also sensibly interpret proportions between the values (the age of 24 is *twice* as great as the age of 12). We may do arithmetic on the values of a ratio variable, which includes not just addition and subtraction, but also division and multiplication. Here, as well, it is possible to calculate an average, e.g., the average age at which the individuals in the sample obtained their first mobile phone.

4.6 Ordering of levels of measurement

In the above, we have discussed the levels of measurement in order of increasing informativeness or strength. A nominal variable contains the least amount of information and is considered the lowest level of measurement, while a ratio variable contains the greatest amount of information and is considered the highest level of measurement.

It is always possible to reinterpret data measured at a high level of measurement as if they had been measured at a lower level. For instance, if, for each individual in a sample, we had measured their monthly income at a ratio level (in €), we would be able to make an ordinal variable out of this with no problem (e.g. *less than average, average to twice the average, more than twice the average*). This would mean discarding information: the original measurements in terms of € contain more information than the classification into three ordered categories derived from it.

Of course, the opposite is not possible: a variable at a low level of measurement cannot be reinterpreted at a higher level. We would have to add, after the fact, information that we did not collect during the original measurement of this variable. It is therefore imperative to observe the relevant variables at the

[2]The zero point is the moment in time when the event begins that the participant is to respond to.

correct level of measurement. Supposed we wanted to compare body height in adult men and women. If we measure body height at an ordinal level (having defined three categories, *short, medium, and tall*, equally for all individuals), this means that we cannot calculate the average body length, and we can also not use any statistical test that would refer to the average body length. This does not have to be a problem, but it is a good idea to think through the consequences of using a particular level of measurement in advance of the actual measurement.

Chapter 5

Validity

5.1 Introduction

The goal of experimental research is to test a hypothesis. Hypotheses may also be tested in other, non-experimental research, but in the following, we will restrict ourselves to experimental research, i.e., research that uses the experiment as its methodology, for the sake of clarity. In experimental research, we attempt to argue for the plausibility of a causal relationship between certain factors. If an experiment study has results that confirm the research hypothesis (i.e., the null hypothesis is rejected), it is plausible that a change in the independent variable is the *cause* of a change, or *effect*, in the dependent variable. In this manner, experimental research allows us to conclude with some degree of certainty that, for instance, a difference in medical treatment after a stroke is the cause, or an important cause, of a difference in patients' language ability as observed 6 months post-stroke. The experiment has made it plausible that there is a causal relationship between the method of treatment (independent variable) and the resulting language ability (dependent variable).

5.2 Causality

A causal relationship between two variables is distinct from 'just' a relationship or correlation between two variables. If two phenomena correlate with one another, one does not have to be the cause of the other. One example can be seen in the correlation between individuals height and their weight: tall people are, in general, heavier than short people (and vice versa: short people are generally lighter than tall people). Does this mean that we can speak of a causal relationship between height and weight? Is one property (partially) cause by the other? No: in this example, there is, indeed, a correlation, but no causal

relationship between the properties: both height and weight are "caused" by other variables, including genetic properties and diet. A second example is the correlation between motivation and language ability in second language learners: highly motivated students learn to speak a new second language better and more fluently than those with low motivation do, but here, also, it is unclear which is the cause and which is the effect.

A causal relationship is a specific type of correlation. A causal relationship is a correlation between two phenomena or properties, for which there are also certain additional requirements (Shadish et al., 2002). Firstly, the cause has to precede the effect (it is after treatment that improvement occurs). Secondly, the effect should not occur if the cause is not there (no improvement without treatment). Moreover, the effect – at least, in theory – should always occur whenever the cause is present (treatment always results in improvement). Thirdly, we cannot find any plausible explanation for the effect's occurrence, other than the possible cause we are considering. When we know the causal mechanism (we understand why treatment causes improvement), we are better able to exclude other plausible explanations. Unfortunately, however, this is very rarely the case in the behavioural sciences, which include linguistics. We do determine that a treatment results in improvement, but the theory that ties cause (treatment) and effect (improvement) together is rarely complete and has crucial gaps. This means that we must take appropriate precautionary measures in setting up our research methodology in order to exclude any possible alternative explanations of the effects we find.

5.3 Validity

A statement or conclusion is *valid* whenever it is *true* and *justified.* A true statement corresponds to reality: the statement that *every child learns at least one language* is true, because this statement appropriately represents reality. A justified statement lends its validity from the empirical evidence upon which is it based: every child observed by us or by others is learning or has learned a language (except for certain extraordinary cases, for which we need a separate explanation). A statement's justification becomes stronger with an increasingly stronger and more reliable method of (direct or indirect) observation. This also means that a statement's validity is not a categorical property (valid/not valid) but a gradual property: a statement can be relatively more or less valid.

Three aspects of a statement's validity may be distinguished:

1. To which degree are the conclusions about the relationships between the dependent and independent variable valid? This question pertains to *internal validity.*

2. To which degree are the operationalizations of the dependent and independent variable (the ways in which they are worked out) adequate? This

question pertains to *construct validity*.

3. To which degree can the conclusions reached be generalized to other test subjects, stimuli, conditions, situations, observations? This question pertains to *external validity*.

These three forms of validity will be further illustrated in the following sections.

5.4 Internal validity

As you already know, it is our aim in an experimental study to exclude as many alternative explanations of our results as possible. After all, we must demonstrate that there is a causal relationship between two variables, X and Y, and this means keeping any confounding factors under control as much as possible. Let us take a look at example 5.1.

Example 5.1: Verhoeven et al. (2004) investigated (among others) the hypothesis that older individuals (above 45 years old) speak more slowly than younger individuals (under 40 years old). To do this, they recorded speech from 160 speakers, divided equally between both age groups, in an interview that lasted about 15 minutes. After a phonetic analysis of their articulation rate, it turned out that the younger group spoke relatively fast at 4.78 syllables per second, while the older group spoke remarkably slower at 4.52 syllables per second (Verhoeven et al., 2004, p.302). We conclude that the latter group's higher age is the *cause* of their lower rate of speaking – but is this conclusion justified?

This question of a conclusion's justification is a question about the study's internal validity. Internal validity pertains to the relationships between variables that are measured or manipulated, and is not dependent on the (theoretical) constructs represented by the various variables (hence the name 'internal validity'). In other words: the question of internal validity is one of possible alternative explanations of the research results that were found. Many of the possible alternative explanations can be pre-empted by the way in which the data are collected. In the following, we will discuss the most prominent threats to internal validity (Shadish et al., 2002).

1. **History** is a threat to internal validity. The concept of 'history' includes, among others, events that took place between (or during) pretest and posttest; here, 'history' refers to events that are not a part of an experimental manipulation (the independent variable), but that might influence the dependent variable. For instance, a heat wave can influence test subjects' behaviour in a study.

In a laboratory, 'history' is kept under control by isolating test subjects from outside influences (such as a heat wave), or by choosing dependent variables that could barely be influenced by external factors. In research performed outside of a laboratory, including field research, it is much more difficult and often even impossible to keep outside influences under control. The following example makes this clear.

Example 5.2: A study compares two methods to teach students at a school to speak a second language, in this case, Modern Greek. The first group is to learn Greek words and grammar in a classroom over a period of several weeks. The second group goes on a field trip to Greece for the same period of time, during which students have to converse in the target language. The total time spent on language study is the same for both groups. Afterwards, it turns out that the second groups' language ability is higher than that of the first group. Was this difference in the dependent variable's value indeed caused by the teaching method (independent variable)?

---

2. **Maturation** stands for participants' natural process of getting older, or maturing, during a study. If the participants become increasingly older, more developed, more experienced, or stronger during a study, then, unless this maturation was considered in the research question, maturation forms a threat to internal validity. For instance, in experiments in which reaction times are measured, we usually see that a test subject's reaction times become faster over the duration of the experiment as a consequence of training and practice. In such cases, we can protect internal validity against this learning effect by offering stimuli in a separate random order for each test subject.

In most cases, maturation occurs because participants perform the same task or answer the same questions many times in a row. However, maturation can also happen when participants are asked to provide their answers in a way they are not used to, e.g., because of an unexpected way of asking the question, or

through an unusual type of multiple choice question. During the first few times a test subject answers a question in such a study, the method of answering may interfere with the answer itself. Afterwards, we could compare, e.g., the first and the last quarter of a test subject's answers to check whether there might have been an effect of experience, i.e., maturation.

3. The **instrumentation** or instruments used for a study may also form a threat to internal validity. Different instruments that are deemed to be measuring the same construct should always produce identical measurements. Conversely, the same instrument should always produce identical measurements under different circumstances. For experiments administered by a computer, this is usually not a problem. However, in the case of questionnaires, or the assessment of writing assignments, internal validity may, indeed, be under threat.

In many studies, observations are made both before a treatment and after. Identical tests could be used for this, but that might lead to a learning effect (see above). For this reason, researchers often use different tests between pretest and posttest. However, this might lead to an instrumentation effect. The researcher has to consider the possible advantages and disadvantages of each option.

> *Example 5.3*: Rijlaarsdam (1986) investigated the effect of peer evaluation on the quality of writing. The setup of his study was as follows (with some simplifications): first, students write an essay on topic A, followed by writing instruction that includes peer evaluation, after which the students write another essay – this time, on topic B. The writing done in the pretest and posttest is assessed, after which the researchers test whether average performance differs between both measurements.
>
> In this study, it is not only the intervention (writing instruction) that forms a clear difference between the pretest and posttest: the writing assignment's topic (A or B) differs, as well. It is doubtful whether both writing assignments measure the same thing. This difference of instrumentation threatens internal validity because it may well be that, at different moments, a (partially) different aspect of writing ability was measured. Instrumentation (here: the difference between the writing assignments' topics) provides a plausible alternative explanation for the difference in writing ability, which may add to or replace the explanation given by the independent variable (here: the writing instruction provided between measurements).

---

4. An additional threat to internal validity is known as the effect of **regression to the mean**. Regression to the mean may play a role whenever the study is focussed on special groups, for instance, bad readers, bad writers, but to an equal extent: good readers, good writers, etc. Let us first give an example, since the phenomenon is not immediately clear from an intuitive point of view.

Example 5.4: There is some controversy about the use of illustrations in children's books. Some argue that books used to teach children how to read should contain no illustrations (or as few as possible): illustrations distract the child from features of words they should be learning. Others argue that illustrations may provide essential information: illustrations serve as an additional source of information.

Donald (1983) investigated how illustrations influenced the understanding of a text. The researcher selected 120 students (of a student body of 1868) from the 1st and 3rd year of primary/elementary education; 60 from each year. According to their performance on a reading test administered earlier, it turned out that, of the 60 students in each year, 30 could be classified as strong readers, and 30 could be classified as less strong readers. Each student was shown the same text, either with or without illustrations (independent variable), see Table 5.1.

The results turned out to mainly support the second hypothesis: illustrations improve the understanding of a text, even with inexperienced readers. The illustrated text was better understood by the less strong readers, and younger readers, too, showed improvement when illustrations were added.

Table 5.1: Conditions in the study by Donald (1983).

group	reading ability	condition	n
1	weak	without	15
1	weak	with	15
1	strong	without	15
1	strong	with	15
3	weak	without	15
3	weak	with	15
3	strong	without	15
3	strong	with	15

So, what is wrong with this study? The answer to this question can be found in how students were selected. Readers were classified as 'strong' or 'less strong' based on a reading ability test, but their performance on this test are always influenced by random factors that have nothing to do with reading ability: Tom was not feeling well and did not perform well on the test, Sarah was not able to concentrate, Nick was having knee troubles, Julie was extraordinarily motivated and outdid herself. In other words: the assessment of reading ability was not entirely reliable. This means that (1) the less strong readers who happened to have performed above their level were unjustly classified as strong readers instead of as less strong readers; and, conversely, (2) strong readers who happened to have performed below their level were unjustly deemed to be less strong readers. Thus, the group of less strong readers will always contain some readers that are not that bad at all, and the group of strong readers will always contain a few that are not so strong, after all.

When the actually-strong readers that were unjustly classified as non-strong readers are given a second reading test (after having studied a text with or without illustrations), they will typically go back to performance at their usual, higher level. This means that a higher score on the second test (the posttest) might be an artefact of the method of selection. The same is true, with the necessary changes, for the actually-less-strong readers that have unjustly been selected as strong readers. When these students are given a second reading test, they, too, will typically go back to performance at their usual (lower) level. Thus, their score on the posttest will be lower than their score on the pretest.

For the study by Donald (1983) used as an example here, this means that the difference found between strong and less strong readers is partially due to randomness. Even if the independent variable has no effect, the group of 'strong' readers will, on average, perform worse, while the group of 'less strong' readers will, on average, perform better. In other words: the difference between the two groups is smaller during the posttest than during the pretest, as a consequence of random variation: regression to the mean. As you may expect, research results may be muddled by this phenomenon. As we saw above, an experimental effect may be weakened or disappear as a consequence of regression to the mean; conversely, regression to the mean can be misidentified as an experimental effect (for extensive discussion, see Retraction Watch (2018)).

Generally speaking, regression to the mean may occur when a classification is made based on a pretest whose scores are correlated with the scores on the posttest (see Chapter **??**)). If there is no correlation at all between pretest and posttest, regression to the mean even plays the main role: in this case, any difference between pretest and posttest must be the consequence of regression to the mean. If there is a perfect correlation, regression to the mean does not play a role, but the pretest is also not informative, since (after the fact) it turned out to be completely predictable from the posttest.

Regression to the mean may offer an alternative explanation for an alleged substantial score increase between pretest and posttest for a lower performing group

(e.g., less strong readers) compared to a smaller increase for a higher performing group (e.g., strong readers). Conversely, it might also offer an alternative explanation for an alleged score decrease between pretest and posttest for a higher performing group (e.g., strong readers) compared to a lower performing group (e.g., less strong readers).

It is better when groups are *not* composed according to one of the measurements (pretest or posttest), but, instead, on the basis of on some other, independent criterion. In the latter case, the test subjects in both groups will have a more or less average score on the pretest, which minimizes the effect of regression to the mean. Each group will have about equal numbers of test subjects whose scores fell out too high by accident and those whose scores fell out too low, both on the pretest and the posttest.

5. A fifth threat to internal validity comes in the form of **selection**. This refers (mainly) to a distribution of test subjects between various conditions under which the groups are not equivalent at the beginning of the study. For instance, when the experimental condition contains all the smarter students, while the control condition contains only the less bright ones, any effect that is found may no longer be attributed to manipulation of the independent variable. The difference in initial levels (here: in intelligence) will provide a plausible alternative explanation that threatens internal validity.

Example 5.5: To make a fair comparison between schools of the same type[1], we must consider differences between schools regarding their students' level at entry. Imagine that school A has students that start at level 50, and perform at level 100 on their final exams (we are using an arbitrary scale here). School B has students that start at level 30, and perform at level 90 on their final exams (on the same scale). Is school B worse than A (because of lower final performance), or is school B better than A (because final performance shows a smaller difference)?

Research in education often does not provide the opportunity to randomly assign students in different classes to various conditions, because this may lead to insurmountable organizational problems. These organizational problems involve more than just (randomly) splitting classes, even though the latter may

[1] The secondary school system in the Netherlands distinguishes three major types (VMBO, HAVO, VWO), which differ in the length of the curriculum and in whether they are geared more towards practical or academic learning.

already be difficult to put into practice. The researcher also has to account for possible transfer effects between conditions: students will talk to one another, and might even teach each other the essential characteristics of the experimental condition(s). This is at least one way in which the absence of an effect could be explained. Because of the problems sketched out here, it often occurs that entire classes are assigned to one of the conditions. However, classes consist of students that go to the same school. When students (and their parents) choose a school, self-selection takes place (in the Dutch education system), which leads to differences between conditions (that is, between classes within conditions) in terms of students' initial performance. This means that any differences we find between conditions could also have been caused by students' self-selection of schools.

In the above, we have already indicated the most straightforward way to give different conditions the same initial level: assign students to conditions by chance, or, at random. This method is known as randomization (Shadish et al., 2002, p.294 ff). For instance, we might randomize test subjects' assignment to conditions by giving each student a random number (see Appendix A)), and then assigning 'even students' to one condition and 'odd students', to the other. When test subjects are randomly assigned to conditions, all differences between test subjects within the various conditions are based on chance, and are thus averaged out. In this way, it is most likely that there are no systematic differences between the groups or conditions distinguished. However, this is only true if the groups are sufficiently large.

Randomization, or random assignment of test subjects to conditions, is to be distinguished from random sampling from a population (see §**??**) below). In the case of random sampling, we randomly select test subjects from the population of possible test subjects to be included in our sample; our goal in this case is that the sample(s) resemble the population from which they are drawn (it is drawn). In the case of randomization, we randomly assign the test subjects within the sample to the various conditions in the study; our goal in this case is that the samples resemble each other.

One alternative method to create two equal groups is *matching*. In the case of matching, test subjects are first measured on a number of relevant variables. After this, pairs are formed that have an equal score on these variables. Within each pair, one test subject is assigned to one condition, and the other, to the other condition. However, matching has various drawbacks. Firstly, regression to the mean might play a role. Secondly, matching is very labour-intensive when test subjects have to be matched on multiple variables, and it requires a sizeable group of potential test subjects. Thirdly matching only reckons with variables that the researcher deems relevant, but not with other, unknown variables. Randomization does not only randomize these relevant variables, but also other properties that might potentially play a role without the researcher's realizing this. In short, randomization, even if it is relatively simple, is far preferable to matching.

6. **Attrition** of respondents or of participants is the final threat to internal validity. In some cases, a researcher will start with a sizeable number of test subjects, but, as the study continues, test subjects drop out. As long as the percentage of drop-out (attrition) remains small, there is no problem. However, a problem does arise when attrition is selective to one of the conditions distinguished. In the latter case, we will not be able to say much about this condition at all. The problem of attrition is mainly relevant to longitudinal research: research in which a small number of respondents is followed over a longer period of time. In this case, we might be confronted with people's moving house, or passing away over the course of the experiment, or participants that are no longer willing to be a part of the study, etc. This may lead to a great reduction in the number of respondents.

In the preceding paragraphs, we discussed a number of frequently occurring problems that may threaten a study's internal validity. However, this list is by no means an exhaustive one. Each type of research has problems of its own, and it is the researcher's task to remain aware of possible threats to internal validity. To this end, always try to think of plausible explanations that might explain a possible effect to the same extent as, or maybe even better than, the cause you are planning to investigate. In this manner, the researcher must adopt the mindset of their own greatest sceptic, who is by no means convinced that the factor investigated is truly the cause of the effect found. Which possible alternative explanations might this sceptic come up with, and how would the researcher be able to eliminate these threats to validity through the way the study is set up? This requires an excellent insight into the logical relationships between the research questions, the variables that are investigated, the results, and the conclusion.

5.5 Construct validity

In an experimental study, an independent variable is manipulated. Depending on the research question, this may be done in many different ways. In the same manner, the way in which the dependent variable(s) is/are measured may take different shapes. The way in which the independent and dependent variables are formulated is called these variables' *operationalization*. For instance, students' reading ability may be operationalized as (a) their score on a reading comprehension test with open-ended questions; (b) their score on a reading comprehension test with multiple choice questions; (c) their score on a so-called cloze test (fill in the missing word); or (d) as the degree to which students can execute written instructions. In most cases, there are quite a few ways to operationalize a variable, and it is rarely the case that a theory would entail just one possible description for the way the independent or dependent variables must be operationalized. *Construct validity*, or *concept validity*, refers

to the degree to which the operationalization of both the dependent variable(s) and the independent variable(s) adequately mirrors the theoretical constructs that the study focuses on. In other words: are the independent and dependent variables properly related to the theoretical concepts the study focuses on?

Example 5.6: Infants' and toddlers' language development is difficult to observe, especially in the case of auditory and perceptual development in these test subjects, who can barely speak, if at all. One often used method is the Head Turn Preference Paradigm (Johnson and Zamuner, 2010). In this method, each trial starts by having the infant look at a green flashing light straight ahead. Once a child's attention has thus been captured, the green light is extinguished, and a red light starts flashing at the test subject's left or right hand side. The child turns their head to be able to see the flashing light. A sound file containing speech is then played on a loudspeaker placed next to this peripheral flashing light. The dependent variable is the period of time during which the child keeps looking to the side (with less than 2 seconds of interruption). After this, a new trial is started. The time spent looking at the light is interpreted as indicating the degree to which the child prefers the spoken stimulus.

However, interpreting the looking times obtained is difficult, because children sometimes prefer new sound stimuli (e.g., sentences in an unknown language) and sometimes prefer familiar stimuli (e.g., grammatical vs. ungrammatical sentences). Even when the stimuli have been carefully adjusted to the test subject's level of development, it is still difficult to relate the dependent variable (looking time) to the intended theoretical construct (the child's preference).

Example 5.7: As indicated above, the concept of reading ability may be operationalized in various ways. Some argue that reading ability cannot be properly measured by multiple choice questions (Houtman 1986, Shohamy 1984). In multiple choice questions, answers are very strongly influenced by other notions, such as general background, aptitude at guessing, experience with earlier tests, and the way the question itself is asked, as is illustrated in the following question:

Who of the following individuals published an autobiography within the last few years?
a. Joan of Arc *(general background)*

b. my neighbour *(way the question is asked, experience)*
c. Malala Yousafzai *(general background)*
d. Alexander Graham Bell *(general background)*

This question is clearly lacking in construct validity for measuring knowledge on autobiographies.

Of course, the problems with construct validity mentioned above arise not only for written questions or multiple choice questions, but also for questions one might ask test subjects orally.

Example 5.8: If we orally ask parents the question, How often do you read to your child?, this question in itself suggests to them that it is desirable to read to one's child, and parents might overestimate how often they do this. This means that we are not only measuring the construct of 'behaviour around reading to one's child', but also the construct of 'propensity towards socially desirable answers' (see below).

A construct that is notably difficult to operationalize is that of *writing ability*. What is a good or bad product of writing? And what exactly is writing ability? Can writing ability be measured by counting relevant content elements in a text, should we count sentences or words, or perhaps mainly connectives (*therefore, because, since, although*, etc.), should we collect multiple judgments that readers have about the written text (regarding goal-orientedness, audience-orientedness, and style), or should we collect a single judgment from readers regarding the text's global quality, should we count spelling errors, etc? The operationalization problems arise from the lack of a theory of writing ability, from which we might derive a definition of the quality of writing products (Van den Bergh and Meuffels, 1993). This makes it easy to criticize research into writing quality, but makes it difficult to formulate alternative operationalizations of the construct.

Another difficult construct is the *intelligibility* of a spoken sentence. Intelligibility may be operationalized in various ways. The first option is that the researcher speak the words or sentences in question, and the test subject repeat them out loud, with errors in reproduction being counted; one disadvantage of this is that there is hardly any control over the researcher's model pronunciation.

A second option is that the words or sentences be recorded in advance and the same procedure be followed for the rest; one disadvantage that remains is that responses are influenced by world knowledge, grammatical expectations, familiarity with the speaker or their use of language, etc. The most reliable method is that of the so-called 'speech reception threshold' (Plomp and Mimpen, 1979), which is described in the next example. However, this method does have the disadvantages of being time-consuming, being difficult to administer automatically, and requiring a great amount of stimulus material (speech recordings) for a single measurement.

Example 5.9: We present a list of 13 spoken sentences masked with noise. The speech-to-noise ratio (SNR) is expressed in dB. A SNR of 0 dB means that speech and noise are equally loud, a SNR of +3 dB means that the speech is 3 dB *louder* than the noise, while a SNR of -2 dB means that the speech is 2 dB *less loud* than the noise, etc. After each sentence, the listener has to repeat the sentence he or she just heard. If this is done correctly, then the SNR for the next sentence is decreased by 2 dB (less speech, more noise); if the response had a error, the SNR for the next sentence is increased by 2 dB (more speech, less noise). After a few sentences, we see little variation in the SNR, which starts swinging back and forth around an optimal value. The average SNR over the last 10 sentences played to the test subject is the so-called 'speech reception threshold' (SRT). This SRT may also be interpreted as the SNR under which half of the sentences was understood correctly.

So far, we have only talked about problems around the construct validity of the dependent variables. However, the operationalization of the *in*dependent variable is also often questioned. After all, the researcher has had to make many choices while operationalizing their independent variable (see §2.6)), and the choices made can often be contested.

A study is not construct valid, or concept valid, if the operationalizations of the independent variables cannot withstand the test of criticism. A study is also not construct valid if the independent variable is not a valid operationalization of the theoretical-concept-as-intended. If this operationalization is not valid, we are, in fact, manipulating something different from what we intended. In this case, the relationship between the dependent variable and the independent variable-as-intended that was manipulated is no longer unambiguous. Any observed differences in the dependent variable are not necessarily caused by the independent variable-as-intended, but could also be influenced by other factors. One well-known effect of this kind is the so-called Hawthorne effect.

Example 5.10: Management at the Hawthorne Works Factory (Western Electric Company) in Cicero, Illinois, USA was alarmed by the company's bad performance. A team of researchers scrutinized the way things were done, investigating more or less anything one can think of: working hours, salary, breaks, lighting, heating, staff and management meetings, management style, etc. This results of this study (from 1927) showed that productivity had increased by leaps and bounds – but there was no correlation with any of the independent variables. In the end, the increase in productivity was attributed to the increased attention towards the employees.

Thus, we observe the Hawthorne effect when a change in behaviour does not correlate with the manipulation of any independent variable, but this change in behaviour is the consequence of a psychological phenomenon: participants who know they are being observed are more eager to show (what they think is) the desired behaviour.

Example 5.11: Richardson et al. (1978) compared the effectiveness of two methods for improving reading ability in less strong readers. Students were selected based on their scores on three tests. The 72 students selected were randomly assigned to one of two method conditions (structured teaching of reading skills versus programmed instruction). In the first condition, the structured teaching was delivered by four instructors, who taught a small group (of four students). This, in fact, leads to a student-teacher of $1 : 1$. In the second condition (programmed instruction), the teachers left the students to their own devices as much as possible. The experiment ran for 75 sessions of 45 minutes each. After the second observation, it turned out that the students who were taught according to the first (structured) method had made more progress than the students taught using the second method (programmed instruction).

So far, there are no problems with this study. However, a problem does arise if we concluded that the structured method is better than the programmed instruction. An alternative explanation, one that cannot be excluded in this study, is that the effect found does not (exclusively) follow from the method used, but (also) from the greater individual attention in the first condition (structured teaching).

Just like for internal validity, we can also mention a number of validity-threatening factors for construct or concept validity.

1. One threat to concept validity is **mono-operationalization**. Many studies operationalize the dependent variable in one way only. The test subjects are only asked to perform one task, e.g., a single auditory task with measurement of reaction times (over multiple trials), or a single questionnaire (with multiple questions). In this case, the study's validity rests entirely on this specific operationalization of the dependent variable, and no further data are available on the validity of this specific operationalization. This means that the researcher leaves room for doubt: strictly speaking, we have nothing but the researcher's word as evidence for the validity of their way of operationalizing the variable. There is a much better way to carry out this kind of research, namely, by considering different operationalizations of the construct to be measured. For instance, this can be done by having test subjects perform multiple auditory tasks, while counting erroneous responses in addition to measuring reaction times; or by not only having test subjects fill out a questionnaire, but also observing the construct intended through other tasks and methods of observation. When test subjects' performance on the various types of response is highly correlated, this can be used to demonstrate that all these tests represent the same construct. This is called *convergent validity*. We speak of convergent validity when performance on instruments that represent the same theoretical construct is highly correlated (or, converges).

However, it is not sufficient to demonstrate that tests meant to measure the same concept or construct are, indeed, convergently valid. After all, this does not show what construct they refer to, or whether the construct measured is actually the intended construct. Did we actually measure a speaker's 'fluentness' using multiple methods, or did we, in reality, measure the construct of 'attention' or 'speaking rate' each time? Did we actually measure 'degree of reading comprehension' using different converging methods, or did we, in reality, measure the construct of 'performance anxiety' each time? To ensure construct validity, we really have to demonstrate that the operationalizations are *divergently valid* compared to operationalizations that aim to measure some other aspect or some other (related) skill or ability. In short, the researcher must be able to show that performance on instruments (operationalizations) that represent a single skill or ability (construct) is highly correlated (is convergent), while performance on instruments that represent different skills or abilities is hardly correlated, if at all (is divergent).

2. The **researcher's expectations** – which are manifested in both conscious and unconscious behaviour – may also threaten a study's construct

validity. The researcher is but human, and therefore by no means immune to the influence their own expectations might have on the outcome of their study. Unfortunately, it is difficult to ascertain after the fact how the researcher might have influenced an experiment.

Example 5.12: Clever Hans (in German: Kluger Hans) was a horse with alleged arithmetic skills. When Clever Hans was asked, *how much is 4 + 4?*, the horse stomped its right front hoof 8 times, and when asked, *how much is 3 – 1?*, Hans stomped his front hoof twice. Clever Hans caused quite a stir and became the object of various studies. In 1904, a committee determined that Clever Hans was, indeed, able to do arithmetic (and communicate with humans). Later, however, Carl Stumpf, a member of the research committee, together with his assistant, Oskar Pfungst, established that "the horse fails to solve the problem posed when the solution is not known to any of those present" (Pfungst, 1907, p.185, vert. AN), or when the horse cannot see the person who does know the solution. "Thus, [the horse] required optical help" (idem). After careful observation, it turned out that Clever Hans' owner (and any other people present) showed very slight signs of relaxation as soon as Hans had stomped his right front leg the correct number of times. This unintentional sign was a sufficient incentive for Clever Hans to stop stomping (i.e., to keep his front hoof down), in order to receive his reward of carrots and bread (Pfungst, 1907) (Watzlawick, 1977, p.38–47).

A more recent, perhaps comparable case is that of Alex, a parrot with extraordinary cognitive skills, see, a.o., Boswall (zj) and Ale (2015).

This famous example illustrates how subtle a researcher's or experimenter's[2] influence on the object of study can be. It goes without saying that this influence threatens construct validity. For this reason, it is better when the researcher does not also function as the experimenter. Studies in which the researcher also is the one who administers the treatment or teaches the students or judges performance may be criticized, because researcher (and their expectations) may influence the outcome, which threatens the independent variable's construct validity. Researchers may, however, defend themselves against this 'experimenter

[2]The experimenter is the person who administers an experiment to a participant or informant. The experimenter may be a person distinct from the researchers who devised the research hypotheses, constructed stimuli, and/or recruited participants.

bias'. For instance, in the Head Turn Preference Paradigm (example 5.6), it is customary that the experimenter does not know which group a test subject belongs to, and does not hear which sound file is being played (Johnson and Zamuner, 2010, p.74).

3. Another threat to construct validity may be summarized by the term **motivation**. There are at least two facets to the validity threat of motivation. If (at least) one of the conditions in a study is very taxing or unpleasant, test subjects may lose motivation and put in less effort into their tasks. Their performance will be less strong, but this is an effect of (a lack of) motivation, rather than a direct effect of the independent variable (here: condition). This means that the effect is not necessarily caused by manipulation of the intended construct, but may (also) be caused by unintentional manipulation of test subjects' *motivation.* The opposite situation could, of course, also be a threat to construct validity. If one of the conditions is particularly motivating for the test subjects, any potential effect may be attributed to matters of motivation. In this case, we may also be looking at an effect of an unintentionally manipulated variable.

4. Yet another threat to validity has to do with the choice of the range of values of an independent variable, i.e., its **'dosage'**, that will be considered. If the independent variable is 'the number of times test subjects are allowed to read a poem silently before reading it aloud', the researcher has to determine which values of the variable will be included: one time, two, three times, more times? If the independent variable is 'the time test subjects may spend studying', the researcher must choose how long each group of subjects will be allowed to study: five minutes, fifteen minutes, two hours? The researcher makes a choice out of the possible dosages of the independent variable, 'study time'. On the basis of this dosage, the researcher might conclude that the dependent variable is not influenced by the independent variable. In fact, however, the researcher should conclude that there seems to be no correlation between the dependent variable and the *chosen dosage of* the independent variable. A possible effect might be concealed by the choice of dosage (values) of the independent variable.

Example 5.12: If a passenger car and a pedestrian collide, there is a risk of this being fatal to the pedestrian. This risk of pedestrian fatality is relatively small (less than 20%) when the speed of collision is smaller than about 50 km/h (about 31 mph). If we limited our research into the relationship between speed of collision and risk of pedestrian fatality to such small 'dosages' of collision speed, we might conclude that collision speed has no influence on the risk of pedestrian fatality. This would be an erroneous conclusion (which

type of error?), because, at higher speeds of collision, the risk of pedestrian fatality increases to almost 100% (Rosén et al., 2011; SWOV, 2012).

---

5. A further threat to construct validity is caused by the **guiding influence of pretests**. In many studies, the independent variable is measured repeatedly, both before and after manipulation of the dependent variable: the so-called pretest and posttest. However, the nature and content of the pretest can leave an imprint upon test subjects. In this manner, a test subject may lose their naïveté, which lessens the effect of the independent variable (e.g., treatment). Any difference in scores between experimental conditions can thus be explained in several ways. This is because we may be purely dealing with an effect of the independent variable, but we may also be dealing with an effect of the pretest and the independent variable combined. Moreover, sometimes we can explain the lack of an effect by the fact that a pretest has been performed (see the Solomon four group design, in Chapter 6, for a design that takes this possibility into account).

Example 5.14: We can compare the effects of two treatments in an experiment in which participants are divided into two groups by random assignment. The first group (E) is first given a pretest, then treatment, then a posttest. The second group (C) is given no pretest and no treatment, only a posttest, which, for this group, is the only measurement.

If we find a difference between the two groups during the posttest, this may not automatically be attributed to the difference in treatment. The difference may also be (partially) caused by the pretest's guiding influence, e.g., as a consequence of a guiding choice of words or sentence structure in the questions or tasks in the pretest. Perhaps the participants in group E have learnt something during the pretest, i.e., not during treatment, which makes them perform better or differently on the posttest compared to the participants in group C.

---

6. Another problem that may influence construct validity is participants' tendency to answer in a **socially desirable** way. This is simply people's inclination to give an answer that is desirable in a given social situation, and will therefore not lead to problems or loss of face. An example may clarify this.

Example 5.15: In opinion polls before elections, respondents are prone to giving socially desirable answers, which is also true for the question of whether the respondent is planning on actually casting their vote (Karp and Brockington, 2005). Respondents show a stronger inclination towards the socially desirable answer ("yes, I will vote") with increasing level of education, which leads to overestimation of the turnout rate for higher-educated voters compared to lower-educated ones. This, in turn, has consequences for the poll results for the various parties, because political parties' popularity differs between voters of different levels of education.

This effect was partially responsible for the overestimation of the number of Clinton votes and underestimation of the number of Trump votes in the opinion polls prior to the 2016 US presidential election.

7. One last problem regarding construct validity concerns **limited generalizability**. When research results are presented, we regularly hear remarks such as, 'I do agree with your conclusion that X influences Y, but how about…' The dots may be filled out with all types of things: applicability to other populations, or other genres, or other languages. Whereas these aspects are important, they do not play a direct role in the study itself: after all, we carried out our study using a specific choice of population, genre, language(s), etc.

Nevertheless, we still recommend facing such questions of generalizability. Are the conclusions reached also applicable to another population or language, and why (not)? Which other factors might influence this generalizability? Could it be that a favourable effect for one population or language turns to an unfavourable effect for some other population or language that was outside the scope of the study?

5.6 External validity

Based on the data collected, a researcher – barring any unexpected problems – may draw the conclusion: *in this study, XYZ is true*. However, it is rarely a researcher's goal to draw conclusions that are true just for one study. A researcher would not just like to show that being bilingual has a positive influence on language development *in the sample of children studied*. A researcher would like to draw conclusions such as: being bilingual has a positive influence on

language development *in children*. The researcher would like to generalize. The same holds for daily life: we might taste a single spoonful of soup from an entire pot, and then express a judgment on the entire pot of soup. We assume that our findings based on the one spoonful may be generalized to the entire pot, and that it is not necessary to eat the entire pot before we can form a judgment.

The question of whether a researcher may generalize their results is the question of a study's external validity (Shadish et al., 2002). The aspects of a study generalization pertains to include:

- units: are the results also true for other elements (e.g., schools, individuals, texts) from the population that did not take part in the study?
- treatment: are the results also true for other types of treatment that are similar to the specific conditions in this study?
- situations: are the results also true outside the specific context of this study?
- time: are this study's results also true at different times?

For external validity, we distinguish between (1) generalization *to* a specific intended population, situation, and time, and (2) generalization *over* other populations, situations, and times. Generalizing *to* and *over* are two aspects of external validity that must be carefully separated. Generalizing *to* a population (of individuals, or often, of language material) has to do with how representative the sample used is: to which extent does the sample properly mirror the population (of individuals, words, or relevant possible sentences)? Thus, "generalizing to" is tied directly to the goals of the study; a study's goals cannot be reached unless it is possible to generalize to the populations defined. Generalizing *over* populations has to do with the degree to which the conclusions we formulate are true for sub-populations we may recognize. Let us illustrate this with an example.

Example 5.16: Lev-Ari and Keysar (2010) looked into whether listeners found speakers with a foreign accent in their English pronunciation to be less credible. The stimuli were made by having speakers with no accent, a light accent, or a strong accent pronounce various sentences (e.g., *A giraffe can hold more water than a camel*). Listeners (all native speakers of English) indicated to which extent they thought the sentence was true. The results showed that the listeners judged the sentences to be true to a lesser extent when the sentence had been spoken by a speaker with a stronger foreign accent.

We may assume that this outcome can be generalized *to* the intended population, namely, all native listeners of American English. This generalization can be made despite the possibility that various listeners were perhaps influenced by the speaker's foreign accent to different degrees.

Perhaps a later analysis might show that there is a difference between female and male listeners. It is not impossible that women and men might differ in their sensitivity to the speaker's accent. Such an (imagined) outcome would show that we may not generalize *over* sub-populations within our population, even though we may still generalize *to* the target population.

In (applied) linguistic research, researchers often attempt to *simultaneously* generalize *to* two populations of units, namely, *individuals* (or schools, or families) and *stimuli* (words, sentences, texts, etc.). We want to show that the results are true not just for the language users we studied, but for other language users, as well. At the same time, we also want to show that the results are true not just for the stimuli we investigated, but also for other, similar language material in the population from which we drew our sample of stimuli. This simultaneous generalization requires studies to have a complex design, because we see repeated observations both within test subjects (multiple judgments from the same test subject) and within stimuli (multiple judgments on the same stimulus). After the observations have been made, the stimuli, test subjects, and conditions are combined in a clever way to protect internal validity as best as possible. Naturally, generalization to other language material does require that the stimuli be randomly selected from the (sometimes infinitely large) population of all possible language material (see Chapter **??**).

Chapter 6

Design

6.1 Introduction

Many of the problems around validity discussed in Chapter 5 can be avoided by properly collecting high-quality data. A study's *design* indicates which schema or plan will be used to collect the required data. Using a proper and strong design allows us to neutralize many of the possible threats to validity, which increases our study's strength. Therefore, it is a good idea to spend a good amount of time thinking through your study's design in advance! Naturally, a study's design must be closely coordinated with the research question at hand: after all, the data obtained in the study must allow the researcher to give a valid answer to the research question.

The research designs discussed in this chapter are but a limited selection of all possible designs. Some designs will be discussed predominantly to indicate what can go wrong with a "weak" design; conversely, other designs are popular because they enable us to make our research relatively "strong".

A research design is composed of various elements:

- *time*, usually depicted as passing in the direction of reading. Temporal order is important to be able to establish a causal relationship: the cause comes first, its effect comes after (§ (§5.2)). However, temporal order is a necessary but not a sufficient condition to establish a causal relationship. Put differently, even when the effect (e.g., recovery) does actually happen after the cause (e.g., treatment), this does not entail that the treatment actually cause the recovery. Perhaps the recovery happened spontaneously, or the recovery is the effect of some other cause not considered in the study.

Example 6.1: Imagine Gus: whenever someone has a nettle sting, an insect bite, eczema, or a bruise, Gus sprays some Glassex (Windex) on it – after a few days, the problem vanishes. Gus is convinced his Glassex treatment is the cause of recovery. However, this is a misconception known as "post hoc ergo propter hoc" ("after this, therefore because of this"; also known as "post hoc fallacy"). It is most probable that the problem would have healed properly even without the Glassex treatment. This means that recovery does not prove that the Glassex treatment is necessary. (This example is taken from the 2004 feature film, *My Big Fat Greek Wedding*).

- *groups* of units (e.g., participants); a group will usually correspond to a line in the design.
- *treatment*, normally depicted as X (as in X-ray). Types of treatment may also include a lack of treatment ("control"), or non-experimental usual care.
- *observation*, normally depicted as O (as in Oscar).
- *assignment* of participants to groups or conditions of treatment may happen in various ways. Most often, we will do this at random (indicated below as "R"), because this usually leads to the best protection of the study's validity.

6.2 Between or within ?

For the study's design it is important whether an independent variable is manipulated between participants or within participants. In many studies in linguistics, in which multiple texts, sentences, or words are offered as stimuli, the same is true of distinctions between stimuli versus within stimuli. Variables that are individual to the participant, such as sex (boy/girl) or whether they are multilingual, may normally only vary between subjects: the same participant may not participate in both sex groups in a study, and monolingual participants may not participate in the group of multilingual participants. However, with other variables, which have to do with the way in which stimuli are processed, this is, indeed, possible. The same participant might write with their left and their right hand, or may be observed preceding and following treatment. In this case, the researcher must choose in their researcher design how treatment and observations are combined. We will return to this in §**??**.

6.3 The one-shot single-case design

This is a weak design, in which observations are made one single time: after treatment. This research design may be schematized as follows:

```
X   O
```

For instance, we might count for all final projects written by students in a particular cohort of a particular programme how many errors (of a certain type) occur in these final projects. This may generate some interest, but these data have very little scientific value. There is no way to compare this to other data (for other students and/or for other projects by the same students). It is not possible to draw a valid conclusion about possible effects that "treatment" (studying in the programme, X) might have on the observations (number of errors, O).

Sometimes, data from a one-shot single-case study are forcibly compared to other data, for instance, with normative results for a large control group. Imagine we would like to investigate whether a new method of teaching languages would lead to better language ability in the target language. After a course that uses the new teaching method, we measure language ability and compare it to previously published results for a control group that used the traditional teaching method. This approach is frequently used, but, notwithstanding, there are various factors that threaten its validity (see §5.4), including history (the new participants have different histories and biographies compared to the control group in the past), maturation (the new participants might have undergone more extensive or less extensive development compare to the control group), instrumentation (the test might not be equally suitable for individuals taught by the new teaching method as it is for those taught by the traditional method), and attrition (attrition of participants prior to the observation is not known, neither for the traditional method, nor for the new method).

Example 6.2: An interviewer may ask so-called 'closed' questions, which only have a few possible answers (*which of these three kinds of vegetables do you like best: peas, green beans, or broccoli?*), or 'open' questions, in which the way the question is asked does not limit the possible answers (*what kind of vegetables do you like best?*). There is also a third category, namely, open question with example answers (*what kind of vegetables do you like best, for instance, peas, or green beans, or…?*). However, it is not clear whether these example answers do or do not have a guiding effect, i.e., whether such questions are rather comparable with closed questions, or with open ones. Houtkoop-Steenstra (1991) studied recorded conversations between doctors and their patients. The doctors would frequently ask

> open questions with example answers. Most of the time, patients turned out *not* to have interpreted the question as a guiding one; they primarily interpreted the question as a prompt for narration.
>
> This study can be seen as a one-shot single-case design, without any comparison with data from other conditions. While the conclusions drawn are, indeed, based on empirical observations, we do not know what answer the interviewee would have given if the question had been posed differently.

Despite all these drawbacks, a one-shot single-case study may be useful during the observation phase within the empirical cycle, when the objective is to get some ideas and to formulate (global) hypotheses, which may be tested later.

6.4 The one-group pretest-posttest design

In a one-group pretest-posttest design, data are collected for one group. At the first point in time (usually indicated as `T1`, but sometimes as `T0`), a first measurement is carried out (pretest, `O1`), after which the group is exposed to an experimental treatment, following which, at a later point in time (`T2`), a second measurement is taken (posttest, `O2`). A one-group pretest-posttest design may be schematized as follows:

```
O1   X   O2
```

As shown in the diagram, the treatment, `X`, does not vary: everyone gets the same treatment, because there is just one group. The time of measurement, usually indicated as a pretest, `T0` (`O1`), and a posttest, `T1` (`O2`), varies within participants.

This design is generally better than the previous one-shot one-case design, and it is definitely better than having no data at all. Despite this, we still consider it a weak research design, because it fails to address various threats to validity (see §5.4)). Any difference between `O2` and `O1` may not exclusively be attributed to treatment `X` that was carried out in between: this effect might also be the consequence of maturation (improvement follows from participants' maturation) or of history (improvement follows from one or several events other than `X` that occurred between the time of `O1` and that of `O2`). If the treatment, `X`, or the posttest, `O2`, is dependent on participants' scores on the pretest, `O1`, then regression to the mean may also threaten validity. In short, this research design has various drawbacks because the hypothesis about the independent variable's effect cannot always be answered in a valid way.

6.5 The pretest-posttest-control group design

The problems mentioned above can partially be solved by adding a control group to the design, which results in a pretest-posttest-control group design. This means that the study involves two groups of elements (participants), which, in a diagram, is shown as two lines. This design is used very often. Whenever possible, researchers try to make the two groups as closely comparable as possible by assigning subjects to the two groups at random. This model can be schematized as follows (`R` stands for random assignment to the two groups):

```
R   O1   X   O2
R   O3       O4
```

This research design is popular because it can neutralize many threats to internal validity (see §5.4). The effect of the manipulation or treatment (`X`) is evaluated by comparing two differences, `(O2 - O1)` and `(O4 - O3)`. Strictly speaking, this research design has not one but two independent variables that may influence measurements: (1) the manipulation or treatment, `X` or not `X`, varying between subjects, and (2) the time of measurement, usually indicated as a pretest, `T0`, and a posttest, `T1`, varying within subjects.

This design does take effects of history into account, at least, to the degree that such effects may have occurred equally for both groups. It does not take into account events that might have influenced just one of the groups (conditions). This does means that, if such an event has occurred for one group and not the other, this difference in history might also be responsible for an unequal difference between pretest and posttest in one group compared to the other.

Threats to internal validity coming from maturation can be easily eliminated in this research design. After all, we expect any effect of maturation to be equally manifested in both groups, which is why it cannot be of any influence on the difference between `(O2 - O1)` and `(O4 - O3)`. Of course, this does presume that the pretest was administered to both groups at the same time, and that the same holds for the posttest.

Any disruptive effect of instrumentation is likewise neutralized, as long as the requirements on comparable conditions of measurement are satisfied, and measurements are taken with the same instrument, for instance, the same device, computer program, or printed test. However, if observers or raters are recruited, like in the case of research into writing ability, instrumentation becomes a more complicated factor. In this case, it is highly important that these raters do *not* know which participants produced the fragments or responses to be judged, or under which condition this happened. Otherwise, the raters may (unwittingly and unintentionally) allow their expectations to play a role in the formation of their judgment. In this latter case, we would not be showing an effect of the independent variable, but an effect of the raters' bias.

The problem of regression to the mean also plays a smaller role in this design. In case the participants have been randomly assigned to one of two groups, and all participants' data are entered into an analysis at the same time, regression to the mean does not play any role, since regression to the mean is expected to take place at equal rates in both groups, so that it does not have any influence on our analysis of the difference between `(O2 - O1)` and `(O4 - O3)`.

The problem of participant selection is excluded in this design by randomly choosing the sample of participants from of the entire population, and by then randomly assigning participants to one of both groups or conditions. Naturally, the law of large numbers is at work here: if a larger sample is randomly split into two groups, there is a greater chance that the two groups will be equivalent compared to when the same is done for a smaller sample.

However, attrition can actually be a reason for a difference between `(O2 - O1)` and `(O4 - O3)`. This threat to validity is difficult to control. After all, we cannot force a participant to keep participating in a study, and we cannot stop them from moving or passing away. Therefore, attrition may form a problem, especially when attrition rates differ between the two groups or conditions. It is best practice to report any attrition in the research report, and to discuss any potential consequences it may have.

Summarizing, we can say that this pretest-posttest control group design allows us to control the various factors that threaten internal validity reasonably well. But how about construct validity (see §5.5))? These threats were not touched upon earlier in our discussion of the one-shot one-case design and the one-group pretest-posttest design, because these designs already generated serious doubt regarding their internal validity.

It must be said that not all aspects of construct validity have repercussions for a study's design. Some aspects that concern manner of operationalization, such as convergent and divergent validity, are irrelevant to choosing a research design. However, other aspects are, indeed, relevant: the researcher's expectations, attention, motivation, and the guiding influence of pretests.

The pretest-posttest control group research design does not provide adequate guarantees for any of these four threats to construct validity. The researcher's *expectations* may play a role both in the experimental and control conditions, even if it is a different role. This because there are two measurements at two points in time. Moreover, any difference between `(O2 - O1)` and `(O4 - O3)` might be attributable to the (additional) *attention* given to the experimental condition: the so-called Hawthorne effect (see example 5.10 in Chapter 5). This effect plays a role predominantly if participants in one condition (group) receive more attention compared to the other condition (group), as is the case in the Hawthorne effect.

A third threat to construct validity is formed by *motivation.* Sometimes, one of the conditions can be so demoralizing that participants in this condition stop seriously participating in the study. Just like in the case of attention, the crucial

factor is not so much the appeal of a particular condition, but any differences between conditions in terms of their appeal.

In addition, construct validity in the pretest-posttest control group design can be threatened by the guiding influence of *pretests.* Because of a pretest (`O1` and `O3`), participants may develop a (greater) awareness of certain aspects of the study, which means that they will no longer behave like naïve participants. In such cases, the pretest can be considered to be a type of manipulation (see example 6.3 below).

6.6 The Solomon-four-groups design

The Solomon four-group design is used much less often than the pretest-posttest control group design. Despite this, the former design is clearly preferable to the pretest-posttest control group design: in particular, it allows for better control of threats to construct validity.

In the Solomon four-group design, four conditions are distinguished, to which participants are assigned at random. In the first two conditions, a pretest is administered first, after which one of both groups is given the experimental treatment. Then, both groups undergo a posttest. Up until this point, the Solomon four-group design is identical to the pretest-posttest control group design. However, no pretest is administered in the third and fourth conditions. In one of the two conditions, participants are given the experimental treatment, but not in the other condition. Finally, both of these groups are given a posttest. We may schematize the Solomon four-group design as follows:

```
O1   X   O2
O3       O4
     X   O5
         O6
```

Summarizing, we can say that the Solomon four-group design is an expansion of the pretest-posttest control group design by two groups, which do not participate in the pretest. Because of these two additional conditions with no pretest, we can take into account the guiding influence of pretests (see example 5.14), since this guiding influence is absent from the third and fourth groups. In addition, the effect of manipulating the independent variable, `X`, is tested several times, namely, in the four comparisons of `O2` versus `O1`, `O2` versus `O4`, `O5` versus `O6`, and `(O2 - O1)` versus `(O4 - O3)`. The effect of the possibly guiding pretest is tested in the two comparisons of `O2` versus `O5`, `O4` versus `O6`. Thus, we can show effects of both treatment and pretest in the same study. However, this does mean we must employ two additional groups of participants (compared to the pretest-posttest control group design).

Example 6.3: A study done by Ayres et al. (2000) examined the effect that habituation training (`X`) has on fear of public speaking. Fear of public speaking was measured by having the participant first hold a speech, and then fill out two questionnaires on fear of (public) speaking. These together form one measurement. One group received habituation training by watching a training video that lasted about 20 minutes; the second group was given a break of the same duration, instead. The study used a Solomon four-group design to allow for studying a possible guiding influence of the pretest. After all, it is possible that the pretest (of which the talk they gave was a part) itself forms an instance of training for the participants, so that any positive effects after "treatment" `X` (habituation training) may not be attributed to said treatment, but (also) to the pretest. However, the results showed that the habituation training, indeed, did have a strongly favourable effect on fear of public speaking, and that the pretest alone (with no treatment) had no effect at all on participants' degree of fear of public speaking.

6.7 The posttest-only control group design

A great amount of studies feature a pretest, because researchers want to demonstrate that the two (or more) groups researched do not differ from one another at the beginning of the study. Nevertheless, an adequate research design does not have to feature a pretest. If the groups are of sufficient size, and if participants (or other units of interest) have been assigned to groups in a completely random way, statistical analysis alone is sufficient to show that the groups are quite comparable. For instance, if we divide 100 participants between 2 groups in a completely random way, there is an extraordinarily small chance of the two groups' showing a difference on the pretest. Therefore, for many cases like this, a posttest control group design is sufficient. This design may be schematized as follows:

```
X   O5
    O6
```

However, this design is only adequate if the groups are large enough, and if participants have been randomly assigned to the conditions. If these demands cannot be met, this design is also insufficient.

Example 6.4: Following up on the study by Houtkoop-Steenstra (1991) (see example 6.2), Wijffels et al. (1992) investigated to which extent questions with or without example answers are interpreted as guiding the listener in an oral (phone) interview. To this purpose, five questions on crime were constructed. Two versions of each question were made: one with example answers, and one without. Each respondent (in a sample of 50) was asked two to three questions with example answers, and two to three questions without example answers. The division of questions between the two types (with or without example answers) was randomized, which means that we may assume that the group of respondents that heard a particular question with example answers does not differ from the group of respondents that heard the exact same question without example answers. If both groups answer the same or similarly, we may assume that example answers have no guiding effect, but if respondents often use example answers to respond to a question, we may assume that the example answers do have a guiding effect. Analysis revealed that such a guiding effect did, indeed, occur for 4 out of 5 questions.

These two studies, Houtkoop-Steenstra (1991) and Wijffels et al. (1992), illustrate the gradual progress of scientific knowledge. Houtkoop-Steenstra (1991) establishes that the professional literature has predominantly looked at written interviews, and asks whether the same effects are seen in oral face-to-face interviews. She concludes that, in face-to-face conversations, example answers do not have a guiding influence. Wijffels et al. (1992) investigate the same hypothesis in an experiment that uses oral interviews over the phone, and conclude that example answers do have a guiding influence in these phone conversations.

6.8 Factorial designs

So far, we have talked about experimental designs in which one single independent variable is manipulated. However, many researchers are (also) interested in the effect of simultaneously manipulating multiple independent variables. Designs in which several factors are varied at the same time are called factorial designs. We already saw an example of this in our discussion of the pretest-posttest control group design (§6.5), in which both time and treatment were varied.

Example 6.5: Drake and Ben El Heni (2003) investigated the perception of musical structure. We may indirectly measure this perception by asking listeners to tap along with the music. If a listener does not understand or recognize the structure of a musical fragment, they will tend to tap every beat (analytical listening). The better a listener understands and recognizes a fragment's structure, the more they will tend to tap along with higher-level units (synthetic or predictive listening): they will not tap once every beat, but once every measure or once per musical phrase. The time interval between taps (called the inter-onset interval or IOI) thus forms an indication of the perceived musical structure. Two groups of listeners participated in the study: one in France and one in Tunisia[1]. All participants listened to 12 pieces of music, of which 6 derived from French musical culture, and the other 6, from Tunisian musical culture (the pieces of music differed in terms of time signature, tempo, and degree of popularity). Results are summarized visually in Figure 6.1.

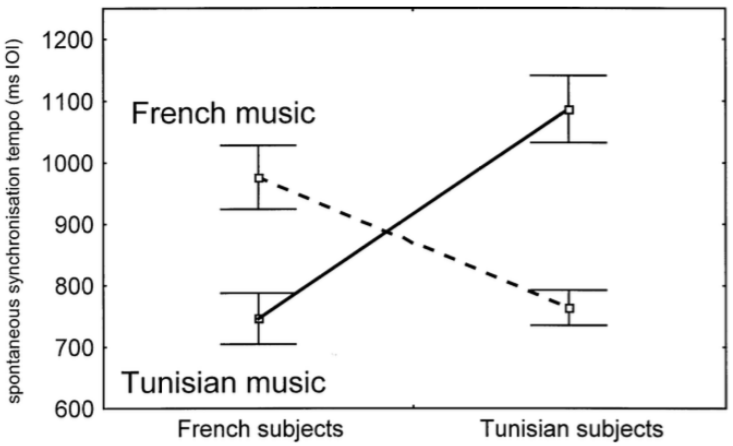


Figure 6.1: Average time interval between taps (IOI, in ms) for two groups of listeners and two types of music (from Drake and Ben El Heni, 2003, Fig.2).

These results show that there is no difference between both groups

[1]Note that participants are not randomly assigned to one of these two groups, which, strictly speaking, makes this a quasi-experimental study (see Chapter 1).

(French vs. Tunisian listeners; both groups have the same IOI on average), and that there is also no difference between both types of music (French vs. Tunisian music; both types of music result in the same IOI on average). Does this mean that the two independent variables have no effect at all? They absolutely do: it turned out that French listeners produced longer IOIs between taps when listening to French music, while, on the other hand, Tunisian listeners produced longer IOIs when listening to Tunisian music. Thus, all listeners produced longer IOIs when listening to a type of music they knew, and shorter IOIs when listening to a type of music they did not know. Drake and Ben El Heni (2003) conclude that listeners are better able to recognize and understand musical structure in music from their own musical culture compared to music from another culture. This pattern is a classic crossover interaction effect, in which one independent variable's effect is exactly opposite in the various conditions defined by the other independent variable.

If it turns out that there is an interaction effect, we cannot sensibly interpret any main effects. This was already illustrated in example 6.5: we cannot conclude that there is no difference between the types of music. However, the size (and direction) of the difference depends on the other independent variable(s), in this case, group/nationality of listeners. Many studies are specifically aimed at demonstrating interaction effects: it is not main effects that are the topic of research, but their interaction, precisely as in example 6.5 above.

It is difficult to schematize a factorial research design, because it features multiple independent variables (with multiple levels each). We could schematically represent these by indexing the manipulation, which was previously shown simply as `X`. The first index (subscript) indicates the level for the first independent variable or factor, while the second index indicates the second factor's level. Following this system, we can schematize the design from example 6.5 as follows:

```
R    X_{1,1}   O1
R    X_{1,2}   O2
R    X_{2,1}   O3
R    X_{2,2}   O4
```

Combining many factors into one big factorial design may often seem seductive: why not investigate how all these factors interact with one another? However, the most sensible option is not to do this, and, instead, limit the number of factors. Firstly, as we will see later, the number of observation has to keep up with the number of possible combinations of factors. Adding more factors means

that many more participants (or other units) are needed. Secondly, it is more difficult to guarantee that all combinations of factors are perfectly comparable (Shadish et al., 2002, p.266): may we reasonably compare Tunisian participants listening to Tunisian music in Tunisian with French participants listening to French music in France? The more combinations of factors are featured in the study, the trickier it becomes to ensure that these combinations are comparable. Thirdly, interactions are notoriously difficult to interpret, which also becomes trickier as interactions become more complex and span a greater number of factors. For all of these reasons, it is better to study the effects of multiple factors in separate individual studies (Quené, 2010).

We will come back later to the analysis and interpretation of data from factorial experimental designs (Chapter **??**). In the meantime, we will concentrate on designs that have just one independent variable.

6.9 Within-subject designs

At the outset of the chapter, we spoke about manipulating an independent variable either between or within subjects ((§6.2). In most designs discussed above, a separate group was formed for each value of the independent variable(s); we call this a between subjects design. The independent variable's value differs between participants.

However, some independent variables may also be manipulated within participants. In such cases, we take repeated measures for (within) the same participants from the same group, switching out different conditions of the independent variable. In the example below, the independent variable, 'language' (native or non-native), is varied within participants. We call this a within subjects design.

Example 6.6: De Jong et al. (2015) investigated the fluency of participants' speech in their native language (Turkish) and in a non-native language (Dutch). The participants first performed a number of speech production tasks in their native language, a few weeks after which they did the same for Dutch. One of the dependent variables was the number of filled pauses (e.g., *uh, uhm*) per second of speech: the greater the prevalence of pauses, the lesser the degree of fluency. As we might expect, the speakers did turn out to produce more pauses (i.e., speak less fluently) in the non-native language compared to their native language. However, one of the goals of this study was to investigate to which extent we may trace back individual fluency differences in the non-native language to individual fluency differences in the native language. These two measurements turned out to be highly correlated ($r = 0.73$; see Chapter **??** for

more on this). Speakers that have many pauses in the non-native language also have many pauses in their native language. The researchers argue that we must take this correlation into account when teaching and testing speaking ability in a non-native language.

The research design described here can be schematized as follows:

```
X1   O1   X2   O2
```

Despite the many threats it poses to internal validity (including history, maturation, guiding influence of pretests), such a design is often useful. In the example above, it is essential that it is the *same* participants that carry out speaking tasks in both languages (conditions) – no other method will be adequate for answering the research questions.

6.10 Designing a study

A researcher who intends to carry out a study has to settle on a way of collecting data: they have to choose a particular design for their study. Sometimes, a standard designed may be chosen, for instance, one of the designs discussed above. In other cases, the researcher will have to devise their own design. Naturally, the design chosen should fit well with the research question (Levin, 1999), and it should exclude as many disruptive, potentially validity-threatening variables as possible. Designing a study is a skill that researchers hone with practice. In the example below, we will try to show to you which reasoning and arguments play a role in developing a design for a study.

Suppose that we would like to investigate whether the way in which test questions are asked, as open vs. closed questions, influences students' scores on the relevant test. If we use a simple design, we will first administer a test with open questions to a group of respondents, and then, a comparable test with closed questions to the same respondents. If the resulting scores are sufficiently correlated, we conclude that both tests measure the same thing, and that performance on the test is not significantly influenced by the way questions are asked. This design can be schematized as follows:

```
Open   O1   Closed   O2
```

However, this research design does have various weaknesses. Firstly, it is not prudent to first administer all open question tests at the first time point, and

leave all closed question tests for the second time point. This is because performance on the second test will always be influenced by effects of ordering (transfer effects): respondents remember and, thus, learn something from the first measurement. However, this transfer always works in one direction, which means that we expect relatively higher performance on the test with closed questions (at the second time point). Because of this, it is better to randomly distribute the open question and close question tests between the first and second time point.

Secondly, all respondents might have been influenced by any events that took place between the two time points (history), for instance, by some instruction relevant to the test's subject matter. Because there is no control group, we cannot take this type of effect into account.

A third problem lies in the way in which the reasoning from findings to conclusions is constructed. For the current example, we defined this reasoning as follows: if scores on both tests are sufficiently correlated, both tests measure the same thing. If you stop and think about it, you might agree with us that this is a strange bit of reasoning. The underlying research question really seems to be whether the correlation between performance on different tests with different types of questions is the same as the correlation between performance on different tests with the same types of questions, since we do assume that the latter group of tests measure the same thing. This, in itself, defines a control group: respondents who, at both points in time, write tests with the same types of questions. Just to be sure, let us add not one, but two control groups: one with open questions at both time points, and one with closed questions at both time points.

By doing this, we have improved the design in at least two ways: (1) tests are randomized between times of measurement, and (2) relevant control groups have been added. At this point, we may schematize our design as follows:

Exp. group 1	Open	O1	Closed	O2
Exp. group 2	Closed	O3	Open	O4
Control group 1	Open	O5	Open	O6
Control group 2	Closed	O7	Closed	O8

For all four groups, we may now determine the correlation between their performance at the first and the second time point. We can subsequently compare these correlation results between the four groups, and use this to answer the research question. This example shows us that the conclusions that can be draw from research results are directly dependent on the design that was chosen (Levin, 1999). In the first design, a low correlation would lead to a conclusion that the two types of testing investigated do *not* test for the same intellectual skills in our respondents. However, in the second design, the same low correlation in the first group (experimental group 1) does not have to lead to the same conclusion! This is because the conclusion also depends on the degree of correlation that was found in the other groups.

6.11 In conclusion

Despite all the books, manuals, websites, and other instructional materials that are available, it is still much too often that we encounter studies with methodological problems in their research questions, operationalization, design, drawing of samples, and/or data collection. Not only do these problems cause a waste of time, money, and energy, but they also yield knowledge that is less reliable, valid, and robust than would otherwise have been possible. The following checklist for good research practice (partly taken from https://www.linkedin.com/groups/4292855/4292855-6093149378770464768) may preempt many problems during a study's later stages.

1. Give your research questions plenty of thought, and formulate them fully into the smallest detail. If the questions have not been formulated clearly, or if there are many sub-questions, keep working on the questions.

2. Arrange the research questions according to their priority. This will help in making good choices regarding design, sampling, operationalization, etc.

3. Think long and hard about your study's design. According to an informal rule of thumb, each hour spent thinking about your study's design will save you 10 hours of additional data analysis and interpretation in the future. Put differently: spending an hour less on thinking about your design will cost you 10 hours of work down the road.

4. Think of various alternative designs for your study, and think about each possible design option's advantages and disadvantages.

5. Imagine the future: you have completed your research project, analysed your data, and written your report or thesis. Which message would you like to impart upon the readers of your report? How does your study's design contribute to this message? What might you change in your design to make this message even clearer? Think of the direction you would like to take, not just of where you are now.

6. Write a research plan in which you describe the various methodological aspects of your study. Explain the details of and the reasoning behind your research questions, design, sample, method of measurement, data collection, instruments of measurement (e.g., questionnaire, software), other requirements (e.g., laboratory environment, transportation), and statistical processing. You will be able to reuse parts of this research plan in your report. When writing your plan, make sure to include a schedule: when will which milestone be reached?

7. Write out what statistical analyses you will use on your data before your actually start collecting any data. Again, be as explicit as possible (using

a script, step-by-step plan, or similar). Make up a miniature collection of fake observations, or real observations from the pilot phase of your study, and analyse these data as if this were your definitive collection of data. Make adjustments to your research plan as needed.

8. Once you are collecting data, do *not* make any changes to your research plan. Keep to this plan and the schedule you made. Analyse your data in the way specified in the (previously adjusted) research plan. Do discuss in your report any problems that arose during the study. If serious problems occur, halt your project, and consider starting anew with an improved version of your study.

Chapter 7

Samples

In generalizing the outcome of a study to the population or the sample, the quality of the sample is all-important. Does the sample adequately reflect the population? To give an extreme example of this: if a sample consists of girls in the last year of primary education, we cannot properly generalize the results to the population of students in primary education, because the sample does not form a good reflection of this population (which consists of boys and girls in all years of the curriculum).

Depending on the method used by the researchers to select participants, many kinds of samples may be distinguished. In this chapter, we make a rough distinction between: (1) convenience samples, (2) systematically drawn samples, and (3) samples drawn at random. For further discussion of the way in which samples may be drawn and the problems that play a role in this, we refer the reader to standard reference works on this topic (Cochran, 1977; Thompson, 2012).

7.1 Convenience samples

Work in the social sciences often uses samples that happen to present themselves to the researcher, so-called *convenience samples*. The researcher carries out the experiment with individuals that happen to be available to them more or less by chance. Some studies use paid or unpaid volunteers. In other studies, students are recruited, who are required to log some number of hours as participants as a part of their studies, or, sometimes, a colleague of the researcher"s sends their own students to participate in the study. A sample of this kind is not without its dangers. The researcher has no control whatsoever over the degree to which results can be generalized to the population. Of course, the researcher does have a population in mind, and will exclude participants that do not form a part of

the intended population (such as non-native speakers) from the study, but the researcher cannot say anything about how representative the sample is.

It is especially in psychology that this convenience sampling has led to heated discussion. For instance, a survey showed that 67% of samples used in published studies in psychology performed in the US was exclusively composed of undergraduate students enrolled in Psychology courses at American universities (Henrich et al., 2010). Naturally, samples like this are hardly representative. As a consequence, the theories based on these data have but a limited scope: they are likely to apply predominantly to the type of individuals (first world, young, highly educated, white) that are also highly represented in the samples (Henrich et al., 2010). Research in linguistics often also uses a convenience sample. Children that participate as participants often have highly educated parents (who often tend to have a linguistics background themselves, which likely means that they have above-average verbal skills), and adult participants are often students from the researchers' environment, who, therefore, also have above-average levels of education and verbal skill.

Despite the valid objections raised against this type of sample, practical considerations often force researchers to use a convenience sample that presents itself. In such cases, we recommend keeping track of the extent to which this convenience sample distinguishes itself from the population over which the researcher would like to generalize. To conclude this discussion of samples that present themselves naturally, we provide an example of the dangers this type of sample carries.

Example 7.1: Some years ago, there was a televised contest in which nine candidates competed on their singing skills. Viewers were invited to announce their preference by phone. For each of the nine candidates, a separate phone line had been opened. For each call, the corresponding candidate received one point. The person with the greatest number of points within a set time limit would win. The audience's response was overwhelming: large swaths of the Dutch phone network were over capacity. Very soon, one of the candidates turned out to have a considerable lead over the others. However, in the course of the evening, this lead became smaller and smaller. In the end, there was only a few calls' difference between the top two candidates. It was striking to see that, as the evening progressed, the relative differences between participants gradually diminished.

We may see this voting procedure as drawing a sample of callers or voters. However, this sample is far from representative. If many voters would like to

vote for the same candidate, the phone line dedicated to this candidate will reach and exceed its capacity. This means that singers who drew many callers will receive relatively fewer votes than singers who draw few callers, because the latter singers' phone lines will not be over capacity. It is precisely for the most popular candidates that a voter is most likely to be unable to cast their vote. Because of this, the real difference in the number of calls per candidate will be far greater than what the organizers measured. The organizers themselved caused this systematic distortion of the results (bias) by opening a separate phone line for each of the nine candidates. The data could have been much more representative if the organizers had opened nine phone lines accessible through one single phone number. In such a scenario, the sample of callers who were able to cast their vote would have been representative for the population of all callers, which was not the case in reality.

7.2 Systematic samples

When the elements in the *sampling space* (i.e., the set of possible elements in a sample) are systematically ordered in some way, a reasonably representative sample can be obtained using a *systematic sampling procedure*. Ordering may, for instance, involve a list of names.

Example 7.2: Let us assume for the moment that we would like to make study of language ability in students in the third year of secondary education. However, the entire population of third year students is far too great to measure all third year students' language ability (reading, writing, speaking, and listening): this group contains about 200,000 students. Consequently, we need to draw a sample. The Dutch Ministry of Education, Culture, and Science has a system in which a list of all schools with third year students is included. An obvious way of proceeding would be to take this list and include each 100th school on the list into the sample. This procedure will presumably result in a reasonably representative sample.

However, two factors may muddle the waters in drawing such a systematic sample, the first of which is the *response rate*. If a considerable proportion of schools that were contacted do not cooperate, we are actually dealing with self-selection (see §5.4 point 5) and, thus, with a convenience sample that presents itself (see §7.1). This is an unwanted situation, since the schools that did cooperate presumably have a greater 'sense of duty' than the schools that refused

participation or than the average school. Moreover, students in the responding and non-responding schools may differ from one another (see §5.4 point 5). This means that the eventual sample may perhaps be no longer representative of the population of all third year students. This, in turn, has as a consequence that the results measured cannot be properly generalized to other third year students at other schools.

The second factor that may influence whether a systematic sample is representative is the presence of a *disruptive trend effect*. We speak of a disruptive trend effect when elements of the population have a greater chance of ending up in the sample if they have a certain characteristic, compared to population elements that do not have this characteristic. In our example of measuring language ability in third year students, we are dealing with a disruptive trend effect. This is because not all students have an equal chance of being in the sample. After all, it is each individual *school* (not: each individual student) that has an equal chance of being in the sample. The consequence of this is that the sample will contain relatively many third year students from small schools with relatively few students, while, conversely, there will be relatively few third year students from large schools with relatively many students. Thus, third year students from large schools will be underrepresented. Is this a bad thing? It might be, because language ability (dependent variable) is partially influenced by the type of instruction, and type of instruction is influenced by the size of a school. This means that the sample described above is not representative for the population of third year students. Once again, this means that the results measured cannot be properly generalized to other third year students at other schools.

7.3 Random samples

The disruptive trend effect described above can be avoided by *random sampling*. Random sampling may happen in various ways, of which we will discuss three.

The first type is simple random sampling: in this procedure, all elements of the population have an equal chance of being drawn. This may, for instance, be realized by giving all elements a *random* number and, depending on the size of the sample, selecting each n-th element. For choosing random numbers, researchers can make use of tables of random numbers (see Appendix A). Random numbers can also be generated by calculators, computers, spreadsheet programs, etc. (Using this type of random numbers is advisable, since a "random" order created by humans is not truly random.) However, one condition for applying this method is that the elements of the population (sampling space) are registered in advance, so that they may all be given numbers in some way.

Example 7.3: We would like to draw a sample of n = 400 primary schools, which is about 4% of the population of primary schools in

the Netherlands. To do this, we request from the Dutch Ministry of Education, Culture, and Science a list of all 9,000 primary schools; this list is the sampling space. After this, we number all schools with subsequent numbers $(1, 2, 3 \ldots, 9000)$. Finally, we select all primary schools whose number happen to end in 36, 43, 59, or 70 (see Appendix A, first column, last two digits). Using this procedure, we randomly select 4 of 100 possible last-two-digit combinations, or 4% of all schools.

The second type of random sampling is *stratified random sampling*. We are dealing with this type of sampling when we know the value of a particular characteristic (e.g., religious denomination) for each element of the population, and we make sure that elements within the sample are divided equally according to this characteristic. To do this, we divide the sample into so-called 'strata' or layers (Lat. *stratum*, 'cover, layer,' related to English *street*, originally meaning 'paved road'). Let us return to our primary school example to clarify a few things. Suppose that, for whatever reason, we are now interested in making the sample (still 4% of the population of primary schools) such that public, catholic, and protestant schools are represented in equal amounts. We therefore devise three lists, a separate one for each type of school. Within each list, we proceed just like for simple random sampling. Eventually, our three sub-samples from the three strata are combined.

Quota sampling goes one step further compared to stratified random sampling: we now also take advantage of the fact that we know the distribution of a certain characteristic (e.g., denomination) within the population. From the list of primary schools, we might have gleaned that 35% of schools is public, 31% is catholic, 31% is protestant, and 3% has some other denomination. From this sampling space, we now draw multiple 'stratified' random samples such that the proportion of schools in each stratum correctly reflects the proportions of this characteristic in the sampling space $(35 : 31 : 31 : 3)$.

7.3.1 SPSS

In order to create a column containing random numbers;

```
Transform > Compute...
```

Select an existing variable (drag to Variables panel) or enter the name of a new variable. From the panel "Function Group", choose "Random numbers", and choose `RV.UNIFORM`. This function samples **r**andom **v**alues from a flat or **uniform** probability distribution, meaning that each number between the lower

and upper limit has an equal chance of being sampled. Enter `0` as lower limit and `9999` as upper limit, or use other limits as appropriate. Confirm with `OK`. This results in a (new or overwritten existing) column with random numbers.

If you wish to sample random numbers from a normal density distribution (see Chapter **??**), then use the function `RV.NORMAL(mean,stdev)`.

We may provide a starting value for the random number generator, in order to make reproducible analyses (and examples):

```
Transform > Random Number Generators...
```

In the panel "Active Generator Initialization", check the option `Set Starting Point`, and enter a starting value, such as your favourite number. Confirm with `OK`.

You can use the resulting random numbers for randomly selecting units (e.g. participants, stimuli) for a sample, and also to randomly assign the selected units to conditions, treatments, groups, etc.

7.3.2 R

In R we may generate random numbers using the predefined function `runif`. This function samples **r**andom values from a flat or **unif**orm probability distribution, meaning that each number between the lower and upper limit has an equal chance of being sampled. The default limits are $(0, 1)$. You may round off the resulting random values to integer numbers, as was done in Appendix A.

If you wish to sample random numbers from a normal density distribution (see Chapter **??**), then use the function `rnorm(n,mean,sd)`.

We may provide a starting value (called a "seed") for the random number generator, in order to make reproducible analyses (and examples), using the predefined function `set.seed`:

```
set.seed(20200912) # reproducible example, number is date on which this chunk was adde
round ( runif( n=5, min=0, max=9999 ) ) # similar to Appendix A
```

```
## [1] 8193 7482 4206 1684 5653
```

You can use the resulting random numbers for randomly selecting units (e.g. participants, stimuli) for a sample, and also to randomly assign the selected units to conditions, treatments, groups, etc.

7.4 Sample size

When you read various research articles, one of the first things that catches the eye is the enormous variation in the number of respondents. In some studies, several thousands of participants are involved, while others only have several multiples of 10, or even fewer. Here, we will discuss two aspects that influence the required size of one's sample: the population's relative homogeneity, and the type of sampling. In the chapters that follow, we will discuss two more aspects that influence the desired sample size: the desired precision (effect size, §13.8) and the desired likelihood to demonstrate an effect if it is present in the population (power, §14.2).

Example 7.4: When cars are tested (for magazines or television), only one car of each type is tested. The results of this tested token are generalized without reservation to all cars of the same type and make. This is possible because the population of cars to which generalization is made is especially homogenous, since the manufacturer strives to make the various tokens of a car type they sell maximally identical.

Firstly, the required sample size depends on the population's homogeneity. If a population is *homogeneous*, like the cars in example 7.4, a small sample will suffice. Things are different when, for instance, we would like to analyse conversation patterns in pre-schoolers. When looking at pre-schoolers' conversation patterns, we come across great differences; conversation patterns exhibit a very high degree of variation. (Some children speak in full sentences, others mainly remain silent. Moreover, there are great individual differences in children's linguistic development.) This means that, to obtain a reasonable picture of language development in pre-schoolers, we need a much bigger sample. Thus, the required sample size increases as the population to which we would like to generalize is less homogeneous (more heterogeneous).

Secondly, the required sample size also depends on the nature of the sample. If a population contains clear strata, but – for whatever reason – we do not apply stratified or quota sampling, then we will need a larger sample compared to a situation where we had, indeed, applied one of these two methods. This is because, in these two latter methods, the researcher actively ensures that strata are represented in the sample either to equal extents, or according to the correct proportions; in simple random sampling, this is left to chance. We must then appeal to the "law of large numbers" to make sure that a sufficient number of elements from each stratum makes its way into the sample, in order to justify

generalization of the results to these various strata. Obviously, this law only works with a sufficiently large sample. When the sample is small, we can in no way be sure that the various strata are represented in the sample to a sufficient extent.

Returning to our primary school example, if we selected three primary schools according to simple random sampling, the chance that this would lead to exactly one public, one catholic, and one protestant school is, no doubt, present. However, other outcomes are quite likely, as well, and even much more likely. If we use stratified or quota sampling, we are guaranteed to have one element (school) of each denomination in our sample. This improves our grounds for generalization, and strengthens external validity.

After all these recommendations that are worth taking to heart, it is now time to discuss how we can describe and analyse research data to properly answer our research questions. This will be done in the next part of this book.

Part II: Descriptive statistics

Chapter 8

Frequencies

8.1 Introduction

When analysing data, a distinction is often made between qualitative and quantitative methods. With the first method, observations (e.g. answers in interviews) are represented in words, and with the second method, observations (e.g. speech pauses in interviews) are represented in numbers. In our opinion, the difference between qualitative and quantitative methods lies in how observations are represented, and how arguments are made on the basis of these observations. Sometimes it is also possible to analyse the very same data (e.g. interviews) both qualitatively and quantitatively. The major advantages of quantitative methods are that the data can be summarised relatively straightforwardly (this is the subject of this part of the syllabus), and that it is relatively simple to draw meaningful conclusions on the basis of the observations.

8.2 Frequencies

Quantitative data can be reported in various different ways. The most straightforward way would be to report the raw data, preferably sorted according to the observed variable's value. The disadvantage of this is that a potential pattern in the observations will not be easily visible.

Example 8.1: Students ($N = 50$) in a first year course reported the following values for their shoe size, a variable of an interval level of measurement:
36, 36, 37, 37, 37, 37, 37, 37, 38, 38, 38, 38, 38, 38, 39, 39, 39, 39,

Table 8.1: Frequency distribution of the phonological class of speech sounds in the *Corpus of Spoken Dutch* (C=consonant, V=vowel; lang=long vowel, kort=short vowel).

main.class	sub.class	count
C	plos	585999
C	fric	426097
C	liq	249275
C	nas	361742
C	glide	146344
V	lang	365887
V	kort	428832
V	schwa	341260
V	diph	61638
V	rest	1146

39, 39, 39, 39, 39, 39, 39, 39, 39, 39, 39, 39, 39, 39, 39, 40, 40, 40, 40, 40, 40, 41, 41, 41, 41, 41, 42, 42, 43, 43, 44, ??.
One of the students did not provide an answer; this missing answer is shown here as ??.

It is usually more insightful and efficient to summarise observations and report them in the form of a *frequency* for each value. This frequency indicates the *number* of observations which have a certain value, or which have a value in a certain interval or class. In order to get the frequencies, we thus *count* the number of observations with a certain value, or the number of observations in a certain interval. These frequencies are reported in a table. We call such a table a frequency distribution.

As a first example, Table 8.1 provides a frequency distribution of a discrete variable of *nominal* level of measurement, namely the phonological class of sounds in Dutch (Luyckx et al., 2007). #52-56

As a second example, Table 8.2 provides a frequency distribution of a continuous variable of *interval* level of measurement, namely the aforementioned shoe size of first year students (Example 8.1).

Table 8.2: Frequency distribution of the self-reported shoe sizes of $N = 50$ students in a first year course (see Example 8.1 above).

Shoe size	36	37	38	39	40	41	42	43	44	??
Number	2	6	6	19	6	5	2	2	1	1

Nevertheless, when a numerical variable is able to assume a great many different values, the frequency distribution thus consequently becomes large and confusing. We then add together values in a certain interval, and afterwards make a frequency distribution on the smaller number of intervals or classes.

Example 8.2: When Queen Beatrix of the Netherlands was giving her last Queen's Speech, on 18th September 2012, she paused some

Table 8.3: Frequency distribution of the length of speech pauses (seconds) in the Queen's Speech of 18th September 2012, given by Queen Beatrix of the Netherlands ($N = 305$).

Interval	Number
4.50–4.99	1
4.00–4.49	0
3.50–3.99	2
3.00–3.49	7
2.50–2.99	4
2.00–2.49	25
1.50–1.99	32
1.00–1.49	16
0.50–0.99	67
0.00–0.49	151

8.2.1 Intervals

For a variable of nominal and ordinal level of measurement, we generally use the original categories to make the frequency distribution (see Table 8.1), although it is possible to add categories together. For a variable of interval or ratio level of measurement, a researcher can choose the number of intervals in the frequency distribution themself. Sometimes that is not necessary, for instance because the variable has a clear number of different discrete values (see Table 8.2). However, sometimes, as a researcher you have to decide for yourself how many intervals you should distinguish, and how to determine the interval boundaries (see Table 8.3). In this instance, the following are recommended (Ferguson and Takane, 1989, Ch.2):

- Ensure that all observations (i.e. the entire range) fall into roughly 10 to 20 intervals.
- Ensure that all intervals are equally wide.
- Make the lower limit of the first or second interval the same as the width of the intervals (see Table 8.3: every interval is 0.50 s wide, and the second interval's lower limit is also 0.50).
- Order the intervals in a frequency distribution from bottom to top in increasing order (i.e. from top to bottom in descending order), see Table **??**).

The wider we make the intervals, the more information we lose about the precise distribution within each interval.

8.2.2 SPSS

`Analyze > Descriptive Statistics > Frequencies...`

Select variable (drag to the "Variable(s)" panel).
Tick: `Display frequency tables`.
Choose `Format`, choose: `Order by: Descending values`.
Confirm with `OK`.

8.2.3 R

```
enq2011 <- read.table(
    file=url("http://www.hugoquene.nl/R/enq2011.txt"),
    header=TRUE )
table( enq2011$shoe, useNA="ifany" )
```

The output of the above `table` command is shown in Table 8.2. The code `NA` (Not Available) is used in R to indicate missing data.

```
table( cut( troon2012, breaks=seq(from=0,to=5,by=0.5) ) )
```

Parse this task from the innermost brackets outwards: (i) `seq`: make a sequence from 0 to 5 (units, here: seconds) in increments of 0.5 seconds, (ii) `cut`: cut up the dependent variable `length` in intervals based on this sequence, (iii) `table`: make a frequency distribution of these intervals.

This task's output is shown (in edited form) in Table 8.3.

8.3 Bar charts

A bar chart is the graphical representation of the frequency distribution of a discrete, categorical variable (of nominal or ordinal level of measurement). A bar chart is constructed of rectangles. All rectangles are equally wide, and the rectangle's height corresponds with the frequency of that category. The surface area of each rectangle thus also corresponds with that category's frequency. In contrast to a histogram, the rectangles are *not* joined up to each other along the horizontal axis, to show that we are dealing with discrete categories.

A bar chart helps us to determine at a glance the most important distributional characteristics of a discrete variable: the most characteristic (most frequently

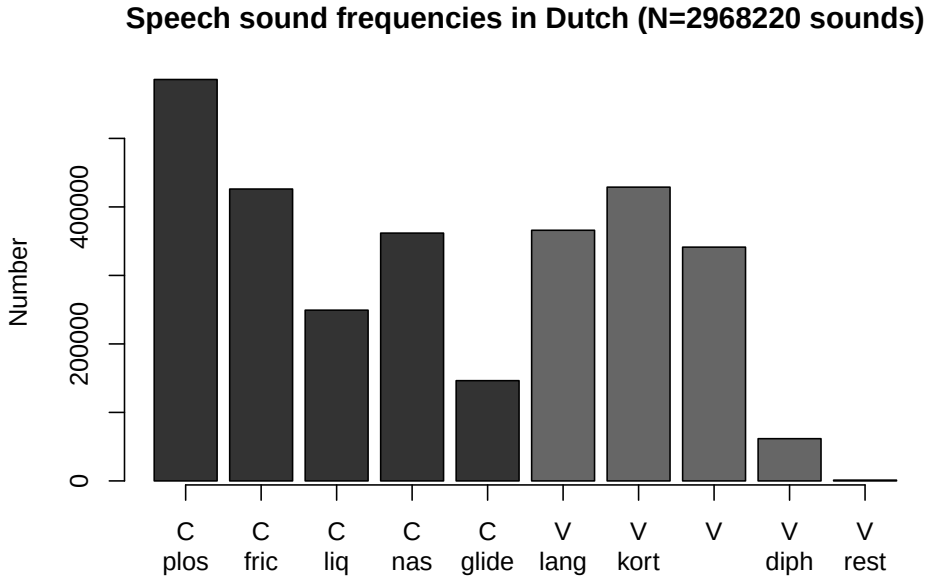


Figure 8.1: Bar chart of the frequency distribution of phonological class of speech sounds in the Corpus of Spoken Dutch (C=consonant, V=vowel).

occurring) value, and the distribution across categories. For the sound frequencies in Dutch (Figure 8.1), we see that amongst the consonants the plosives occur the most, that amongst the vowels the short vowels occur the most, that diphthongs are not used much (the sounds in Dutch *ei, ui, au*), and that more consonants are used compared with vowels.

Tip: Avoid shading and other 3D-effects in a bar chart! These make the width and height of a rectangle less readable, and the visible surface area of a shaded rectangle or of a bar no longer corresponds well with the frequency.

8.4 Histograms

A histogram is the graphical representation of a frequency distribution of a continuous, numerical variable (of interval or ratio level of measurement). A histogram is constructed of rectangles. The width of each rectangle corresponds with the interval width (a rectangle can also be one unit wide) and the height corresponds with the frequency of that interval or value. The surface area of each rectangle therefore corresponds with the frequency. In contrast to a bar chart, the rectangles do join up to each other along the horizontal axis.

A histogram helps up to determine at a glance the most important distributional characteristics of a continuous variable: the most characteristic (most frequently occurring) value, the degree of dispersion, the number of peaks in the frequency

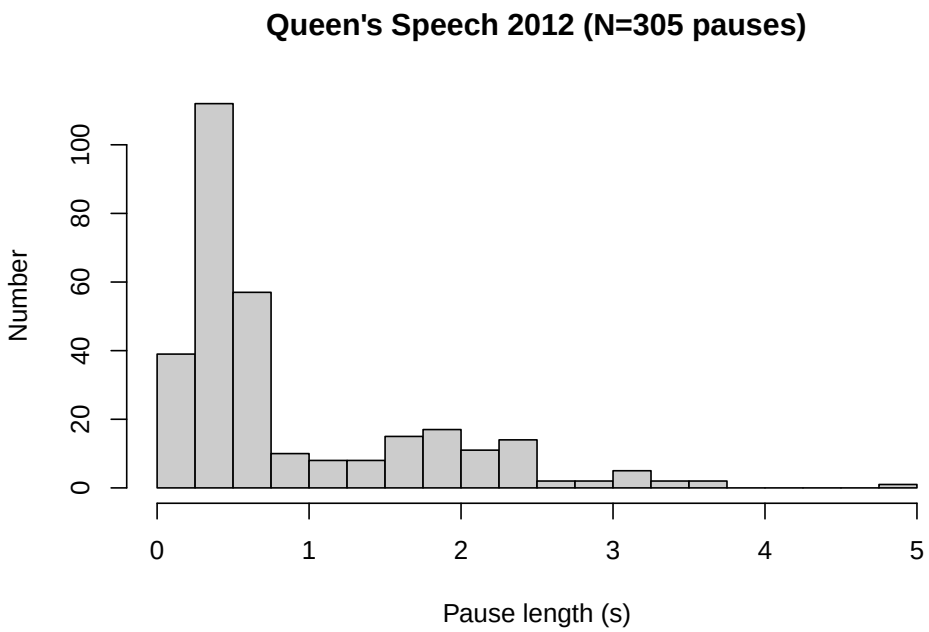


Figure 8.2: Histogram for the lengths of pauses (in seconds) in the Queen's Speech of 18 September 2012, read by Queen Beatrix (N=305).

distribution, the position of the peaks, and potential outliers. (see §9.4.2). For the pauses in the Queen's Speech of 2012 (Figure 8.2), we see that the majority of pauses last between 0.25 and 0.75 s (these are presumably pauses for breath), that there are two peaks in the distribution (the second peak is at 2 s), and that there is one extremely long pause (with a duration of almost 5 s).

Tip: Avoid shading and other 3D-effects in a histogram! These make the width and height of a rectangle less readable, and the visible surface area of a shaded rectangle or of a bar no longer correspond well with the frequency.

8.4.1 SPSS

`Analyze > Descriptive Statistics > Frequencies...`

Select variable (drag to the "Variable(s)" panel).
Choose `Charts`, then pick `Chart type: Bar chart` for a bar chart or `Chart type: Histogram` for a histogram (see the above text for the difference between these options).
Confirm with `OK`.

8.4.2 R

You can make a bar chart like Figure 8.1 in R with the following commands:

```
# read data
klankfreq <- read.table( file="data/klankfreq.txt", header=T )
# 20201130 column names in English
dimnames(klankfreq)[[2]] <- c("main.class","sub.class","count")
# make barplot from column `count` in dataset `klankfreq`
with( klankfreq, barplot( count, beside=T,
                          ylab="Frequency",
                          main="Frequencies of speech sounds in Dutch (N=2968220)",
                          col=ifelse(klankfreq[,1]=="V","grey40","grey20") ) ) -> klankfreq_barpl
# make labels along the bottommost horizontal axis
axis(side=1, at=klankfreq_barplot, labels=klankfreq$main.class)
axis(side=1, at=klankfreq_barplot, tick=F, line=1, labels=klankfreq$sub.class )
# or simpler: with(klankfreq, barplot(count) ) # all defaults
```

You can make a histogram like in Figure 8.2 in R with the follow commands:

```
# read dataset
load(file="data/pauses6.Rda")
# extract pause lengths (columns 12) for the year 2012, into a separate dataset `troon2012`
troon2012 <- pauses6[ pauses6$jaar==2012, 12 ] # save col_12 as single vector
# make histogram
hist( troon2012,
      breaks=seq(0, 5, by=0.25),
      col="grey80",
      xlab="Length of pause (s)", ylab="Frequency",
      main="Queen's Speech 2012 (N=305 pauses)" ) -> troonrede2012pauzes_hist
```

Chapter 9

Centre and dispersion

9.1 Introduction

In the preceding chapter, we learnt to count and classify observations. These allow us to summarise a variable's observations, for example in a table, a frequency distribution, or in a histogram. We can often summarise the observations even further, in characteristics which indicate the manner in which the observations are distributed. In this chapter we will acquaint ourselves with a number of such characteristics. Some of these characteristics are applicable to variables of all levels of measurement (e.g. mode), others only to variables of interval or ratio level (e.g. mean). After an introduction on using symbols, we will firstly discuss how we can describe the centre of a distribution, and how we can describe the dispersion.

9.2 Symbols

In descriptive statistics, much work is done with symbols. The symbols are abbreviated indications for a series of actions. You already know some of these symbols: the exponent ^{2} in the expression x^2 is a symbol which means "multiply x with itself", or $x^2 = x \times x$ (where $\times$ is also again a symbol).

Often a capital letter is used to indicate a variable (X), and a lower case letter is used to indicate an individual score of that variable. If we want to distinguish the individual scores, we do so with a subscript index: x_1 is the first observation, x_2 is the second observation, etc. As such, x_i indicates the score of participant number i, of variable X. If we want to generalise over all the scores, we can omit the index but we can also use a dot as an "empty" index: in the expression $x_\cdot$ the dot-index stands for any arbitrary index.

We indicate the number of observations in a certain group with a lower case n, and the total number of observations of a variable with the capital letter N. If there is only one group, like in the examples in this chapter, then it holds that $n = N$.

In descriptive statistics, we use many addition operations, and for these there is a separate symbol, $\sum$, the Greek capital letter Sigma, with which an addition operation is indicated. We could say "add all the observed values of the variable X to each other", but we usually do this more briefly:

$$\sum_{i=1}^{n} x_i, \text{or even shorter} \sum x$$

This is how we indicate that all x_i scores have to be added to each other, for all values from i (from $i = 1$, unless indicated otherwise) to $i = n$. All n scores of the variable x therefore have to be added up.

When brackets are used then pay good attention: actions described within a pair of brackets have priority, so you have to execute them first. Also when it is not strictly necessary, we will often use brackets for clarity, like in $(2 \times 3) + 4 = 10$.

9.3 Central tendencies

9.3.1 mean

The best known measure for the centre of a distribution is the mean. The mean can be calculated straightforwardly by adding all scores to each other, and then dividing the sum by the number of observations. In symbols:

$$\overline{x} = \frac{\sum x}{n} = \frac{1}{n} \sum_{i}^{n} x_i \tag{9.1}$$

Here we immediately encounter a new symbol, $\overline{x}$, often named "x-bar", which indicates the mean of x. The mean is also often indicated with the symbol M (mean), amongst others in articles in the APA-style.

Example 9.1: In a shop, it is noted how long customers have to wait at the checkout before their turn comes. For $N = 10$ customers, the following waiting times are observed, in minutes:
1, 2, 5, 2, 2, 2, 3, 1, 1, 3.
The mean waiting time is $(\sum X)/N = 22/10 = 2.2$ minutes.

The mean of X is usually expressed with one decimal figure more than the scores of X (see also §9.6.1 below about the number of significant figures with which we represent the mean).

The mean can be understood as the "balance point" of a distribution: the observations on both sides hold each other "in equilibrium", as illustrated in Figure 9.1, where the "blocks" of the histogram are precisely "in equilibrium" at the "balance point" of the mean of 2.2. The mean is also the value relative to which the N observations together differ the least, and therefore forms a good characteristic for the centre of a probability distribution.

The mean can only be used with variables of the interval or ratio level of measurement.

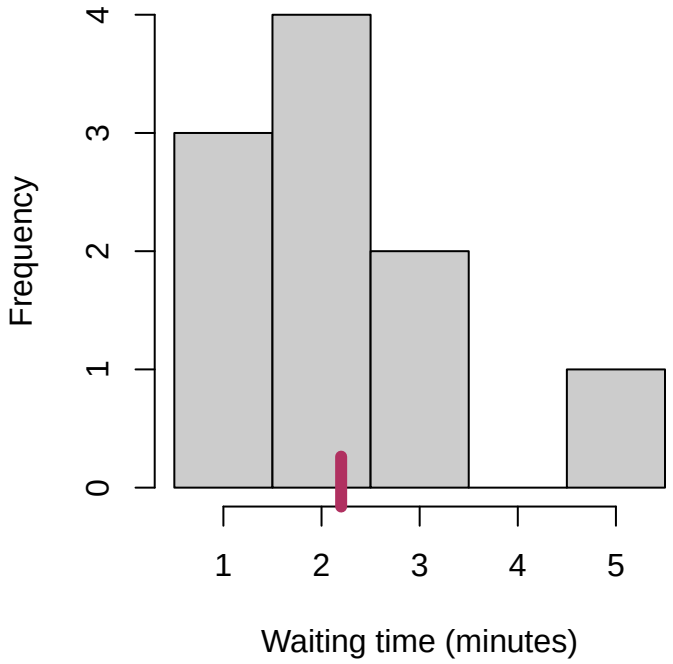


Figure 9.1: Histogram of N=10 waiting times, with the mean marked.

9.3.2 median

The median (symbol Md or $\tilde{x}$) is the observation in the middle of the order of observations [1]. When we sort the scores of X from smallest to largest, the

[1] In American English, the strip of ground in the middle of a road is called the "median (strip)" (British English: "central reservation"); this strip splits the road into two equally large halves.

median is the midpoint of the sorted sequence. Half of the observations are smaller than the median, and the other half is larger than the median.

For an odd number of observations, the middlemost observation is the median. For an even number of observations, the median is usually formed from the mean of the two middlemost observations.

Example 9.2: The waiting times from Example 9.1 are ordered as follows:
1, 1, 1, 2, *2, 2*, 2, 3, 3, 5.
The median is the mean of the two middlemost (italicised) observations, so 2 minutes.

The median is less sensitive than the mean to extreme values of x. In the above example, the extreme waiting time of 5 minutes has a considerable influence on the mean. If we remove that value, then the mean changes from 2.2 to 1.9 but the median is still 2. Extreme values thus have less great an influence on the median then on the mean. Only if the ordering of the observations changes, may the median also change.

The median can be used with variables of ordinal, interval or ratio level of measurement.

9.3.3 mode

The mode (adj. 'modal') is the value or score of X which occurs the most frequently.

Example 9.3: In the waiting times from Example 9.1 the score 2 occurs the most often (4×); this is the mode.

Example 9.4: In 2018, the mean income per household in the Netherlands was €29,500. The modal income (per household) was between €18,000 and €20,000[2]. As such, in 2018, most households in the Netherlands fell within this income class.

[2] https://www.cbs.nl/nl-nl/visualisaties/inkomensverdeling

The mode is even less sensitive than the mean to extreme values of x. In the Example 9.2 above, it does not matter what the value of the longest waiting time is: even if that observation has the value 10 or $1,000$, the mode remains invariably 2 (check it for yourself).

The mode can be used with variables of all levels of measurement.

9.3.4 Harmonic mean

If the dependent variable is a fraction or ratio, like the speed with which a task is conducted, then the (arithmetic) mean of formula (9.1) does not actually provide a good indication for the most characteristic or central value. In that case, it is better for you to use the harmonic mean:

$$H = \frac{1}{\frac{1}{n}\sum_{i}^{n}\frac{1}{x_i}} = \frac{n}{\sum_{i}^{n}\frac{1}{x_i}} \tag{9.2}$$

Example 9.5: A student writes $n = 3$ texts. For the first text (500 words) (s)he takes 2.5 hours, for the second text (1,000 words) (s)he takes 4 hours, and for the third text (300 words) (s)he takes 0.6 hours. What is this student's mean speed of writing? The speeds of writing are respectively 200, 250 and 500 words per hour, and the "normal" (arithmetic) mean of these is 317 words per hour. Nevertheless, the "actual" mean is $(500 + 1000 + 300)/(2.5 + 4 + 0.6) = 1800/7.1 = 254$ words per hour. The high writing speed of the short text counts for $1/n$ parts in the arithmetic mean, even though the text only contains $300/1,800 = 1/6$ of the total number of words.

Since the dependent variable is a fraction (speed, words/hour), the harmonic mean is a better central tendency. We firstly convert the speed (words per time unit) into its inverse (see (9.2), in denominator, within sum sign), i.e. to *time* per word: 0.005, 0.004, and 0.002 (time units per word, see footnote[3]). We then average these times, to a mean of 0.00366 hours per word, and finally we again take the inverse of this. The harmonic mean speed of writing is then $1/0.00366 = 273$ words per hour, closer to the "actual" mean of 254 words per hour.

[3]This is comparable with sports like rowing, swimming, cycling, ice skating, etc., where the time over an agreed distance is measured and compared, rather than the speed over an agreed time.

9.3.5 winsorized mean

The great sensitivity of the normal (arithmetic) mean for outliers can be restricted by changing the most extreme observations into less extreme, more central observations. The mean of these (partially changed) observations is called the *winsorized* mean.

Example 9.6: The waiting times from Example 9.1 are ordered as follows:
1, 1, 1, 2, 2, 2, 2, 3, 3, 5.
For the 10% winsorized mean, the 10% of smallest observations (by order) are made to equal the first subsequent larger value, and the 10% of largest observations are made to equal the last preceding smaller value (changed values are italicised here):
1, 1, 1, 2, 2, 2, 2, 3, 3, *3*.
The winsorized mean over these changed values is $\overline{x}_w = 2$ minutes.

9.3.6 trimmed mean

An even more drastic intervention is to remove the most extreme observations entirely. The mean of the remaining observations is called the *trimmed* mean. For a 10% trim, we remove the lowermost 10% *and* the uppermost 10% of the observations; as such, what remains is then only $(1 - (2 \times (10/100)) \times n$ observations (Wilcox, 2012).

Example 9.7: The waiting times from Example 9.1 are again ordered as follows:
1, 1, 1, 2, 2, 2, 2, 3, 3, 5.
For the 10% trimmed mean, the 10% of smallest observations (by order) are removed,
and likewise the 10% of largest observations are removed:
1, 1, 2, 2, 2, 2, 3, 3.
The trimmed mean over these $10 - (.2)(10) = 8$ remaining values here is $\overline{x}_t = 2$ minutes.

9.3.7 comparison of central tendencies

Figure 9.2 illustrates the differences between the various central tendencies, for asymmetrically distributed observations.

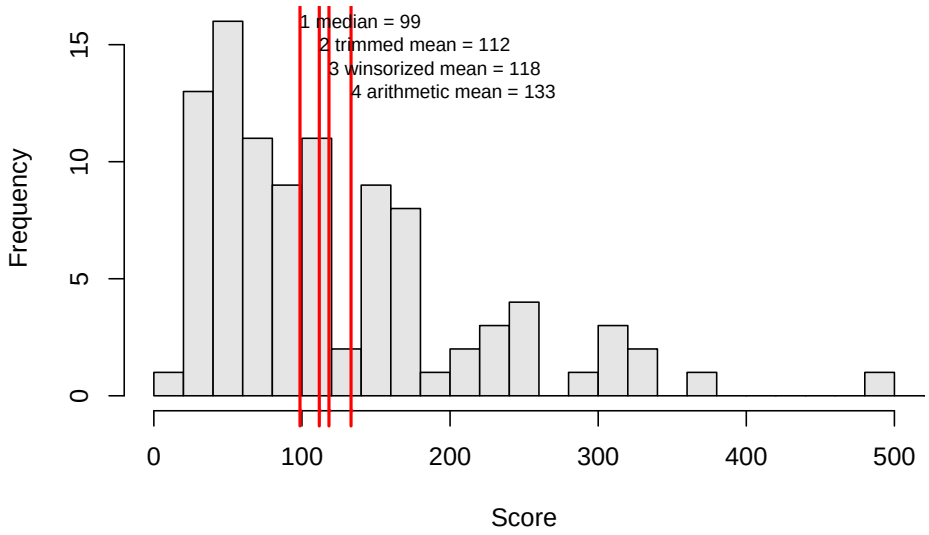


Figure 9.2: Histogram of a variable with positively skewed (asymmetric) frequency distribution, with (1) the median, (2) the 10% trimmed mean, (3) the 10% winsorized mean, and (4) the arithmetic mean, indicated. The observed scores are marked along the horizontal axis.

The arithmetic mean is the most sensitive to extreme values: the extreme values "pull" very hard at the mean. This influence of extreme values is tempered in the winsorized mean, and tempered even more in the trimmed mean. The higher the trim factor (the percentage of the observations that have been changed or removed), the more the winsorized and trimmed means will look like the median. Indeed, with a trim factor of 50%, out of all the observations, only one (unchanged) observation remains, and that is the median (check it for yourself). In §9.7 we will look further into the choice for the appropriate measure for the centre of a distribution.

9.4 Quartiles and boxplots

The distribution of a variable is not only characterised by the centre of the distribution but also by the degree of dispersion around the centre, i.e. how large the difference is between observations and the mean. For instance, we not only want to know what the mean income is but also how large the *differences* in income are.

9.4.1 Quartiles

Quartiles are a simple and useful measure for this (Tukey, 1977). We split the ordered observations into two halves; the dividing line between these is the median. We then halve each of these halves again into quarters. The quartiles are formed by the dividing lines between these quarters; as such, there are three quartiles. The first quartile Q_1 is the lowermost half's median, Q_2 is the median of all n observations, and the third quartile Q_3 is the uppermost half's median. Half of the observations (namely the second and third quarters) are between Q_1 and Q_3. The distance between Q_1 and Q_3 is called the "interquartile range" (IQR). This IQR is a first measure which can be used for the dispersion of observations with respect to their central value.

To illustrate, we use the fictive reading test scores shown in Table 9.1.

Table 9.1: The scores of N=10 pupils on three sections of the CITO test, taken in the final year of primary school in the Netherlands.

Pupil	Reading	Arithmetic	Geography
1	18	22	55
2	32	36	55
3	45	34	38
4	25	25	40
5	27	29	48
6	23	20	44
7	29	27	49
8	26	25	42
9	20	25	57
10	25	27	47
$\sum x$	270	270	475
$\overline{x}$	27.0	27.0	47.5

Example 9.8: The scores for the reading section in Table 9.1 are ordered as follows:
18, 20, 23, 25, 25, 26, 27, 29, 32, 45.
The median is $Q_2 = 25.5$ (between the 5th and 6th observation in this ranked list). The median of the lowermost half is $Q_1 = 23$ and that of the uppermost half is $Q_3 = 29$. The interquartile range is $\text{IQR} = 29 - 23 = 6$.

9.4.2 Outliers

In the reading scores in Table 9.1, we encounter one extreme value, namely the score 45, which differs markedly from the mean. A marked value like this is referred to as an "outlier". The limit for what we consider to be an outlier generally lies at $1.5 \times \text{IQR}$. If a value is more than $1.5 \times \text{IQR}$ above or under Q_1, we consider that observation to be an outlier. Check these observations again (recall the principle of diligence, see §3.1).

Example 9.9: For the aforementioned reading scores in Table 9.1, we found $Q_1 = 23$, $Q_3 = 29$, and $\text{IQR} = Q_3 - Q_1 = 29 - 23 = 6$. The uppermost limit value for outliers is $Q_3 + 1.5 \times \text{IQR} = 29 + 1.5 \times 6 = 29 + 9 = 38$. The observation with the score 45 is above this limit value, and is therefore considered to be an outlier.

9.4.3 Boxplots

We can now show the frequency distribution of a variable with five characteristics, the so-called "five-number summary", namely the minimum value, Q_1, median, Q_3, and maximum value. These five characteristics are represented graphically in a so-called "boxplot", see Figure 9.3 for an example (Tukey, 1977, §2C).

The box spans (approximately) the area from Q_1 to Q_3, and thus spans the central half of the observations. The thicker line in the box marks the median. The lines extend to the smallest and largest values *which are not outliers* [4]. The separate outliers are indicated here with a distinct symbol.

9.5 Measures of dispersion

9.5.1 Variance

Another way to show the dispersion of observations would be to look at how each observation deviates from the mean, thus $(x_i - \overline{x})$. However, if we add up all the deviations, they always total zero! After all, the positive and negative deviations cancel each other out (check that out for yourself in Table @#ref(tab:cito)).

[4] In a classic boxplot, the lines extend to the minimum and maximum (Tukey, 1977) and outliers are not indicated separately.

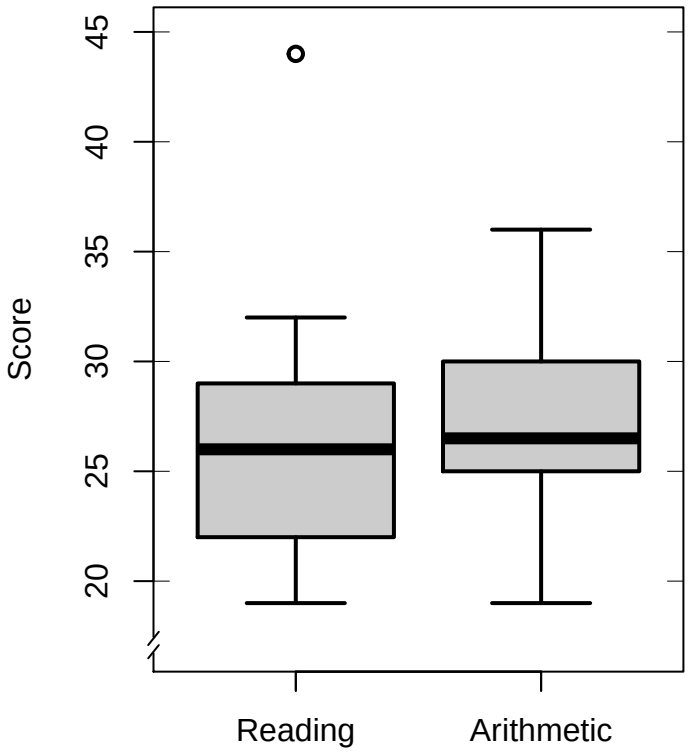


Figure 9.3: Boxplots of the scores of $N = 10$ pupils on the Reading and Arithmetic sections of the CITO test (see Table 9.1), with outliers marked as open circles. The observed scores are marked along the vertical axes.

Instead of calculating the mean of the deviations themselves, we thus calculate the mean of the squares of those deviations. Both the positive and negative deviations result in positive squared deviations. We then calculate the mean of all those squared deviations, i.e. we add them up and divide them by $(n-1)$, see Footnote[5]. We call the result the *variance*, indicated by the symbol s^2:

$$s^2 = \frac{\sum(x_i - \overline{x})^2}{n-1} \tag{9.3}$$

The numerator of this fraction is referred to as the "sum of squared deviations" or "sum of squares" (SS) and the denominator is referred to as the number of "degrees of freedom" of the numerator (d.f.; see §**??**).

Nowadays, we always calculate the variance with a calculator or computer.

9.5.2 standard deviation

To calculate the above variance, we squared the deviations of the observations. As such, the variance is a quantity which is not expressed in the original units (e.g. seconds, cm, score), but in squared units (e.g. s^2, cm^2, $score^2$). In order to return to the original units, we take the square root of the variance. We call the result the *standard deviation*, indicated by the symbol s:

$$s = \sqrt{s^2} = \sqrt{\frac{\sum(x_i - \overline{x})^2}{n-1}} \tag{9.4}$$

Example 9.10: The mean of the previously stated reading scores in Table 9.1 is 27.0, and the deviations are as follows:
-9, 5, 18, -2, 0, -4, 2, -1, -7, -2.
The squared deviations are 81, 25, 324, 4, 0, 16, 4, 1, 49, 4.
The sum of these squared deviations is 508, and the variance is $s^2 = 508/9 = 56.44$. The standard deviation is the root of the variance, thus $s = \sqrt{508/9} = 7.5$.

The variance and standard deviation can only be used with variables of the interval or ratio level of measurement. The variance and standard deviation can also be based again on the winsorized or trimmed collection of observations.

[5]We divide by $n-1$ and not by n, to get a better estimation of the dispersion in the *population*. In this way, we take into account the fact that we are using a characteristic of the sample (namely the mean) to determine the dispersion. If you are only interested in the dispersion in your *sample* of observations, and not in the population, divide it by n.

We need the standard deviation (a) when we want to convert the raw observations to standard scores (see §9.8 below), (b) when we want to describe a variable which is normally distributed (see §10.3, and (c) when we want to test hypotheses with the help of a normally distributed variable (see §13.2 et seq.).

9.5.3 MAD

Besides standard deviation, there is also a robust counterpart which does not use the mean. This measure is therefore less sensitive for outliers (robuster), which is sometimes useful.

For this, we look for the deviation of every observation from the median (not the mean). We then take the absolute value of these deviations[6] (not the square). Finally, we determine again the median of these absolute deviations (not the mean). We call the result the "median absolute deviation" (MAD):

$$\text{MAD} = k \;\; Md(|x_i - Md(x)|) \tag{9.5}$$

We normally use $k = 1.4826$ as a constant here; with this scale factor the MAD usually roughly matches the standard deviation s, if x is normally distributed (§10.3).

Example 9.11: The median of the previously mentioned reading scores in Table 9.1 is 25.5, and the deviations from the median are as follows:
-7.5, 6.5, 19.5, -0.5, 1.5, -2.5, 3.5, 0.5, -5.5, -0.5.
The ordered absolute deviations are
0.5, 0.5, 0.5, 1.5, *2.5, 3.5*, 5.5, 6.5, 7.5, 19.5.
The median of these 10 absolute deviations is 3, and MAD = $1.4826 \times 3 = 4.4478$. Notice that the MAD is smaller than the standard deviation, amongst others because the MAD is less sensitive for the extreme value $x_3 = 45$.

9.6 On significant figures

9.6.1 Mean and standard deviation

A mean result is shown in a limited number of significant figures, i.e. a limited number of figures, counted from left to right, ignoring the decimal place.

[6] Positive deviations remain unchanged, negative deviations are reversed.

The mean result's number of significant figures must be equal to the number of significant figures of the *number of observations* from which the mean is calculated. (Other figures in the mean result are not precisely determined.) The mean result must firstly be rounded to the appropriate number of significant figures, before the result is interpreted further, see Table 9.2.

Table 9.2: The number of significant figures in the reported mean is equal to the number of significant figures of the number of observations.

Num.obs.	Num.signif.figures	example mean	reported as
1 ... 9	1	21/8 = 2.625	3
10 ... 99	2	57/21 = 2.714286	2.7
100 ... 999	3	317/120 = 2.641667	2.64
1000 ... 9999	4	3179/1234 = 2.576175	2.576

The number of significant figures in the reported standard deviation is the same as in the mean, in accordance with Table 9.2.

9.6.1.1 Background

Let us assume that I have measured the distance from my house to my work along a fixed route a number of times. The mean of those measurements supposedly amounts to 2.954321 km. By reporting the mean with 7 figures, I am suggesting here that I know precisely that the distance is 2954321 millimetres, and at most 1 mm more or less: the last figure is estimated or rounded off. The number of significant figures (in this example 7) indicates the degree of precision. In this example, the suggested precision of 1 mm is clearly wrong, amongst other reasons because the start point and end point cannot be determined within a millimetre. It is thus usual to report the mean of the measured distance with a number of significant figures which indicates the precision of those measurements and of the mean, e.g. 3.0 km (by car or bike) of 2.95 km (by foot).

The same line of thought is applicable when measuring a characteristic by means of a survey question. With $n = 15$ respondents, the average score might be $43/15 \approx 2.86667$. However, the precision in this example is not as good as this decimal number suggests. In fact, here one deviant answer already brings about a deviation of ± 0.06667 in the mean. Besides, a mean score is always the result of a division operation, and "[for] quantities created from measured quantities by multiplication and division, the calculated result should have as many significant figures as the measured number with the least number of significant figures" [7]. In this example, the mean's numerator (43) and its denominator (15) both

[7] https://en.wikipedia.org/wiki/Significant_figures

consist of 2 significant figures. The mean score should be reported as 2.9 points, with only one figure after the decimal point.

9.6.2 Percentages

A percentage is a fraction, multiplied by 100. Use and report a rounded off percentage (i.e. two significant figures) only if the fraction's numerator is larger than 100 (observations, instances). If the numerator is smaller than 100 (observations, instances), then percentages are misleading, see Table 9.3.

Table 9.3: The number of significant figures in the reported proportion (or percentage) is related to the number of significant figures of the number of observations in the denominator of the fraction.

num.obs.(denominator)	num.signif.figures	example fraction	report as
1 ... 9	1	3/8 = 0.4	3/8
10 ... 99	2	21/57 = 0.36	21/57
100 ... 999	3	120/317 = 0.378	38%
1000 ... 9999	4 12	34/3179 = 0.3882	38.8%

9.6.2.1 Background

The rules for percentages arise from those in §9.6.1 applied to division operations. If the denominator is larger than 100, the percentage (with two significant figures) is the result of a scaling "down" (from a denominator larger than 100 to a denominator of precisely 100 percentage points). The percentage scale is less precise than the original ratio; the percentages are rounded off to two significant figures; the percentage's last significant figure is thus secured.

However, if the denominator is smaller than 100, then the percentage (with two significant figures) is the result of a "scaling upwards" (from a denominator smaller than 100 to a denominator of exactly 100 percentage points). The percentage scale then suggests a pseudo-precision which was not present in the original fraction, and the precision of the percentage scale is false. As such, if the denominator is smaller than 100, percentages are misleading.

Example 9.12: In a course of 29 students, 23 students passed. In this case, we often speak of a course return of 23/29 = 79%. However, a rendering as a percentage is misleading in this case. To see this, let us look at the 6 students who failed. You can reason that the number of 6 failed students has a rounding error of 1/2 student(s); when

converted to the percentage scale this rounding error is also thereby increased so that the percentages are less precise than the whole percentages (2 significant figures) suggest. Or put otherwise: the number of 6 failed students (i.e. a number with one significant figure) means we have to render the proportion with only one significant figure, and thus not as a percentage. It is preferable to report the proportion itself (23/29), or the "odds" (23/6 = 4) rounded off to the correct number of significant figures[8].

On the basis of the same considerations, a percentage with one decimal place (i.e. with three significant figures, e.g. "36.1%") is only meaningful if the ratio or fraction's denominator is larger than 1000.

Example 9.13: In 2013, 154 students began a two-year research master's degree. After 2 years, 69 of them had graduated. The nominal return for this cohort is thus 69/154 = 0.448052, which should be rounded off and reported as 44% (not as 44.81%).

9.7 Making choices

You can describe the distribution of a variable in various manners. If variable X is measured on the interval or ratio level of measurement, always begin with a histogram (§8.4) and a boxplot (§9.4.3).

The centre measures and dispersion measures can be arranged as in Table 9.4.

Table 9.4: Overview of discussed centre measures and dispersion measures. For assumptions abbreviated to *(a & b & c)*, see text below table.

Distribution	Centre measure	Dispersion measure
all	median	quartiles, IQR, MAD
…	trimmed or wins. mean	trimmed or wins.std.dev.
(a & b & c)	mean	standard deviation

[8]These "odds" indicate that there are 23 successful students to 6 failed students, i.e., rounded up, 4 successful students for every failed student.

The most **robust** measures are at the top (median, quartiles, IQR, MAD). These measures are robust: they are less sensitive for outliers or for potential *a*symmetry in the frequency distribution, as the examples in this chapter show.

The most **efficient** measures are at the bottom of Table 9.4: mean and standard deviation. These measures are efficient: they represent the centre and the dispersion the best, they have themselves the smallest standard deviation, and they need the (relatively) smallest number of observations for this. The other measures occupy a between position: the trimmed measures are somewhat more robust, and the winsorized measures somewhat more efficient.

However, the most efficient measures also demand the furthest reaching assumptions (and the most robust measures demand the fewest assumptions). These efficient measures are only meaningful if the distribution of X satisfies three assumptions: (a) the distribution is more or less symmetrical, i.e. the left and right halves of the histogram and the uppermost and lowermost halves of the boxplot look like each other's mirror image, (b) the distribution is unimodal, i.e. the distribution has a unique mode, and (c) the distribution contains no or hardly any outliers. Inspect these assumptions in the histogram and the boxplot of X. If one of these assumptions is not satisfied, then it is better to use more robust measures to describe the distribution.

9.8 Standard scores

It can sometimes be useful to compare scores which are measured on different scales. Example: Jan got an 8 as his final grade for
maths at Dutch secondary school, and his IQ is 136. Is the deviation of Jan with respect to the mean as large on both of the scales? To answer a question like this, we have to express the scores of the two variables on the same measurement scale. We do so by converting the raw scores to standard scores, or z-scores. For this, we take the deviation of every score with respect to the mean, and we divide the deviation by the standard deviation:

$$z_i = \frac{(x_i - \overline{x})}{s_x} \tag{9.6}$$

The standard score or z-score thus represents the distance of the i'the observation to the mean of x, expressed in units of standard deviation. For a standard score of $z = -1$, the observed score is precisely $1 \times s$ below the average $\overline{x}$. For a standard score of $z = +2$, then the observed score is precisely $2 \times s$ above the mean[9].

Z-scores are also useful for comparing two variables which are in fact measured on the same scale (for example, a scale of $1 \ldots 100$), but which nevertheless have

[9] Check: $z = +2 = \frac{(x_i - \overline{x})}{s_x}$, thus $2s = (x_i - \overline{x})$, thus $x_i = \overline{x} + 2s$.

different means and/or standard deviations, like the scores in Tabel 9.1. In Chapter 10, we will work more with z-scores.

The standard score or z-score has two useful characteristics which you should remember. Firstly, the mean is always equal to zero: $\overline{z} = 0$, and, secondly, the standard deviation is equal to 1: $s_z = 1$. (These characteristics follow from the definition in formula (9.6); we omit the mathematical proof here.) Thus, transformation from a collection of observations to standard scores or z-scores always yields a distribution with a mean of zero and a standard deviation of one. Do remember that this transformation to standard scores is only meaningful, provided that and to the extent that the mean and the standard deviation are also meaningful measures to describe the distribution of x (see §9.7).

9.9 SPSS

For **histogram, percentiles and boxplot**:

`Analyze > Descriptive Statistics > Explore...`

Select variable (drag to Variable(s) panel)
Choose `Plots`, tick: `Histogram`, and confirm with `Continue`
Choose `Options`, tick: `Percentiles`, and confirm with `Continue` and afterwards with `OK`.
The output comprises descriptive statistics and histogram and boxplot.

For **significant figures**:

`Analyze > Descriptive Statistics > Descriptives...`

Select variable (drag to Variable(s) panel)
Choose `Options`; tick: `Mean, Sum, Std.deviation, Variance, Minimum, Maximum`, and confirm with `Continue` and afterwards with `OK`.
The output comprises the requested statistical characteristics of the variable's distribution.

For **median**:

`Analyze > Compare Means > Means...`

Select variable (drag to Variable(s) panel)
Choose `Options`; tick: `Mean, Number of cases, Standard deviation, Variance, Minimum, Maximum` and also `Median`, and confirm with `Continue`

and afterwards with `OK`.
The output comprises the requested statistical characteristics of the variable's distribution.

Calculate and save **Standard scores** in a new column:

`Analyze > Descriptive Statistics > Descriptives...`

Select variables (drag to Variable(s) panel)
Tick: `Save standardized values as variables` and confirm with `OK`.
The new variable(s) with z-scores are added as new column(s) to the data file.

9.10 R

For **quartiles and boxplot** like Figure 9.3, we use the commands `fivenum`, `quantile`, and `boxplot`:

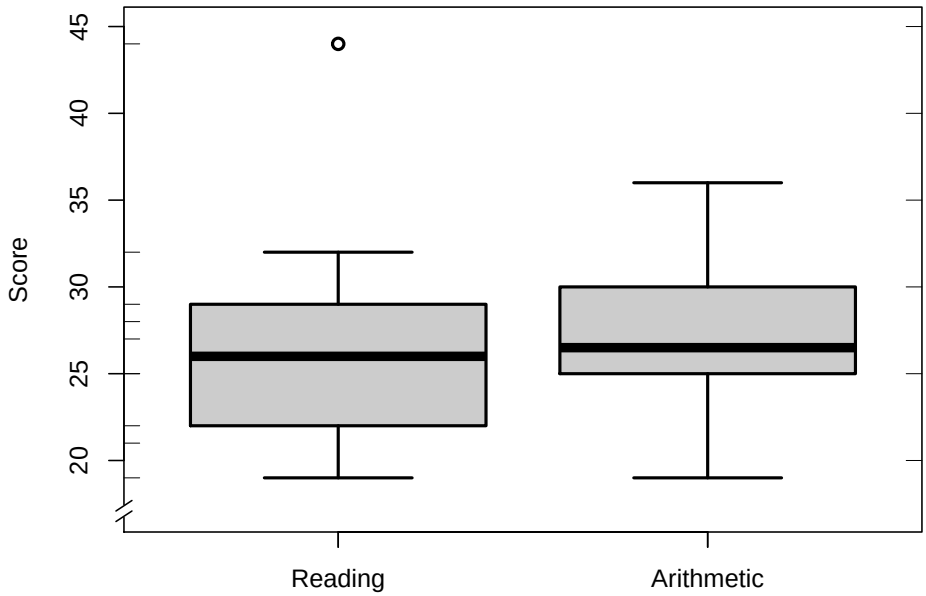
```
require(foreign) # for foreign::read.spss
cito <- read.spss("data/cito.sav")
# Columns in `cito.sav` have Dutch names:
# in Dutch: Leerling Lezen Rekenen Wereldorientatie stadplat rek.f
# in English: Pupil Reading Arithmetic World UrbRural Arith.factor
fivenum(cito$Lezen) # minimum, Q1, median, Q3, maximum
```

```
## [1] 19 22 26 29 44
```

```
quantile(cito$Lezen, c( 1/4, 3/4 ) ) # Q1 and Q3, calculated differently
```

```
##   25%   75%
## 22.75 28.75
```

```
op <- par(mar=c(4,4,1,2)+0.1) # smaller margins
with(cito,
  boxplot(Lezen, Rekenen, col="grey80", lwd=2, lty=1, ylab="Score", ylim=c(17,45) )
)
axis(side=1, at=c(1,2), labels=c("Reading","Arithmetic") )
plotrix::axis.break(axis=2) # break in left Y-aixs
rug(cito$Lezen, side=2) # markings on left Y-axis
rug(cito$rekenen, side=4) # markings on right Y-axis
```



Many **central tendencies** are pre-programmed as functions in R:

```
mean(cito$Lezen) # mean
```

```
## [1] 27.2
```

```
psych::winsor.mean(cito$Lezen, trim=.1) # winsorized mean, from psych package
```

```
## [1] 26.3
```

```
mean(cito$Lezen, trim=.1) # trimmed mean
```

```
## [1] 26.125
```

```
median(cito$Lezen)   # median
```

```
## [1] 26
```

Various **dispersion measures** are also pre-programmed:

```
var(cito$Lezen)  # variance
```

```
## [1] 50.17778
```

```
sd(cito$Lezen)   # standard deviation, sd(x) = sqrt(var(x))
```

```
## [1] 7.083627
```

```
mad(cito$Lezen)  # MAD
```

```
## [1] 5.1891
```

In contrast, we have to calculate **standard scores** ourselves, and save them ourselves as a new variable, called here `zReading` (note the parentheses in the first line):

```
zReading <- (cito$Lezen-mean(cito$Lezen)) / sd(cito$Lezen) # standardized (z) reading
head(zReading) # first few observations of variable zReading
```

```
## [1] -1.1575990  0.6776189  2.3716662 -0.3105753  0.1129365 -0.7340872
```

Chapter 10

Probability distributions

10.1 Probabilities

Calling behind the wheel increases the chance of an accident (Bhargava and Pathania, 2013). The average chance of precipitation in the Netherlands is 7%. My order has a 10% chance of being delivered a day later than promised. Chances and probabilities play an important role in our daily lives, and also in academic research. After all, many hypotheses are probabilistic in nature (see Chapter 2): hypotheses make statements about a difference in the *chances* of outcomes. To be able to draw conclusions with respect to these probabilistic hypotheses, we need to know something about probabilities and probability distributions. This is the subject of the present chapter.

As an introduction, let us take a look at a Dutch *Scrabble* game. The game contains a bag with 102 tiles inside, each of which has a letter on it [1]. Of the 102 tiles, 6 have the letter A on them. If I take one tile from a full and well-mixed bag, what is the chance that I draw the letter A? The probability-of-the-outcome-A is referred to as $P(\text{A})$, with the P of *Probabilitas* (Lat. "chance, probability"), and can be determined as

$$P(\text{A}) = \frac{\text{number of A's}}{\text{total number of tiles}} = \frac{6}{102} = 0.0588 \tag{10.1}$$

The probability of an event is expressed as a proportion, a number between 0 and 1, or as a percentage, i.e. a proportion in units of 1/100. A probability can never be smaller than 0 and can never be larger than 1: after all, the probability is the proportion between the number of specific outcomes (numerator) and the total number of possible outcomes (denominator) (see formula (10.1)),

[1] However, two of the tiles are blank (without any letter); later in this section we will remove these blank tiles from the bag.

where the numerator can never be larger than the denominator (Schuurman and De Kluiver, 2001).

When two outcomes mutually exclude each other, as is the case for the outcomes A or B in our Scrabble example, then we may *sum up* these outcomes (rule of sum, or addition principle, or *OR* rule). The probability of outcome A *or* outcome B (where outcomes A and B exclude each other), is the *sum* of $P(\text{A})$ and $P(\text{B})$:

$$P(\text{A or B}) = P(\text{A}) + P(\text{B}) \tag{10.2}$$

Example 10.1: In our Scrabble example, $P(\text{A}) = \frac{6}{102}$ and $P(\text{B}) = \frac{2}{102}$. As such, the probability of outcome A-or-B is $P(\text{A or B}) = P(\text{A}) + P(\text{B}) = 6/102 + 2/102 = 8/102 = .0784$.

If I take one tile from a full and well-mixed bag, then two complementary outcomes are possible: Either I draw an A, or I do *not* draw an A. The outcomes again mutually exclude each other so we may sum up the probabilities too. Moreover, the outcomes are complementary, i.e the outcome can only have one of these two possible outcomes. The respective probabilities of these complementary events are also complementary, i.e. these respective probabilities sum up to precisely 1 =100% (complement rule). After all, there is a 100% probability that the outcome is one of the two possible outcomes of the draw. If we already know $P(\text{A})$, we can easily calculate the probability of the complementary outcome:

$$P(\text{A}) + P(\text{not-A}) = 1 \tag{10.3}$$

$$P(\text{A}) = 1 - P(\text{not-A}) \tag{10.4}$$

$$P(\text{not-A}) = 1 - P(\text{A}) \tag{10.5}$$

Example 10.2: In our Scrabble example, $P(\text{A}) = \frac{6}{102}$. As such, the probability of the not-A outcome is $P(\text{not-A}) = 1 - P(\text{A}) = 1 - \frac{6}{102} = \frac{96}{102} = .9412$.

As a thought experiment, let us now take a second Scrabble game, and, from it, take a second tile bag which is equally full and well-mixed. Without looking, we

will now take a letter tile from each bag. There are now two events or outcomes, namely the outcome of the first draw (from the first bag), and the outcome of the second draw (from the second bag). These two outcomes do not mutually exclude each other, since they have no mutual influence on each other. After all, the second bag's outcome is not influenced by the first bag's outcome, and vice versa. As such, we say that these outcomes are *independent* of each other. When the outcomes are indeed independent of each other, we calculate the probability of a combination of the outcomes by *multiplication* (multiplication principle, or product rule, or *AND* rule).
The probability of the combination of outcome A *and* outcome B (where outcomes A and B are independent of each other), is the *product* of $P(\text{A})$ and $P(\text{B})$:

$$P(\text{A and B}) = P(\text{A}) \times P(\text{B}) \tag{10.6}$$

Example 10.3: In our Scrabble example, $P(\text{A}) = \frac{6}{102}$ and $P(\text{B}) = \frac{2}{102}$. The probability of outcome A with the first bag *and* B with the second bag is $P(\text{A and B}) = P(\text{A}) \times P(\text{B}) = \frac{6}{102} \times \frac{2}{102} = .0012$.

Example 10.4: In our Scrabble game, $P(\text{vowel}) = \frac{38}{102}$. The probability of drawing a vowel (A, E, I, O, U, Y) from the first bag *and* a vowel from the second bag is $P(\text{vowel-and-vowel}) = P(\text{first vowel}) \times P(\text{second vowel}) = \frac{38}{102} \times \frac{38}{102} = (\frac{38}{102})^2 = .1388$.

10.2 Binomial probability distribution

For the remainder of this chapter, we will adopt two changes to the Scrabble game. Firstly, we will remove the 2 blank, letter-less titles from the bag. There are now precisely 100 tiles left, of which 38 have a vowel (V) and 62 have a consonant (C). Accordingly, there are only two possible outcome categories left and these mutually exclude each other. We call such a variable of the nominal level of measurement, with precisely two categories, binomial ('two-named'). We regard the vowels as hits, and the consonants as misses. These two possible outcomes are complementary: $P(V) = .38$ (abbreviated as p) and $P(C) = .62$ (abbreviated as $q = 1 - p$).

Secondly, from now on, we will put the drawn letter tile back into the bag, once we have noted down the drawn letter. We also mix the bag again. In this

way, we do not require multiple complete letter bags, but only one letter bag which, after each draw with replacement, is once again complete and mixed. We consider the outcomes of consecutive draws to be independent.

An aside: The outcome of a certain draw is thus independent of the outcome of previous draws. If a vowel has just been drawn 100× in a row, then that has no influence at all on (the outcome of) the next draw from the letter bag. After all, the letter bag, or the hand of the person drawing tiles does not have any memory. At *each* draw, the probability of a hit is $p = .38$, even if a vowel has just been drawn 100× or \$1000×. The same is the case for consecutive outcomes with roulette: in *each* round, the probability of a hit is 1/37, even if the ball has just landed on the same number 100×[2].

With the aforementioned changes, let us now conduct $n = 3$ draws (with replacement, see above), and for each possible outcome determine the probability of the outcome, see Table 10.1.

Table 10.1: Probabilities of possible outcomes of $n = 3$ vowel draws, $p = .38$, with replacement (see text).

Outcome	Number of vowels	Probabilitiy
CCC	0	$qqq = q^3$
VCC	1	$pqq = pq^2$
CVC	1	$qpq = pq^2$
CCV	1	$qqp = pq^2$
VVC	2	$ppq = p^2q$
VCV	2	$pqp = p^2q$
CVV	2	$qpp = p^2q$
VVV	3	$ppp = p^3$

The number of hits (vowels) in the $n = 3$ draws has the *probability distribution* summarised in Table 10.2 (first and last column) and Figure 10.1 (horizontal and vertical axes). In such a probability distribution, we can see, for each possible outcome of x (here: number of vowels), how high the probability of the outcome is.

Table 10.2: Probability distribution of a binomial variable with $n = 3$ and $p = .38$.

Number of vowels	Probability	Probability
0	$1q^3$	$= .2383$
1	$3pq^2$	$= .4383$
2	$3p^2q$	$= .2686$

[2]Roulette players can gamble on 36 of the 37 possible outcomes, so in the long term the casino receives a 1/37 share of all bets.

Number of vowels	Probability	Probability
3	$1p^3$	$= .0549$
total	$(p+q)^3$	$= 1.0000$

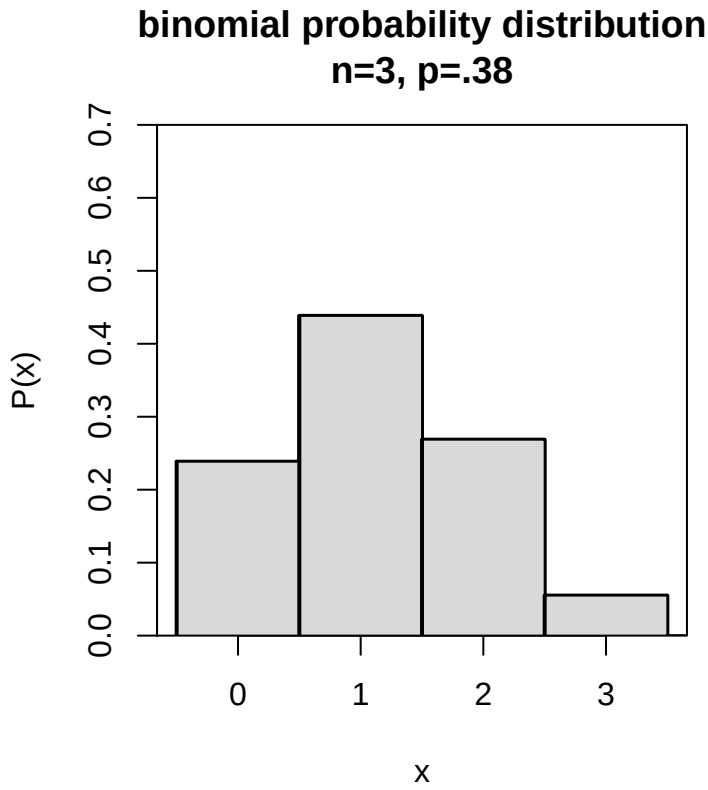


Figure 10.1: Probability distribution of a binomial variable with $n = 3$ and $p = .38$.

We call the probability distribution of a binomial variable the binomial probability distribution, also referred to as the binomial distribution. You can calculate the precise probabilities of the binomial probability distribution with the formula (10.7) below.

10.2.1 formulas

The probability of an x number of hits in n draws is given as

$$P(x\,\text{hits}) = \binom{n}{x} p^x (1-p)^{n-x} \tag{10.7}$$

in which n is the number of draws or attempts, x is the number of hits (between 0 and n), and p is the probability of a hit.

The coefficient $\binom{n}{x}$ indicates the number of different orderings in which we can choose a combination (syllable) of x elements from
n. With $x = 1$ vowel from $n = 3$ draws, there are three possibilities: one vowel might have been drawn in the first draw, or the second draw, or the third draw, see Table 10.1. The number of different possible orderings is indicated as

$$\binom{n}{x} = \frac{n!}{x!(n-x)!} \tag{10.8}$$

in which $x! = x(x-1)(x-2)\cdots \times 2 \times 1$, thus $4! = 4 \times 3 \times 2 \times 1 = 24$.

Example 10.5: There are 4 chairs for 2 persons. A maximum of 1 person is allowed to sit down on one chair. How many different orderings of $x = 2$ persons are possible over $n = 4$ chairs?

Answer: There are $\binom{4}{2} = \frac{4\times3\times2\times1}{2\times1\times2\times1} = \frac{24}{4} = 6$ possible orderings, namely 1100, 1010, 1001, 0110, 0101, and 0011.

These binomial coefficients indicating the number of different possible orderings can quickly be retrieved from Pascal's so-called triangle, depicted in Table 10.3. We can find the number of different orderings of $x = 2$ persons over $n = 4$ chairs in row $n = 4$. The uppermost row is that for $n = 0$. The fifth row is that for $n = 4$ and we can see the binomial coefficients for $x = 0, 1, 2, 3, 4$ there one after another. For $\binom{4}{2}$, we find there the binomial coefficient 6. Every coefficient is the total of the two coefficients above[3], and every coefficient can be understood as the number of possible routes descending from the top of the triangle to the cell.

Table 10.3: Pascal's triangle: Binomial coefficients for the number of possible orderings for a combination of x elements from n (see text).

$n = 0$:					1				
$n = 1$:				1		1			
$n = 2$:			1		2		1		
$n = 3$:		1		3		3		1	
$n = 4$:	1		4		6		4		1

[3]Thus, $\binom{n}{x} = \binom{n-1}{x} + \binom{n-1}{x-1}$ (Weisstein, 2015).

Table 10.3: Pascal's triangle: Binomial coefficients for the number of possible orderings for a combination of x elements from n (see text).

$n = 5$:			1		5		10		10		5		1	
$n = 6$:		1		6		15		20		15		6		1
$n = 7$:	1		7		21		35		35		21		7		1

The mean and the standard deviation of the binomial probability distribution are

$$\begin{aligned} \mu &= np \\ \sigma &= \sqrt{np(1-p)} \end{aligned}$$

Example 10.6: The binomial probability distribution for x hits from $n = 3$ draws with $p = .38$ probability of a hit is shown in Figure 10.1. This binomial probability distribution has an average $\mu = n \times p = 3 \times .38 = 1.14$, and a standard deviation $\sigma = \sqrt{n \times p \times (1-p)} = \sqrt{3 \times .38 \times .62} = 0.84$>

10.2.2 R

```
dbinom( 0:3, size=3, prob=.38 )
```

```
## [1] 0.238328 0.438216 0.268584 0.054872
```

The output is shown in Table 10.2 below.

```
matrixcalc::pascal.matrix( 10 ) # left under diagonal is Pascal's triangle
```

```
##       [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
##  [1,]    1    0    0    0    0    0    0    0    0     0
##  [2,]    1    1    0    0    0    0    0    0    0     0
##  [3,]    1    2    1    0    0    0    0    0    0     0
##  [4,]    1    3    3    1    0    0    0    0    0     0
##  [5,]    1    4    6    4    1    0    0    0    0     0
##  [6,]    1    5   10   10    5    1    0    0    0     0
##  [7,]    1    6   15   20   15    6    1    0    0     0
```

```
##  [8,]    1    7   21   35   35   21    7    1    0    0
##  [9,]    1    8   28   56   70   56   28    8    1    0
## [10,]    1    9   36   84  126  126   84   36    9    1
```

Pascal's triangle can be found on the left under the diagonal of this matrix.

10.3 Normal probability distribution

The more the sample size n increases, the less gradually the binomial probability distribution will move up, and the more fluid the probability distribution will become, as is shown in Figure 10.2.

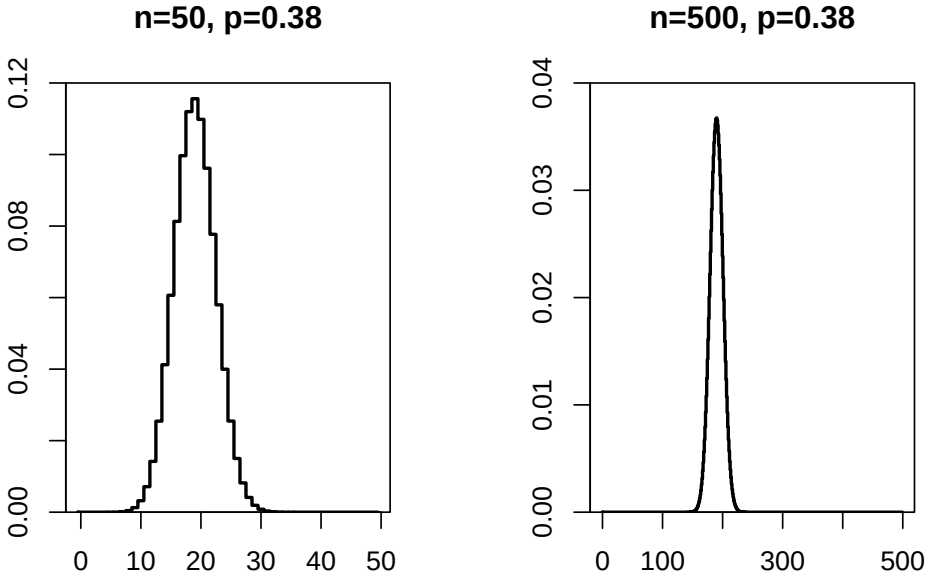


Figure 10.2: Probability distribution of a binomial variable x with n=50 (left) and n=500 (right) and p=.38.

With an even larger sample, the probability distribution becomes a fluid line. This probability distribution occurs so often, that it is called the *normal probability distribution* or 'normal distribution'. The distribution is also referred to as the Gaussian distribution (named after the mathematician Carl Friedrich Gauss, 1777–1855), or the 'bell curve' (after the shape). Many variables approximately follow this probability distribution: birth weight, body length, vocabulary size, IQ, contents of a 1 litre ℮ carton of milk, length of a telephone conversation, etc. etc. For all of these variables, observations close to the average have a high probability of occurring, and observations which deviate greatly from the average are relatively rare (low probability).

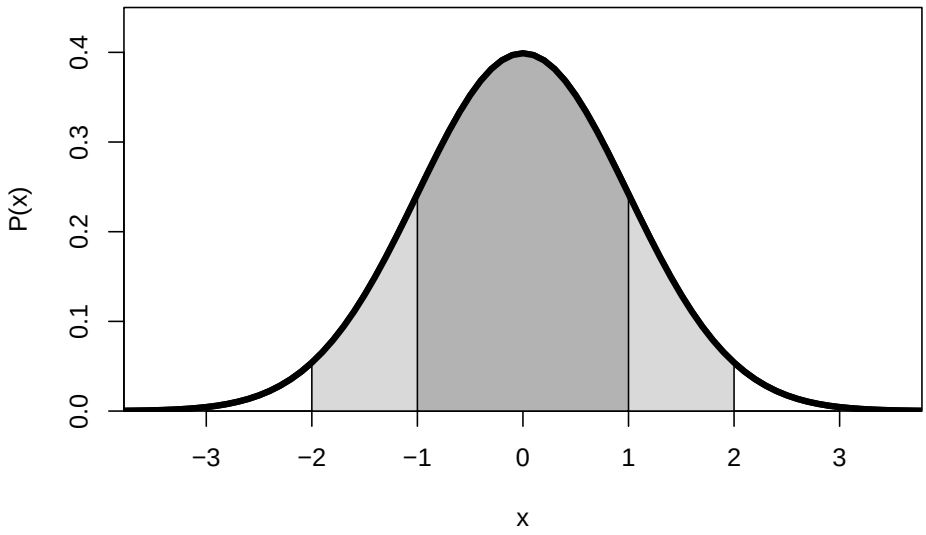


Figure 10.3: Normal probability distribution of a variable x with average 0 and standard deviation 1.

The normal probability distribution of a variable X with average μ and standard deviation σ has the following characteristics (see Figure 10.3):

- the distribution is symmetrical around the average μ,
- the distribution is asymptotic, i.e. the tails go on infinitely,
- the average, the median and the mode coincide,
- the total area under the curve, i.e. the total probability of one of the possible outcomes, is equal to 1,
- the area under the curve indicates the probability of a value of X within a certain interval,
- the inflection points of the curve (from concave to convex and vice versa) are at $X = \mu - \sigma$ and $X = \mu + \sigma$,
- around 2/3's of the observations are between $X = \mu - \sigma$ and $X = \mu + \sigma$ (dark grey area; $P(-1 < x/\sigma < 1) = .6827$ or 68%) and around 95% of the observations are between -2σ and $+2\sigma$ (dark grey plus light grey areas; $P(-2 < x/\sigma < 2) = .9546$), this is known as the Empirical Rule.

A normal probability distribution with $\mu = 0$ and $\sigma = 1$ is referred to as the standard normal probability distribution. Just as we saw earlier (§9.8), we can standardise a normally distributed variable x, i.e. transform the observations

to a standard score or z-score: $z = (x - \overline{x})/s$. The probability distribution in Figure 10.3 is that of the standard normal probability distribution of Z, or the probability distribution of $(X - \mu)/\sigma$.

You could calculate the probability distribution of a normally distributed variable X yourself with the help of the formula (10.10) below. However, it is more convenient to use a table for it; this can be found in Appendix B. Explained in graphical form, the tables provide you, for different areas or probabilities p on the right-hand side under the curve, the positive value of Z^* which constitutes the left-hand limit of the area. This means that you have precisely probability p of finding a value Z which is as large as or larger than this lower limit Z^* (provided of course that the variable is indeed normally distributed).

Example 10.7: On the right-hand side of Figure 10.3, we can see a small white area under the curve. This area renders the probability that $Z > 2$. The area has a size of 0.0228. The probability of finding a value of $Z > 2$ is thus 0.0228 or a little less than 2.5%. (Tip: relate this probability to the aforementioned Empirical Rule).

In Appendix B, you can find for convenience not one but two tables, each consisting of several column designations and and a row of cells. The first table provides you, for different 'rounded' probabilities p (columns), the critical values Z^* (cells), for which the probability p of finding a value of Z which is as large as or larger than the critical value Z^*, is precisely equal to the value p at the top of the column. The second table works the same, but in the case the 'rounded' values of Z^* are in the cells, and precise probabilities p are in the column designations.

What is the probability p that $Z > 1$? In the second subtable, second column, we find $p = 0.1587$. Based on this, we also know that $P(Z < 1)$ must be $1 - 0.1587 = 0.8413$. Moreover, we know that the distribution is symmetric (see above), so we know that $P(Z < -1)$ must also be .1587. What is the probability p that $Z > 3$? In the second subtable, we find for boundary value $Z^* = 3$ the p-value $p = 0.0013$. Thus, for a normally distributed variable Z there is a p-value $p = 0.0013$ of finding a value of Z which is at least three standard variations above the average.

We often want to know the opposite: when we choose a certain p-value, what should the boundary value Z^* be? Which boundary value distinguishes the highest 5% of observations from the lowest 95% ($p = 0.05$)? In the first subtable, we find for p-value $p = 0.05$ the boundary value $Z^* = 1.645$. This boundary value, calculated back to the original variable, is often referred to as the 95th

percentile or 'P95' of the distribution. Whoever has achieved at least this score, is part of the top 5% and has thus performed better than 95% of the participants (provided again that the variable is indeed normally distributed).

Example 10.8: By definition, extreme values occur infrequently with a normally distributed variable. But what is the limit for an extreme value. Let us assume that we want to consider no more than 5% of all observations as extreme. The normal probability distribution is symmetric, thus from this 5% we can expect that one half (2.5%) is at the left extremity of the distribution, and the other 2.5% is on the right-hand side. Which boundary value Z^* corresponds with this p-value $p = 0.025$?

In Appendix B, we take the first subtable. In the column for p-value $p = 0.025$, we find boundary value $Z^* = 1.960$. If we find an observation with $Z \geq 1.960$ or with $Z \leq -1.960$, then we consider that to be an extreme, rare observation.

Example 10.9: Intelligence is expressed as an IQ score, a variable with a normal probability distribution with $\mu = 100$ and $\sigma = 15$. "Membership of Mensa is open to persons who have achieved a score within the upper two percent of the general population on an approved intelligence test that has been properly administered and supervised" (www.mensa.org). What is the minimum IQ score you must achieve to become a member?

Answer: The 98th percentile from a standard normally distributed variable is at $Z^* = +2.0537$, and thus with $x = \overline{x} + 2.0537s = 100 + 30.8 = 130.8$. Rounded upwards, you thus have to achieve an IQ score of 131 points or higher.

Example 10.10: Verify the aforementioned Empirical Rule with the help of Appendix B.

10.3.1 formulas

If variable X has a normal probability distribution, with average μ and standard deviation σ, then this is shown as

$$X \sim \mathcal{N}(\mu, \sigma) \tag{10.9}$$

The normal probability distribution of variable X with average μ and standard deviation σ is

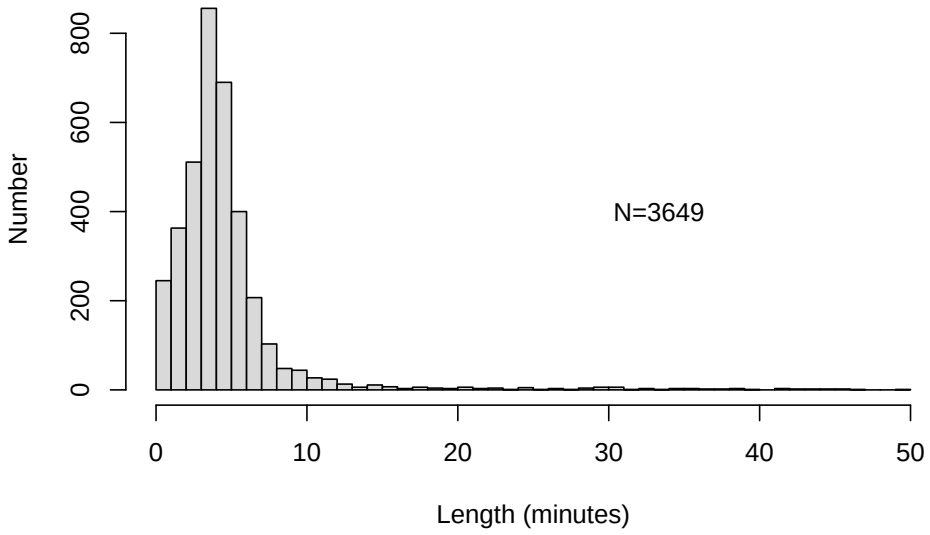
$$P(X) = \frac{1}{\sigma\sqrt{2\pi}} \mathrm{e}^{\frac{-(X-\mu)^2}{2\sigma^2}}. \tag{10.10}$$

The standard normal probability distribution of variable Z with average $\mu = 0$ and standard deviation $\sigma = 1$ is

$$P(Z) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{\frac{-Z^2}{2}} \tag{10.11}$$

10.4 Does my variable have a normal probability distribution?

The longest song in my digital music library lasts around 50 minutes (it's a piece of classical Indian music, a 'morning raga'). A histogram of all music number lengths is shown in Figure **??**.



This histogram shows that these lengths clearly do *not* follow a normal probability distribution: the distribution is not symmetric, and the lowest tail does not go on infinitely (there are no music numbers with negative lengths).

The average $\bar{x} = 4.698$ and standard deviation $s = 5.11$ also point to a non-normal probability distribution: with a normal distribution, we expect that only $(68/2) + 50 = 84\%$ of the lengths last longer than $\bar{x} - s \approx 0$ minutes, but in reality 100% last longer than 0 minutes (thus a larger proportion than expected).

A frequently used technique to inspect whether or not a variable X has a normal probability distribution, is to make a graph with the observed values along one of the axes (here the horizontal axis), and the corresponding z-scores along the other axis. A figure like this is called a quantile-quantile plot or Q-Q plot; the Q-Q plot for the lengths in my music library are shown in Figure 10.4.

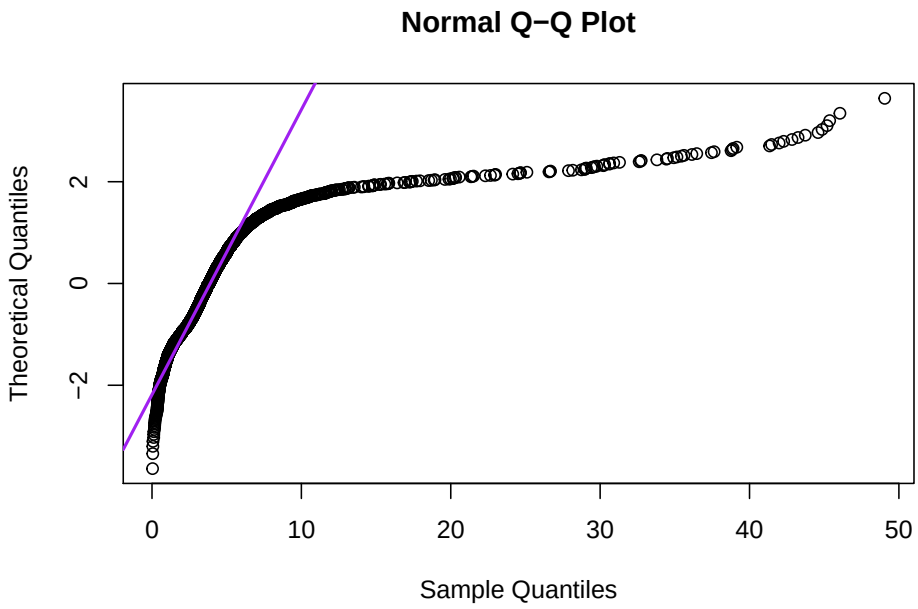


Figure 10.4: Quantile-quantile plot of the lengths of music numbers in my digital library.

If the lengths had a normal probability distribution (were normally distributed), then the points would cluster around the purple straight line. And there would have to be a number of negative lengths… The deviations from the purple straight line in Figure 10.4 thus indicate that the observed lengths do not follow a normal probability distribution, as we already saw in the histogram (Figure **??**).

There are also different statistical tests to investigate whether or not a variable has a normal probability distribution. The two most used are the Shapiro-Wilk test (with test statistic W)
for normality, and the Kolmogorov–Smirnov test (with test statistic D) for normality. Both tests investigate the H0:$X \sim \mathcal{N}(\bar{X}, s)$ (see formula (10.9)).

10.4.1 SPSS

`Analyze > Descriptive Statistics > Explore...`

Select the variable Time (drag to the Dependent List panel).
Choose the button `Plots`, and tick `Normality plots with tests`, which means 'if you make a QQ-plot or Normality plot, you should also then conduct tests on normality'. Confirm with `Continue` and afterwards with `OK`. The output contains firstly the results of the Shapiro-Wilks test and the Kolmogorov-Smirnov test. According to both tests, the probability of finding this distribution, if H0 is true, is almost null – see however the warning in §@ref(#sec:plargerthannull)! We thus reject H0 and conclude that the lengths of music numbers are not normally distributed. After these test results, there is, amongst others, a Q-Q plot.

10.4.2 R

```
itunes <- read.table( file="data/itunestimes20120511.txt", header=TRUE )
# Size in bytes, Time in ms
qqnorm(itunes$Time/60000, datax=T, plot.it=FALSE) # normally we'd use plot.it=TRUE
# qqline(itunes$Time/60000, datax=T, col="purple", lwd=T) # see QQ-plot above
```

```
shapiro.test(itunes$Time/60000)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  itunes$Time/60000
## W = 0.50711, p-value < 2.2e-16
```

According to this test, the probability of finding a distribution, if H0 is zero, is almost null, namely smaller than 2.2×10^{-16} (i.e. smaller than the smallest number that the analysis packet can render). We therefore reject H0 and conclude that the length of music numbers is not normally distributed.

10.5 What if my variable is not normally distributed?

In Part III, we will discuss various statistical tests. The tests which we discuss in Chapters 13 and 14 and **??** however require that the independent variable has

a normal probability distribution. If a variable does *not* have a (approximately) normal probability distribution, then the variable cannot simply be used for statistical testing with the statistical tests there, or to be more precise, the conclusions from such a statistical testing are not valid then. What can be done? There are then two possibilities.

Firstly, it is possible to transform the dependent variable y, i.e. to apply an arithmetic operation to it. If all is well, that results in a variable y' which is actually normally distributed. Much used transformations are: to take the logarithm ($y' = \log y$), take the square root, or invert ($y' = 1/y$). Then, the transformed dependent variable y' is used for the statistical testing. Of course, it is imperative to check whether the new dependent variable y' is indeed (approximately) normally distributed. When interpreting the results of the analysis, you should also take into account the transformation performed!

Secondly, it is sometimes possible to use another statistical test which does not require that the dependent variable is normally distributed. Those are called nonparametric tests. We will look at these in more detail in the chapters **??** and **??**. A disadvantage of those tests is nevertheless that they have less statistical power (for a discussion of power, see Chapter 14): they are less sensitive, and thus require larger samples to establish an effect.

10.6 Probability distribution of average

In this section, we consider the music numbers in my digital music library as a *population*. We now take a random sample of $n = 50$ numbers, and determine the average length of these 50 music numbers in the sample: let us say $\bar{x} = 4.401$ minutes. Surprisingly enough, the average of this sample is close to the average of the population ($\mu = 4.615$, see above). We repeat this operation 250×: in this way, we get 250 sample averages. The frequency distribution of these 250 sample averages are shown in Figure 10.5.

Surprisingly enough, these *averages* from (the dependent variables X in) the samples *do* show a more or less normal probability distribution, regardless of whether the variable X in the population is normally distributed or not. Put otherwise, the probability distribution of a sample average *always* approximates the normal probability distribution, regardless of the probability distribution of the variable in question in the population, provided that the sample was sufficiently large. (This is known as the Central Limit Theorem). Reread the above sentences again carefully. As a rule of thumb, the size of the sample, n, should be at least 30. The larger the sample is, the less the probability distribution of the sample averages deviates from the normal distribution.

The normal probability distribution of the sample averages has its own average, $\mu_{\bar{X}}$, and its own standard deviation, $s_{\bar{X}}$. For this, the following applies:

$$\mu_{\bar{X}} = \mu_X \tag{10.12}$$

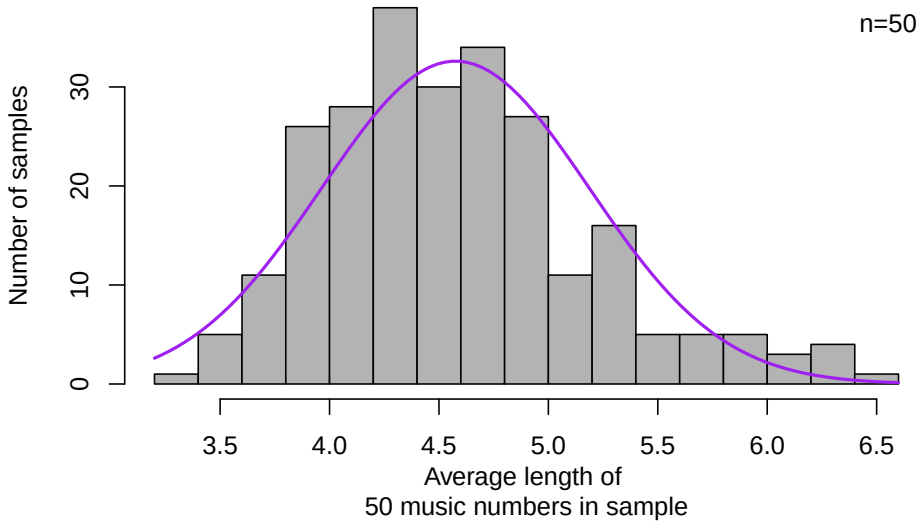


Figure 10.5: Frequency distribution of 250 averages, each over a random sample of $n = 50$ music numbers (the dependent variable is the length of a music number, in minutes). The matching normal distribution is shown as a fluid curve.

and

$$s_{\bar{X}} = \frac{s}{\sqrt{n}} \tag{10.13}$$

The standard deviation of the mean, $s_{\bar{X}}$, is also known as the 'standard error of the mean'. The same averages $\bar{X}$ have less dispersion than the separate observations X, and the averages also vary less when taken over a larger sample, as seems to be the case from formula (10.13). You can consider this standard error of the mean as the 'margin of error' in the estimation of the population average out of the sample average.

What is special now is that we do not have to draw and analyse 250 repeated random samples. After all, we know that the sample averages have a normal probability distribution with $\mu_{\bar{X}} = \mu_X$ and $s_{\bar{X}} = \frac{s}{\sqrt{n}}$. We can derive the probability distribution of the mean from only one sample of n observations, with a sample mean $\bar{X}$ and one standard deviation (Cumming, 2012). Reread this paragraph carefully.

10.7 Confidence interval of the mean

As explained above, we can use the mean of the sample, $\bar{X}$, as a good estimate of the unknown mean in the population, μ. On the basis of the Central Limit Theorem (§**??**), we also know that the means of repeated samples (of

n observations) follow a normal distribution: $\mu_{\bar{X}} \sim \mathcal{N}(\mu_X, \sigma/\sqrt{n})$, and thus that 95% of these repeated sample means will lie between $\mu_X - 1.96\sigma/\sqrt{n}$ and $\mu_X + 1.96\sigma/\sqrt{n}$. This interval is called the 95% confidence interval. We know with 95% confidence that the population mean μ lies in this interval — provided that n is sufficiently large, and provided that the standard deviation, σ, is known in the population.

In practice, this last condition is rarely or never satisfied. The standard deviation in the population is usually not known and this σ is thus also estimated from the sample. We use the sample of n observations not only to estimate μ_X but also to estimate σ_X. We can then no longer determine the confidence interval on the basis of the standard normal probability distribution. Instead, we use an adapted version of it, the so-called t-distribution (Figure 10.6). This probability distribution of t is somewhat broader, i.e. with a somewhat lower peak and with somewhat thicker tails than the standard normal probability distribution of Z in Figure 10.3. The thought behind this is that the estimation of μ is a bit more uncertain (thus the probability distribution is wider) since not only μ but also the standard error of the mean ($s/\sqrt{n}$) are estimated on the basis of the sample. In both estimations, there can be deviations which mean that there is somewhat more probability of finding a mean which deviates from the population mean. As we have already seen, the larger n is, the better the estimation of μ: the t-distribution then approximates the standard normal probability distribution.

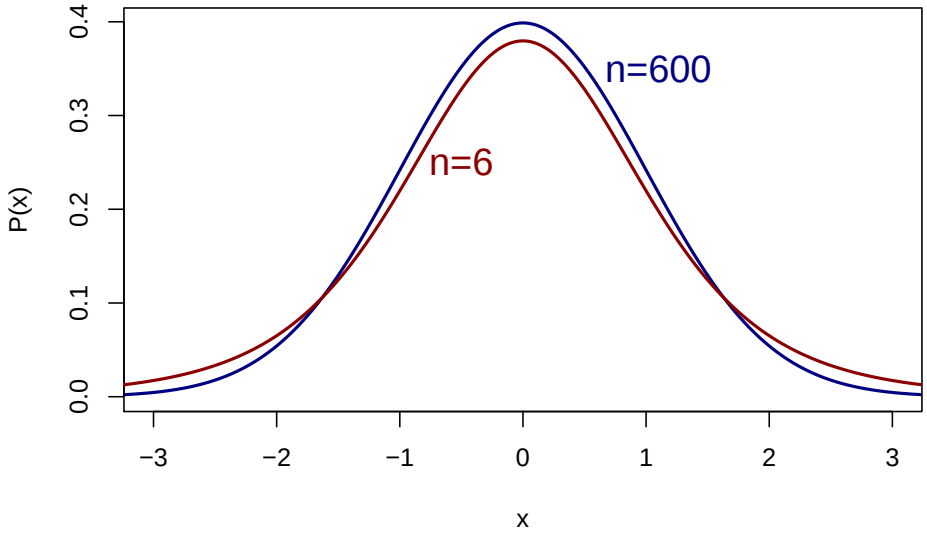


Figure 10.6: Probability distribution according to the t-distribution of a variable x with mean 0 and standard deviation 1, for n=600 and n=6.

For the t-distribution, we thus have to know how large the sample was; after all, this n determines the precise probability distribution of t, and with it the

critical value $t*$. We will go into more detail on that in §13.2.1. Here, a detailed example will suffice.

Example 10.11: Sometimes a researcher wants to know the speed or tempo with which Dutch is actually spoken, and how much variation in this speech rate or tempo there is between speakers. This variable, speech rate, is expressed as the number of seconds a syllable lasts (typically about 0.2 second or 200 milliseconds). Although Quené (2008) estimates that $\mu = 0.220$ s and $\sigma = 0.0225$ s, we act as if we do not know these population parameters — just like real researchers who usually do not know the population parameters.

For a sample of $n = 30$ speakers, we find $\overline{x} = 0.215$ and $s = 0.0203$ seconds. From this, we estimate [4] $\hat{\mu} = 0.215$ and $\hat{\sigma} = 0.0203$. Since σ is not known, we use the t-distribution to determine the confidence interval. We use the t-distribution for $n = 30$ and find a critical value $t^* = 2.05$ (see Appendix C, for $B = 95\%$). According to formula (13.2), we know with 95% confidence that the unknown population mean μ lies between $\overline{x} - 2.05 \times s_{\bar{x}}$ and $\overline{x} + 2.05 \times s_{\bar{x}}$, or between $0.215 - 2.05 \times 0.0037$ and $0.215 + 2.05 \times 0.0037$, or between 0.208 and 0.223 seconds. If the true value in the population is within these limits, then the observed sample value has a probability of 95% of occurring (Spiegelhalter, 2019, p.241).

In Figure 10.7, you can see the results of a computer simulation to illustrate this. We have taken 100× imaginary samples of $n = 30$ native speakers of Standard Dutch, and established the speech tempo of these speakers. For each sample, we have drawn the 95% confidence interval. For 95 of the 100 samples, the population mean $\mu = 0.220$ indeed falls within the interval. But for 5 out of 100 samples, from a population with $\mu = 0.220$, the sample's confidence interval does not contain the population mean (these are marked along the right hand side).

10.7.1 formulas

The two-sided $B\%$ confidence interval for the population average is

$$\bar{X} \pm t^*_{n-1} \times \frac{s}{\sqrt{n}} \tag{10.14}$$

in which t^* with $n - 1$ degrees of freedom is found with the help of Appendix C, see §13.2.1 for more explanation about this.

[4] Estimations of parameters are indicated with a "circumflex" or "hat" above them.

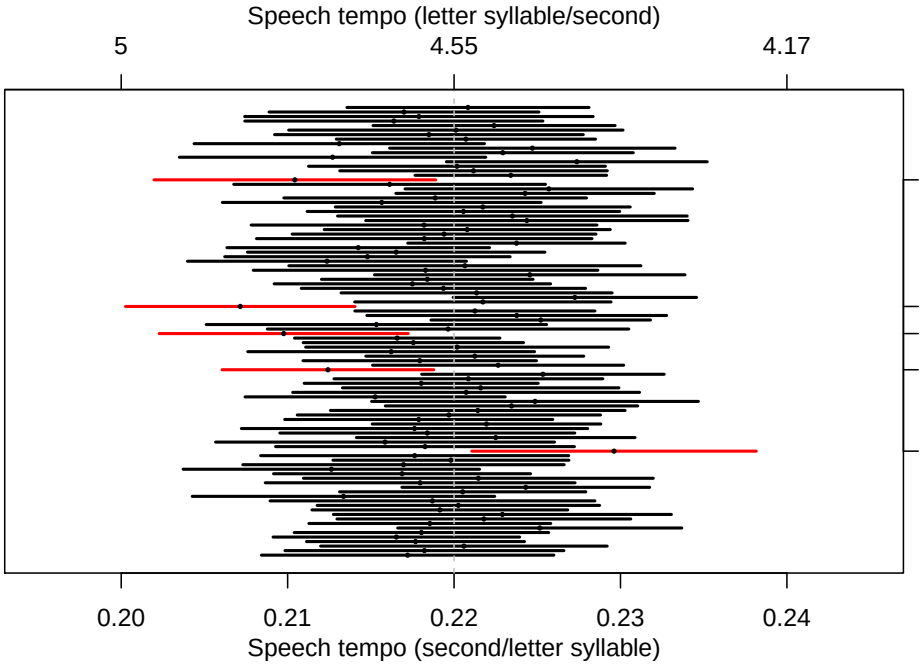


Figure 10.7: 95% confidence intervals and sample means, over 100 simulated samples (n=30) out of a population with mean 0.220 and s.d. 0.0225; see text.

Chapter 11

Correlation and regression

11.1 Introduction

Most empirical research is focused on establishing associations between variables. In experimental research, this primarily concerns associations between independent and dependent variables. In the coming section, we will look in more detail at the distinct ways of establishing whether a "significant" (meaningful, non-accidental) relation exists between the independent and dependent variables. In addition, the researcher might be interested in the associations between several dependent variables, for example the associations between the judgements of several raters or judges (see also Chapter 12).

In quasi-experimental research, the difference between independent and dependent variables is usually less clear. Several variables are observed and the researcher is particularly interested in the associations between the observed variables. What, for instance, is the association between the scores for reading, arithmetic, and geography in the CITO study (see Table 9.1)? In this chapter, we will look in more detail into the ways of expressing the association in a number: a correlation coefficient. There are different correlation coefficients depending on the variable's levels of measurement, which we will examine more in this chapter.

It is advisable to always first make a graphic representation of an association between the variables, in the form of a so-called *scatter plot*, like in Figure 11.1. Each point in this scatter plot corresponds with a pupil (or more generally, with a unit from a sample). The position of each point (pupil) is determined by the observed values of two variables (here X is the score for the reading test, Y is the score for the arithmetic test). A scatter plot like this helps us to interpret a potential correlation, and to inspect whether the observations indeed satisfy the preconditions for calculating a correlation from the observations. In any case, look at (a) the presence of a potential correlation, (b) the form of that

correlation (linear, exponential,...), (c) potential outliers (extreme observations, see §9.4.2), and (d) the distribution of the two variables, see §9.7.

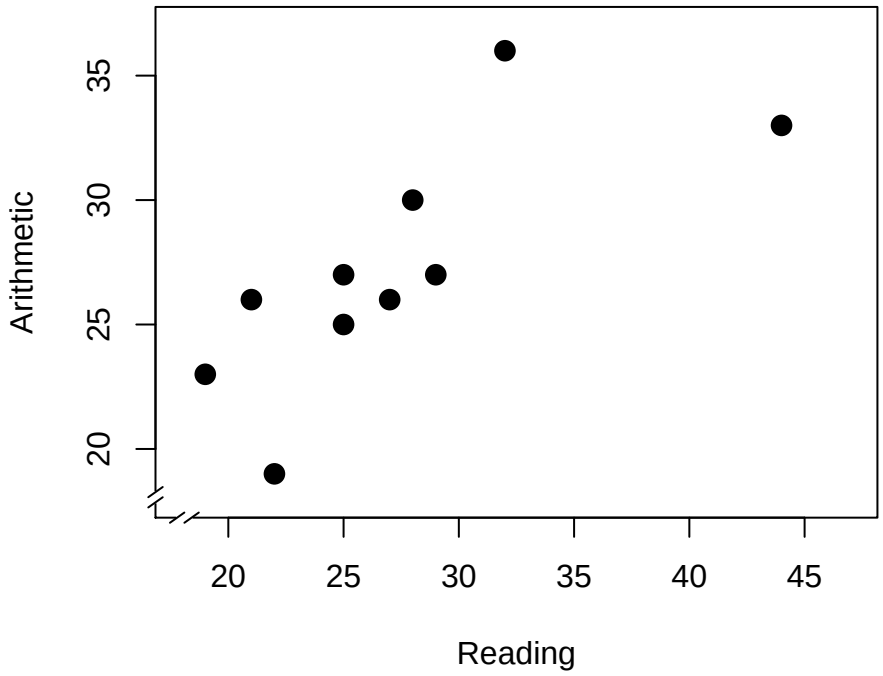


Figure 11.1: Scatter plot of the scores of a reading test and an arithmetic test; see text.

This scatter plot shows (a) that there is a relation between the scores for reading and arithmetic. The relation is (b) approximately linear, i.e. can be described as a straight line; we will return to this in §11.3. The relation also helps us to explain the dispersion in the two variables. After all, the dispersion in the arithmetic scores can be partially understood or explained from the dispersion in the reading test: pupils who achieve a relatively good score in reading, also achieve this in arithmetic. The observations from the two variables thus not only provide information about the two variables themselves, but moreover about the association between the variables. In this scatter plot, we can moreover see (c) that the highest reading score is an outlier (see also Fig.9.3); such outliers can have a disproportionately large influence on the correlation found.

11.2 Pearson product-moment correlation

The Pearson product-moment correlation coefficient is referred to with the symbol r (in the case of two variables). This coefficient can be used if both variables are observed on the interval level of measurement (§4.4), and if both variables

are approximately normally distributed (§10.3). Nowadays, we do this calculation by computer.

For the observations in the scatter plot in Fig.11.1, we can find a correlation of $r = +.79$. The correlation coefficient is a number that is by definition between -1 and $+1$. A positive correlation coefficient indicates a positive relation: a larger value of X corresponds with a larger value of Y. A negative correlation indicates a negative relation: a larger value of X corresponds with a smaller value of Y. A value of r which is close to zero indicates a weak or absent correlation: the dispersion in X is not related to the dispersion in Y; there is no or only a weak correlation. We call a correlation of $.4 < r < .6$ (or $-.6 < r < -.4$) moderate. A correlation of $r > .6$ (or $r < -.6$) indicates a strong association. If $r = 1$ (or $r = -1$), then all observations are precisely on a straight line. Figure 11.2 shows several scatter plots with the accompanying correlation coefficients.

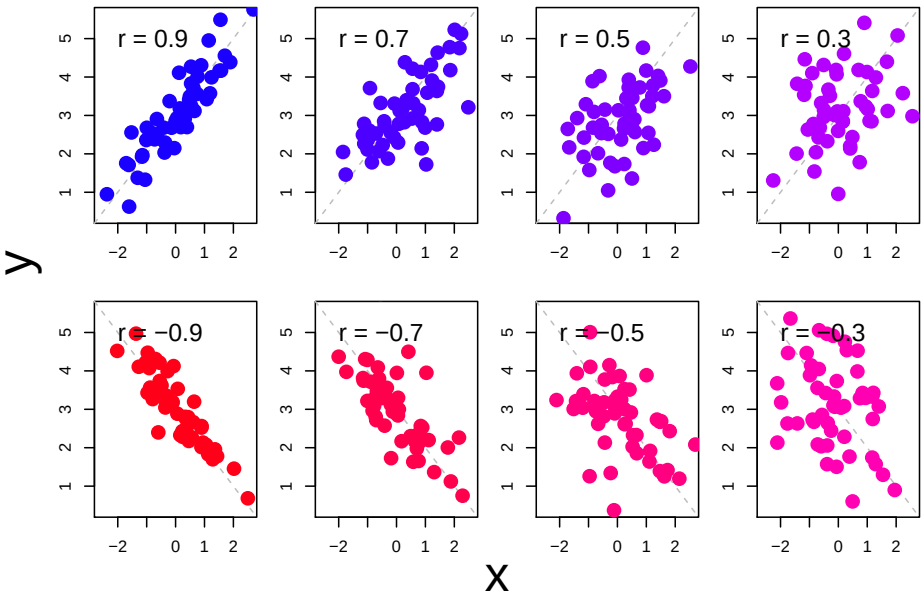


Figure 11.2: Some scatter plots of observations with accompanying correlation coefficients.

The correlation we see between the scores of the two variables (like $r = .79$ between scores for the reading test and arithmetic test, Fig.11.1) might also be the result of chance variations in the observations. After all, it is possible that the pupils who have a good score on the reading test achieve a good score on the arithmetic test purely by chance — also when there is actually *not* a correlation between the two variables in the population. We refer to the unknown correlation in the population with the Greek letter ρ ("rho"); as such, it is also possible that $\rho = 0$. Even if $\rho = 0$, it is possible to have $n = 10$ pupils in the sample who

by chance combine high scores on one part with high scores on the other part (and by chance not have pupils in the sample who combine high scores on one part with low scores on the other part). We can estimate what the probability p is of finding this correlation of $r = 0.79$ or stronger in a sample of $n = 10$ students, if the association in the population is actually nil (i.e. if $\rho = 0$). We call this probability p the *significance* of the correlation coefficient; in Chapter 13, we will look in more detail at this term 'significance'. In anticipation of this: if this probability p is smaller than .05, then we assume that the correlation found r is *not by chance*, i.e. is *significant*. We often see a small probability p with a strong correlation r. The correlation coefficient r indicates the direction and strength of the relation, and the significance p indicates the probability of finding this relation by chance if $\rho = 0$ in the population. We report these findings as follows[1]:

Example 11.1: The scores of the $n = 10$ pupils on the CITO test subparts in Table 9.1 show a strong correlation between the scores on the Reading and Arithmetic tests: Pearson $r = 0.79, p = .007$. Pupils with a relatively high score on one test generally also achieve a relatively high score on the other test.

In many studies, we are interested in the correlations between more than two variables. These correlations between variables are often reported in a so-called pairwise correlation matrix like Table 11.1, which is a table where the correlations of all pairs of correlations are reported.

Table 11.1: Correlations between the three parts of the CITO test, as summarised in Table 9.1, with the accompanying significance level between brackets.

	Reading	Arithmetic	Geography
Reading	1.00		
Arithmetic	0.79 (.007)	1.00	
Geography	-0.51 (.131)	-0.01 (.970)	1.00

In this matrix, only the lowest (left) half of the complete matrix is shown. This also suffices because the cells are mirrored along the diagonal: after all, the

[1] When the correlation found r is *not* significant, then this can thus be by chance, and then we discount an interpretation of the correlation. We do then state in our report the correlation coefficient found and the established significance for it.

correlation between Reading (column 1) and Arithmetic (row 2) is the same as the correlation between Arithmetic (row 2) and Reading (row 1). In the cells on the diagonal, the pairwise correlation matrix always contains the value 1.00, since a variable always correlates perfectly with itself. We report these findings as follows:

Examples 11.2: The pairwise correlations between scores from the $n = 10$ pupils on the three subparts of the CITO test are summarised in Table 11.1. We can see a strong correlation between the scores for the Reading and Arithmetic tests: pupils with a relatively high score on the Reading test generally also achieve a relatively high score on the Arithmetic test. The remaining correlations were not significant.

11.2.1 Formulas

The simplest formula for the Pearson product-moment correlation coefficient r makes use of the standard normal scores we already used earlier (§9.8):

$$r_{XY} = \frac{\sum z_X z_Y}{n-1} \tag{11.1}$$

Just like when we calculate variance (formula (9.3)), we divide again by $(n-1)$ to make an estimate of the association in the population.

11.2.2 SPSS

For Pearson's product-moment correlation coefficient:

`Analyze > Correlate > Bivariate...`

Choose `Pearsons` correlation coefficient, tick: `Flag significant correlations`. Confirm `OK`. The resulting output (table) does not satisfy the style requirements; as such, you should take the data into or convert it into a table of your own which does satisfy these requirements.

11.2.3 R

```
cito <- read.table(file="data/cito.txt", header=TRUE)
# variable names are Lezen=Reading, Rekenen=Arithmetic, WO=Geography, ...
dimnames(cito)[[2]] <- c( "Pupil", "Reading", "Arithmetic", "Geography",
                          "UrbanRural", "Arith.2cat" )
cor( cito[,2:4] ) # correlation matrix of columns 2,3,4
```

```
##               Reading Arithmetic   Geography
## Reading     1.0000000 0.74921033 -0.50881738
## Arithmetic  0.7492103 1.00000000  0.06351024
## Geography  -0.5088174 0.06351024  1.00000000
```

```
with( cito, cor.test( Reading, Arithmetic ) )
```

```
##
##  Pearson's product-moment correlation
##
## data:  Reading and Arithmetic
## t = 3.1994, df = 8, p-value = 0.01262
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2263659 0.9368863
## sample estimates:
##       cor
## 0.7492103
```

11.3 Regression

The simplest relation that we can distinguish and describe is a linear relation, i.e. a straight line in the scatter plot (see Fig.11.2). This straight line indicates which value of Y_i is predicted, on the basis of the value of X_i. This predicted value of Y_i is noted as $\widehat{Y_i}$ ("Y-hat"). The best prediction $\widehat{Y_i}$ is based on both the value of X_i and the linear relation between X and Y:

$$\widehat{Y_i} = a + bX_i \tag{11.2}$$

The straight line is described with two parameters, namely the intercept (or constant) a and the slope b [2]. The straight line which describes the linear relation is often referred to as the "regression line"; after all, we try to trace the observed values of Y back to this linear function of the values of X.

[2] In school books, this comparison is described as $Y = aX + b$, with a as the slope and b as the intercept; however, we keep to the conventional international notation here.

The difference between the observed value Y and the predicted value $\widehat{Y}$ $(Y - \widehat{Y})$ is called the *residual* (symbol e). In other words, the observed value is considered to be the sum of two components, namely the predicted value and the residual:

$$Y = \widehat{Y} + e \tag{11.3}$$
$$= a + bX + e \tag{11.4}$$

The above rationale is illustrated in the scatter plot in Figure 11.3. The dashed line indicates the linear relation between the two tests:

$$\widehat{\text{Arithmetic}} = 12.97 + 0.52 \times \text{Reading} \tag{11.5}$$

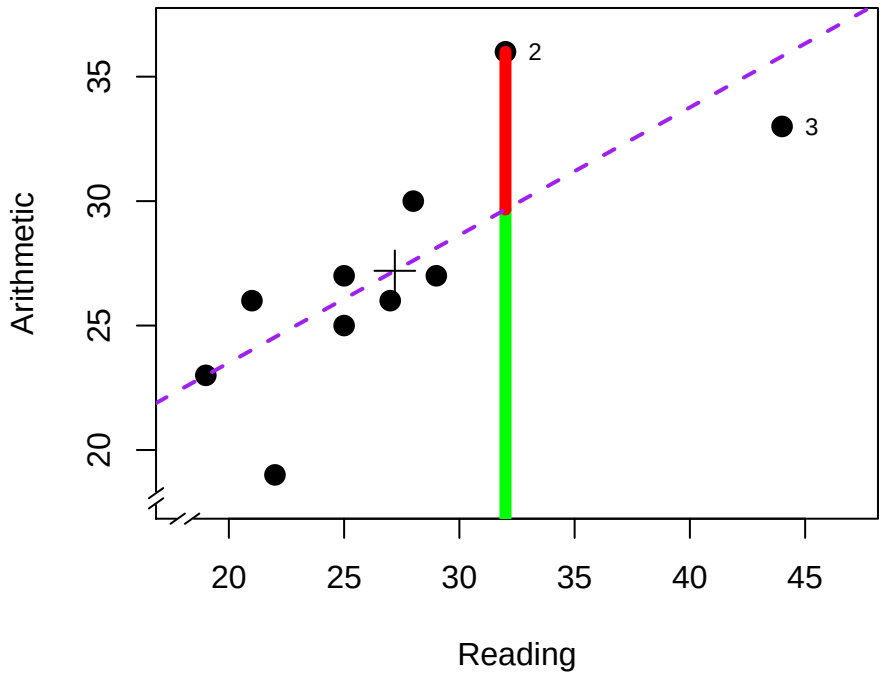


Figure 11.3: Scatter plot of the scores of a reading test and an arithmetic test. The diagram also indicates the regression line (dashed line), the predicted value (green) and residual (red) of the arithmetic test for pupil 2, the average (plus symbol), and markings for pupil 2 and 3; see text.

This dashed line thus indicates what the value $\widehat{Y}$ is for each value of X. For the second pupil with $X_2 = 32$, we thus predict $\widehat{Y_2} = 12.97 + (0.52)(32) = 29.61$ (predicted vowel, green line fragment). For all observations which are not precisely on the regression line (dashed line), there is a deviation between the predicted score $\widehat{Y}$ and the observed score Y (residual, red line fragment). For

the second pupil, this deviation is $e_2 = (Y_2 - \widehat{Y_2}) = (36 - 29.61) = 6.49$ (residual, red line fragment).

As stated, the observed values of Y are considered to be the sum of two components, the predicted value $\widehat{Y}$ (green) and the residual e (red). In the same way, the total *variance* of Y can be considered to be the sum of the two *variances* of these components:

$$s_Y^2 = s_{\widehat{Y}}^2 + s_e^2 \tag{11.6}$$

Of the total variance s_Y^2 of Y, one part ($s_{\widehat{Y}}^2$) can be traced back to and/or explained from the variance of X, via the linear relation described with parameters a and b (see formula (11.2)), and the other part (s_e^2) cannot be retraced or explained. The second part, the non-predicted variance of the residuals is also called the residual variance or unexplained variance.

When we are able to make a good prediction Y from X, i.e. when the Pearson product-moment correlation coefficient r is high (Fig. 11.2, left), then the residuals e are thus relatively small, the observations are close around the regression line in the scatter plot, and then the residual variance s_e^2 is also relatively small. Conversely, when we are *not* able to predict Y well from X, i.e. when the correlation coefficient is relatively low (Fig. 11.2, right), then the residuals e are thus relatively large, the observations are widely dispersed around the regression line in the scatter plot, and then the residual variance s_e^2 is thus also relatively large. The square of the Pearson product-moment correlation coefficient r indicates what the relative size of the two variance components is, with respect to the total variance:

$$r^2 = \frac{s_{\widehat{Y}}^2}{s_Y^2} \tag{11.7}$$

$$= 1 - \frac{s_e^2}{s_Y^2} \tag{11.8}$$

This statistic r^2 is referred to as the "proportion of explained variance" or as the "coefficient of determination".

The values of the linear parameters a and b in formula (11.2) are so chosen that the collective residuals are as small as possible, i.e. that the residual variance s_e^2 is as small as possible ("least squares fit"), and thus r^2 is as large as possible (see §11.3.1). In this way, we can find a straight line which best fits the observations for X and Y.

A linear regression can also be reported as follows:

Example 11.3: Based on a linear regression analysis, it appears that the score for Arithmetic is related to that for Reading: $b = 0.51, r = .79, p_r = .007$, over $n = 10$ pupils. This linear regression model

explains $r^2 = .51$ of the total variance in the arithmetic scores (the residual standard deviation is $s_e = \sqrt{82.803/(n-1-1)} = 3.217$).

11.3.1 Formulas

For linear regression of y on x, we try to estimate the coefficients a and b such that (the square of) the deviation between the predicted value $\hat{y}$ and the observed value y is as small as possible, in other words that the square of the residuals $(y-\hat{y})$ is as small as possible. This is called the "least squares" method (see http://www.itl.nist.gov/div898/handbook/pmd/section4/pmd431.htm).

The best estimation for b is

$$b = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

The best estimation for a is

$$a = \overline{y} - b\overline{x}$$

11.3.2 SPSS

For linear regression:

`Analyze > Regression > Linear...`

Choose `Dependent variable: Arithmetic` and choose `Independent variable: Reading`. Under the button `Statistics`, tick `Model fit`, tick `R squared change`, choose `Estimates`, and afterwards `Continue`.
Under the button `Plot`, tick `Histogram` and tick also `Normal probability plot`; these options are required to a get a numerical (!) summary over the residuals. (already check once) Under the button `Options`, choose `Include constant` to also have the constant coefficient a calculated. Confirm all choices with `OK`.

The resulting output includes several tables and figures; you cannot transfer these directly into your report. The table titled *Model Summary* contains the correlation coefficient, indicated here with capital letter $R = .749$.
The table titled *Coefficients* contains the regression coefficients. The line which has the designation `(Constant)` states coefficient $a = 13.25$; the line titled `Reading` states coefficient $b = 0.51$.
The table titled *Residual Statistics* provides information about both the predicted values and the residuals. Check whether the mean of the residual is indeed null. In this table, we can also see (line 2, column 4) that the standard variation of the residuals is 3.212.

11.3.3 R

```
summary( m1 <- lm( Arithmetic~Reading, data=cito ) )
```

```
##
## Call:
## lm(formula = Arithmetic ~ Reading, data = cito)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.5332 -1.1167 -0.5332  1.7168  6.3384
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  13.2507     4.4910   2.950   0.0184 *
## Reading       0.5128     0.1603   3.199   0.0126 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.406 on 8 degrees of freedom
## Multiple R-squared:  0.5613, Adjusted R-squared:  0.5065
## F-statistic: 10.24 on 1 and 8 DF,  p-value: 0.01262
```

The command `lm` specifies a linear regression model, with `Arithmetic` as the dependent variable and `Reading` as the predictor. This model is saved as an object called `m1`, and that is used again directly as an argument (input) for reporting. In the reporting of model `m1` the constant coefficient a is referred to as the `Intercept`.

```
sd( resid( m1 ) ) # s.d. of residuals according to `m1`
```

```
## [1] 3.211533
```

11.4 Influential observations

In the previous section, we saw that the aim in a correlation analysis or regression analysis is for a minimal residual variance. Earlier we also already saw that outliers or extreme observations, by definition, make a relatively large contribution to variance. Together, this means that outliers or extreme observations can have a large influence on the size of the correlation or on the regression found (the linear relation found). Pupil 3 has an extremely high score for Reading (see also Fig.9.3). If we discount pupil 3, then this would not greatly change

the correlation ($r_{-3} = .79$) but it would change the slope of the regression line ($b = 0.84$, more than one and a half times as large as if pupil 3 were in fact included). This observation thus "pulls" hard on the regression line, precisely because this observation has an extreme value for X and therefore has much influence.

Non-extreme observations can, however, also have a large influence on the correlation and regression, if they are far away from the regression line and thus make a large contribution to the residual variance. This too can be seen in Fig.11.3. Pupil 2 has no extreme scores but does have the largest residual. If we discounted this pupil 2 then the correlation would be considerably higher ($r_{-2} = .86$) but the slope of the regression line would only change a little ($b = 0.45$).

As such, for a correlation analysis or regression analysis you always have to make and study a scatter plot in order to see to what extent the results have been influenced by one or a small number of observations. When doing so, pay particular attention to observations which are far away from the mean, to each of the two variables, and to observations which are far away from the regression line.

11.5 Spearman's rank correlation coefficient

The variables whose correlation we want to investigate are not always both expressed on the interval level of measurement (§4.4), regardless whether or not the researchers want to and are able to assume that both variables are approximately normally distributed (§10.3). In that case, the product-moment correlation is less suitable for quantifying the association. If the data is indeed expressed on an ordinal level of measurement, then we can use other correlation coefficients to express the association: Spearman's rank correlation coefficient (r_s) and Kendall's τ (the Greek letter "tau"). Both of these coefficients are based on the ranking of the observations; we can thus always compute these correlations when we are able to order the observations. Nowadays, we also perform this calculation by computer. In this chapter, we only discuss the Spearman's rank correlation coefficient.

The Spearman's rank correlation coefficient is equal to the Pearson product-moment correlation coefficient applied to the ranks of the observations. We convert every observation from a variable to a rank number, from the smallest observed value (rank 1) to the largest observed value (rank n). If two or more observations have the same value, then they also receive the same (mean) rank. In Table 11.2, you can see the ranks of the scores for Reading and Arithmetic, ordered here according to the ranks for Reading.

Table 11.2: Ranks for the scores of 10 pupils on parts of a test, as summarised in Table 9.1, with difference v_i between the two ranks per pupil.

Pupil	1	9	6	4	10	8	5	7	2	3
Reading	1	2	3	4.5	4.5	6	7	8	9	10
Arithmetic	2	4	1	4	6.5	4	8	6.5	10	9
Difference v_i	-1	-2	1	0.5	-2	2	-1	1.5	-1	1

The ranking in Table 11.2 makes it clear at a glance that the three pupils with the lowest score for Reading (nos. 1, 9, 6) also almost achieved the lowest scores for Arithmetic. That indicates a positive relation. The two best readers (nos. 2 and 3) are also the two best arithmeticians. That also indicates a positive relation. However, there is also no question of a perfect positive relation (thus here $r_s < 1$), because then the two rankings would match perfectly.

Think how Table 11.2 would look if there were a perfect negative relation ($r_s = -1$) between the scores for Reading and Arithmetic, and how the table would look if there were no correlation whatsoever ($r_s = 0$) between these scores.

11.5.1 Formulas

The association between the rankings of the two variables is expressed in the Spearman's rank correlation coefficient:

$$r_s = 1 - \frac{6 \sum v_i^2}{n(n^2 - 1)} \tag{11.9}$$

in which v_i stands for the difference in rankings on both variables for respondent i. The fraction in this formula gets larger and r_s thus gets smaller, the larger the differences between the ranks are. However, this formula can only be used if there are no "ties" (shared rankings) in the variables; for the dataset in Table 11.2 with "ties" in both variables we have to use another formula.

As can be seen, the Spearman's rank correlation r_s is not equal to the Pearson product-moment correlation r for the scores observed. If the preconditions of the Pearson coefficient are satisfied, then this Pearson product-moment correlation coefficient provides a better estimation of the association than the Spearman's rank correlation coefficient. However, if the preconditions are *not* satisfied, then the Spearman's coefficient should be preferred again. The Spearman's coefficient is, amongst others, less sensitive for influential extreme observations — after all, such outliers have less weighting once the raw scores have been replaced by the ranks.

11.5.2 SPSS

For Spearman's rank correlation coefficient:

`Analyze > Correlate > Bivariate...`

Choose `Spearman` rank correlation coefficient, tick: `Flag significant correlations`. Confirm with `OK`. The resulting output (table) does not satisfy the style requirements; you thus have to take the data into or convert it into a table of your own which does satisfy these requirements, and report according to the usual conventions for correlation analysis.

11.5.3 R

```
with(cito, cor.test( Reading,Arithmetic, method="spearman" ) )
```

```
## Warning in cor.test.default(Reading, Arithmetic, method = "spearman"): Cannot
## compute exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  Reading and Arithmetic
## S = 25.229, p-value = 0.00198
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.8470988
```

11.6 Phi

The two variables for which we want to investigate the association are themselves not always expressed on an ordinal level of measurement (Chapter 4). Even if both of the variables are measured only on a nominal level of measurement, then a correlation can still be calculated, namely the phi correlation coefficient (symbol r_Φ, with Greek letter "Phi"). This correlation coefficient can also be used if one of the two variables is measured on a nominal level of measurement, and the other one is measured on an ordinal level of measurement.

With our CITO test example, let us assume that the first five pupils come from a large city (urban), and the last five from the countryside (rural). The pupil's

place of origin is a nominal variable, with 2 categories, here randomly referred to as `1` resp. `0` (see §4.2; a nominal variable with precisely 2 categories is also called a binomial or dichotomous variable). We now ask ourselves whether there is some association between a pupil's place of origin and their score for the Arithmetic part of the CITO test. The second variable is of interval level of measurement. We convert this to a nominal level of measurement. That can be done in many ways, and it is the researcher's role to make a wise choice when doing so. Here, we choose the often used 'mean split': one of the categories (low, code `0`) consists of scores smaller than or equal to the mean (§9.3.1), and the other category consists of scores larger than the mean (high, code `1`). We summarise the number of pupils in each of the 2×2 categories in a contingency table (Tabel 11.3).

Table 11.3: Contingency table of $n = 10$ pupils, subdivided according to origin (urban=1, rural=0) and according to category of score for the Arithmetic part of the CITO test ('mean split', low=0, high=1), with letter designations for the number of observations; see text.

Origin	Low (0)	High (1)	Total
Rural (0)	5 (A)	0 (B)	5 (A+B)
Urban (1)	2 (C)	3 (D)	5 (C+D)
Total	7 (A+C)	3 (B+D)	10 (A+B+C+D)

The nominal correlation coefficient r_Φ is equal to the Pearson product-moment correlation coefficient applied to the binomial codes (`0` and `1`) of the observations. All 5 pupils from the rural countryside have an Arithmetic score which is equal to or lower than average ($\overline{y} = 27.2$); out of the pupils from the urban city, 2 have a score which is (equal to or) lower than average. There is thus an association
between the binomial codes of the rows (origin) and those of the columns (score categories) in Table 11.3. This association is quantified in the correlation coefficient $r_\Phi = 0.65$ for this data.

11.6.1 Formulas

The nominal correlation coefficient r_Φ is calculated as follows, where the letters refer to the numbers in the cells of a contingency table (see Table 11.3):

$$r_\Phi = \frac{(AD - BC)}{\sqrt{(A+B)(C+D)(A+C)(B+D)}} \tag{11.10}$$

For the example discussed above we then find

$$r_\Phi = \frac{(15-0)}{\sqrt{(5)(5)(7)(3)}} = \frac{15}{22.91} = 0.65$$

11.6.2 SPSS

The dataset `cito` already contains the variable `UrbanRural` which indicates the origin of the pupils. However, for completeness, we will still show how you can construct a variable like this for yourself.

`Transform > Recode into different variables...`

Choose `Pupil` as the old variable and fill in as the new name for the new variable `UrbanRural2`. Indicate that the old values in `Range` from 1 to 5 (old) have to be transformed to the new value 1, and likewise that pupils 6 to 10 have to get the new value 0 for the new variable `UrbanRural2`.

For `Arithmetic` it is a bit more complex. You firstly have to delete your transformation rules (which relate to `UrbanRural`). Then, make a new variable again in the same way as before, named `Arithmetic2`. All values from the lowest value to the mean (27.2) are transformed to the new value 0 for this new variable. All values from the mean (27.2) to the highest value are transformed to the new value 1.

After this preparatory work, we can finally calculate r_Φ.

`Analyze > Descriptives > Crosstabs...`

Select the variables `UrbanRural2` (in the "Rows" panel) and `Arithmetic2` (in the "Columns" panel) for contingency table @ref(tab: cito-contingency-table).
Choose `Statistics…` and tick the option `Phi and Cramer's V`!
Confirm firstly with `Continue` and then again with `OK`.

11.6.3 R

The dataset `cito` already contains a variable `UrbanRural` which indicates the pupils' origins. However, for completeness, let us still see how you can construct such a variable for yourself.

```
UrbanRural2 <- ifelse( cito$Pupil<6, 1, 0) # 1=urban, 0=rural
Arithmetic2 <- ifelse( cito$Arithmetic>mean(cito$Arithmetic), 1, 0 ) # 1=high, 0=low
```

Here we build a new variable `Arithmetic2`, which has the value `1` if the score for Arithmetic is higher than the average, and otherwise has the value `0`.

In R, we also start by making a contingency table (Table 11.3 and we then calculate the r_Φ over the contingency table.

```
print( table(UrbanRural2,Arithmetic2) -> citocontingencytable )
```

```
##            Arithmetic2
## UrbanRural2 0 1
##           0 5 0
##           1 2 3
```

```
# make and store a contingency table
if (require(psych)) { # for psych::phi
  phi(citocontingencytable) # contingency table made earlier is input here!
}
```

```
## Loading required package: psych
```

```
##
## Attaching package: 'psych'
```

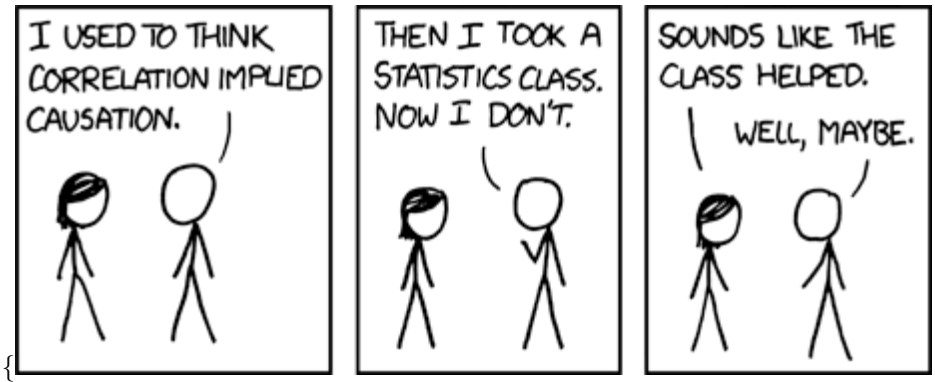
```
## The following objects are masked from '.hqenv':
##
##     harmonic.mean, logit
```

```
## [1] 0.65
```

11.7 Last but not least

At the end of this chapter, we want to emphasise again that an association or correlation between two variables does not necessarily mean that there is a causal relation between the variables, in other words a correlation does not mean that one variable (e.g. treatment) is the consequence of the other variable (e.g. cure). The common saying for this is "correlation does not imply causation," see also Example 6.1 (Chapter 6) and accompanying Figure 11.7.

\begin{figure}



{

}

\caption{*Correlation does not imply causation*, borrowed with permission from http://xkcd.com/552.} \end{figure}

Chapter 12

Reliability

12.1 Introduction

In Chapter 5, we talked about, amongst other matters, construct validity, the distance between the intended (theoretical) concept or construct on the one hand, and the independent or dependent variable on the other hand. In this Chapter, we will look at another very important aspect of the dependent variable, namely its *reliability*. This reliability can be estimated based on the association between observations of the same construct. We will also look at the relations between reliability and construct validity.

Often validity and reliability are mentioned in the same breath, and discussed in consecutive chapters. There is something to be said for this, since both concepts are about how you define and operationalise your variables. Nevertheless, we have chosen a different ordering here. Reliability will only be discussed following our discussion of correlation (Chapter 11), since reliability is based on the relation or correlation between observations.

12.2 What is reliability?

A reliable person is stable and predictable: what he or she does today is consistent with what he or she did last week, you can trust this person — in contrast to an unreliable person, who is unstable and behaves unpredictably. According to Collins English Dictionary, someone or something is reliable when they/it "…can be trusted to work well or to behave in the way that you want them to".[1] Reliable measurements can form the basis for a "justified true

[1] https://www.collinsdictionary.com/dictionary/english/reliable

belief" (see §2.4); conversely, it is not worth giving credence to unreliable measurements.

Measurements always show some degree of fluctuation or variation or inconsistency. This variability can partially be attributed to the variation in the behaviour which is being measured. After all, even if we measure the same construct for the same person, we still see variance as a result of the momentary mental or physical state of the participant, which simply fluctuates. Moreover, there is variation in the measuring device (thermometer, questionnaire, sensor), and there are probably inconsistencies in the manner of measurement or evaluation. With the quantification of such consistencies and inconsistencies, we enter the realms of reliability analysis.

The term 'reliability' has two meanings in academic research, which we will treat separately. Firstly, reliability signifies the *precision* or *accuracy* of a measurement. This aspect concerns the question of the extent to which the measurement is influenced by chance factors (through which the measurement does not exclusively render the construct investigated). If we do *not* measure accurately then we also know what the measurements actually show — perhaps they show the construct investigated but perhaps they also do not. If we do measure accurately then we would expect, if we were to conduct the same measurement again, that we would then measure the same outcome. The less precise a measurement is, the more variation or inconsistency there is between the first measurement and the repeated measurement, and the measurements are thus less reliable.

Example 12.1: If we want to measure the reading ability of pupils in their final examinations, then we present them with a reading comprehension test with a number of accompanying questions. The degree to which the different questions measure *the same* construct, here the construct 'reading ability', is called reliability, precision or homogeneity.

In what follows, to avoid confusion, we will refer to this form of reliability with the term *homogeneity* (vs. heterogeneity). With a heterogenous (non-homogenous) test, the total score is difficult to interpret. With a perfectly homogenous test, people who have the same total score have also answered the same questions correctly. However, when we measure human (language) behaviour, such perfectly homogenous tests never occur: respondents who do achieve the same total score, have not always answered the same questions correctly (e.g. in the final examination reading ability test, Example 12.1). This implies that the questions have not measured exactly the

same capacity. This is also the case: one question was about paraphrasing a paragraph, whilst another was related to a relationship between a referential expression and its antecedent.
As such, the questions or items were not perfectly homogenous!

Secondly, reliability signifies a measurement's *stability*. To measure your weight, you stand on a weighing scale. This measurement is stable: five minutes later, the same weighing scale with the same person under the same circumstances will also yield (almost) the same measurement. Stability is often expressed in a so-called correlation coefficient (a measurement for association, see Chapter 11). This correlation coefficient can assume all values between $+1$ and -1. The more similar the first and second measurement, the higher the correlation is, and the higher the association between the first and second measurement. Conversely, the lower the association between the first and second measurement is, the lower too the correlation is.

Stable measurements nevertheless rarely occur in research on (language) behaviour. If is a test is taken twice, then there is often a considerable difference in scores on the first measurement point and scores on the second measurement point.

Example 12.2: In the final examination for Dutch secondary school, pupils typically have to write an essay, which is assessed by two raters. The raters are stable if, after some time, they give the same judgements to the same essays. Thus: if rater A at first gave a grade 8 to an essay, and for the second evaluation sometime later, he/she also gave the same essay an 8, then this rater is (very) stable. If, however, the same rater A gave this same essay a grade 4 on the second evaluation, then this rater is not stable in his/her judgements.

Now, grading essays is a tricky task: criteria are not precisely described and there is a relatively large amount of room for interpretation differences. Accordingly, the stability of judgements is also low; previously, a stability coefficient of even 0.40 has been reported.

To calculate a test's stability, the same test has to be taken twice; the degree of association between the first and second measurement is called the *test-retest-reliability*. In practice, repeatedly sitting a test like this rarely takes place due to the relatively high costs and relatively low benefits.

Example 12.3: Lata-Caneda et al. (2009) developed a Spanish-language questionnaire consisting of 39 questions, intended for aphasia patients to determine their quality of life. The quality of life is described as "the patient perception about, either the effects of a given disease, or the application of a certain treatment on different aspects of life, especially regarding its consequences on the physical, emotional and social welfare" (Lata-Caneda et al., 2009, p.379). The new questionnaire was taken twice with a sample of 23 Spanish-language patients with aphasia as a result of cerebral haemorrhage. The reported test-retest stability for this questionnaire was 0.95.

Both homogeneity and stability are expressed as a coefficient with a value between 0 and 1 (in practice, negative coefficients do not occur). How should we interpret the reported coefficients? Generally, it is of course the case that the higher the coefficient is, the higher (better) the reliability. But how large should the reliability minimally be before we can call a test "reliable"? There are no clear rules for this. However, when considerations have to be made about people, then the test has to have a reliability of at least 0.90 according to the *Nederlands Instituut van Psychologen* 'Dutch Institute for Psychologists' (NIP). This is, for instance, the case for tests which are used to determine whether or not a child is eligible for a so-called dyslexia declaration. For research purposes, such a strict requirement for the reliability of a test is not required. Often, 0.70 is used as the lower limit of the reliability coefficient.

12.3 Test theory

Classical test theory refers to the measurement of variable x for the i-th element of a sample consisting of random members of the population. Test theory posits that each measurement x_i is composed of two components, namely a true score t_i and an error score:

$$x_i = t_i + e_i \tag{12.1}$$

Imagine that you "actually" weigh $t = 72.0\,\text{kg}$, and imagine also that your measured weight is $x = 71.6\,\text{kg}$, then the error score is $e = -0.4\,\text{kg}$.

A first important assumption in classical test theory is that the deviations e_i neutralise or cancel each other out (i.e. are zero when averaged out, and thus do not deviate systematically from the true score t), and that larger deviations

above or below occur less often than smaller deviations. This means that the deviations are normally distributed (see §10.3), with $\mu_e = 0$ as mean:

$$e_i \sim \mathcal{N}(0, s_e^2) \tag{12.2}$$

A second important assumption in classical test theory is that there is no relation between the true scores t_i and the error scores e_i. Since the component e_i is completely determined by chance, and thus does not have any relation with x_i, the correlation between the true score and the error score is null:

$$r_{(t,e)} = 0 \tag{12.3}$$

The total variance of x is thus[2] equal to the sum of the variance of the true scores and the variance of the error scores:

$$s_x^2 = s_t^2 + s_e^2 \tag{12.4}$$

When the observed variance s_x^2 proportionately contains much error variance (i.e. variance of deviations), then the observed scores have been determined for the most part by chance deviations. That is of course undesirable. In a such instance, we say that the observed scores are not reliable; there is much "noise" in the observed scores.
When the error variance in contrast is relatively small, then the observed scores provide a good reflection of what the true scores are, and then the observed differences are indeed reliable, i.e. they are not much determined by chance differences.

In that case, we can also define reliability (symbol ρ) as the proportion between true score variance and total variance:

$$\rho_{xx} = \frac{s_t^2}{s_x^2} \tag{12.5}$$

However, in practice, we cannot use this formula (12.5) to establish reliability, since we do not know s_t^2. We must thus firstly estimate what the true score variance is — or what the error variance is, which, after all, is the complement of the true score variance (see formula (12.4))[3].

The second assumption (in formula (12.3)), that there is no relation between true score and error score, is, in practice, not always justified. To illustrate, let us look at the results of a test on
a scale from 1.0 to 10.0. Students with scores of 9 or 10 have a high true score too (they master the material very well) and thus usually have a low error

[2] $s_{(t+e)}^2 = s_t^2 + s_e^2 + 2r_{(t,e)}s_t s_e$, with here $r_{(t,e)} = 0$ according to the formula(12.3).

[3] An exception to this is a situation in which $s_x^2 = 0$, and thus $s_t^2 = 0$, thus reliability $\rho = 0$; the dependent variable x has then not been operationalised well.

score. The students with scores of 1 or 2 also have a low true score (they master the material very badly) and thus also usually have a low error score. For the students with scores of 5 or 6, the situation is different: perhaps they master the material fairly well but have just given a wrong answer, or perhaps they master the material poorly but have by chance given a good answer. For these students with an observed score in the middle of the scale, the error scores are relatively larger than for the students with a score at the ends of the scale. In other domains, e.g. for reaction times, we see other relations, e.g. that the error score increases more or less equally with the score itself; there is then a positive relationship between the true score and the error score ($\rho_{(t,e)} > 0$). Nevertheless, the advantages of the classical test theory are so large that we use this theory as a starting point.

From the formulas (12.4) and (12.5) above, it also follows that the standard error of measurement is related to the standard deviation and to the reliability:

$$s_e = s_x\sqrt{1 - r_{xx}} \tag{12.6}$$

This standard error measurement can be understood as the standard deviation of the error scores e_i, assuming still that the error scores are normally distributed (formula (12.2)).

Example 12.4: External inspectors doubt whether teachers mark their students' final papers well. If a student got a 6, should the final paper have perhaps actually been judged as a fail?

Let us assume that the given assessment shows a standard deviation of $s_x = 0.75$, and let us equally assume that an analysis of reliability had shown that $r_{xx} = 0.9$. The standard error measurement is then $s_e = 0.24$ points (rounded up). The probability that the true score t_i is smaller than or equal to 5.4 (the minimum for a fail), with an observed score of $x_i = 6.0$ and $s_e = 0.24$, is only $p = 0.006$ (for explanation, see §13.5 below). The final paper's assessment as a pass is with high probability correct.

12.4 Interpretations

Before we look at the different ways of calculating reliability, it is a good idea to pause on the different interpretations of reliability estimations.

First, reliability can be interpreted as the proportion of true score variance (see formula (12.5)), or as the proportion of variance which is "systematic". This is different from the proportion of variance resulting from the concept-as-defined, the "valid" variance (see Chapter 5). The variance resulting from the concept-as-defined is part of the proportion of true score variance. However, many other factors may systematically influence respondents' scores, such as differences in test experience. If two students i and j possess a concept (let us say: language proficiency) to the same degree, then one of the students can still score more highly because he or she has done (and practiced) language proficiency tests more often than the other student. Then, there is no difference in the concept-as-defined (language proficiency $T_i = T_j$), but there is in another factor (experience), and thus a difference arises between the students in their 'true' scores ($t_i \neq t_j$) which we measure with a valid and reliable language proficiency measurement. When measuring, deviations and measurement errors appear (e_i and e_j), through which the observed differences between students ($x_i - x_j$) can be larger or smaller than their differences in 'true' score ($t_i - t_j$). This is the reason why a reliability estimate always forms the upper limit of the validity.

A second interpretation of reliability (formula (12.5)) is that of the theoretically expected correlation (see §11.2) between measurements, when measurements are replicated many times. For convenience, we assume that memory and fatigue effects have no effect at all on the second and later measurements. If we were to measure the same people with the same measurement three times, without memory or fatigue effects, then the scores from the first and second measurement, and from the first and third measurement, and from the second and third measurement would always show the same correlation ρ. This correlation thus indicates the extent to which the repeated measurements are consistent, i.e. represent the same unknown true score.

In this interpretation, reliability thus expresses the expected association between scores from the same test taken repeated times. We then interpret the reliability coefficient ρ as the correlation between two measurements with the same instrument.

Thirdly, reliability can be interpreted as the loss of efficiency in the estimation of the mean score $\overline{X}$ (Ferguson and Takane, 1989, p.474). Imagine that we want to establish the mean score of a group of $n = 50$ participants, and for this we use a measurement instrument with reliability $\rho_{xx} = 0.8$. In this case, there is uncertainty in the estimation, which come from the chance deviations e_i in the measurements. If the measurement instrument were perfectly reliable ($\rho = 1$), we would only need $\rho_{xx} \times n = 0.8 \times 50 = 40$ participants for the same accuracy in the estimation of $\overline{X}$ (Ferguson and Takane, 1989, p.474). As such, we have, as it were, played away 10 participants to compensate for the unreliability of the measurement instrument.

Above, we spoke about measurements with the help of measurement

instruments, and below we will talk about ratings done by raters. In these situations, the approach to the notion of 'reliability' is always the same. Reliability plays a role in all situations where elements from a sample are measured or assessed by multiple assessors or instruments. Non-final exams and questionnaires can also be such measurement instruments: a non-final exam or questionnaire can be thought of as a composite instrument with which we try to measure an abstract property or condition of the participants. Each question can then be considered as a "measurement instrument" or "assessor" of the respondent's property or condition. For this, all of the above mentioned insights and interpretations concerning test theory, measurement error and reliability are equally applicable.

12.5 Methods for estimating reliability

A measurement's reliability can be determined in different ways. The most important are:

- The *test-retest method*
  We conduct all measurements twice, and then calculate the correlation between the first and the second measurement. The fewer measurement errors and deviations the measurements contain, the higher the correlation and thus also the reliability is. This method is time consuming but can also be applied to a small portion of the measurements. In speech research, this method is indeed used to establish the reliability of phonetic transcriptions: part of the speech recordings are transcribed by a second assessor, and then both transcriptions are compared.

- The *parallel forms method*
  We have a large collection of measurements which are readily comparable and measure the same construct. We conduct all measurements repeatedly, the first time by combining the measurements of several measurement instruments chosen at random (let's say A and B and C) and the second time by using other random instruments (let's say D and E and F). Since the measurement instruments are 'parallel' and the same measurement is considered to be measured, the correlation between the first and the second measurement is an indication of the measurement's reliability.

- The *split-half method*
  This method is similar to the parallel forms method. The k questions or instruments are divided into two halves, after which the score is determined within each half. From the correlation r_{hh} between the scores from the two half tests, the reliability of the whole test can be deduced, $r_{xx} = \frac{2r_{hh}}{1+r_{hh}}$.

12.6 Reliability between assessors

As an example, let us look at language proficiency measurements from students in a foreign language. This construct 'language proficiency' is measured in this example by means of two assessors who each, independently of the other, award a grade between 1 and 100 to the student (higher is better). However, when assessing, measurement errors also arise, through which the judgements not only reflect the underlying true score but also a deviation if it, with all the aforementioned assumptions. Let us firstly look at the judgements by the first and second rater (see Table 12.1). For the time being, the final judgement of a student is the mean of the judgements from the first and second rater.

Table 12.1: Judgements about language proficiency from $n = 10$ students (rows) by $k = 3$ raters (columns).

Student	B1	B2	B3
1	67	64	70
2	72	75	74
3	83	89	73
4	68	72	61
5	75	72	77
6	73	73	78
7	73	76	72
8	65	73	72
9	64	68	71
10	70	82	69
$\overline{x_i}$	71.0	74.4	71.7
s_i	5.6	7.0	4.7

The judgements of only the first and the second assessor show a correlation of $r_{1,2} = .75$. This means (according to the formula (12.5)) that 75% of the total variance in the judgements of these two raters can be attributed to differences between the students rated, and thus 25% of measurement errors (after all, we have assumed that there are no systematic differences between raters). The proportion of measurement errors appears to be quite high. However, we can draw hope from one of our earlier observations, namely that the raters' measurement errors are not correlated. The *combination* of these two raters — the mean score per student over the two raters — thus provides more reliable measurements than each of the two raters can do separately. After all, the measurement errors of the two raters tend to cancel each other out (see formula (12.2)). Reread the last two sentences carefully.

Reliability is often expressed as *Cronbach's Alpha* (Cortina, 1993). This number is a measure for the consistency or homogeneity of the measurements, and thus also indicates the degree to which the two raters have rated the same

construct. The simplest definition is based on $\overline{r}$, the mean correlation between measurements of k different raters[4].

$$\alpha = \frac{k\overline{r}}{1 + (k-1)\overline{r}} \tag{12.7}$$

Filling in $k = 2$ raters and $\overline{r} = 0.75$ provides $\alpha = 0.86$ (SPSS and R use a somewhat more complex formula for this, and report $\alpha = 0.84$). This measurement for reliability is not only referred to as Cronbach's Alpha but also as the Spearman-Brown formula or the Kuder-Richardson formula 20 (KR=20)[5].

The value of Cronbach's Alpha found is a bit tricky to evaluate since it is also dependent on the number of instruments or raters or questions in the test (Cortina, 1993; Gliner et al., 2001). For academic research, a lower limit of 0.75 or 0.80 is often used. If the result of the test or measurement is of great importance to the person concerned, as in the case of medical or psychological patient diagnosis, or when recruiting and selecting personnel, then an even higher reliability of $\alpha = .9$ is recommended (Gliner et al., 2001).

If we want to increase reliability to $\alpha = 0.9$ or higher, then we can achieve that in two ways. The first way is to expand the number of raters. If we combine more raters in the total score, then the measurement errors of these raters also better cancel out each other, and then the total score is thus more reliable. Using the formula (12.7), we can investigate how many raters are needed to improve the reliability to $\alpha = 0.90$ or better. We fill in $\alpha = 0.90$ and again $\overline{r} = 0.75$, and then find an outcome of minimally $k = 3$ raters. The *increase* in reliability levels off, the more raters there are already participating: if $k = 2$ then $\alpha = .84$, if $k = 3$ then $\alpha = .84 + .06 = .90$, if $k = 6$ then $\alpha = .90 + .05 = .95$, if $k = 9$ then $\alpha = .95 + .01 = .96$, etc. After all, if there are already 6 raters who are already readily cancelling out each other's measurement errors, then 3 extra assessors add little to the reliability.

The second way of increasing reliability is by reducing the measurement error. We can try to do this, for example, by instructing the raters as well as possible about how they should rate the students' language proficiency. An assessment protocol and/or instruction can make the deviations between and within raters smaller. Smaller deviations mean smaller measurement errors, and that again means higher correlations between the raters. For an $\overline{r} = 0.8$, we already almost achieve the desired reliability, with only $k = 2$ raters.

A third way of increasing reliability requires a closer analysis of the separate raters. To explain this, we also now involve the third rater in our considerations (see Table 12.1). However, the judgements of the third rater

[4] In our example, there are only $k = 2$ raters, thus there is only one correlation, and $\overline{r} = r_{1,2} = 0.75$.

[5] The so-called 'intra-class correlation coefficient' (ICC) for k is likewise identical to the Cronbach's Alpha.

show low correlations with those of the first and second assessor: $r_{1,3} = 0.41$ and $r_{2,3} = 0.09$. As a consequence, the mean correlation between assessors is now lower, $\overline{r} = 0.42$. As a result of taking this third rater, the reliability has not risen but instead actually lowered to $\alpha = \frac{3 \times 0.42}{1+2 \times 0.42} = 0.68$. We can thus perhaps better ignore the measurements of the third rater. Also if we investigate the reliability of a non-final exam or test or questionnaire it can seem that the reliability of a whole test *increases* if some "bad" questions are removed. Apparently, these "bad" questions measured a construct which differed from what the remaining questions measured.

12.7 Reliability and construct validity

When a measurement is reliable, then "something" has been measured reliably. But this still does not show *what* has been measured! There is a relation between reliability (how measurements are made) and construct validity (what is measured, see Chapter 5), but these two terms are not identical. Sufficient reliability is a requirement for validity, but is not a sufficient condition for it. Put otherwise: a test which is not reliable can also not be valid (since this test also measures noise),
but a test which is reliable does not have to be valid. Perhaps, the test used does measure another construct other than what was intended very reliably.

An instrument is construct valid if the concept measured matches the intended concept or construct. In Example 12.3, the questionnaire is valid if the score from the questionnaire matches the quality of life (whatever that actually is) of the aphasia patients. Only once it has been shown that an instrument is reliable, is it meaningful to speak about a measurement's construct validity. Reliability is a necessary but not a sufficient condition for construct validity. An unreliable instrument can thus not be valid but a reliable instrument does not necessarily have to be valid.

To measure reading proficiency, we get the pupils to write an essay. We count the number of letters *e* in each essay. This is a very reliable measurement: different raters arrive at the same number of *e*'s (raters are homogenous) and the same rater always also delivers the same outcome (raters are stable). The great objection here is that the number of *e*'s in an essay does not or does not necessarily match the concept of reading proficiency. A pupil who has incorporated more *e*'s into his/her essay is not necessarily a better writer.

Whilst researchers know that reliability is a necessary but not sufficient condition for validity, they do not always use these terms carefully. In many studies, it is tacitly assumed that if the reliability is sufficient, the validity is also then guaranteed. In Example 12.3 too, the difference is not made clear and the researchers do not discuss the construct validity of their new questionnaire explicitly.

12.8 SPSS

For a reliability analysis of the $k = 3$ judgements over language proficiency in Table 12.1:

`Analyze > Scale > Reliability Analysis...`

Select the variables which are considered to measure the same construct; here that is three raters. We look at these $k = 3$ assessors as "items" who measure the property "language proficiency" of 10 students. Drag these variables to the Variable(s) panel.
As Scale label, fill in an indication of the construct, e.g. `language proficiency`.
As a method, choose `Alpha` for Cronbach's Alpha (see formula (12.7))
Choose `Statistics...`, tick: Descriptives for `Item, Scale, Scale if item deleted`, Inter-Item `Correlations`, Summaries `Means, Variances`, and confirm with `Continue` and again with `OK`.

The output includes Cronbach's Alpha, the desired inter-item correlations (particularly high between raters 1 and 2), and (in Table Item-Total Statistics) the reliability if we remove a certain rater. This last output teaches us that raters 1 and 2 are more important than rater 3. If we were to replace raters 1 or 2, then the reliability would collapse but if we were to remove rater 3 then the reliability would even increase (from 0.68 to 0.84). Presumably, this rater has rated a different concept to the others.

12.9 R

For a reliability analysis of $k = 3$ language proficiency judgements in Table 12.1:

```
raters <- read.table(file="data/beoordelaars.txt", header=TRUE)
if (require(psych)) { # for psych::alpha
  alpha( raters[,2:4] ) # columns 2 to 4
}
```

```
## Number of categories should be increased  in order to count frequencies.

##
## Reliability analysis
## Call: alpha(x = raters[, 2:4])
##
##   raw_alpha std.alpha G6(smc) average_r S/N  ase mean  sd median_r
```

```
##       0.68      0.68    0.74      0.41 2.1 0.17   72 4.6     0.41
##
##  lower alpha upper     95% confidence boundaries
## 0.35 0.68 1.01
##
##  Reliability if an item is dropped:
##    raw_alpha std.alpha G6(smc) average_r  S/N alpha se var.r med.r
## B1      0.15      0.16   0.088     0.088 0.19    0.497    NA 0.088
## B2      0.58      0.58   0.410     0.410 1.39    0.264    NA 0.410
## B3      0.84      0.85   0.745     0.745 5.84    0.095    NA 0.745
##
##  Item statistics
##     n raw.r std.r r.cor r.drop mean  sd
## B1 10  0.93  0.92  0.91   0.81   71 5.6
## B2 10  0.84  0.78  0.72   0.53   74 7.0
## B3 10  0.56  0.64  0.38   0.25   72 4.7
```

This output includes Cronbach's Alpha (`raw_alpha 0.68`), and the reliability if we were to remove a certain rater. If we were to replace rater 3, then the reliability would even increase (from 0.68 to 0.84). Over all three raters, `average_r=0.41`.

Correlations between k raters or items are not explicitly provided (even if they can be deduced from the above output), thus we still request these:

```
cor( raters[ ,c("B1","B2","B3") ] ) # explicit column names
```

```
##           B1         B2         B3
## B1 1.0000000 0.74494845 0.40979738
## B2 0.7449484 1.00000000 0.08845909
## B3 0.4097974 0.08845909 1.00000000
```

Part III: Inferential statistics

Chapter 13

Testing hypotheses

13.1 Introduction

From this chapter onwards, we will be concerned with the testing of research hypotheses and, in particular, with null hypothesis significance testing (NHST), as explained in Chapter 2.

Over the course of the years, a large number of techniques have been developed for tests like this. The tests with which we will concern ourselves are the most used and can be divided into parametric and non-parametric tests. Parametric tests assume that the dependent variable is (at least) measured on an interval level of measurement (see Chapter 4), and that the measured outcomes or scores are normally distributed (see §10.3 and §10.5). For non-parametric tests, dependent on the technique, fewer assumptions are made over the level of measurement or over the distribution of the observed scores; these are so-called distribution free tests. The consequence is that the testing is a little less 'sensitive' under otherwise equal circumstances, i.e. that the null hypothesis can be rejected less often in otherwise equal circumstances. These tests therefore have less power (see Chapter 14). Researchers thus usually prefer parametric tests.

We already discussed the general principle of testing briefly in §2.4 and §2.5. We will illustrate this again here with an example. We investigate the statement H1:
'Linguistics students master traditional grammar *better* than the average language student'. As a measurement instrument, we use the so-called "grammar test"[1] which is required for most students in language programs at Utrecht University. On the basis of previous year groups, we expect a mean score of 73 on this test; this is the mean number of good answers from 100

[1] We would like to thank Els Rose for making these data available.

questions. We thus operationalise this first as $\mu > 73$, and from this deduce the accompanying null hypothesis which is actually tested: $\mu = 73$. (In §13.4 below, we will go into more detail about whether or not to name the *direction* of the difference in H1).

For the first year Linguistics students ($n = 34$) from a certain year group, we find an average score of 84.4. That is indeed far above the reference value of 73 but that might also be a coincidence. Perhaps, H0 is true, and, wholly by chance, there are many grammar experts in our sample (from the population of possible first year students in Linguistics). We can calculate the probability P of the situation i.e. the probability P of finding a mean score of $\overline{x} = 84.4$, given a random sample of $n = 34$ people and given that H0 is in fact true (i.e. $\mu = 73$): then it appears that $P = .000000001913$. This probability P represents the probability of finding this data, whilst H0 is true: $P(\overline{x} = 84.4|\text{H0}, n = 34)$. In this case, the probability P is very small.

For the argumentation, it is essential that the data is valid and reliable — this is precisely the reason why we discussed validity (Chapter 5) and reliability (Chapter 12). If we have done everything properly, we can, after all, trust the data obtained. We are then *not* reasonably able to attribute the low probability of the data according to H0 to errors in operationalisation, or measurement errors, or other deviations in the data. The logical conclusion then is that the improbable outcome shows that the premise (H0) is probably *not* true: we reject H0; H0 has thus been falsified. Thereby, our knowledge has increased because we can now assume on legitimate grounds that H0 is untrue (and thus that H1 is true).

If we reject H0 on the basis of the reasoning above which in turn is based on probability, then we do have to take into account the small probability P that rejecting H0 is an unjustified decision (Type I error; see §2.5). After all, there is the probability P that we find these data when H0 is in fact true (in this example: when the linguists on average do not score differently than $\mu = 73$).

Figure 13.1 shows the probability of the sample mean ($n = 34$) if H0 is true. We see that the value 73 can have the highest probability, but also 72 or 74 are probable mean scores according to H0. However, a mean of 84.4 is very improbable, the probability P of this mean score (height of the curve) is almost null according to H0.

The boundary value for P, at which we reject H0 is called the significance level, often referred to with the symbol α (see §2.5). Researchers often use $\alpha = .05$, but sometimes other boundary values are also used. In Figure 13.1, you see that the probability of a mean score of 77.7 or more has a probability of $P = .05$ or smaller, according to H0. This can be seen from the area under the curve. The coloured part has precisely an area of 0.05 of the total area under the curve.

The decision about whether or not to reject H0 is based on the probability P of the outcomes, given H0. The decision might also be incorrect. The finding

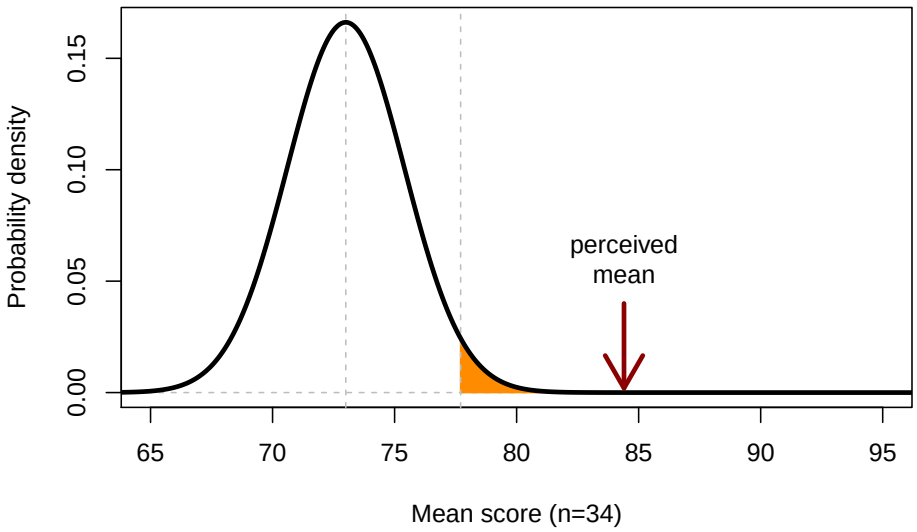


Figure 13.1: Probability distribution of the mean score from a sample (n=34) with a population mean 73 and population s.d. 14. The coloured area covers 5% of the total area under the curve; outcomes along the X-axis of this area thus have a probability of at most 5% of occurring if H0 is true.

that $P < \alpha$ is thus not an *irrefutable* proof that H0 is untrue (and *has to* be rejected); it is also true that H0 is in fact true but that the effect found was a fluke (Type 1 error). Conversely, the finding that $P > \alpha$ is not conclusive evidence that H0 is true. There can be all kinds of other, plausible reasons why an effect which exists (H0 is untrue) can still not be observed. If I do not hear any birds singing, that does not necessarily mean that there are genuinely no birds singing. More generally: "absence of evidence is not evidence of absence" (;). It is thus good to always report the size of the effect found (this is explained in more detail in §@ref(#sec:ttest-effectsize) below).

Example 13.1: Assume H0: 'birds do not sing'. Write down at least 4 reasons why I do not hear birds singing, even if there are in fact birds singing (H0 is untrue). If I do not reject H0, what type of error will I be making?

13.2 One-sample t-test

The Student's t-test is used to investigate a difference between the mean score of a sample, and an a priori assumed value of the mean. We use this test when the standard deviation σ in the population is unknown, and thus has to be estimated from the sample. The line of thought is as follows.

We determine the test statistic t on the basis of the mean and the standard deviation in the sample, and of the assumed mean (according to H0). If H0 is true, then the value $t = 0$ is the most
probable. The larger the difference between the observed sample mean and the assumed sample mean, the more t increases. If the test statistic t is larger than a certain boundary value $t*$, and thus $t > t*$, then the probability of this test statistic, if H0 is true, is very small: $P(t|\text{H0}) < \alpha$. The probability of finding this result if H0 is true is then so small that we decide to reject H0 (see §2.5). We speak then of a *significant* difference: the deviation between the observed and the expected mean is probably not a coincidence.

In the earlier example of the grammar test with Linguistics students (§13.1), we already became acquainted with this form of t-test. If $\overline{x} = 84.4, s = 8.4, n = 34$, then the test statistic is $t = 7.9$ according to formula (13.1) below.

The probability distribution of test statistic t under H0 is known; you can find the boundary value t^* in Appendix C. Put otherwise, if the test statistic t is larger than the boundary value t^* which is stated in the table then $P(t|\text{H0}) < \alpha$. To be able to use the table in Appendix C, we still have to introduce a new term, namely the number of degrees of freedom. That term is explained in §13.2.1 below.

With the number of degrees of freedom, you can look in Appendix C to see which boundary value t^* is needed to achieve a certain p-value for the established test statistic $t = 7.9$. Let us see what the p-value is for the established test statistic $t = 7.9$. We firstly look for the degrees of freedom ('d.f.') in the left column. If the number of degrees of freedom does not occur in the table, then, to err on the side of caution, we should round down, here to 30 d.f. This determines the row which is applicable for us. In the third column, we find $t^* = 1.697$. Our established test statistic $t = 7.9$ is larger than this $t^* = 1.697$, thus the p-value is smaller than the $p = .05$ from the third column. If we go further right on the same line, we see that the stated t^* increases further.
Our established test statistic t is even larger than $t^* = 3.385$ in the last column. The p-value is thus even smaller than $p = .001$ from the title of that last column. (The statistical analysis program usually also calculates the p-value.) We report the result as follows:

The mean score of Linguistics students (class of 2013) is 84.4 ($s =$

8.4); this is significantly better than the assumed population mean of 73 ($t(33) = 7.9, p < .001$).

13.2.1 Degrees of freedom

To explain the concept of degrees of freedom, we begin with an analogy. Imagine that there are three possible routes for getting from A to B: a coast path, a mountain path, and a motorway. It is true that a walker who wants to travel from A to B has three options but there are only two degrees of freedom for the walker: he or she only has to make two choices to choose from the three options. Firstly, the motorway drops out (first choice), and then the mountain path (second choice), and then the chosen route along the coast is the only one left over. There are thus two choices 'free' in order to choose one of the three possible routes in the end. If we know the two choices, then we can deduce from them which route must have been chosen.

Now, we will look at a student who on average has achieved a $\overline{x} = 7.0$ over the $N = 4$ courses from the first introductory track of his or her degree programme. The mean of 7.0 can be arrived at in many ways, e.g. $(8, 7, 7, 6)$ or $(5, 6, 8, 9)$. But if we know the result of three of the courses, and we also know that the mean is a 7.0 then we also know what the value of the fourth observation must be. This last observation is thus no longer 'free' but is now fixed by the first three observations, in combination with the mean over $N = 4$ observations. We then say that you have $N - 1$ degrees of freedom to determine this characteristic of the sample, like the sample mean here, or like the test statistic t. The degrees of freedom is often abbreviated to 'd.f.' (symbol ν, Greek letter "nu").

In practice, the number of degrees of freedom is not difficult to determine. We namely indicate for every test how the degrees of freedom are established — and the number of d.f. is usually also calculated by the statistical analysis program which we use.

For the t-test of a single sample, the number of degrees of freedom is the number of observations $N - 1$. In the above discussed example, we thus have $N - 1 = 34 - 1 = 33$ degrees of freedom.

13.2.2 formulas

$$t = \frac{\overline{y} - \mu}{s} \times \sqrt{N} \tag{13.1}$$

13.2.3 assumptions

The t-test for a single sample requires three assumptions which must be satisfied in order to be able to use the test.

- The data must have been measured on an interval level of measurement (see Chapter 4).
- All the observations have to be independent of each other.
- The scores must be normally distributed (see §10.3).

13.2.4 SPSS

The above discussed data can be found in the file `data/grammaticatoets2013.csv`.

To test our earlier hypothesis, in SPSS, we firstly have to select the observations of the Linguistics students.

`Data > Select cases...`

Choose `If condition is satisfied` and click on the button `If...` to indicate the conditions for selection (inclusion).
Select the variable `progr` (drag to the panel on the right-hand side), pick button `=`, and then type *`TW`* (the Dutch label for "Linguistics"), so that the whole condition is `progr = TW`.

Afterwards, we can test our earlier hypothesis as follows:

`Analyze > Compare Means > One-Sample T Test...`

Select variable (drag to the Test variable(s) panel).
Indicate which value of μ has to be tested: set it as Test Value `73`. Confirm `OK`.

The output contains both descriptive statistics and the results of a *two-sample* t-test.

When transferring this output, take good note of the warning in §13.3 below: SPSS reports as if `p=.000` but that is untrue.

13.2.5 R

Our hypothesis discussed above can be tested with the following exercises:

```
gramm2013 <- read.csv( file="data/grammaticatoets2013.csv",header=F)
dimnames(gramm2013)[[2]] <- c("score","progr")
# program levels have Dutch labels: TW=Linguistics
with( gramm2013,
      t.test( score[progr=="TW"], mu=73, alt="greater" ) )
```

```
##
##  One Sample t-test
##
## data:  score[progr == "TW"]
## t = 7.9288, df = 33, p-value = 1.913e-09
## alternative hypothesis: true mean is greater than 73
## 95 percent confidence interval:
##  81.97599      Inf
## sample estimates:
## mean of x
##  84.41176
```

The notation `1.913e-09` must be read as the number (1.913×10^{-9}).

13.3 p-value is always larger than zero

The p-value p can be very small but it is always larger than zero! In the grammar test example above, we found $P = .000000001913$, a very small probability but one that is larger than zero. This can also be seen from the tails of the corresponding probability distribution which approach zero asymptotically (see Fig.13.1) but never become completely equal to zero. There is always a minimally small probability of finding an extreme value (or an even more extreme value) from you test statistic in a sample — after all, we are investigating the sample precisely because the outcome of the test statistic cannot be established a priori.

In SPSS, however, the p-value is rounded off, and can then appear as `'Sig. .000'` or $p = .000$. This is incorrect. The p-value or significance is not equal to zero, but has been *rounded off* to zero, and that is not the same. Always report the p-value or significance with the correct accuracy, in this example as $p < .001$ or even $p < .0005$ (taking into account the rounding off by SPSS to three decimal places).

13.4 One-sided and two-sided tests

The procedure which we discussed above is valid for one-sided tests. This is to say that the alternative hypothesis does not only put forward that the means will differ but also in which direction that will be: H1: $\mu > 73$, the Linguistics students score *better* than the population mean. If we were to find a difference in the opposite direction, say $\overline{x} = 68$, then we would not even conceive of statistical testing: the H0 simply still stands. It is only when we find a difference in the hypothesised direction that it is meaningful to inspect whether this difference is significant. When you look now at the figure in

Appendix C, then this is also the case. The p-value corresponds with the area of the coloured region.

If the alternative hypothesis H1 does *not* specify the direction of the difference, then a complication arises. Differences in any of the two possible directions are relevant. We speak then of two-sided tests or two-tailed tests. To calculate the two-sided p-value, we multiply the p-value from Appendix C by 2 (because we are now looking at two coloured regions, on the lower and upper sides of the probability distribution).

In the grammar test example, let us now use a two-sided test. We then operationalise the alternative hypothesis as H1: $\mu \neq 73$. Again, there is $\overline{x} = 73, t = 7.9$ with 33 d.f. (rounded off to 30 d.f.). With the one-sample p-value $p = .025$ (fourth column), we find the critical value $t^* = 2.042$. The two-sided p-value
for this critical value is $2 \times .025 = .05$. The test statistic we found $t = 7.9$ is larger than this $t^* = 2.042$, thus the two-sided p-value is smaller than $p = 2 \times .025 = .05$. The test statistic we found is larger even than $t^* = 3.385$ in the last column, thus the two-sided p-value is even smaller than $2 \times .001$. We can report our two-sided testing as follows:

> The mean score of Linguistics students (class of 2013) is 84.4 (s = 8.4); the differs significantly from the hypothesised population mean of 73 ($t(33) = 7.9, p < .002$).

In the majority of studies two-sided tests are used; if the direction of the test is not stated then you may assume that two-sided or two-tailed tests have been used.

13.5 Confidence interval of the mean

This section looks more deeply into a subject that was already discussed in §10.7, and illustrates the confidence interval of the mean with the grammar test scores.

We can consider the sample's mean, $\overline{x}$, as a good estimation of an unknown mean in the population, μ. For this, we can also use the value found for t^* to indicate how reliable the estimation is: the confidence interval. With this, we express with what (un)certainty we know that the sample mean, $\overline{x}$, matches the population mean (Cumming, 2012). We are also familiar with such error margins from election results, where they indicate with what certainty the result of the sample (of respondents) matches the actual election result for the whole population (of voters). An error margin of 2% means that it is 95% certain that x, the percentage voting for a certain party, will lie between $(x - 2)\%$ and $(x + 2)\%$.

In our example with 30 d.f., we find $t^* = 2.042$ for 95% reliability. Via formula (13.2), we arrive at the 95% confidence interval $(81.5, 87.3)$. We know with 95% certainty that the unknown average score on the grammar test, from the population of all possible Linguistics students is larger than 81.5 and smaller than 87.3. We thus also know, with 95% certainty, that the *unknown* population mean μ deviates from the hypothesised value 73 (Cumming, 2012). We report this as follows:

> The mean score of Linguistics students (class of 2013) is 84.4, with 95% confidence interval (81.5, 87.3), 33 d.f.

In Figure 13.2, you can see the results of a computer simulation to illustrate this. This figure is made in the same way as Figure 10.7 in Chapter 10 and illustrates the same point. We have drawn 100× samples from Linguistics students, with $\mu = 84.4$ and $\sigma = 8.4$ (see §9.5.2) and $N = 34$. For each sample, we have drawn the 95% confidence interval. For 95 of the 100 samples, the population mean $\mu = 84.4$ is indeed within the interval, but for 5 of the 100 samples the confidence interval does not contain the population mean (these are marked along the right hand side).

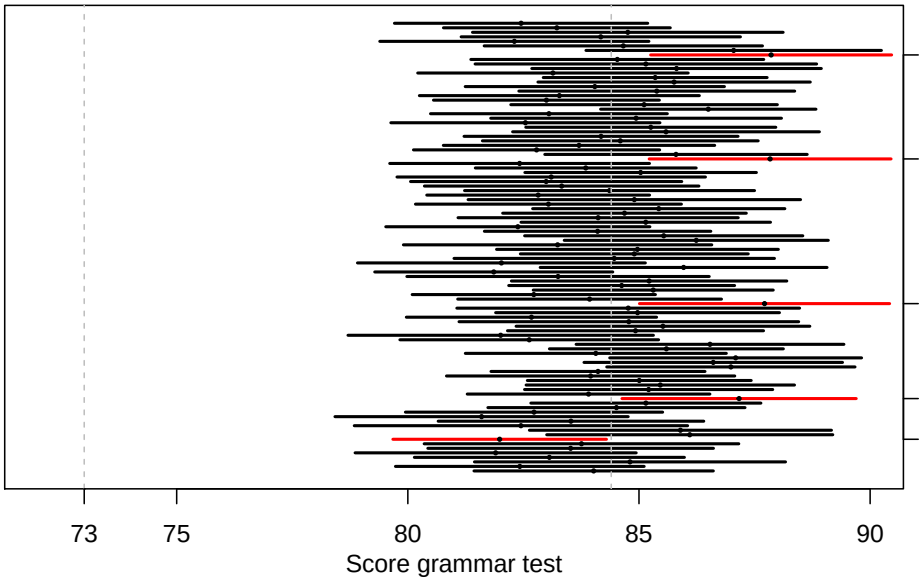


Figure 13.2: 95% confidence interval and sample means, over 100 simulated samples (n=34) from a population with population mean 84.4, population-s.d. 8.4.

13.5.1 formulas

The two-sample confidence interval for $B\%$ reliability for a population mean $\overline{y}$ is

$$\overline{y} \pm t^*_{N-1} \times \frac{s}{\sqrt{N}} \tag{13.2}$$

13.5.2 SPSS

`Analyze > Descriptive Statistics > Explore...`

Select dependent variables (drag to Dependent List panel).
Click on button `Statistics` and tick `Descriptives` with Confidence Interval 95%.
Confirm with `Continue` and with `OK`.
The output contains several descriptive statistic measures, now also including the 95% confidence interval of the mean.

13.5.3 R

R states the confidence interval of the mean (with self-specifiable confidence level) for a t-test. We thus again conduct a t-test and find the confidence interval of the mean in the output.

```
with( gramm2013, t.test( score[progr=="TW"] ) )
```

```
##
##  One Sample t-test
##
## data:  score[progr == "TW"]
## t = 58.649, df = 33, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  81.48354 87.33999
## sample estimates:
## mean of x
##  84.41176
```

13.6 Independent samples t-tests

The Student's t-test is used to allow the investigation of a difference between the mean scores of two independent samples, e.g of comparable boys and girls.

On the basis of the mean and the standard deviations of the two samples, we determine the test statistic t. If H0 is true, then the value $t = 0$ is the most probable. The larger the difference between the two means, the larger t is too. We again reject H0 if $t > t^*$ for the chosen significance level α.

As a first example, we will take a study of the productive vocabulary size of 18-month old Swedish girls and boys (Andersson et al., 2011). We investigate the hypothesis that the vocabulary of girls differs from that of boys, i.e. H1: $\mu_m \neq \mu_j$. We cannot a priori assume that a potential difference can only go in one direction; we thus use a two-sided test, as already appears to be the case from H1. The corresponding null hypothesis which we test is H0: $\mu_m = \mu_j$. In this study, the vocabulary is estimated on the basis of questionnaires from the parents of the children in the sample. Participants were (parents of) $n_1 = 123$ girls and $n_2 = 129$ boys, who were all 18 months old. Based on the results, it seems that the girls have a mean vocabulary of $\overline{x_1} = 95$ words ($s_1 = 82$), and for the boys it is $\overline{x_2} = 85$ words ($s_2 = 98$). With these data, we determine the test statistic t according to the formula (13.4), resulting in $t = 0.88$ with 122 d.f. We look for the accompanying critical value t^* again in Appendix C. In the row for 100 d.f. (after rounding down), we find $t^* = 1.984$ in the fourth column. For two-sample testing we have to double the p-value which belongs to this column (see §13.4), resulting in $p = .05$. The test statistic $t < t^*$, thus $p > .05$. We decide *not* to reject H0, and report that as follows:

> The mean productive vocabulary of Swedish 18-month old Swedish children barely differs between girls and boys ($t(122) = 0.88, p > .4$). Girls produce on average 95 different words ($s = 82$), and boys on average 85 different words ($s = 98$).

As a second example, we take a study of the speech tempo of two groups of speakers, namely originating from the West (first group) and from the North (second group) of the Netherlands. The speech tempo is expressed here as the mean duration of a spoken syllable, in seconds, over an interview of ca. 15 minutes (see Example **??**.1). We investigate H0: $\mu_W = \mu_N$ with two-sample testing. From the results, it appears that those from the West ($n = 20$) have a mean syllable duration of $\overline{x_W} = 0.235\,\mathrm{s}$ ($s = 0.028$), and that for those from the North (also $n = 20$) that is $\overline{x_N} = 0.269\,\mathrm{s}$ ($s = 0.029$). With these data, we again determine the test statistic t according to the formula (13.4), resulting in $t = -3.76$ with 38 d.f. We look for the accompanying critical value again in Appendix C. The correct d.f. are not stated in the table so we round them down (i.e. in the conservative direction) to 30 d.f. In the row, we find $t^* = 2.042$ in the fourth column. For two-sample testing, we have to double the p-value corresponding to these columns (see §13.4), resulting in $p = .05$. The test statistic is $t < t^*$, thus $p < .05$. We thus decide to *indeed* reject H0, and report that as follows:

> The mean duration of a syllable spoken by a speaker from the West of the Netherlands is 0.235 seconds ($s = 0.028$). This is significantly

shorter than from speakers from the North of the Netherlands ($\overline{x} = 0.269\,\text{s},\ s = 0.029$) ($t(38) = -3.76, p < .05$). In the investigated recordings from 1999, the speakers from the West thus speak more quickly than those from the North of the Netherlands.

13.6.1 assumptions

The Student's t-test for two independent samples requires four assumptions which must be satisfied in order to use the test.

- The data has to be measured on an interval level of measurement (see §4.4).
- All the observations must be independent of each other.
- Both groups' scores must be normally distributed (see §10.4).

The variance of the scores has to be equal in both samples. The more the two samples differ in size, the more serious the violation of this assumption is. It is thus prudent to work with equally large, and preferably not too small samples. If the samples are equally large, then the violation of this assumption of equal variances is not so serious.

13.6.2 formulas

13.6.2.1 test statistic

To calculate test statistic t, various formulas are used.

If the samples have about equal variance, then we firstly use the "pooled standard deviation" s_p as an intermediate step. In this, both standard deviations from the two samples are weighted according to their sample size.

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \tag{13.3}$$

Then

$$t = \frac{\overline{x_1} - \overline{x_2}}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{13.4}$$

If the samples do *not* have equal variance, and the fourth assumed sample above is thus violated, then Welch's t-test is used:

$$s_{\text{WS}} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \tag{13.5}$$

Then

$$t = \frac{\overline{x_1} - \overline{x_2}}{s_{\text{WS}}} \tag{13.6}$$

13.6.2.2 degrees of freedom

The t-test is usually conducted by a computer program. There, the following approximation of degrees of freedom is usually used (ν, see §13.2.1). Firstly, $g_1 = s_1^2/n_1$ and $g_2 = s_2^2/n_2$ are calculated. The number of degrees of freedom of t is then

$$\nu_{\text{WS}} = \frac{(g_1 + g_2)^2}{g_1^2/(n_1 - 1) + g_2^2/(n_2 - 1)} \tag{13.7}$$

According to this approximation, the number of degrees of freedom has as its liberal upper limit $(n_1 + n_2 - 2)$, and as its conservative lower limit the smallest of $(n_1 - 1)$ or $(n_2 - 1)$. You can thus always use this conservative lower limit. If the two groups have around the same variance (i.e. $s_1 \approx s_2$), then you can also use the liberal lower limit.

For the second example above, the approximation of formula (13.7) gives an estimation of $37.99 \approx 38$ d.f. The conservative lower limit is $n_1 - 1 = n_2 - 1 = 19$. The liberal lower limit is $n_1 + n_2 - 2 = 38$. (In the table with critical values $t*$, in Appendix C, it is usually advisable to use the row with the first-following smaller value for the number of degrees of freedom.)

13.6.3 SPSS

Here, the second example above is worked out.

`Analyze > Compare Means > Independent-Samples T Test`

Drag the dependent variable `syldur` to the Test Variable(s) panel. Drag the independent variable `region` to the Grouping Variable panel. Define the two groups: value `W` for region group 1 and
value `N` for region group 2. Confirm with `Continue` and `OK`.

As you could see above the calculation of the t-test is dependent on the answer to the question whether the standard deviations of the two groups are around equal. SPSS solves this rather clumsily: you get to see all the relevant outputs, and have to make a choice from them yourself.

13.6.3.1 Test for equality of variances

With Levene's test, you can investigate H0: $s_1^2 = s_2^2$, i.e. whether the variances (and with them the standard deviations) of the two groups are equal. If you find a small value for the test statistic F, and a $p > .05$, then you do not have to reject this H0. You can then assume that the variances are equal. If you find a large value for F, with $p < .05$, then you should indeed reject this H0, and you cannot assume that the variances of these two groups are equal.

13.6.3.2 Test for equality of means

Depending on this outcome from Levene's test, you have to use the first or the second row of the output of the Independent Samples Test (a test which investigates whether the means from the two groups are equal). In this example, the variances are around equal, as the Levene's test also indicates. We thus use the first line of the output, and report $t(38) = -3.765, p = .001$.

13.6.4 R

```
require(hqmisc)
data(talkers)
with(talkers, t.test( syldur[region=="W"], syldur[region=="N"],
            paired=F, var.equal=T ) )
```

```
##
##  Two Sample t-test
##
## data:  syldur[region == "W"] and syldur[region == "N"]
## t = -3.7649, df = 38, p-value = 0.0005634
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.0519895 -0.0156305
## sample estimates:
## mean of x mean of y
##   0.23490   0.26871
```

13.7 t-test for paired observations

The Student's t-test is also used to investigate a difference between the means of two dependent or paired observations. This is the case if we only draw one sample (see Chapter 7), and then collect two observations from the members

of this sample, namely one observation under each of the conditions. The two observations are then paired, i.e. related to each other, and these observations are thus not independent (since they come from the same member of the sample). With this, one of the assumptions of the t-test is violated.

As an example, we take an imaginary investigation of the use of the Dutch second person pronouns *U* (you *formal*) and *je* (you *informal*) as forms of address on a website. The researcher makes two versions of a website, one with *U* and the other with *je*. Each respondent has to judge both versions on a scale from 1 to 10. (For validity reasons, the order of the two versions is varied between respondents; the order in which the pages are judged can thus have no influence on the total score per condition.) In Table 13.1, the judgements of $N = 10$ respondents are summarised.

Table 13.1: Fictional judgements of a webpage with *U* or *je* as the forms of addressed, by $N = 10$ respondents.

ID	*U* Condition	*je* Condition	D
1	8	9	-1
2	5	6	-1
3	6	9	-3
4	6	8	-2
5	5	8	-3
6	4	6	-2
7	4	8	-4
8	7	10	-3
9	7	9	-2
10	6	7	-1
			$\overline{D}$=-2.2

The pair of observations for the i-th member of the sample has a difference score which we can write as:
$D_i = x_{1i} - x_{2i}$ where x_{1i} is the dependent variable score of the i-th member for condition 2. This difference score is also stated in Table 13.1.

This difference score D is then actually analysed with the earlier discussed t-test for a single sample (see §13.2), where H0: $\mu_D = 0$, i.e. according to H0, there is no difference between conditions. We calculate the mean of the difference score, $\overline{D}$, and the standard variance of the difference score, s_D, in the usual manner (see §9.5.2). We use this mean and this standard deviation to calculate the test statistic t via formula (13.8), with $(N - 1)$ degrees of freedom. Finally, we again use Appendix C to determine the critical value. and with it, the p-value p for the value of the sample size t under H0.

For the above example with *U* or *je* as forms of address, we thus find $\overline{D} = -2.2$ and $s_D = 1.0$. If we put this into formula (13.8), we find $t = -6.74$

with $N - 1 = 9$ d.f. We again look for the corresponding critical value t^* in Appendix C. Thereby, we ignore the sign of t, because, after all, the probability distribution of t is symmetric. In the row for 9 d.f., we find $t^* = 4.297$ in the last column. For two-sided testing, we have to double the p-value corresponding to this column (see §13.4), resulting in $p = .002$. The test statistic is $t > t^*$, thus $p < .002$. We decide to *indeed* reject H0, and report that as follows:

> The judgement of $N = 10$ respondents on the page with U as the form of address is on average 2.2 points lower than their judgement over the comparable page with *je* as the form of address; this is a significant difference ($t(9) = -6.74, p < .002$).

13.7.1 assumptions

The t-test for paired observations within a single sample requires three assumptions which must be satisfied, in order to be able to use these tests.

- The data must be measured on an interval level of measurement (see §4.4).
- All *pairs* of observations have to be independent of each other.
- The *difference scores* D have to be normally distributed (see §10.4); however, if the number of pairs of observations in the sample is larger than ca. 30 then the t-test is usually useable.

13.7.2 formulas

$$t = \frac{\overline{D} - \mu_D}{s_D} \times \sqrt{N} \tag{13.8}$$

13.7.3 SPSS

The data for the above example can be found in the file `data/ujedata.csv`.

```
Analyze > Compare Means > Paired-Samples T Test
```

Drag the first dependent variable `cond.u` to the Paired Variables panel under Variable1, and drag the second variable `cond.je` to the same panel under Variable2. Confirm with `OK`.

13.7.4

The data from the above example can be found in the file `data/ujedata.csv`.

```
ujedata <- read.table( file="data/ujedata.csv", header=TRUE, sep=";" )
with(ujedata, t.test( cond.u, cond.je, paired=TRUE ) )
```

```
##
##  Paired t-test
##
## data:  cond.u and cond.je
## t = -6.7361, df = 9, p-value = 0.00008498
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.938817 -1.461183
## sample estimates:
## mean of the differences
##                    -2.2
```

13.8 Effect size

Until now, we have mainly dealt with testing as a binary decision with regards to whether or not to reject H0, in the light of the observations. However, in addition to this, it is also of great importance to know how large the observed effect actually is: the *effect size* ('ES') (Cohen, 1988; Thompson, 2002; Nakagawa and Cuthill, 2007).

In the formulas (13.1) and (13.8), it is expressed that the larger the effect gets, the larger t gets, i.e. for a larger difference $(\overline{x} - \mu)$ or $(\overline{x_1} - \overline{x_2})$ or $(\overline{D} - \mu_D)$], *and.or* the larger the sample gets. Put briefly (Rosenthal and Rosnow, 2008, p.338, formula 11.10):

$$\text{significance test} = \text{size of effect} \times \text{size of study} \tag{13.9}$$

This means that a small, and possibly trivial effect can also be statistically significant if only the sample is large enough. Conversely, a very large effect can be firmly established on the basis of a very small sample.

Example 13.2: In an investigation of the life times of inhabitants from Austria and Denmark (Doblhammer, 1999), it appears that life times differ according to the week of birth. This is presumably

because babies from "summer pregnancies" are (or were) on average somewhat healthier than those from "winter pregnancies". In this investigation, the differences in life times were very small ± 0.30 year in Austria, ± 0.15 year in Denmark), but the number of observations (deceased persons) was very large.

Meanwhile, the difference in body length between dwarfs (shorter than 1.5 m) and giants (taller than 2.0 m) is so large that the difference can be firmly empirically established on the basis of only $n = 2$ in each group.

In our investigation, we are especially interested in important differences, i.e. usually large differences. We have to appreciate that studies also entail costs in terms of money, time, effort, privacy, and loss of naïveté for other studies (see Chapter 3). We thus do not what to want to perform studies on trivial effects needlessly. A researcher should thus determine in advance what the smallest effect is that he/she wants to be able to detect, e.g. 1 point difference in the score of the grammar test. Differences smaller than 1 point are then considered to be trivial, and differences larger than 1 point to be potentially interesting.

It is also important to state the effect size found with the results of the study, to be of service for readers and later researchers. In some academic journals, it is even required to report the effect size. It should be said that this can also be in the form of a confidence interval of the mean (see 13.5), because we can convert these confidence intervals and effect sizes into each other.

The raw effect size is straightforwardly the difference D in means between the two groups, or between two conditions, expressed in units of the raw score. In §13.6, we found such a difference in vocabulary of $D = 95 - 85 = 10$ between boys and girls.

However, we mainly use the standardised effect size (see the formulas below), where we take into account the distribution in the observations, e.g. in the form of "pooled standard deviation" s_p [2]. In this way, we find a standardised effect size of

$$d = \frac{\overline{x_1} - \overline{x_2}}{s_p} = \frac{10}{90.5} = 0.11 \tag{13.10}$$

In the first example below, the standardised effect size of the difference in vocabulary between girls and boys is thus 0.11. In this case, the difference between the groups is small with respect to the distribution within the groups

[2] In this case, we use $s_p = \sqrt{\frac{122 \times 82^2 + 128 \times 98^2}{122+128}} = 90.5$, see formulas (13.3) and (13.12).

— the probability that a randomly selected girl has a larger vocabulary than a randomly selected boy, is only 0.53 (McGraw and Wong, 1992), and that is barely better than the probability of 0.50 which we expect according to H0. It is then no surprise that this very small effect is not significant (see §13.6). We could report the effect size and significance as follows:

> The mean productive vocabulary of Swedish 18-month old children barely differs between girls and boys. Girls produce on average 95 different words ($s = 82$), and boys on average 85 different words ($s = 98$). The difference is very small ($d = 0.11$) and not significant ($t(122) = 0.88, p > .4$).

In the second example above, the standardised effect size of the difference in syllable length is about $(0.235 - 0.269)/0.029 \approx 1.15$. We can report this relatively large effect as follows:

> The average length of a syllable spoken by a speaker from the West of the Netherlands is 0.235 seconds ($s = 0.028$). This is considerably shorter than for speakers from the North of the Netherlands ($\overline{x} = 0.269\,\text{s}$, $s = 0.029$). The difference is ca. 10%; this difference is very large ($d = -1.15$) and significant ($t(38) = -3.76, p < .05$). In the investigated recordings from 1999, the speakers from the West thus speak considerably more quickly than those from the North of the Netherlands.

If d is around 0.2, we speak of a small effect. We call an effect size d of around 0.5 a medium effect, and we call one of around 0.8 or larger a large effect (Cohen, 1988; Rosenthal and Rosnow, 2008).

> *Example 13.3:* Look again at the formula (13.9) and at the Figure 13.3 which illustrate the relation between sample size and effect size. With an effect size of $n_1 = 122$, we can only detect an effect of $d = 0.42$ or more, with sufficiently low probabilities of Type I and II errors again ($\alpha = .05$, $\beta = .10$). To detect the very small effect of $d = 0.11$, with the same small error probabilities α and β, samples of at least 1738 girls and 1738 boys would be needed.

We can also express the effect size as the probability that the difference occurs in the predicted direction, for a randomly chosen element from the population (formulas (13.11) and (13.13)), or (if applicable) for two randomly and

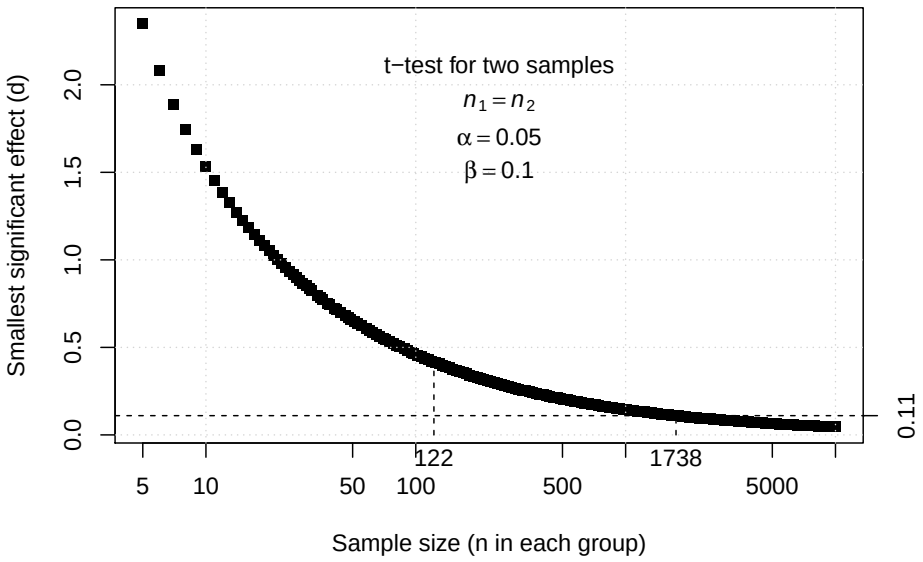


Figure 13.3: Relation between the sample size and the smallest effect (d) that is significant according to a *t*-test for unpaired, independent observations, with errors probabilities alpha=.05 and beta=.10.

independently chosen elements from the two populations (formula (13.12)) (McGraw and Wong, 1992). Let us again return to the grammar test from the Linguistics students (§13.2). The effect which we found is not only significant but also large. Expressed in terms of probability: the probability that a random Linguistics student achieves a score larger than $\mu_0 = 73$ is 0.91. (And a randomly chosen Linguistics student thus still has 9% probability of achieving a lower score than the hypothesised population mean of 73.)

For the fictional judgements about the webpages with *U* or *je* (see Table 13.1), we find a standardised effect size of

$$d = \frac{\overline{D} - \mu_D}{s_D} = \frac{-2.20 - 0}{1.03} = -2.13$$

It is then not surprising that this extremely large effect is indeed significant. We can report this as follows:

> The judgements of $N = 10$ respondents about the pages with *U* or *je* as forms of address differ significantly, with on average −2.2 points difference. This difference has a 95% confidence interval of −2.9 to −1.5 and an estimated standardised effect size $d = -2.13$; the probability that a randomly chosen respondent judges the *je*-version more highly than the *U*-version is $p = .98$.

13.8.1 formulas

For a single sample:

$$d = \frac{\overline{x} - \mu}{s} \tag{13.11}$$

where s stands for the standard variation s of the score x.

For two independent samples (see formula (13.3)):

$$d = \frac{\overline{x_1} - \overline{x_2}}{s_p} \tag{13.12}$$

For paired observations:

$$d = \frac{\overline{x_1} - \overline{x_2}}{s_D} \tag{13.13}$$

where s_D is the standard deviation from the difference D according to the formula (13.13).

13.8.2 SPSS

In SPSS, it is usually easiest to calculate the effect size by hand.

For a simple sample (formula (13.11)), we can simply calculate the effect size from the mean and the standard deviation, taking into account the value μ which we are testing against.

`Analyze > Descriptive Statistics > Descriptives...`

Choose the button `Options` and ensure that `Mean` and `Std.deviation` are ticked. As a result there is the required data in the output:
$d = (84.41 - 73)/8.392 = 1.36$, a very large effect.

For unpaired, independent observations, it is likewise the easiest to calculate the effect size by hand on the basis of the means, standard deviations, and size of the two samples, making use of the formulas (13.3) and (13.12) above.

For a single sample with two paired observations (formula (13.13)), we can again calculate the effect size more simply from the mean and the standard deviation of the difference. The data are in the output of the pairwise t-test (§13.7.3), respectively as `Mean` and `Std.Deviation`:
$d = -2.200/1.033 = 2.130$, a super large effect.

13.8.3 R

In R, it is easier to have the effect size calculated.

For a single sample (formula (13.11)):

```
gramm2013 <- read.csv( file="data/grammaticatoets2013.csv",header=F)
dimnames(gramm2013)[[2]] <- c("score","progr")
# programs have Dutch labels, TW=Linguistics
with(gramm2013, score[progr=="TW"]) -> score.ling
# auxiliary variable
( mean(score.ling)-73 ) / sd(score.ling)
```

```
## [1] 1.359783
```

The probability of a score larger than the population mean (the test value) `73` for a random Linguistics student (of which we assume that $\mu = 84.4$ and $s = 8.4$):

```
1 - pnorm( 73, mean=84.4, sd=8.4 )
```

```
## [1] 0.9126321
```

For unpaired, independent observations, we can calculate the smallest significant effect (see also Fig. 13.3); for which we use the function `power.t.test`. (This function is also used to construct Fig.13.3.) With this function, you have to set the desired `power` as an argument (power $= 1 - \beta$; see §14.1).

```
power.t.test( n=122, sig=.05, power=.90, type="two.sample" )
```

```
##
##      Two-sample t test power calculation
##
##               n = 122
##           delta = 0.4166921
##              sd = 1
##       sig.level = 0.05
##           power = 0.9
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

In the output, the smallest significant effect is indicated by `delta`; see also Example 13.3 above.

For a single sample with two paired observations (formula(13.13)):

```
ujedata <- read.table( file="data/ujedata.csv", header=TRUE, sep=";" )
with( ujedata, mean(cond.u-cond.je) / sd(cond.u-cond.je) )
```

```
## [1] -2.130141
```

13.8.4 Confidence interval of the effect size

Earlier, we already saw (§10.7 and §13.5) that we can estimate a characteristic or parameter of the population on the basis of a characteristic from a sample. This is how we estimated the unknown population mean μ on the basis of the observed sample mean $\overline{x}$. The estimation does have a certain degree of uncertainty or reliability: perhaps the unknown parameter differs in the population somewhat from the sample characteristic, which we use as an estimator, as a result of chance variations in the sample. The (un)certainty and (un)reliability is expressed as a confidence interval of the estimated characteristic. We then know with a certain reliability (mainly 95%) that the unknown parameter will lie within that interval (§10.7 and §13.5).

This reasoning is now valid not only for the mean score, or for the mean or for the variance, but equally for the effect size. After all, the effect size is also an unknown parameter from the population, which we are trying to estimate based on a limited sample. For the fictional judgements about the webpages with the formal *U* or informal *je* pronouns (see Table 13.1), we found a standardised effect size of $d = -2.13$. This is an estimation of the unknown effect size (i.e. of the strength of the preference for the *je*-variant) in the population of assessors, on the basis of a sample of $n = 10$ assessors. We can also indicate the reliability of this estimation here, in the form of a *confidence interval* around the observed sample $d = -2.13$.

The confidence interval of the effect size is tricky to establish though (Nakagawa and Cuthill, 2007; Chen and Peng, 2015). We illustrate it here in a simple manner for the simplest case, namely that of the *t*-test for a single sample, or for two paired observations. For this, we need two elements: firstly, the effect size expressed as a correlation (Rosenthal and Rosnow, 2008, p.359, formula 12.1),

$$r = \sqrt{\frac{t^2}{t^2 + \mathrm{df}}}$$

and secondly the standard error of the effect size d (Nakagawa and Cuthill, 2007, p.600, formula 18):

$$\mathrm{se}_d = \sqrt{\frac{2(1-r)}{n} + \frac{d^2}{2(n-1)}} \tag{13.14}$$

In our earlier example of the $n = 10$ paired judgements about a webpage with *U* or *je* as forms of address we found $d = -2.13$. We also find that $r = .9135$. With these data, we find $\mathrm{se}_d = 0.519$ via formula (13.14).

With this, we then determine the confidence interval for the effect size:

$$d \pm t^*_{n-1} \times \mathrm{se}_d \tag{13.15}$$

(see the correspond formula (13.2)).

After filling in $t^*_9 = 2.262$ (see Appendix C) and $\mathrm{se}_d = 0.519$, we eventually find a 95% confidence interval of $(-3.30, -0.96)$. We thus know with 95% confidence that the unknown effect size in the population is somewhere within this interval, and thus also that it is smaller than zero. On the basis of this last consideration, we can reject H0. But: we now know not only *that* the preference deviates from zero, but also *to what extent* the (standardised) preference deviates from zero, i.e. how strong the preference for the *je*-version is. This new knowledge about the extent or size of the effect is often more useful and more interesting than the binary decision of whether there is an effect or not (whether or not to reject H0) (Cumming, 2012).

Chapter 14

Power

14.1 Introduction

With statistical testing of H0, we determine the probability P of the observed differences or effects (or of even larger differences or effects than observed) if H0 were true, and thus if the observed difference had to be attributed solely to chance (see §?? and Chapter 13). If the probability P is very small, then we have thus found results which are very improbable if H0 were true. We then conclude that H0 is presumably *not true* and we thus reject H0. We then call the difference or effect found, the "significant" (Latin: 'meaning making'). However, there is in fact a probability, P, that the difference found is actually a fluke, and that, by rejecting H0, we are making a Type I error (i.e. wrongly rejecting H0, see §13.1). As we use a certain significance level, with which to compare P, this α is thus also the probability that we are making a Type I error.

At least as important, however, is the opposite error of *not* wrongly rejecting H0, a Type II error. Examples of such errors are: not convicting a suspect even if he is guilty, letting a 'spam' email message through into my mailbox, examining a patient and nevertheless not noticing their illness, concluding that birds are silent when they are in fact singing (Example 13.1), or wrongly concluding that two groups do not differ when an important difference does in fact exist between the two groups. The probability of a Type II error is referred to with the symbol β.

If H0 is in fact not true (there is a difference, the message is 'spam', birds are singing, etc.), then H0 should be rejected, and β should thus be as small as possible. The probability of *rightly* rejecting H0 is then $(1 - \beta)$ (see complement rule (10.5)); this probability $(1 - \beta)$ is called the *power*. Power can be interpreted as **the probability of the researcher being right** (H0 is rejected) **when she is indeed right** (H0 is untrue).

The probabilities of Type I and Type II errors have to be weighed up carefully against each other. In many studies, the p-values $\alpha = .05$ (significance level) and $\beta = .20$ (power=.80) are used. With these, an implicit weighting is made that a Type I error is $4\times$ more grave than a Type II error. For some studies that might be justified but it is also easily conceivable that, under certain circumstances, a Type II error is actually more grave or serious than a Type I error. If we find both types of error more or less equally grave, then we should strive for a smaller β and larger power (Rosenthal and Rosnow, 2008).

The power of a study depends on three factors: (i) the effect size d, which itself is in turn dependent on the measured difference D and the variation s in the observations (formula (13.10)), (ii) the sample size N, and (iii) the significance level α. In the following sections, we will discuss each of these factors separately, and, when doing so, keep the other two factors as constant as possible. For this discussion, we will use the depictions of calculated power (Figures 14.1 and 14.2). The depicted power contours are specifically for a t-test for independent samples (§13.6) with two-sided testing. The relations discussed below also apply to other statistical tests.

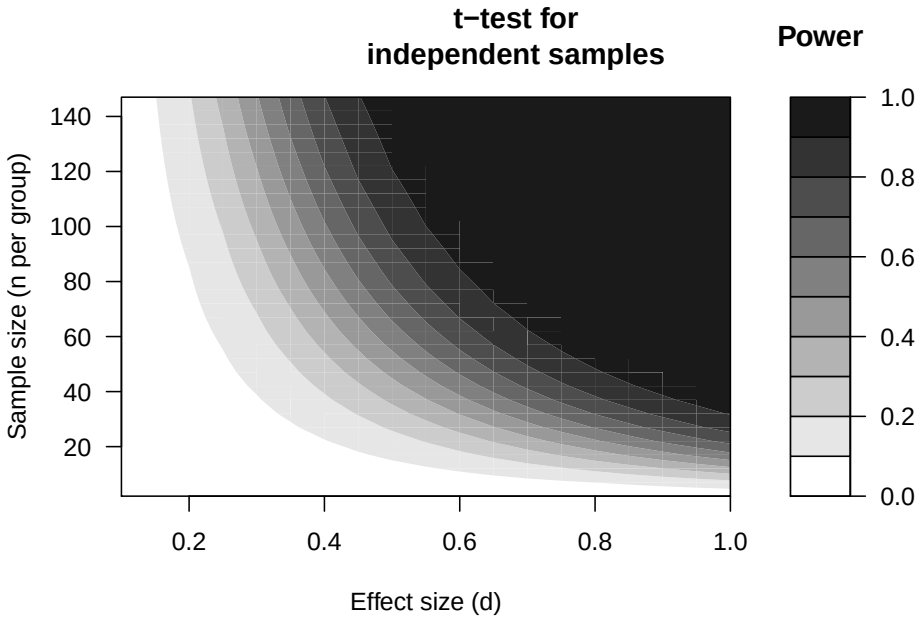


Figure 14.1: Power expressed in contours (see shading), dependent on the standardised effect size (d) and sample size according to a two-sided t-test for unpaired, independent observations, with two-sided significance level alpha=.01.

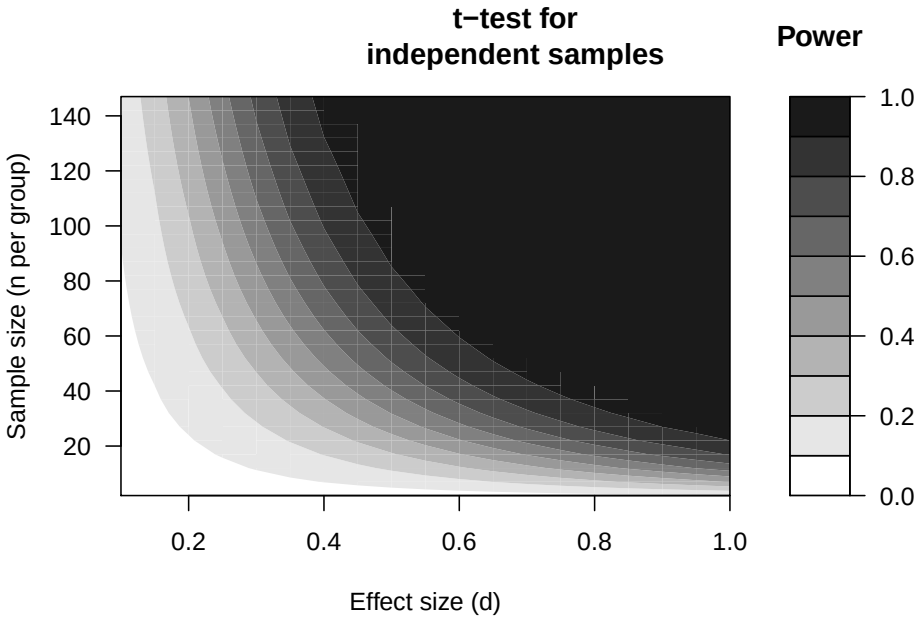


Figure 14.2: Power expressed in contours (see graduation), dependent on the standardised effect size (d) and the sample size (n), according to a two-sided t-test for unpaired, independent observations with significance level alpha=.05.

14.2 Relation between effect size and power

The two figures 14.1 and 14.2 show that, in general, the larger the effect to be tested is (more to the right in each figure), the larger the power. This is also not surprising: a larger effect has a higher probability of being detected in a statistical test, under the same circumstances. A moderately large effect of $d = .5$, with $n = 30$ observations in each group, only has a probability of .48 of being detected (if $\alpha = .05$, Figure 14.2). On the basis of a study with $n = 30$ observations per group is thus actually a gamble whether a researcher will actually detect such an effect, and will reject H0. Put otherwise, the probability of a Type I error is admittedly safely low ($\alpha = .05$) but the probability of a Type II error is more than $10\times$ as large, and thus dangerously high ($\beta = .52$) (Rosenthal and Rosnow, 2008, Ch.12).

A larger effect has a higher probability of being detected. A larger effect of $d = .8$, for example, results in a power of .86 with this testing. The probability of a Type II error $\beta = .14$ here is admittedly also larger than the probability of a Type I error, but the proportion β/α is considerably less skewed.

As researchers, we only have an indirect influence on effect size. We of course have no influence on the true raw difference D in the population. For the power, however, the raw difference D is not important, but rather the standardised difference $d = D/s$ (§13.8). Thus, if we ensure that the standard deviation s decreases, then d will increase, and then the power will also increase (figures 14.1 and 14.2), and we thus have a higher probability of actually detecting an effect! With this goal in mind, researchers always strive to neutralise disrupting influences from all kinds of other factors as much as possible. After all, the disrupting influences produce extra variability in the observations, and, with this, a lower power with the statistical testing.

In a well-designed study, we want to determine in advance what the power will be, and how large the sample should be (see below). For this we need an estimation of the smallest effect size d which we still want to detect (§13.8) (Quené, 2010). To estimate the effect size, firstly, an estimate of the raw difference D between the groups or conditions is needed. Secondly, an estimation is needed of the variability s in the observations. These estimations can be largely deduced from earlier publications, in which the standard deviation s is usually reported. If no earlier research reports are available, then s can be roughly estimated from some informal 'pilot' observations. Take the difference between the highest and the lowest (range) of these, divide this range by 4, and use the outcome of this as a rough estimation for s (Peck and Devore, 2008).

14.3 Relation between sample size and power

The relation between the sample size N and the power of a study is illustrated in Figure 14.1 for a strict significance level $\alpha = .01$, and in Figure 14.2 for the most used significance level $\alpha = .05$. The figures show that, in general, the larger the sample gets (further upwards), the larger the power is. The increase is steeper (power increases more quickly) with larger effects (right-hand side) than with smaller effects (left-hand side). Put differently: with small effects, the sample is actually always too small to detect these small effects with sufficient power. We already saw that in Example 13.3 (Chapter 13).

The two figures 14.1 and 14.2 are based on the comparison between two groups which are equally large, each with precisely half of the observations ($n_1 = n_2 = N/2$). That is also most efficient. The power is based on the *harmonic mean* of n_1 and n_2 (see §9.3.4), and that harmonic mean is always smaller than the arithmetic mean of those two numbers. It is thus advisable to ensure that the groups or samples which you compare are approximately equally large.

Example 14.1: In a study, two groups of participants are compared, with $n_1 = 10$ and $n_2 = 50$ ($N = n_1 + n_2 = 10 + 50 = 60$). The harmonic mean of $n_1 = 10$ and $n_2 = 50$ is $H = 17$. This study thus has the same power as a smaller study with two groups, each of 17 participants, thus 34 participants in total. For this study, thus, 26 participants more have been investigated (and bothered) than necessary (see also §3.2) (Aron et al., 2011, p.295).

14.4 Relation between significance level and power

The relation between the significance level α and the power is illustrated by the difference between the two figures 14.1 and 14.2. For each combination of effect size and sample size, the power is lower in Figure 14.1 for $\alpha = .01$ than in Figure 14.2 for $\alpha = .05$. If we choose a higher significance level α, then the probability of rejecting H0 increases, and thus also the power of correctly rejecting H0 when H0 is untrue (see Table 2.2). However, unfortunately, with a high significance level α, the probability of wrongly rejecting H0 (i.e. of making a Type I error) also increases. The investigator must make a

well-considered decision between Type I errors (with probability α) and Type II errors (with probability β); as said earlier, this decision has to depend on the seriousness of (the consequences of) these two types of errors.

14.5 Disadvantages of insufficient power

Unfortunately, many examples can be found of 'underpowered' research in the domain of language and communication. This research has a too small probability of rejecting H0 when the investigated effect indeed exists (H0 is not true). Why is that bad (Quené, 2010)?

Firstly, the Type II error which occurs here can have serious consequences: a treatment method which is actually better is not recognised as such, a patient is not or wrongly diagnosed, a useful innovation is wrongly pushed aside. This error hinders the growth of our knowledge and our insight, and hinders scientific progress (see also Example 3.2 in Chapter 3).

The outcomes of simulated experiments with different sample size, and thus with different power, are summarised in Figure 14.3. We explain the second disadvantage on the basis of the somewhat complex figure. In the left panel of Figure 14.3, we can see that the different (simulations of) 'underpowered' studies show a mixed picture. Some of these studies do show a significant effect (dark symbols), and many other studies do not (light symbols). The mixed picture then usually leads to follow up research, in which people try to find out *why* the effect does occur in some studies, and not in others. Might the difference in results be attributable to differences in stimuli? participants? tasks? instruments? All that follow up research is *superfluous* though: the mixed picture from these studies can be explained by the small power of each study. The needless and superfluous follow up research costs much time and money (and indirect costs, see Chapter 3), and comes at the cost of other, more useful research (Schmidt, 1996, p.118). Put otherwise: one well designed study with power which is more than sufficient can avoid many needless follow up studies.

The third disadvantage is based on the experience that studies in which a significant effect is found (dark symbols) have a higher probability of being reported; this phenomenon is called 'publication bias' or the 'file drawer problem'. After all, a positive result often does get published, whilst a negative result often disappears into a file drawer. With small power, that leads to the third disadvantage, namely an overestimation or 'bias' of the reported effect size. In an underpowered study, after all, an effect must be quite large to be found. In the leftmost panel, we can see that a significant effect has only been found 31×. The average effect size of these 31 significant outcomes is $\overline{d_{\text{signif}}} = 0.90$ (black dashed line), i.e. a distortion or 'bias' of 0.40 relative to the actual $d = 0.50$ (grey dashed line)[1]. In the rightmost panel, we can see

[1] A replication study which does have sufficient power, thus typically finds a smaller effect

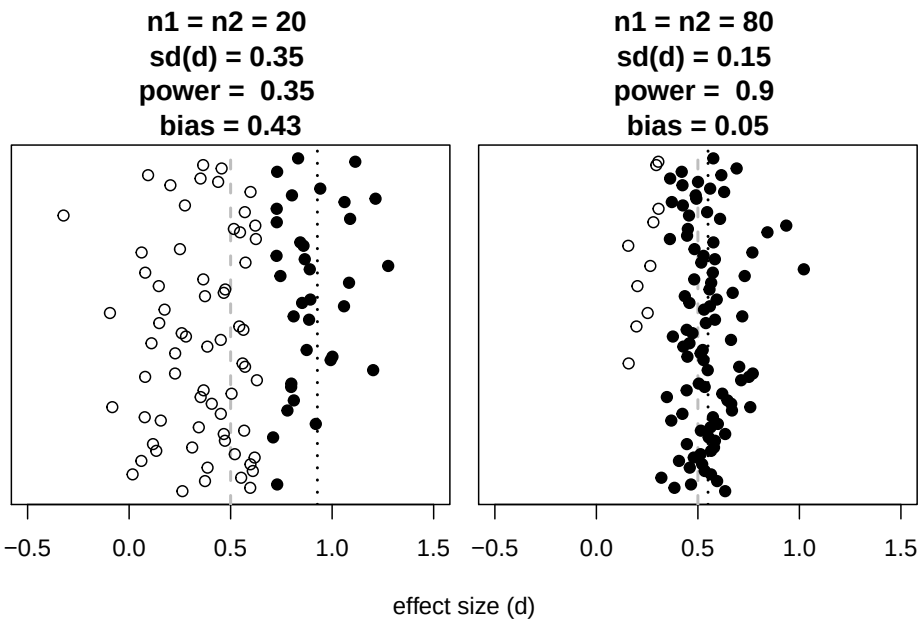


Figure 14.3: Effect sizes (along the horizontal axis) found in simulated experiments (two-sided t-test for independent observations, alpha=.05), broken down according to sample size (left $n = 20$, right $n = 80$) and according to testing result (dark symbols: significant; light symbols: not significant). The true effect size between groups is always $d = 0.5$, indicated by the grey dashed line. The mean effect size found from the significant outcomes is referred to with the black dashed line. For each sample size, 100 simulations have been carried out (long vertical axis).

that a significant effect has been found 91× (thus the power here is sufficient). The mean effect size of these 91 significant outcomes hardly deviates from the actual d. Moreover, the standard deviation of the reported effect size is smaller, and that is again important for later research, meta-studies, and systematic reviews.

Fourthly, the mixed picture from the different studies, sometimes with significant outcomes and sometimes without, and with great variation in the reported effect size, carries the danger that these outcomes are taken less seriously than 'consumers' of scientific knowledge (practitioners, health insurers, developers, policy makers, etc.). In this way, these consumers get the impression that the scientific evidence for this investigated effect is not strong, and/or that the researchers are in disagreement about whether the effect exists and if it does, how large it then is (Van Kolfschooten, 1993) (Figure 14.3). This, also hinders scientific progress, and it hinders the use of scientific insights in societal applications.

To avoid all these objections, researchers have to take into account the desired power of a study in an early stage. Designing and conducting a study with insufficient power is after all in opposition with the earlier discussed ethical and moral principles of diligence and responsibility (§3.1).

---

than the original 'underpowered' study. The smaller effect found in the replication study is then typically also not significant. We then say that the replication study "fails to replicate" the effect that was significant in the original study - but that was actually a spurious finding.

Appendix A

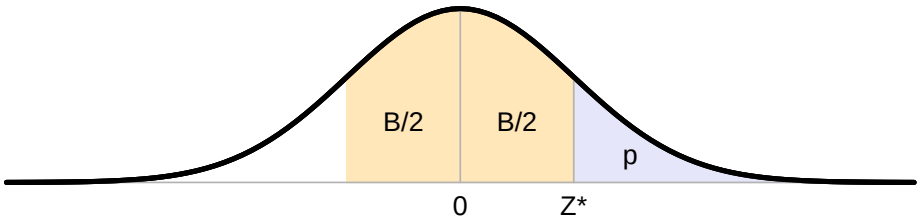
Random numbers

Table A.1: The table below contains 200 random numbers between 0 and 9999.

2836	264	6789	1483	3459	9200	4996	3761	699	5622
1943	6034	8838	1349	8750	3181	8799	4525	6536	5111
7259	8030	5709	8334	3526	2768	6296	8335	6350	6192
570	8266	9050	7771	3	7983	1871	3927	5549	1487
1241	2273	505	8816	4786	533	9347	888	3728	4135
6688	9456	2880	4616	7698	2955	9597	9188	8932	5605
1325	1294	8001	1814	5020	9470	8702	4083	6452	2863
6196	5085	9961	5306	1660	1809	8405	2019	2710	1368
1577	5112	874	6909	4126	8473	2065	1511	4778	4440
5778	1207	3337	1888	1420	6917	4160	2682	5263	5926
6635	1887	8836	2940	2404	7017	3119	3699	2529	8663
6813	5759	3314	6929	5238	6008	5900	8485	5938	5642
5208	2391	8324	6888	9449	2577	7859	176	1650	8389
5446	4412	9857	9535	2794	7883	4119	6439	8082	7918
2984	2126	9506	2188	9762	9775	4213	7624	4520	1086
371	4559	12	718	8403	8150	6533	3741	6279	8546
4669	1053	3343	4889	9088	9188	8093	9496	8806	923
4070	3408	8102	3012	9706	771	8296	3094	148	7244
4867	6267	1225	6539	7958	7217	7833	728	1610	5284
4665	1912	5320	8563	1365	3834	1818	7791	7704	2460

Appendix B

Standard normal probability distribution



The critical value Z^* given below has a probability of p under H_0, i.e., $P(Z > Z^*|H_0) = p$ (the blue area), and it has a probability of B to have a value in the interval $(-Z^*, +Z^*)$ (the yellow area). The Z distribution is symmetrical around $Z = 0$, hence $P(Z < -Z^*) = P(Z > Z^*)$.

The first table reports the critical boundary values Z^* for some frequently used probabilities of p and frequently used confidence intervals of B:

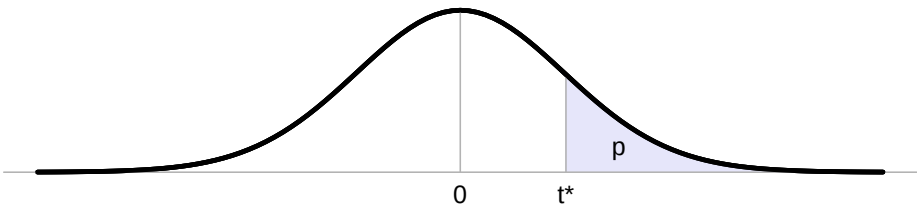
p	0.2	0.1	0.05	0.025	0.01	0.005	0.0025	0.001
B	60%	80%	90%	95%	98%	99%	99.5%	99.8%
Z*	0.8416	1.282	1.645	1.960	2.326	2.576	2.807	3.090

The second table reports the probabilities p and confidence intervals B for some frequently used critical values of Z^*:

p	0.3085	0.1587	0.0668	0.0228	0.0062	0.0013	0.0002
B	38.29%	68.27%	86.64%	95.45%	98.76%	99.73%	99.95%
Z*	0.5	1	1.5	2	2.5	3	3.5

Appendix C

Critical values for t-distribution



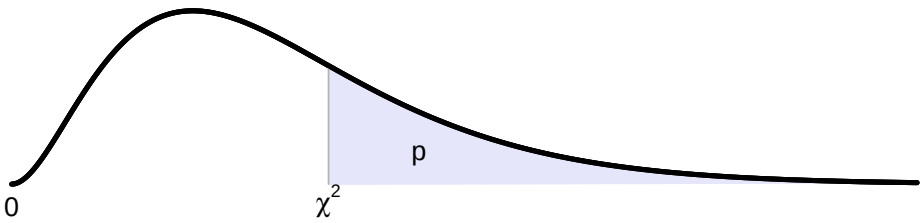
The critical boundary value t^* given below has a critical probability p under H_0, i.e. $P(t \geq t^*|H_0) = p$, and has probability B of a value between $(-t^*, +t^*)$. The t-distribution is symmetric around $t = 0$, thus $P(t < -t^*) = P(t > t^*)$.

The table below provides the critical boundary values t^* for much used critical probabilities p and confidence intervals B, for the degrees of freedom indicated in the first column.

	p	0.2	0.1	0.05	0.025	0.01	0.005	0.0025	0.001
	B	60%	80%	90%	95%	98%	99%	99.5%	99.8%
1		1.376	3.078	6.314	12.706	31.821	63.657	127.321	318.309
2		1.061	1.886	2.920	4.303	6.965	9.925	14.089	22.327
3		0.9785	1.638	2.353	3.182	4.541	5.841	7.453	10.215
4		0.941	1.533	2.132	2.776	3.747	4.604	5.598	7.173
5		0.9195	1.476	2.015	2.571	3.365	4.032	4.773	5.893
6		0.9057	1.440	1.943	2.447	3.143	3.707	4.317	5.208
7		0.896	1.415	1.895	2.365	2.998	3.499	4.029	4.785
8		0.8889	1.397	1.860	2.306	2.896	3.355	3.833	4.501
9		0.8834	1.383	1.833	2.262	2.821	3.250	3.690	4.297
10		0.8791	1.372	1.812	2.228	2.764	3.169	3.581	4.144
11		0.8755	1.363	1.796	2.201	2.718	3.106	3.497	4.025
12		0.8726	1.356	1.782	2.179	2.681	3.055	3.428	3.930
13		0.8702	1.350	1.771	2.160	2.650	3.012	3.372	3.852
14		0.8681	1.345	1.761	2.145	2.624	2.977	3.326	3.787
15		0.8662	1.341	1.753	2.131	2.602	2.947	3.286	3.733
16		0.8647	1.337	1.746	2.120	2.583	2.921	3.252	3.686
17		0.8633	1.333	1.740	2.110	2.567	2.898	3.222	3.646
18		0.862	1.330	1.734	2.101	2.552	2.878	3.197	3.610
19		0.861	1.328	1.729	2.093	2.539	2.861	3.174	3.579
20		0.860	1.325	1.725	2.086	2.528	2.845	3.153	3.552
21		0.8591	1.323	1.721	2.080	2.518	2.831	3.135	3.527
22		0.8583	1.321	1.717	2.074	2.508	2.819	3.119	3.505
23		0.8575	1.319	1.714	2.069	2.500	2.807	3.104	3.485
24		0.8569	1.318	1.711	2.064	2.492	2.797	3.091	3.467
25		0.8562	1.316	1.708	2.060	2.485	2.787	3.078	3.450
30		0.8538	1.310	1.697	2.042	2.457	2.750	3.030	3.385
40		0.8507	1.303	1.684	2.021	2.423	2.704	2.971	3.307
50		0.8489	1.299	1.676	2.009	2.403	2.678	2.937	3.261
100		0.8452	1.290	1.660	1.984	2.364	2.626	2.871	3.174
200		0.8434	1.286	1.653	1.972	2.345	2.601	2.839	3.131
400		0.8425	1.284	1.649	1.966	2.336	2.588	2.823	3.111
∞		0.8416	1.282	1.645	1.960	2.326	2.576	2.807	3.090

Appendix D

Critical values for χ^2-distribution



The critical value $(\chi^2)^*$ given below has a critical probability p under H_0, i.e. $P(\chi^2 \geq (\chi^2)^* | H_0) = p$.

The table below provides the critical boundary values $(\chi^2)^*$ for much used critical probabilities p, for the degrees of freedom indicated in the first column.

p	0.2	0.1	0.05	0.025	0.01	0.005	0.0025	0.001
1	1.64	2.71	3.84	5.02	6.63	7.88	9.14	10.83
2	3.22	4.61	5.99	7.38	9.21	10.60	11.98	13.82
3	4.64	6.25	7.81	9.35	11.34	12.84	14.32	16.27
4	5.99	7.78	9.49	11.14	13.28	14.86	16.42	18.47
5	7.29	9.24	11.07	12.83	15.09	16.75	18.39	20.52
6	8.56	10.64	12.59	14.45	16.81	18.55	20.25	22.46
7	9.80	12.02	14.07	16.01	18.48	20.28	22.04	24.32
8	11.03	13.36	15.51	17.53	20.09	21.95	23.77	26.12
9	12.24	14.68	16.92	19.02	21.67	23.59	25.46	27.88
10	13.44	15.99	18.31	20.48	23.21	25.19	27.11	29.59
11	14.63	17.28	19.68	21.92	24.72	26.76	28.73	31.26
12	15.81	18.55	21.03	23.34	26.22	28.30	30.32	32.91
13	16.98	19.81	22.36	24.74	27.69	29.82	31.88	34.53
14	18.15	21.06	23.68	26.12	29.14	31.32	33.43	36.12
15	19.31	22.31	25.00	27.49	30.58	32.80	34.95	37.70
16	20.47	23.54	26.30	28.85	32.00	34.27	36.46	39.25
17	21.61	24.77	27.59	30.19	33.41	35.72	37.95	40.79
18	22.76	25.99	28.87	31.53	34.81	37.16	39.42	42.31
19	23.90	27.20	30.14	32.85	36.19	38.58	40.88	43.82
20	25.04	28.41	31.41	34.17	37.57	40.00	42.34	45.31
21	26.17	29.62	32.67	35.48	38.93	41.40	43.78	46.80
22	27.30	30.81	33.92	36.78	40.29	42.80	45.20	48.27
23	28.43	32.01	35.17	38.08	41.64	44.18	46.62	49.73
24	29.55	33.20	36.42	39.36	42.98	45.56	48.03	51.18
25	30.68	34.38	37.65	40.65	44.31	46.93	49.44	52.62
30	36.25	40.26	43.77	46.98	50.89	53.67	56.33	59.70
40	47.27	51.81	55.76	59.34	63.69	66.77	69.70	73.40
50	58.16	63.17	67.50	71.42	76.15	79.49	82.66	86.66
100	111.67	118.50	124.34	129.56	135.81	140.17	144.29	149.45
200	216.61	226.02	233.99	241.06	249.45	255.26	260.74	267.54
400	423.59	436.65	447.63	457.31	468.72	476.61	483.99	493.13

Bibliography

(2015). Alex Foundation.

American Psychological Association (2010). *Publication Manual of the American Psychological Association*. American Psychological Association, Washington, D.C., 6th edition.

Andersson, I., Gaudin, J., Graca, A., Holm, K., Öhlin, L., Marklund, U., and Ericsson, A. (2011). Productive vocabulary size development in children aged 18-24 months: Gender differences. *KTH Speech, Music and Hearing: Quarterly Progress and Status Report*, 51(1).

Aron, A., Coups, E. J., and Aron, E. N. (2011). *Statistics for the Behavioral and Social Sciences: A Brief Course*. Pearson, Boston, 5th edition.

Ayres, J., Hopf, T., and Will, A. (2000). Are reductions in CA an experimental artifact? A Solomon four-group answer. *Communication Quarterly*, 48(1):19–26.

Bhargava, S. and Pathania, V. (2013). Driving under the (cellular) influence. *American Economic Journal: Economic Policy*, 5(3):92–125.

Boswall, J. (z.j.). Alex, the talking parrot.

Chen, L.-T. and Peng, C.-Y. J. (2015). The sensitivity of three methods to nonnormality and unequal variances in interval estimation of effect sizes. *Behavior Research Methods*, 47(1):107–126.

Cochran, W. (1977). *Sampling Techniques*. Wiley, New York, 3e edition.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum, Hillsdale, N.J., 2e edition.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, 78(1):98–104.

Cumming, G. (2012). *Understanding The New Statistics*. Routledge, New York.

Dalgaard, P. (2002). *Introductory Statistics with R*. Springer.

De Groot, A. (1961). *Methodologie: Grondslagen van onderzoek en denken in de gedragswetenschappen*. Mouton, 's-Gravenhage.

De Jong, N. H., Groenhout, R., Schoonen, R., and Hulstijn, J. H. (2015). Second language fluency: Speaking style or proficiency? Correcting measures of second language fluency for first language behavior. *Applied Psycholinguistics*, 36(2):223–243.

Deutsch, D. (2006). The enigma of absolute pitch. *Acoustics Today*, 2:11–19.

Dingemanse, M., Torreira, F., and Enfield, N. (2013). Is "huh?" a universal word? conversational infrastructure and the convergent evolution of linguistic items. *PLOS One*, 8(11):e78273.

Doblhammer, G. (1999). Longevity and month of birth: Evidence from Austria and Denmark. *Demographic Research*, 1(3).

Donald, D. (1983). The use and value of illustrations as contextual information for readers at different progress and developmental levels. *British Journal of Educational Psychology*, 53(2):175–185.

Drake, C. and Ben El Heni, J. (2003). Synchronizing with music: Intercultural differences. *Annals of the New York Academy of Sciences*, 999(1):429–437.

Ferguson, G. A. and Takane, Y. (1989). *Statistical Analysis in Psychology and Education*. McGraw-Hill, New York, 6e edition.

Gelman, A. and Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, Cambridge.

Gliner, J. A., Morgan, G. A., and Harmon, R. J. (2001). Measurement reliability. *Journal of the American Academy of Child and Adolescent Psychiatry*, 40(4):486–488.

Henrich, J., Heine, S. J., and Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, 33(2-3):61–83.

Houtkoop-Steenstra, H. (1991). Hoe een gesloten vraag toch open kan zijn. *Tijdschrift voor Taalbeheersing*, 13(3):185–196.

Hume, D. (1739). *A Treatise on Human Nature*.

Johnson, E. K. and Zamuner, T. (2010). *Using infant and toddler testing methods in language acquisition research*, chapter 4, pages 73–93. John Benjamins, Amsterdam.

Karp, J. A. and Brockington, D. (2005). Social desirability and response validity: A comparative analysis of overreporting voter turnout in five countries. *Journal of Politics*, 67(3):825–840.

Kerlinger, F. N. and Lee, H. B. (2000). *Foundations of Behavioral Research*. Harcourt College Publishers, Fort Worth, 4th edition.

Koring, L., Mak, P., and Reuland, E. (2012). The time course of argument reactivation revealed: Using the visual world paradigm. *Cognition*, 123(3):361–379.

Lata-Caneda, M., Piñeiro-Temprano, M., García-Fraga, I., García-Armesto, I., Barrueco-Egido, J., and Meijide-Failde, R. (2009). Spanish adaptation of the stroke and aphasia quality of life scale-39 (SAQOL-39). *European Journal of Physical and Rehabilitation Medicine*, 45(3):379–384.

Lev-Ari, S. and Keysar, B. (2010). Why don't we believe non-native speakers? The influence of accent on credibility. *Journal of Experimental Social Psychology*, 46(6):1093–1096.

Levin, I. P. (1999). *Relating Statistics and Experimental Design: An introduction*. Sage University Papers Series on Quantitative Applications in the Social Sciences; 07-125. Sage, Thousand Oaks, CA.

Luyckx, K., Kloots, H., Coussé, E., and Gillis, S. (2007). *Klankfrequenties in het Nederlands*. Academia Press.

MacFarlane, J. (2020). *Pandoc: a universal document converter*.

McGraw, K. O. and Wong, S. P. (1992). A common language effect size statistic. *Psychological Bulletin*, 111(2):361–365.

Morton, A. (2003). *A Guide through the Theory of Knowledge*. Blackwell, Malden, MA, 3e edition.

Nakagawa, S. and Cuthill, I. C. (2007). Effect size, confidence interval and statistical significance: a practical guide for biologists. *Biological Reviews*, 82(4):591–605.

Office of Research Integrity (2012). Responsible conduct of research training.

Peck, R. and Devore, J. (2008). *Statistics: The exploration and analysis of data*. Thomsom/Cole, Belmont, CA, 6e edition.

Pfungst, O. (1907). *Das Pferd des Herrn von Osten (Der kluge Hans): Ein Beitrag zur experimentellen Tier- und Menschen-Psychologie*. J. A. Barth, Leipzig.

Plomp, R. and Mimpen, A. M. (1979). Improving the reliability of testing the speech reception threshold for sentences. *International Journal of Audiology*, 18(1):43–52.

Popper, K. (1935). *Logik der Forschung. Zur Erkentnistheorie der modernen Naturwissenschaft*. Julius Springer, Wien.

Popper, K. (1959). *The logic of scientific discovery*. Routledge, London.

Popper, K. (1963). *Conjectures and Refutations: The Growth of Scientific Knowledge*. Routledge and Kegan Paul, London.

Quené, H. (2010). *How to design and analyze language acquisition studies*, pages 269–287. Benjamins, Amsterdam.

Quené, H., Semin, G. R., and Foroni, F. (2012). Audible smiles and frowns affect speech comprehension. *Speech Communication*, 54(7):917–922.

Quené, H. (2008). Multilevel modeling of between-speaker and within-speaker variation in spontaneous speech tempo. *Journal of the Acoustical Society of America*, 123(2):1104–1113.

Retraction Watch (2018). The "regression to the mean project:" what researchers should know about a mistake many make. Technical report.

Richardson, E., DiBenedetto, B., Christ, A., Press, M., and Winsberg, B. G. (1978). An assessment of two methods for remediating reading deficiencies. *Reading Improvement*, 15(2):82.

Rijlaarsdam, G. (1986). *Effecten van leerlingrespons op aspecten van stelvaardigheid*. PhD thesis.

Rosenthal, R. and Rosnow, R. L. (2008). *Essentials of Behavioral Research: Methods and Data Analysis*. McGraw Hill, Boston, 3e edition.

Rosén, E., Stigson, H., and Sander, U. (2011). Literature review of pedestrian fatality risk as a function of car impact speed. *Accident Analysis and Prevention*, 43(1):25–33.

Sanders, E. (2011). *Eerste Hulp bij e-Onderzoek voor studenten in de geesteswetenschappen: Slimmer zoeken, slimmer documenteren*. Early Dutch Books Online.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods*, 1(2):115–129.

Schuurman, W. and De Kluiver, H. (2001). *Kop of munt: Kansrekening in het dagelijks leven*. Bert Bakker, Amsterdam.

Shadish, W. R., Cook, T. D., and Campbell, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Wadsworth, Belmont, CA.

Spiegelhalter, D. (2019). *The Art of Statistics: Learning from Data*. Pelican.

SWOV (2012). De relatie tussen snelheid en ongevallen.

Thompson, B. (2002). "Statistical," "practical," and "clinical": How many kinds of significance do counselors need to consider? *Journal of Counseling and Development*, 80(1):64–71.

Thompson, S. K. (2012). *Sampling*. Wiley series in probability and statistics. John Wiley, Hoboken, NJ, 3e edition.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley, Reading, MA.

Universiteitsbibliotheek, Vrije Universiteit Amsterdam (2015). Webcursus informatievaardigheden - algemeen - niveau b.

Van den Berg, M., Amuzu, E. K., Essizewa, K., Yevudey, E., and Tagba, K. (2017). Crosslinguistic effects in adjectivization strategies in Suriname, Ghana and Togo. In Cutler, C., Vrzić, Z., and Angermeyer, P., editors, *Language Contact in Africa and the African Diaspora in the Americas: in honor of John V. Singler*, pages 343–362. Benjamins, s.l.

Van den Bergh, H. and Meuffels, B. (1993). Schrijfvaardigheid. In Braet, A. and Van de Gein, J., editors, *Taalbeheersing als tekstwetenschap: terreinen en trends*. ICG, Dordrecht.

Van Kolfschooten, F. (1993). *Valse Vooruitgang: Bedrog in de Nederlandse wetenschap*. L.J. Veen, Amsterdam.

Verhoeven, J., De Pauw, G., and Kloots, H. (2004). Speech rate in a pluricentric language: A comparison between Dutch in Belgium and the Netherlands. *Language and Speech*, 47(3):297–308.

VSNU (2018). Nederlandse gedragscode wetenschappelijke integriteit. Technical report, VSNU.

Watzlawick, P. (1977). *Is 'werkelijk' waar? Spraakverwarring, zinsbegoocheling en onvoorstelbare werkelijkheid*. Van Loghum Slaterus, Deventer.

Weisstein, E. W. (2015). Pascal's formula.

Wijffels, J., van den Bergh, H., and van Dillen, S. (1992). Het sturend effect van vragen met voorbeeldantwoorden. *Tijdschrift voor Taalbeheersing*, 14(2):136–147.

Wilcox, R. R. (2012). *Introduction to Robust Estimation and Hypothesis Testing*. Elsevier, Burlington, 3rd edition.

Xie, Y. (2020). *bookdown: Authoring Books and Technical Documents with R Markdown*. R package version 0.21.